Automatic Text Summarization of Patent Documents

ELIN GUSTAFSSON
MASTER´S THESIS
DEPARTMENT OF ELECTRICAL AND INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING | LTH | LUND UNIVERSITY

Automatic Text Summarization of Patent Documents

Elin Gustafsson
mhi12egu@student.lu.se

Department of Electrical and Information Technology
Lund University and
AWA Sweden

Supervisor: Fredrik Edman, fredrik.edman@eit.lth.se
Anders Fredriksson, anders.fredriksson@awa.com

Examiner: Erik Larsson, erik.larsson@eit.lth.se

December 7, 2020

Printed in Sweden
Tryckeriet i E-huset, Lund

Abstract

This thesis investigates Automatic Text Summarization and how it can be used to generate summaries of patent documents. The purpose of the generated summary was to convey the patent document's subject so that the reader could decide whether the patent document is relevant and should be read in full or could be discarded. This, in order to reduce the time patent attorneys need to spend reading full patent documents in their daily work.

A summarizing tool using Extraction-based summarization, and a Graph-based ranking method was implemented and tested. Summaries were generated from ten patent descriptions using the implemented tool. Each summary was also evaluated using human evaluation. The method showed very promising results in summarizing patent descriptions. In addition, the results highlighted some areas of improvement for future work.

Keywords: *automatic text summarization*, *patent documents*, *natural language processing*, *extraction-based summarization*, *PageRank*, *legal tech*.

Acknowledgements

There are a couple of people I want to thank for their contributions to this thesis. First, thanks to Anders Fredriksson, my supervisor at AWA, for giving me this opportunity and support during the project. Second, thanks to Fredrik Edman, my supervisor at LTH, for guiding me through the project, providing me with good feedback, ideas, and support.

Thank you to all the evaluators taking the time to evaluate the generated summaries for me.

Lastly, and most importantly, thank you to my family and friends for supporting me and cheering me on through the though times. Thank you for your ideas and feedback, and for proofreading all my drafts.

Popular Science Summary

INSTITUTIONEN FÖR ELEKTRO- OCH INFORMATIONSTEKNIK | LUNDS TEKNISKA HÖGSKOLA
PRESENTERAD 2020-12-03
EXAMENSARBETE Automatic Text Summarization of Patent Documents
STUDENT Elin Gustafsson
HANDLEDARE Fredrik Edman (LTH), Anders Fredriksson (AWA)
EXAMINATOR Erik Larsson (LTH)

Automatisk Textsammanfattning av Patentdokument

POPULÄRVETENSKAPLIG SAMMANFATTNING **Elin Gustafsson**

Går det att generera sammanfattningar automatiskt eller måste det göras manuellt? Jag har undersökt om Automatisk Textsammanfattning kan vara användbart inom patentbranschen. Mitt verktyg genererar sammanfattningar baserat på patentbeskrivningar för att minska tiden ett patentombud behöver lägga på att läsa hela patentdokument.

En stor del av arbetet för ett patentombud består av att läsa patentdokument. Detta är en tidskrävande uppgift då ett patentdokument kan vara uppemot hundra sidor långt. För att minska denna tid har jag tagit fram ett verktyg som automatiskt genererar sammanfattningar av patentbeskrivningar. Målet med sammanfattningen är att den ska förmedla ämnet i det fulla patentdokumentet så att läsaren kan avgöra om det fulla patentdokumentet är relevant eller inte.

Mitt verktyg använder extraktionsbaserad sammanfattning. Det innebär att man extraherar meningar från det dokument man vill sammanfatta och dessa meningar bygger upp den genererade sammanfattningen. För att identifiera vilka meningar som ska extraheras används en grafbaserad metod. Varje mening representeras av en nod i grafen. Länken mellan två noder representerar hur lika meningarna är varandra, ju fler gemensamma ord desto större likhet. Genom att använda en PageRank algoritm, något som oftast används för att ranka webbsidor, kan man identifiera de mest centrala noderna, alltså de meningar som liknar många andra meningar i texten. Dessa innehåller förmodligen mycket av textens information. Meningar plockas ut från texten, sorteras så att de hamnar i den ordning de är skrivna och sätts samman till en sammanfattning. För att illustrera har jag kört denna populärvetenskapliga sammanfattning genom mitt verktyg. Följande textruta är den genererade sammanfattningen. Resultaten från studien visar att den valda metoden fungerar bra. Sju av tio sammanfattningar lyckades förmedla vad patentdokumentet handlade om. De tre som misslyckades gav tydlig information om vad som kan förbättras i metoden. Den främsta förbättringsmöjligheten är att begränsa hur många meningar, som är väldigt lika varandra, som får extraheras till sammanfattningen. Detta för att sammanfattningen inte ska domineras av meningar som upplevs repetitiva.

Automatisk Textsammanfattning av Patentdokument. Mitt verktyg genererar sammanfattningar baserat på patentbeskrivningar för att minska tiden ett patentombud behöver lägga på att läsa hela patentdokument. För att minska denna tid har jag tagit fram ett verktyg som automatiskt genererar sammanfattningar av patentbeskrivningar. Det innebär att man extraherar meningar från det dokument man vill sammanfatta och dessa meningar bygger upp den genererade sammanfattningen. För att identifiera vilka meningar som ska extraheras används en grafbaserad metod.

Table of Contents

List of Figures

List of Tables

Notation

w	Word
N	Total number of words in document
$f(w)$	Number of occurrences of word w in document
S	Sentence
|S|	Length of sentence in terms of number of words in sentence
$\mathit{Weight}(w)$	Weight for word w
$\mathit{Weight}(S)$	Weight for sentence S
G	Graph
ν	Nodes in graph
ε	Links in graph
W	Weight matrix
k	Out-degree
k^-	In-degree
z	Centrality measure

Chapter 1
Introduction

1.1 Overview

Gathering and disseminating huge amounts of digital information has been a growing trend in society for the last decade. The International Data Corporation (IDC) projects that the total amount of digital data circulating annually around the world would hit 180 zettabytes in 2025 [8]. Although the data is available digitally, a large portion of managing and analyzing its content is still done manually by people. This is especially true for people working within academia and in law - such as researchers and patent attorneys. Because of this, a new field of technology called Legal Tech has emerged. Legal Tech refers to the use of software and technology to provide legal service and support the legal industry, such as the patent industry [17]. In this thesis, I have investigated whether the area of Natural Language Processing (NLP), specifically Automatic Text Summarization, can be used to produce summaries of good enough quality to reduce the time a patent attorney needs to spend reading the complete patent document. In this chapter the problem is presented in more detail, outlining the objective of the master thesis and relating it to previous work.

1.2 The problem

Working as a patent attorney involves reading and analyzing many patent documents that sometimes can be over 100 pages each. The main purpose of this is to search for prior art (already existing inventions that prevent a patent from being granted due to lack of novelty of the invention). This work is usually very tedious, time-consuming, and error-prone as a patent attorney may have 30-40 patent documents to read through from each search. There is therefore a great need to reduce the manual workload in every way possible. Having access to an abstract or a summary of the patent document would be a tremendous help in providing an overview of the patent document and help in deciding whether the patent document is relevant or not and if it should be read in full or could be discarded. The abstract and/or the summary of the patent document is however not always representative of the detailed technical content that is needed for deciding the relevance. A solution to this problem is to make an additional summary of the patent document, which takes all relevant legal and technical aspects of the

document into account. Such qualitative summaries are very expensive (since they are usually done by hand) and only available to order from companies having professional patent information tools and is therefore not a readily available solution to the problem.

Creating a summary in a straightforward and fast way that can capture the information in the patent document in a sufficient way is therefore desirable. One way to do it is by using Natural Language Processing techniques. The problem to be studied here is if one can create an Automatic Text Summarizer that can solve the task with sufficient quality, and within a reasonable time, i.e not be too computational heavy.

1.2.1 Objectives

The aim of this master thesis project is to

1. investigate to which extent one can use automatic text summarization to create a qualitative summary of a patent document,
2. to implement a prototype summarizer for producing summaries of patent documents, and
3. to run the summarizer on a number of patent documents and evaluate the result.

1.2.2 Limitations

Patent documents can be written in several languages. The European Patent Office acknowledges English, French, and German as official languages for patent documents. Many countries' local patent registration offices also acknowledge the main language in the country. Natural language processing techniques are most developed in the English language and this thesis will therefore focus its efforts on patent documents written in English.

1.3 Previous/related work

There are several papers on Automatic Text Summarization. *A survey on automatic text summarization* (2019) by Nazari and Mahdavi [11], *Mining Text Data, A survey of text summarization techniques* (2012) by Nenkova and McKeown [1], and *Text summarization techniques: A brief survey* (2017) by Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez, and Kochut [2] all provide a good overview of the subject.

Regarding automatic text summarization of patent documents, two relevant articles have been found.

Trappey and Trappey presented a paper in 2008 on patent document summarization called *An R&D knowledge management method for patent document summarization* [14]. Their objective was to create a summary for engineers who

in their work need to analyze and categorize patents as part of research, development, and design processes. The method consists of three parts. The first part is key phrase recognition technology. The second part is significant information density which uses paragraph similarities for information concept clustering. The third part is a summary template and domain-specific patent rules to be used to improve the result of the summary. The method was tested on 111 patents from the IPC classification B25, which is the category of hand tools, portable power-driven tools, and handles for hand implement. The automatically generated summaries were evaluated by using an automatic classification tool. If the summary was classified in the same patent class as the full patent was then it was deemed accurate. The method showed promising results in being able to correctly classify patents based on the automatically generated summary. As future work, the authors suggest a focus on creating summarization on multiple patents from a common domain and clustering of patents.

In 2009 Trappey, Trappey, and Wu presented a new method in their paper *Automatic patent document summarization for collaborative knowledge systems and services* [15]. Using domain concepts and semantic relationships they could identify key words and key phrases. The domain concepts and ontology needed to be defined in advance. The retrieved key words and phrases were then used to identify high information density paragraphs in the patent document to be used for the summary. In the categories of hand tools and chemical mechanical polishing 200 patents were summarized using this method and the results were compared to the previous method presented by Trappey and Trappey (2008). The results were promising here as well, and the new method performed better than the previous.

Both methods performed well in accurately classifying the patents based on the summaries. From the articles, it is however not clear whether the automatically generated summaries were able to capture information from the full patent for a reader to understand what the patent is about, which is what this thesis will investigate.

Chapter 2
Theory

2.1 Background

Automatic text summarization is the task of using software to produce a concise and fluent summary of one or multiple text documents while preserving the original documents' key information and overall meaning [2]. The first ever method on automatic text summarization was introduced in 1958 by Baxendale in the article *"Machine-made index for technical literature - An experiment"* [12]. Baxendale's method focused on the position of sentences in the input document and concluded that in 85% of the cases the first sentence in a paragraph was a topic sentence and in 7% of the cases the last sentence was a topic sentence. Based on this he proposed that the first and the last sentence of a paragraph should be extracted to form the summary of the paragraph. For the time and type of documents he was working on it was a very reasonable approach but with today's computing possibilities it is deemed too simple and naive, and not used anymore. Many other methods have been presented since, and this chapter will detail some of the main approaches.

2.2 General types of summarization

Automatic text summarization is divided into two methods: extraction-based summarization, and abstraction-based summarization. The main difference between them is that extraction-based summarization extracts objects from the original text without modifying them to form the summary while abstraction-based summarization generates new sentences for the summary by interpreting and examining the input text.

2.2.1 Extraction-based summarization methods

Extraction-based summarization produces summaries by extracting objects from the original texts without modifying them. These objects can be key words, phrases, or whole sentences. The extracted objects will then be used to form the summary. Even though extraction-based summaries are not the way a human would create a summary, most research has focused on this method as it is straightforward to compute and has produced good results. There are several ways

to identify the sentences to be extracted. The most common are frequency-based, and machine learning [2, 6]. In Figure 2.1 an example of how an extraction-based summarizer would work is shown. The highlighted sentences are the ones selected for the summary. These were chosen by running the text passage through the implemented program explained in Chapter 3.

Extraction-based summarization produces summaries by extracting objects from the original texts without modifying them. These objects can be key words, phrases or whole sentences. The extracted objects will then be used to form the summary. Even though extraction-based summaries are not the way a human would create a summary, most research has focused on this method as it is straightforward to compute and has produced good results. There are several ways to identify the sentences to be extracted. The most common are frequency-based, and machine learning [2, 6]

Figure 2.1: An example of how an extraction-based summarizer works if it is asked to select a couple of sentences to summarize a passage of text. The highlighted sentences are the ones selected for the summary.

2.2.2 Abstraction-based summarization

Abstraction-based summarization generates summaries which are closer to what a human might generate [10]. The method interprets and examines the text using advanced natural language techniques and creates new sentences to convey the most critical information from the original text. In general, abstraction-based methods can create a more condensed summary than extraction-based methods. The challenge with abstraction-based methods is to handle semantic representation, inference, and natural language generation which is computationally more taxing and often requires deep domain knowledge of the documents [2, 16]. Figure 2.2 shows examples of how abstraction-based summarization would work if it was used to summarize books in just a couple of sentences.

This thesis will focus on the extraction-based method since they do not require domain knowledge of the document and are more straightforward to implement. The summarization methods described in this chapter will therefore only relate to the extraction-based method.

2.3 Machine Learning

Machine learning methods for automatic text summarization are modeled as a classification problem, where sentences are classified into two categories, summary sentences and non-summary sentences, based on their features. A corpus of training data is used to train a statistical classifier to be able to sort sentences into the two categories. An advantage of using machine learning approaches is the freedom it offers in terms of the number of indicators of importance one can use to classify

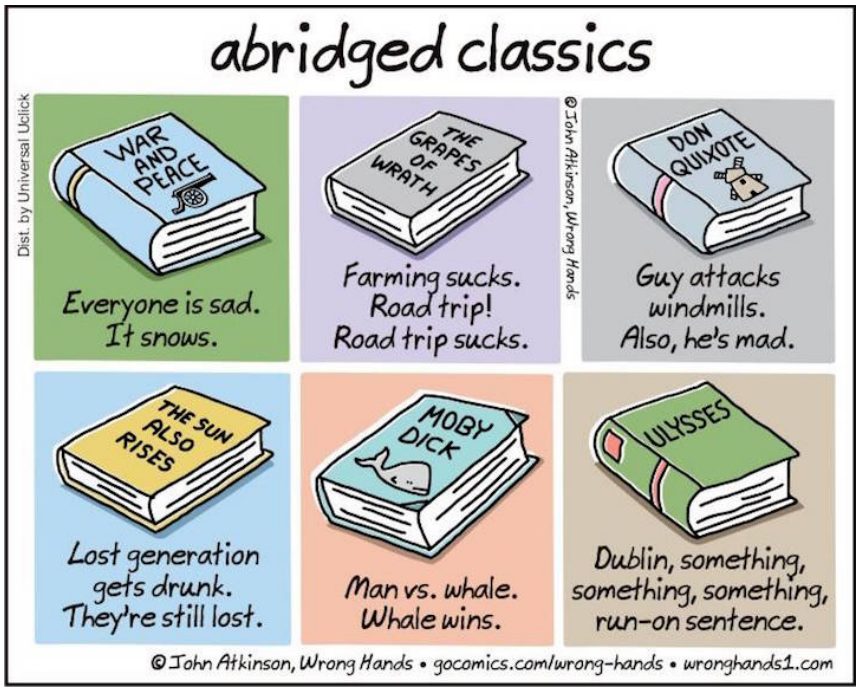


Figure 2.2: Examples of how abstraction-based summarization works.

the sentences. Examples of indicators of importance are, e.g. position in text, weight of words in a sentence, and sentence length.

Machine learning methods have proven successful in both single and multi-document summarization, especially when the classifiers are trained to locate a particular type of information in class specific documents such as scientific papers. The need for training data is however the most significant drawback with using supervised machine learning classifiers. To create training data one common approach is to ask annotators to select summary sentences in documents. This however is time-consuming and not fully reliable as different people tend to choose different sentences for an extractive summary.

Another option is to use a semi-supervised approach. In this approach, two classifiers are trained on a small set of examples for summary and non-summary sentences. After they are trained one classifier is run on unannotated data and its most confident predictions are added as examples to continue to train the second classifier, repeating this setup until one is satisfied with the training of the second classifier. Then the second classifier is used for the task of choosing summary sentences [1, 2, 11].

2.4 Frequency-driven methods

Frequency-driven methods work by assigning weights to words, segments, and/or sentences depending on how frequently they appear in the original document. A problem with frequency-driven methods is that words that are repeated and carry little information are very common in texts and they will therefore have a high frequency without adding value to the overall meaning of the text. Another problem is that a word can appear in many forms, e.g. its root form or inflected forms. The algorithms treat these two forms as two different words even though the difference might be as small as one letter. To address these problems a pre-processing step is often added before using the frequency-driven methods. The pre-processing step usually contains stopping, splitting, lemmatization, and stemming which is going to be further discussed in the next section [14, 19].

2.4.1 Pre-processing

Stopping is the process of removing the words that carry little information from the text. By having a predefined list of the so-called stop words one can choose which words to remove [14].

Splitting is the process of splitting the text into sentences, segments, or individual words [14].

Lemmatization does a morphological analysis of the words (ex. rocks → rock, better → good, corpora → corpus). Running lemmatization on the text makes it easier to compare sentences to each other since grouping together the inflected forms of a word enables it to be analyzed as a single item [14, 18].

Stemming is the process of reducing words to their stem. The stem does not necessarily need to be a word. The Porter Stemmer [19], for example, reduces the words *argue, argued, argues, arguing,* and *argus* to the stem *argu*. There are two sources of error one need to be aware of when stemming; over-stemming and under-stemming [14, 19].

Over-stemming is when two words are stemmed to the same root even though they should not have been. For example, the Porter stemmer stems *universal, university,* and *universe* to *univers*. This poses a problem when comparing sentences and similarities between the sentences is found even though it should not. The problem with under-stemming is the opposite. Here the stemmer does not stem two words to the same stem even though they are the same root word, for example the Porter stemmer stems *alumnus* → *alumnu*, *alumni* → *alumni*, *alumna/alumnae* → *alumna*. This leads to a similarity not being found between two sentences even though it is there [19].

Running these pre-processing steps facilitates the calculation of weights of words and sentences since words are now written in their root form, and the text is structured in the same manner throughout, making it analyzable. After the pre-processing steps the weight of words, and/or sentences are calculated. The two most commonly used methods for calculating weights of words and/or sentences are Word Probability and Term Frequency - Inverse Document Frequency.

2.4.2 Word Probability

The simplest way to calculate the weights or importance of a word is to use word probability. The probability (weight) of a word w, $p(w)$, is calculated as the number of occurrences, $f(w)$, in the input divided by the number of all the words in the input, N, hence

$$Weight(w) = p(w) = \frac{f(w)}{N} \tag{2.1}$$

[1, 2].

2.4.3 Term Frequency - Inverse Document Frequency (TF-IDF)

Deciding which words to include in a stop list is not a trivial task. There are Natural Language Processing libraries with pre-prepared lists that cover the most common words, but sometimes domain specific words need to be added, a task that requires domain knowledge. Term Frequency - Inverse Document Frequency, TF-IDF, is a statistical method that is an alternative to removing stop words and assigning weights to words. The method relies on a large collection of background documents. Normally, the background collection consists of documents from the same genre as the document that is to be summarized. The background collection serves as an indication of how often a word is expected to be found in a similar text to the one being summarized [1, 2].

TF-TDF assigns high weights to words appearing often in a document but are not very common in other documents according to the following equation where we can see that Term Frequency is calculated the same way as Word Probability

$$Weight(w) = TF * IDF = \frac{f(w)}{N} * \log \frac{D}{d(w)}. \tag{2.2}$$

Here D is the number of documents in the background collection, and $d(w)$ is the number of documents in the background collection that contain the word w. The IDF will be low for words that have a high frequency in the background collection, and even zero for words that occur in all documents.

TF-IDF weights are proven to be good indicators of importance, and the method is easy to compute. It is, therefore, a very popular method and is incorporated in many existing summarizers today where a representative collection of background documents is available [1, 2, 14].

The frequency-driven method favors sentences that contain words that appear with high frequency in the input text. This is an advantage if the goal is to identify the main subject in a text, but a drawback if one wants a general summary of all the subjects from the input text [2, 11].

2.5 Clustering

The similarity between sentences or paragraphs is a relevant feature to consider when summarizing text. Clustering is the method of gathering similar text units, e.g. sentences, paragraphs, etc. together to identify the common information

between them. The more sentences in a cluster the more important the information in the cluster is to the summary. Similarities between sentences are calculated to create the clusters. This can be done in many ways but using the co-sine similarity is the most common way.

For the clustering methods the sentences need to be represented as numerical vectors. The most straightforward way to do this is to set each vector index to represent a word in the document. A vector representing a sentence will then have a value at the indices of the words that are present in the sentence and 0 for those words that are not. The value corresponding to the word present in the sentence is often the weight of the word calculated using one of the frequency-driven methods [1, 11].

2.5.1 Co-sine similarity

Co-sine similarity is a way to measure similarity between two non-zero vectors. It is defined as the cosine of the angle between the vectors and can be derived from the dot product between a vector u and a vector v as

$$\mathit{Similarity} = \cos\theta = \frac{u \cdot v}{|u||v|} \tag{2.3}$$

A value of 1 corresponds to identical vectors and 0 corresponds to orthogonal vectors [4].

2.5.2 Latent semantics analysis

Latent semantics analysis (LSA) is a method for extracting representation of text semantics based on observed co-occurrences of words. It clusters sentences into topics using singular value decomposition and was first presented by Gong and Liu in 2001 [1]. It starts by building a term-sentence matrix A, $m \times n$, where each row corresponds to a word from the input and each column corresponds to a sentence. The entries in matrix A are the weight of the word in the sentence, i.e. entry a_{ij} corresponds to the weight of word i in sentence j. The weights are calculated using the TF-IDF method, see Equation (2.2), and if the sentence does not contain the word the weight is zero.

Singular value decomposition (SVD) is used on matrix A to represent it as the product of three matrices: $A = U\Sigma V^T$. Where

- U is an $m \times n$ matrix where each column can be interpreted as a topic (a specific combination of words) with the weight of each word in the topic given by a real number.
- Σ is a diagonal $n \times n$ matrix where each diagonal element corresponds to the weight of a topic in U sorted in descending order.
- V^T is an orthonormal $n \times n$ matrix. One sentence per row, where each sentence, one sentence per row, is expressed in terms of the topics given in U.

The matrix $D = \Sigma V^T$ describes to what extent the sentence convey the topic, i.e. d_{ij} indicate the weight for topic i in sentence j.

When the LSA method was originally proposed by Gong and Liu [1] their idea was to select the sentence with the highest weight for each of the most important topics to form the summary. This selection strategy suffers from the drawback that more than one sentence may be required to convey the information pertinent to that topic. Several extensions have been presented to improve the method.

One extension generating good results is to use the weight of each topic to decide the relative size of the summary that should cover the topic. Another extension is to try and identify sentences that discuss several of the important topics, as these sentences are good candidates to include in the summary. To identify these sentences the weight of a sentence, S_i, is calculated as

$$Weight(S_i) = \sqrt{\sum_{j=1}^{n} d_{i,j}^2} \tag{2.4}$$

and sentences with high weights are selected [1, 3, 2].

After clustering the clusters are sorted according to size or weight depending on which clustering method is used and one or more sentences from either each cluster, or each of the n highest ranked clusters, are extracted to form the summary.

The disadvantage with clustering is that sentences can only be assigned into one cluster even though a sentence can express more than one topic [1, 11].

2.6 Graph-based methods

Graph-based methods exploit the same idea as clustering, i.e. to identify sentences with similarities, but in a more flexible way. In this method sentences are not restricted to being assigned to only one topic (cluster) but can be linked to multiple topics at the same time. Finding similarities between sentences can help sort out topics from the input text and give a good indication of which topics are most important. If the graph representation of the document consists of sub-graphs it indicates several discrete topics covered in the document.

The approach of graph-based methods is to convert the input document's sentences into a graph using similarity measures, for example co-sine similarity, and then use the graph structure to calculate the importance of the sentences. Sentences form the nodes and the links between nodes are present if there is a similarity between the sentences. Usually, a threshold value is set for the link values to simplify the graph resulting in only links between nodes that have similarity measure higher than the threshold [1, 2, 3, 11].

To understand how the graph-based method works we need to understand the basics of graph theory.

2.6.1 Graph theory

Key aspects and definitions in a graph

A graph is a mathematical structure used to model pairwise relations between objects. It consists of nodes, ν, links between the nodes, ε, and The three main aspects in a graph are the following:

- A set of nodes, ν, which in this case represent sentences from the input document. ν is a countable set.
- Pairwise connections between nodes by a set of links, ε, where in this case a link $e = (i, j)$ represents similarity between sentences i, $j \in \nu$. $\varepsilon \subseteq \nu \times \nu$ is the set of links.
- Positive, scalar value, W_{ij} for each link $(i, j) \in \varepsilon$ to be referred to as the link's weight, with the aim of quantifying the strength of the connection. $W \in \mathbb{R}_{+}^{\nu \times \nu}$ is the weight matrix and has the property that $W_{ij} > 0$ if and only if $(i, j) \in \varepsilon$, i.e. if (i, j) is a link, otherwise it is 0. Here the weight represents the degree of similarity between the sentences. In this case the similarity is equal in both ways, i.e. links (i, j) and (j, i) are either both present with the same weight $W_{ij} = W_{ji} > 0$ or both absent $W_{ij} = W_{ji} = 0$. Graphs with this feature are called undirected.

Given these aspects we define a a graph as

$$G = (\nu, \varepsilon, W). \tag{2.5}$$

To further illustrate these definitions an example of an undirected graph is shown in Figure 2.3 and the corresponding weight matrix presented below.

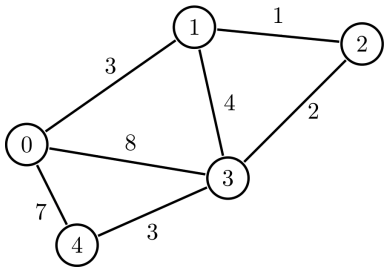


Figure 2.3: Example of an undirected graph with weights assigned to the links.

The corresponding weight matrix for the undirected graph in Figure 2.3:

$$W = \begin{bmatrix} 0 & 3 & 0 & 8 & 7 \\ 3 & 0 & 1 & 4 & 0 \\ 0 & 1 & 0 & 2 & 0 \\ 8 & 4 & 2 & 0 & 3 \\ 7 & 0 & 0 & 3 & 0 \end{bmatrix}.$$

Two other important aspects in graph theory are out-degree, k_i, and in-degree, k_i^- of a node i. They give a measure of the strength of the links going into and out from a node and they are defined as

$$k_i = \sum_{j \in \nu} W_{ij} \quad \text{and} \quad k_i^- = \sum_{j \in \nu} W_{ji}. \tag{2.6}$$

In an undirected graph $k_i = k_i^-$. Using the out-degree we can calculate the normalized weight matrix P, see Equation (2.7). The normalization is done so that the sum of each row equals 1, i.e. the sum of the out-degree for each node is 1. P is needed for the PageRank algorithm

$$P = \text{diag}(k)^{-1} W \tag{2.7}$$

where k is a vector containing the out-degree for all nodes. For the graph presented in Figure 2.3 P is calculated as follows [5].

$$P = \begin{bmatrix} 18 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 17 & 0 \\ 0 & 0 & 0 & 0 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 3 & 0 & 8 & 7 \\ 3 & 0 & 1 & 4 & 0 \\ 0 & 1 & 0 & 2 & 0 \\ 8 & 4 & 2 & 0 & 3 \\ 7 & 0 & 0 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 3/18 & 0 & 8/18 & 7/18 \\ 3/8 & 0 & 1/8 & 4/8 & 0 \\ 0 & 1/3 & 0 & 2/3 & 0 \\ 8/17 & 4/17 & 2/17 & 0 & 3/17 \\ 7/10 & 0 & 0 & 3/10 & 0 \end{bmatrix}$$

It is now clear that the sum of each row in the matrix is 1, i.e. the out-degree for each node now sums to 1.

2.6.2 Centrality and PageRank

Centrality is the measure that captures the importance of a node's position in a graph. The higher the centrality measure the more important is the node. The most simple centrality measure is degree-centrality, whereby the importance of a node i is simply by its degree, k.

A more sophisticated centrality measure is the PageRank centrality, z. The PageRank algorithm calculates the centrality measure iteratively, by doing a random walk on the graph and iteratively updating the centrality measure for each node. The centrality measure is determined by using the normalized weight matrix transposed, P', see Equation (2.7), and an intrinsic centrality $\boldsymbol{\mu}$, which is a non-negative vector, combined with $\beta \in (0, 1]$, which is a parameter that measures the weight of the intrinsic centrality relative to the network topology. The PageRank centrality vector for all nodes is calculated as Equation (2.8)

$$\mathbf{z} = (1 - \beta) P' \mathbf{z} + \beta \boldsymbol{\mu}. \tag{2.8}$$

For PageRank typical values of β are around 0.15, and the standard choice for $\boldsymbol{\mu}$ is $\boldsymbol{\mu} = \mathbf{1}$ so that all nodes have identical intrinsic centrality [5].

2.6.3 LexRank and TextRank

The most popular graph-based clustering methods that are inspired by the PageRank algorithm are called LexRank and TextRank. In both methods sentences from the input document are represented in the graph as nodes and links between nodes are representing a similarity score between the sentences. The similarity score in LexRank is given by co-sine similarity, see Section 2.5.1, and in TextRank one determines the similarity between two sentences, S_i and S_j based on the content that both share. This overlap is calculated as the number of common words, w_k, between the sentences, divided by the length, $|S|$ of each sentence to avoid promoting long sentences. The calculation is presented in Equation (2.9)

$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \ \ \& \ \ w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)}. \tag{2.9}$$

In both methods, PageRank is used to rank the sentences and the top ranked sentences are extracted for the summary. The drawback with this method is the same as with the frequency-driven methods, the method favors sentences containing words that are frequent in the input text. The advantage is, however, that more topics can be extracted using these methods than with the frequency-driven [1, 3, 4].

2.7 The impact of context in summarization

To know the structure of the text that is going to be summarized can help a lot when choosing the summarization method and constructing the summarizer. In many cases, summarizers can use additional information to determine the most important topics to be summarized. For example, if one wants to summarize a blog post the comments following the post can be indicative of what parts of the post are most important. If summarizing a scientific paper, later papers that cite the paper in question to be summarized, and in particular the citation sentences, can indicate what sentences in the original document are important [1, 2].

This thesis is summarizing patent documents and therefore wants to look into the structure of a patent document to see if there are any features that can be used in the summarization process.

2.7.1 The Structure of a Patent Document

A patent document covers documents ranging from the first piece of paper sent in as a patent application to a finished and approved patent. There is no universal template for how a patent document should be structured. The structure varies between countries and authors. There are, however, some common parts that are either mandatory or accepted as common practice to include by most patent attorneys. These parts are:

- **Front page**

  In the first page, bibliographic information about the title, inventors, and filing date can be found. One can usually also find an abstract describing the invention in short. The abstract is most often a rewrite of the first claim in the patent and can therefore lack the information needed for the reader to fully understand the topic of the patent.

- **Description**

  After the first page, the section titled "Description" follows. It contains a description of the technical field, background, summary of the invention, description of the drawings, and detailed description of the embodiments. The description provides sufficient disclosure of how to reduce the invention into practice.

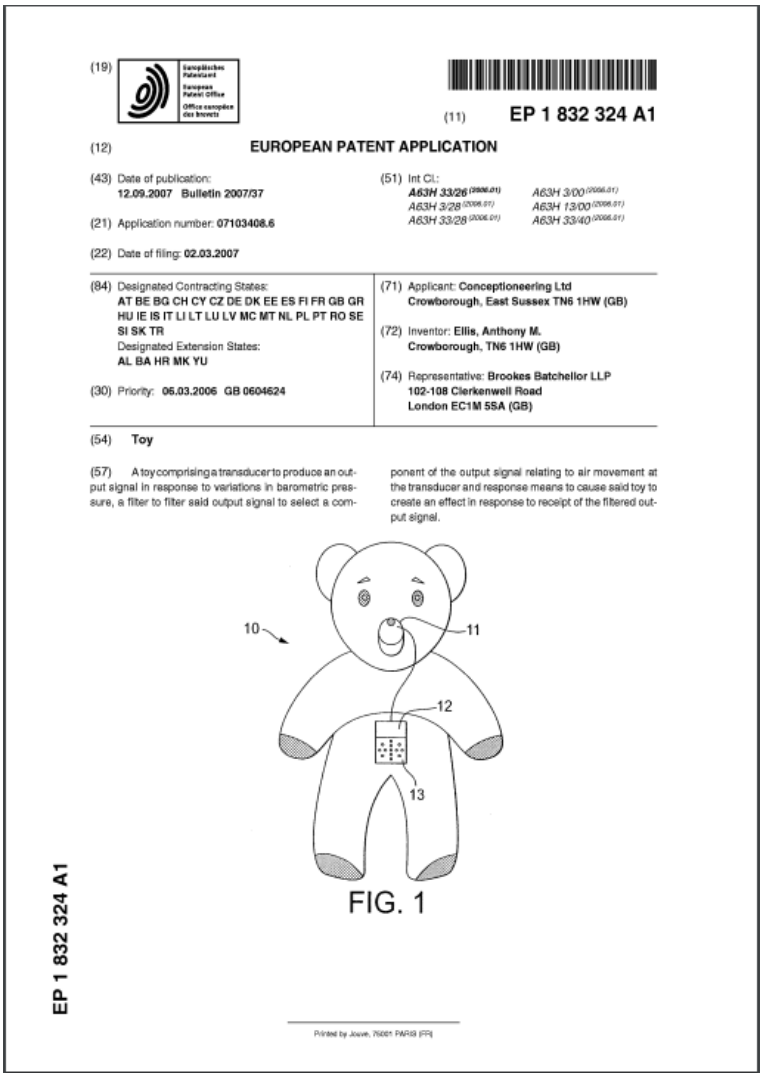
- **Claims**

  The claims follow the description. The claims define the scope of protection. A patent can contain several claims, each describing a certain part of the invention. In general, the first claim gives a general description of the invention and the further down the list one goes the more detailed the description becomes. Reference to earlier claims in the list is common. As each claim can only consist of one sentence, these sentences are usually very long and contain many subordinate clauses.

- **Figures**

  For European patent documents, the claims are followed by figures of the invention. In US patent documents the figures are located after the front page. The figures are always in black and white.

In Figure 2.4 an example of the the first page of a European patent application can be seen.



(19) Europäisches Patentamt / European Patent Office / Office européen des brevets

(11) EP 1 832 324 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
12.09.2007 Bulletin 2007/37

(51) Int Cl.:
A63H 33/26 (2006.01) A63H 3/00 (2006.01)
A63H 3/28 (2006.01) A63H 13/00 (2006.01)
A63H 33/28 (2006.01) A63H 33/40 (2006.01)

(21) Application number: 07103408.6

(22) Date of filing: 02.03.2007

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IS IT LI LT LU LV MC MT NL PL PT RO SE SI SK TR
Designated Extension States:
AL BA HR MK YU

(30) Priority: 06.03.2006 GB 0604624

(71) Applicant: Conceptioneering Ltd
Crowborough, East Sussex TN6 1HW (GB)

(72) Inventor: Ellis, Anthony M.
Crowborough, TN6 1HW (GB)

(74) Representative: Brookes Batchellor LLP
102-108 Clerkenwell Road
London EC1M 5SA (GB)

(54) Toy

(57) A toy comprising a transducer to produce an output signal in response to variations in barometric pressure, a filter to filter said output signal to select a component of the output signal relating to air movement at the transducer and response means to cause said toy to create an effect in response to receipt of the filtered output signal.

FIG. 1

EP 1 832 324 A1

Printed by Jouve, 75001 PARIS (FR)

Figure 2.4: First page of a European patent application.

A characteristic of patent documents is that the text is consistent in the use of words. Patent attorneys chose to use the same word choice in the description as in the claims. Synonyms are rarely used since the description needs to be clear

in what the invention is describing. Information is often repeated. Clarity is more important than a varied language. This is a great advantage if one uses the frequency-based methods and/or clustering for the summarization. It is, however, a drawback if the information needed is not repeated.

Sometimes patent applications can contain several inventions within the same area. The purpose is to get an application date for all by using divisional patent applications or continuation applications. This is a fairly common practice and therefore something to keep in mind when reading and summarizing patent documents [13].

For a reader to understand the general concept of the patent document, the bulk of information is located in the description section as this section should describe the invention sufficiently clear for enabling a skilled person to reduce this into practice. The description section is also the section that contains the most text. Thus, the description seems to be a good part of the patent document to summarize.

2.8 Evaluation methods

To evaluate the quality of a summary is a difficult task. There is no objectively correct summary to any document. If one, for instance, wants to summarize a book it is important that the story is conveyed in the summary. Details of names, clothing, etc. are less important. To understand a patent document one needs to understand the technical area of the invention as well as details of how the invention is constructed. The purpose related to the generated summary in this thesis is, therefore, to convey enough information of the technical area and the invention so that the summary can be used to determine whether the patent document is relevant to be read in full or not. If the patent document consists of several inventions the purpose is to be able to convey information about all of the inventions. This is important to keep in mind when choosing the evaluation method.

A summary can be used for information retrieval, question answering, executing instructions, and relevance assessment to mention a few, and one can evaluate structure, coherence, grammar, and more. If the summary is able to convey the content of the document then its structure, coherence, and grammar might not be as important. But if the goal is to create a high quality text such criteria play an important part. In general, there are two ways to evaluate an automatically produced summary: intrinsic evaluation and extrinsic evaluation. In Figure 2.5 the separation of these is illustrated. Intrinsic evaluation evaluates on content and extrinsic evaluation can beside content also evaluate coherence, structure, grammar etc. [2, 3, 9].

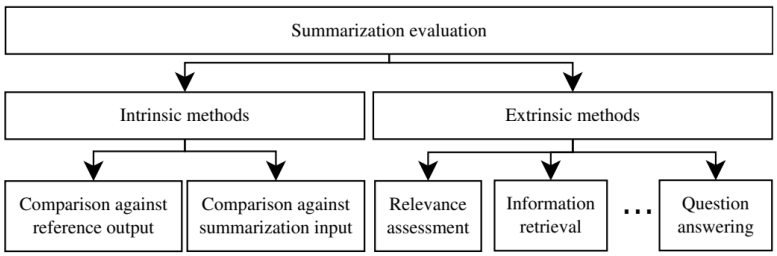


Figure 2.5: Classification of evaluation methods [3].

2.8.1 Intrinsic evaluation

Intrinsic evaluation relies on comparing the generated summary with either a reference summary or the original document. An easy evaluation method consists of the generated summary, GS, and the reference summary, RS which consists of pre-selected sentences from the input document.

We define recall, R as the quantity of right information recovered by the system compared to what it should recover (see Equation (2.10)), i.e. the intersection

between the reference summary and the generated summary divided by the size of the generated summary

$$R = \frac{RS \cap GS}{|GS|}. \tag{2.10}$$

The precision, P, is defined as the quantity of right information recovered by the system compared to what it has recovered (see Equation (2.11)), i.e the intersection between the reference summary and the generated summary divided by the size of the reference summary [3]

$$P = \frac{RS \cap GS}{|RS|}. \tag{2.11}$$

The F_1-score is a measure of a test's accuracy. It considers both the precision and the recall of the test to measure the accuracy. The F_1-score is computed as in Equation (2.12) [11].

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{2.12}$$

This evaluation method does only gives credit to the generated summary if it has chosen the exact same sentences as in the reference summary. Similar sentences, therefore, get no credit at all, which is why it is good to use if the generated summary must contain specific content from the input text. If it is sufficient that the summary conveys the content by using similar sentences, then this method is not as reliant [3].

ROUGE

The most widely used metric for intrinsic evaluation is called ROUGE, Recall-Oriented Understudy for Gisting Evaluation. It evaluates summary quality by comparing it to a human-generated summary. There are five evaluation metrics in the ROUGE tool: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU. The two most commonly used are the following:

- ROUGE-N is based on a comparison of n-grams. A n-gram is a continuous sequence of n items from a given sample of text. The items can be phonemes, syllables, letters, words, or base pairs. A series of n-grams are elicited from both the reference summary and the generated summary. The number of common n-grams between the two summaries is calculated and the score of the generated summary is the number of common n-grams divided by the number of extracted n-grams from the reference summary [2].
- ROUGE-L measures the longest common sub-sequence between the reference summary and the generated summary. The longer the common sub-sequence is the more similar the summaries are [2].

The main drawback with intrinsic evaluation is that it requires a reference summary to score the automatically generated summary. This summary must be created by a human, and, depending on the summarization method chosen, need to be extraction-based or abstraction-based. Creating these summaries is a tedious job and often require domain specific knowledge [9].

2.8.2 Extrinsic evaluation

Extrinsic evaluation evaluates the effectiveness of the summary's purpose and does therefore not need a reference summary. It usually relies on human evaluation. Automatic tools are also available, but since having a human assess the summary's quality is the simplest way to extrinsic evaluation they are not as common. The evaluation methods vary depending on the purpose of the summary [2, 3, 9].

- *Ad-hoc task* is intended for evaluating how well a text relates to a specific topic. The assessors are given a text and a description of a topic and are then asked to determine if the given text is pertinent to the topic. Some assessors are given a full text, while others are given a summary, but they are not told which one they will evaluate [3, 9].
- *Categorization task* aims at measuring the effectiveness of a summary in terms of containing enough information for an analyst to as quickly and correctly as possible categorize the text. The analyst is given either a full text or a summary and must then choose from a number of categories which one is pertinent to the text or else choose "None of the above" [3, 9].
- *Reading comprehension* evaluates the informativeness of a text. It uses a query-based evaluation to score the summary. The assessor is asked to answer some multiple choice-questions first based on the summary and then the original document. The number of correct answers based on the summary is defined to be the score of the summary [3, 9].

In addition to evaluating the task performance, one can also ask the assessors to evaluate coherence, structure, grammar, etc. The most straightforward way to do it is to ask the assessors to score the text for each category given a predefined (often five-point) scale [9].

The main drawback with extrinsic evaluation is its need for human assessors to do the evaluation. Even though the evaluation itself is easier to perform than intrinsic evaluation, human evaluation takes time and the results can become dependent on the assessors, in particular if the group of assessors is very homogeneous [9].

Chapter 3
Approach

3.1 Choice of method

Based on the literature study in Chapter 2 abstraction-based summarization was deemed not having sufficient evidence of its success rate and to difficult too implement, hence extraction-based summarization was chosen.

The lack of training data meant that machine learning could not be used either.

As mentioned in Section 2.7.1 a characteristic of patent text is the consistency in choice of words and phrasing. The same words and information is often repeated in several sentences.

The first, consistent use of words, is an advantage in the frequency driven methods, latent semantics analysis, clustering, and graph-based methods.

The second, information is repeated, is an advantage when using the latent semantics analysis, clustering method, and the graph-based methods.

Thirdly we need to consider that patent applications can contain several inventions and it is key that the summary can capture all of them and not only the one which has most written about it. For this, either the clustering method or the graph-based methods can be used. Comparing these two the graph-based methods can both identify different topics as well as links between them. The clustering method assigns sentences into clusters and cannot accommodate if a sentence may contain information about two or more topics. For this, the graph-based methods are considered a better approach.

The method that was chosen to implement was the graph-based method with TF-IDF as the frequency-based method for calculating weights of words to input in the weight matrix for the graph.

3.2 Software setup

The implementation was done in Jupyter Lab Notebook, i.e. Python, on a Lenovo Yoga C930 i7 core 8th generation. Processing of the examples and evaluation was done on the same computer. Python is used both in academia and by the industry. It is a versatile and powerful programming language and was chosen because it also contains several Natural Language Processing libraries with predefined functions to build upon.

The following packages were used in Python:

- Natural Language Toolkit (NLTK) - most widely used library for Natural Language Processing tools
- Numpy - Contains functions to be implemented when creating matrices
- Networkx - Contains functions to be implemented when creating graphs
- Pandas - Contains functions which are used when pre-processing the text
- Matplotlib - Contains plotting functions used to visualize graphs
- Time - Contains functions used for calculating the run time of the program
- Scikit-learn (Sklearn) - Contains mathematical functions such as cosine similarity which are used in the similarity calculations
- Fpdf - Contains functions to write out text as a pdf

3.3 Detailed description of the used method

The chosen summarization method will now be described in further detail. The summarizer can be divided into five steps.

1. Reading of document
2. Pre-processing of text
3. Creating a similarity matrix and graph
4. Ranking sentences
5. Extracting summary sentences

These steps are illustrated in Figure 3.1.

A simple text example will be used to illustrate the different steps. The chosen example is text taken from the Harry Potter Wikipedia-page. The full text which the program was run on can be found in Appendix A.

The first part of the Harry Potter Wikipedia example text can be seen in Figure 3.2. A few interesting words have been highlighted and will be commented later on.

3.3.1 Step 1: Reading of document

To read a document one needs a function to read files. The initial intention was for the function to be able to handle both text-files (.txt) and pdf-files (.pdf) as patent descriptions can be downloaded both as docx-files and pdf-files. Both docx-files and pdf-files contain text as well as information about formatting. For the summarizer, only the text needs to be extracted. There are tools for extracting text from both types of documents, but for pdf-files it is more difficult. Several pdf packages and functions were tested, but none worked fully, and it was therefore decided to leave this feature for future work. For text-files the Python function *read* was used. It extracts a string containing all characters in the file. *Read* requires the user to define the name of the file to be read. To handle this the user is asked to write the name of the file to be read.

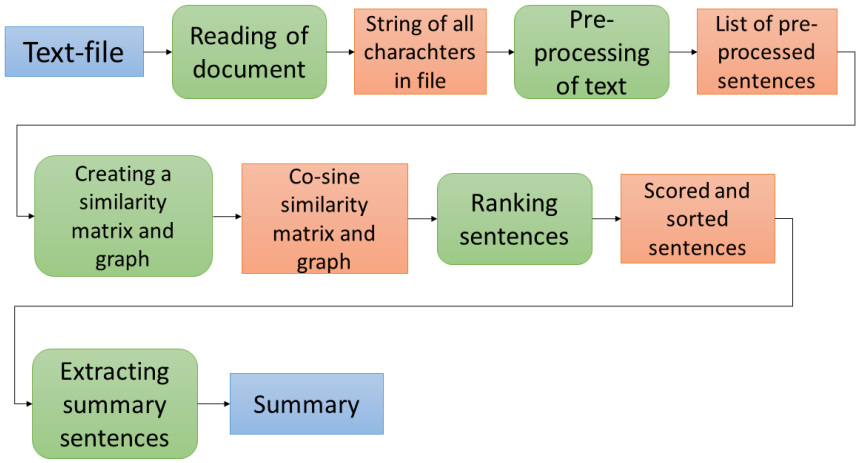


Figure 3.1: A visual overview of the summarizer's steps. The blue boxes are input and output, the green boxes are the steps in the method, and the orange boxes are the output generated after each step.

3.3.2 Step 2: Pre-processing of text

Pre-processing of text is an important step in all natural language processing methods. It is implemented to make the text analyzable. Pre-processing was done with the *Natural Language Toolkit* package (NLTK), and *Pandas* as well as some self-written steps and include:

(a) Removing dots after abbreviations so that they are not misinterpreted as the end of a sentence (own).

(b) Splitting the text into sentences (NLTK) and creating a list containing all sentences.

(c) Adding patent specific common words to NLTK's predefined list of stop words (own).

(d) Removing numbers, punctuation, and special characters (Pandas).

(e) Setting all letters to lower case (Pandas).

(f) Removing stop words (NLTK & own).

(g) Lemmatization (NLTK).

(h) Stemming using the *Porter Stemmer* (NLTK).

In step (c) the word "figure" was added to the list of stop words to be removed as it is one of the most common words in a patent description.

The output after the pre-processing was a list containing the pre-processed sentences as strings. The pre-processing steps made it possible to count the number

Harry Potter is a series of fantasy novels written by British author J K Rowling. The novels chronicle the lives of a young wizard, Harry Potter, and his friends Hermione Granger and Ron Weasley, all of whom are students at Hogwarts School of Witchcraft and Wizardry. The main story arc concerns Harry's struggle against Lord Voldemort, a dark wizard who intends to become immortal, overthrow the wizard governing body known as the Ministry of Magic and subjugate all wizards and Muggles (non-magical people). Since the release of the first novel, Harry Potter and the Philosopher's Stone, on 26 June 1997, the books have found immense popularity, critical acclaim and commercial success worldwide. They have attracted a wide adult audience as well as younger readers and are often considered cornerstones of modern young adult literature.[2] As of February 2018, the books have sold more than 500 million copies worldwide, making them the best-selling book series in history, and have been translated into eighty languages.[3] The last four books consecutively set records as the fastest-selling books in history, with the final installment selling roughly eleven million copies in the United States within twenty-four hours of its release.

Figure 3.2: First seven sentences of text used in example.

of words and sentences in the original text as well as after all pre-processing steps. The user will therefore also receive output with this information.

We will now use the example text to demonstrate the pre-processing step. First, in Figure 3.3 the same part of the text as in Figure 3.2 is presented, but here the pre-processing steps (a) through (g) have been run. We can see that the sentences are now separated with apostrophes and commas, i.e. marked as strings and in the list separated by commas. All letters are in lower case. All stop words have been removed. All numbers and special characters have been removed.

To see the effect of the lemming process (g) we need to compare it with the original text in Figure 3.2. On the highlighted words "books" (yellow) and "copies" (blue) we can see that in Figure 3.2 they are in an inflected form. In the lemmatization process, the words have been changed to their base form (the lemma of the word) which can be seen in the highlighted words in Figure 3.3.

To see the effect of the stemming process (h) the result presented in Figure 3.4 shows the text after all pre-processing steps were executed. If one compares this result with the result in Figure 3.3 and with the original text in Figure 3.2 the stemming becomes evident. As explained in Section 2.4.1 the stem to which the words are reduced to does not need to be a word as long as the words are consistently stemmed in the same way. It is clear by looking at Figure 3.4 that most of the words have been stemmed to a stem that is not a word and therefore look a bit strange. A clear example is the word "roughly" highlighted in pink. The original word stem for "roughly" is "rough". The word may also have inflections as "roughlier" and "roughliest". In the Porter Stemmer it is here evident that the chosen stem for "roughly" is "roughli", and one can assume that "roughlier" and "roughliest" would be stemmed to the same stem.

In our case, we will in the next step (Step 3: Creating a similarity matrix and

'harry potter series fantasy novel written british author j k rowling ', 'novel chronicle life young wizard harry potter friend hermione granger ron weasley student hogwarts school witchcraft wizardry ', 'main story arc concern harry struggle lord voldemort dark wizard intends become immortal overthrow wizard governing body known ministry magic subjugate wizard muggles non magical people ', 'since release first novel harry potter philosopher stone june book found immense popularity critical acclaim commercial success worldwide ', 'attracted wide adult audience well younger reader often considered cornerstone modern young adult literature ', 'february book sold million copy worldwide making best selling book series history translated eighty language ', 'last four book consecutively set record fastest selling book history final installment selling roughly eleven million copy united state within twenty four hour release '

Figure 3.3: First seven sentences of text used in example after step (a)-(g) of the pre-processing.

graph) compare all sentences to each other and find if they have words in common. As long as all sentences have been stemmed in the same way the information we need is still there even though the words look a bit strange.

'harri potter seri fantasi novel written british author j k rowl ', 'novel chronicl life young wizard harri potter friend hermion granger ron weasley student hogwart school witchcraft wizardri ', 'main stori arc concern harri struggl lord voldemort dark wizard intend becom immort overthrow wizard govern bodi known ministri magic subjug wizard muggl non magic peopl ', 'sinc releas first novel harri potter philosoph stone june book found immens popular critic acclaim commerci success worldwid ', 'attract wide adult audienc well younger reader often consid cornerston modern young adult literatur ', 'februari book sold million copi worldwid make best sell book seri histori translat eighti languag ', 'last four book consecut set record fastest sell book histori final instal sell roughli eleven million copi unit state within twenti four hour releas '

Figure 3.4: First seven sentences of text used in example after step (a)-(h) of the pre-processing.

3.3.3 Step 3: Creating a similarity matrix and graph

Similarity matrix

By using the *Scikit-learn* package, and specifically the *CountVectorizer* function, the sentences were vectorized and a matrix with each sentence as a row and a column for each word that appears in the text was created. Element $a_{i,j}$ in the matrix will therefore have a value if and only if word j is in sentence i. The weights for all words were calculated using the function *Tfidftransformer*, and the entries

in the matrix were updated to correspond to the TF-IDF weights.

Using the *cosine similarity* function also from the *Scikit-learn* package the co-sine similarity between each row in the matrix, i.e. between each sentence in the document, was calculated. The values were stored in a new matrix, the similarity matrix.

To calculate the PageRank centrality on a similarity graph requires quite a lot of computing power. A way to reduce this is to set a threshold for the lowest similarity value accepted. In this case, the threshold was set based upon the number of words the sentences at least should have in common to get a link. In the average English language sentences consist on average of 15-20 words [7]. After removing stop words sentences are reduced to around 10-15 words. With no threshold it means that as long as the sentences have at least one word in common the co-sine similarity value will be larger than zero. If one uses a threshold it means that one will sacrifice some accuracy in favor of reducing the computational complexity. If a threshold of 0.1 for the co-sine similarity value is used it means that sentences need to have two or more words in common for there to be a co-sine similarity value. For a threshold of 0.15 the sentences need to have on average three or more words in common. The higher the threshold, the more words the sentences need to have in common for there to be a value in the similarity matrix.

A test with different values on the threshold was made. When there was no threshold the PageRank calculations took a significant time (36 seconds for a file of 28 kilobytes). When using a threshold of 0.1 the calculations were considerably faster. It almost reduced the time by half (19 seconds for a file of 28 kilobytes). Comparing the results between not having a threshold and having a threshold of 0.1 showed a very small difference. In the output only one or two out of ten summary sentences differed meaning the summary consisted of 80-90% of the same sentences when not using a threshold and using a threshold of 0.1.

When increasing the threshold value further the results started to differ more. Comparing no threshold with a threshold of 0.2 the results were only overlapping with 60-70%, and using a threshold of 0.3 the results were only overlapping with 30-40%. If assumed that the absolute true result is the result by not having a threshold at all we can argue that we almost get the same (true) result by using a threshold of 0.1 for half the computing time.

As efficiency also is an important measure of an automatic summarizer a threshold value of 0.1 was set, based on the argument that it reduces the computing time while still giving an accurate result. All values below 0.1 were therefore removed from the similarity matrix.

Graph

Based on the similarity matrix a graph was created for visualization of the similarities between the sentences in the document. The graph was created using the package *networkx*. Each node in the graph represents a sentence and the link between nodes is the similarity measure the two sentences have. In Figure 3.5 the graph for the Harry Potter-text can be seen. In the graph we can see that four sentences have no links to other sentences. This evidently means that they do not have two or more words in common with any other sentence in the whole text.

For the interested reader, these are sentences number 4, 65, 84, and 106.

In Figure 3.6 the right corner of the graph in Figure 3.5 has been zoomed in for a clearer view of the graph.

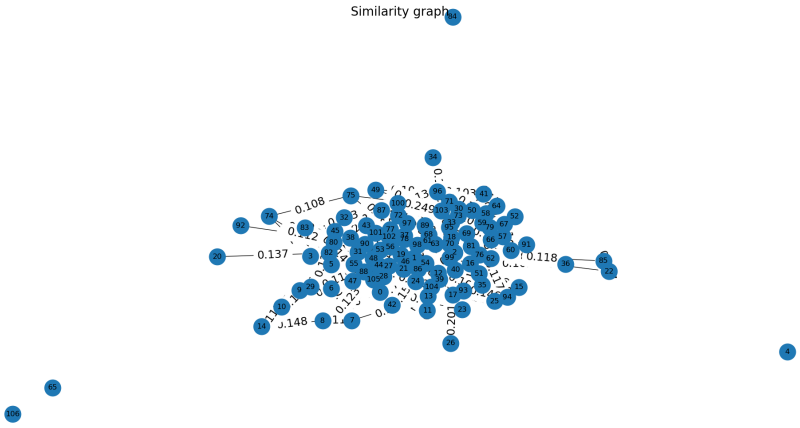


Figure 3.5: Similarity graph for the Harry Potter-text, where node numbers is the sentence number in the document and the link values represent the cosine similarity between the sentences.

The graph is also a visual aid in identifying clusters in the text. For example, if the text covers many topics these will show in the graph. In the Harry Potter-example there were no clear clusters to see, so to illustrate this a new example was made. A new document was created containing the Harry Potter-text found in Appendix A amended with the same amount of text this time taken from the Lord of the Rings-Wikipedia page. The idea with these two texts is that although they are different topics both cover fantasy novels and movies, meaning they should have some things in common even if they are different subjects. In Figure 3.7 the graph for this new document is presented. As expected we can perceive two clusters roughly marked with a yellow and a purple circle in the figure.

3.3.4 Step 4: Ranking sentences

Using the *PageRank* function in *networkx* the PageRank scores were calculated for all sentences in the Harry Potter-text. The sentences were then sorted into a list based on their score.

To visualize the PageRank scores a new graph was drawn. In this graph the nodes (sentences) were colored based on their PageRank score. The warmer the color (orange, yellow) the higher the score, and the colder the color (purple, blue) the lower the score. This resulted in a color map of the PageRank score which can be seen in Figure 3.8.

To further illustrate the two clusters we could see in Figure 3.7 a color map was drawn for the Harry Potter and Lord of the Rings-text as well. It can be seen in Figure 3.9. In the figure, one can see two centers (the orange nodes) with higher PageRank scores in the middle of each of the previously marked clusters. This

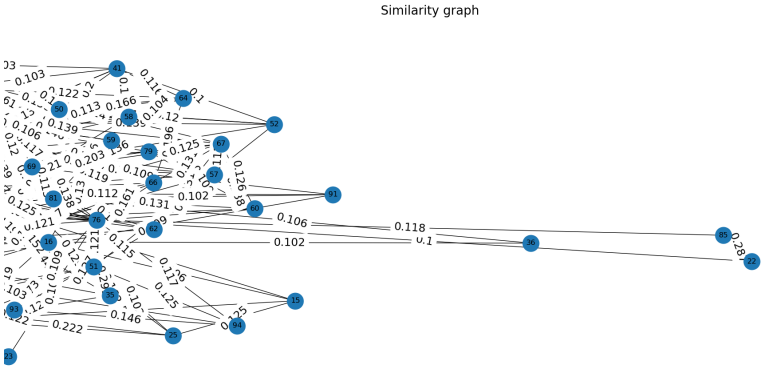


Figure 3.6: Zoomed in similarity graph for the Harry Potter-text, where node numbers is the sentence number in the document and the link values represent the cosine similarity between the sentences.

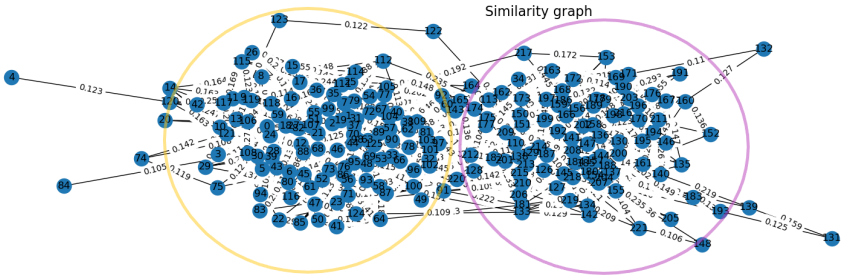


Figure 3.7: Graph for second example with Harry Potter and Lord of the Rings-text. The two clusters, marked with a yellow and a purple circle, corresponds to the two different texts.

indicates that there are two separate sets of important sentences corresponding to the two texts respectively.

3.3.5 Step 5: Extracting summary sentences

Texts can be of very varying lengths. This was taken into account when deciding the length of the summary, i.e. how many sentences that should be extracted. It was decided that the summary length should correspond to 10% of the original text. From Step 2: Pre-processing of text we know the number of sentences in the document and the number of sentences to be extracted (n) was calculated by $n = 0.1 \times total\ number\ of\ sentences\ in\ document$and rounding up if the result was not an integer.

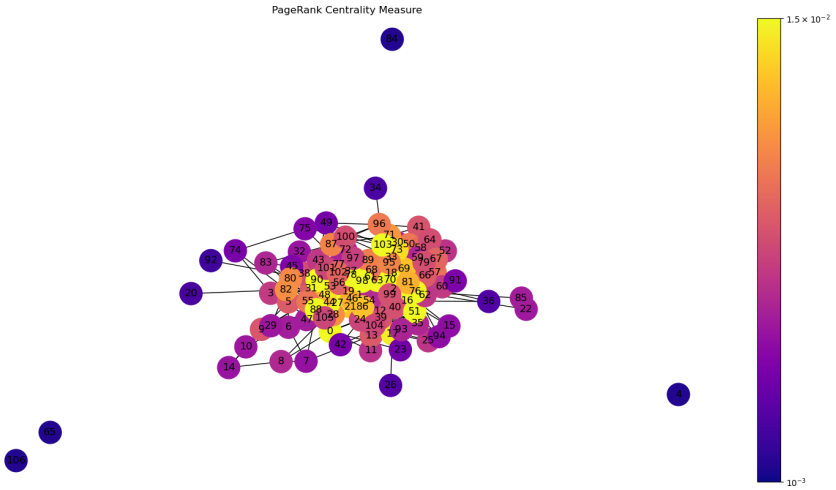


Figure 3.8: Example of drawn graph with PageRank color map. The warmer the color the higher PageRank score the sentence has. The graph is from the Harry Potter example.

The top n scored sentences were identified and extracted from the sentence list created in Step 2 which handles the pre-processing of text.

In patent description documents the first sentence is the headline of the patent document, i.e. the patent number as well as the name of the patent document. As this is good information to keep when doing a summary it was added to the extracted sentences.

Creating the summary

For the summary, the extracted sentences were ordered so that they would appear in the same order they appear in the original patent description document. Using the *fpdf* package a pdf-file was created where the summary was written. The headline, i.e. the first sentence, was marked in bold to separate it from the other extracted sentences.

To make it easier for the user to identify the extracted sentences in the graph-structure a new graph was made where the extracted sentences' corresponding nodes are marked in green. The new graph for the Harry Potter-example can be seen in Figure 3.10.

For our Harry Potter-example the resulting summary is presented in Figure 3.11.

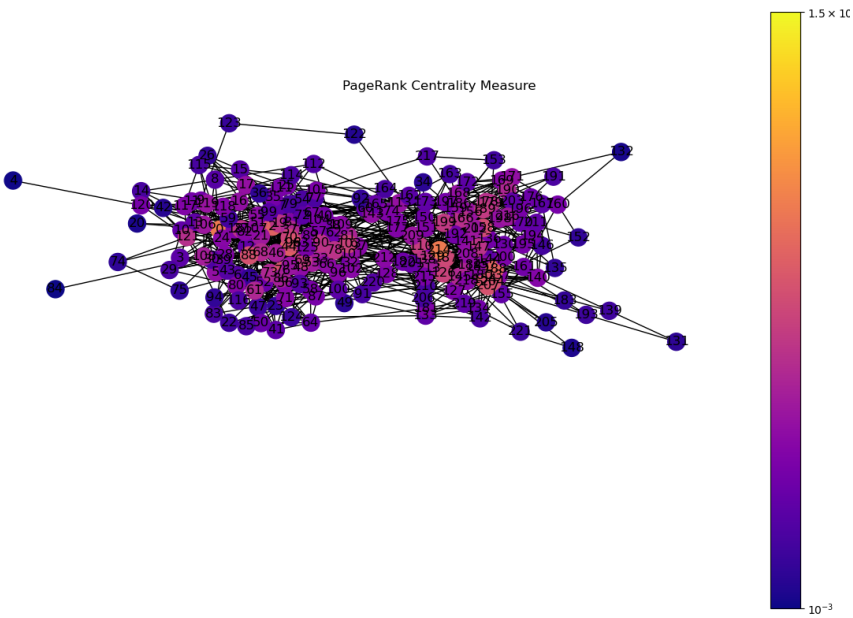


Figure 3.9: Example of drawn graph with PageRank color map. The warmer the color the higher PageRank score the sentence has. The graph is from the Harry Potter and Lord of the Rings example.

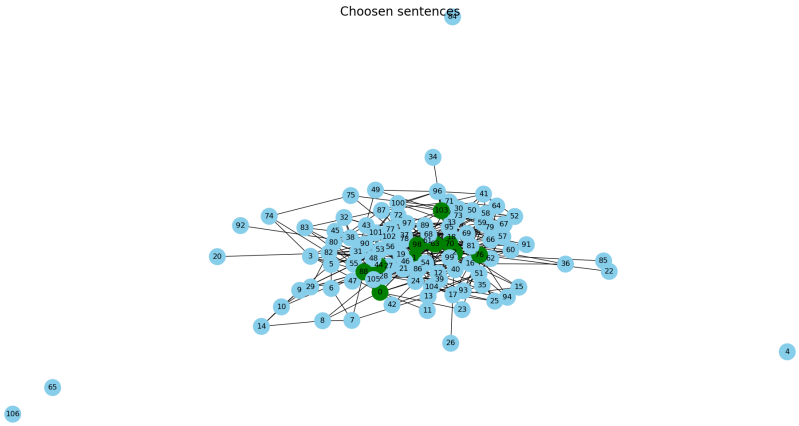
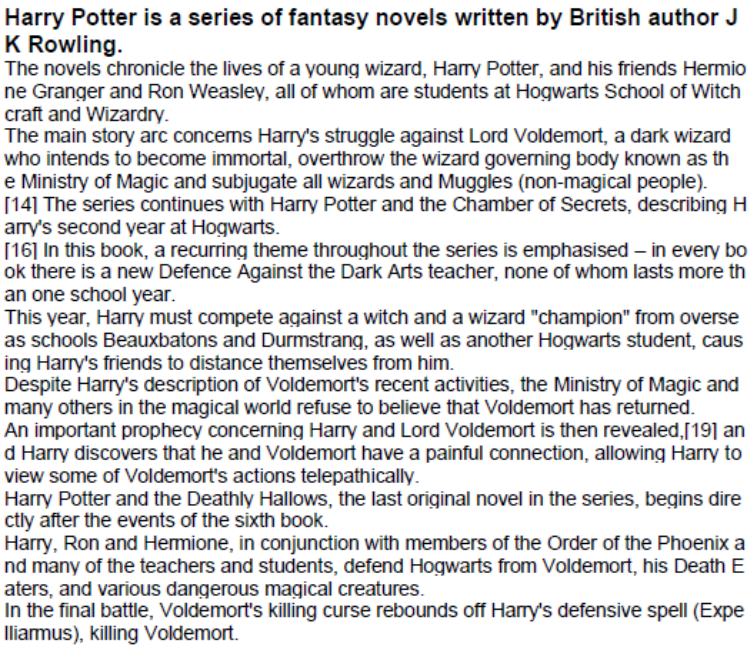


Figure 3.10: Example of drawn graph with extracted sentences marked in green. The graph is from the Harry Potter example.



Harry Potter is a series of fantasy novels written by British author J K Rowling.
The novels chronicle the lives of a young wizard, Harry Potter, and his friends Hermione Granger and Ron Weasley, all of whom are students at Hogwarts School of Witchcraft and Wizardry.
The main story arc concerns Harry's struggle against Lord Voldemort, a dark wizard who intends to become immortal, overthrow the wizard governing body known as the Ministry of Magic and subjugate all wizards and Muggles (non-magical people).
[14] The series continues with Harry Potter and the Chamber of Secrets, describing Harry's second year at Hogwarts.
[16] In this book, a recurring theme throughout the series is emphasised – in every book there is a new Defence Against the Dark Arts teacher, none of whom lasts more than one school year.
This year, Harry must compete against a witch and a wizard "champion" from overseas schools Beauxbatons and Durmstrang, as well as another Hogwarts student, causing Harry's friends to distance themselves from him.
Despite Harry's description of Voldemort's recent activities, the Ministry of Magic and many others in the magical world refuse to believe that Voldemort has returned.
An important prophecy concerning Harry and Lord Voldemort is then revealed,[19] and Harry discovers that he and Voldemort have a painful connection, allowing Harry to view some of Voldemort's actions telepathically.
Harry Potter and the Deathly Hallows, the last original novel in the series, begins directly after the events of the sixth book.
Harry, Ron and Hermione, in conjunction with members of the Order of the Phoenix and many of the teachers and students, defend Hogwarts from Voldemort, his Death Eaters, and various dangerous magical creatures.
In the final battle, Voldemort's killing curse rebounds off Harry's defensive spell (Expelliarmus), killing Voldemort.

Figure 3.11: The resulting pdf for the Harry Potter example.

3.4 Evaluation procedure

Reference summaries are generally not available for patent documents, and to generate human written evaluations for patent documents takes a lot of time. Therefore, the extrinsic evaluation method was chosen. It was based on humans evaluating the summaries by answering questions in a form produced for this task. The form consisted of questions about the respondent's view of summaries in general (importance of length, grammar, structure, and coherence) as well as how much experience they have working with patent documents. The respondent was then asked to read the summary and determine what the patent description is about by answering a multiple-choice question. The choice "None of the above/Can't tell" was provided if one could not choose one of the subjects given. The respondent was also asked to score the summary based on length, grammar, structure, and coherence. The last step of the evaluation was allowing the respondent to read the full description of the patent description and then answer the exact same multiple choice-question about what the patent description is about. The respondent was also asked to score how well the summary had captured the information from the full description. The evaluation form for one summary is attached in Appendix B.

3.4.1 Chosen patent descriptions

Ten patent descriptions were set to be a reasonable amount since the evaluation was to be done manually. These patent descriptions were chosen from five different Swedish companies (two from each company); Axis, Ericsson, Husqvarna, Tetra Pak, and Volvo, operating in different fields. For each patent description a summary was generated from the description using the summarizer detailed above. All patent descriptions were downloaded from Espacenet and links to them can be found in Appendix D.

Here is a list of the patents from which the patent descriptions chosen to summarize was retrieved:

- EP0205073A2 An opening arrangement for packing containers (Tetra Pak)
- EP3035664A1 Method for processing a video stream (Axis)
- EP3439458A1 Multi-purpose can (Husqvarna)
- EP3627321A1 Mobile terminal with middleware security access manager (Ericsson)
- EP3696022A1 Vehicle interior lightning system (Volvo)
- US6216772B1 Device for filtering and cooling (Volvo)
- US2016094765A1 Method and image processing device for image stabilization of a video stream (Axis)
- WO9200640A1 A hands-free module (Ericsson)
- WO2020098963A1 Cutting tool (Husqvarna)
- WO2020120499A1 A Centrifugal separator (Tetra Pak)

3.4.2 Evaluating groups

Ten employees at AWA were asked to evaluate three summaries each, making sure the summaries were distributed between them so that each summary was evaluated three times by a patent attorney or similar.

Some people with engineering backgrounds and/or some knowledge of patent documents were asked to evaluate one to three summaries each.

Employees at LU Innovation, Lund University, were also asked to evaluate one to three summaries each.

3.4.3 Score of the summary

The choice of asking the same multiple-choice question regarding the subject of the patent description for both the summary and the full description was for the purpose of calculating a score for the summary. For all of those who could determine the patent description's subject based on the full description, the summary was given a score if they could also determine the subject of the patent description solely based on the summary. If the subject could not be determined the summary did not receive a score.

Chapter 4
Results

The ten generated summaries are included in Appendix C. The summaries were sent to patent attorneys at AWA, employees at LU Innovation, and people with engineering backgrounds for evaluation through a predefined evaluation form (see Appendix B).

30 evaluators answered the evaluation forms for the generated summaries. 21 answers from patent attorneys at AWA, 7 answers from people with engineering backgrounds, and 2 answers from employees at LU Innovation, Lund University. This resulted in each summary being evaluated 2-4 times. The results from these evaluations are presented in this chapter.

4.1 Run time

The run time for the implemented summarizer is dependent on the size of the file one wants to summarize. For each kilobyte in file size the summarizer takes on average one second. The chosen patent descriptions varied between 10-40 kilobyte, hence, the summarizer took 10-40 seconds to run depending on the patent description. This run time is important since the objective was to create a summarizer which was not too computationally complex and could be used in daily work with reasonable waiting time. These run times are the result of running the summarizer on a commonly used laptop. If one wants to reduce the run time one might consider running the summarizer on a more powerful computer or dedicated hardware.

4.2 General opinions of summaries

The first question (see Appendix B) in the evaluation form concerned the preferred maximum length of a summary. The evaluators' answers are depicted in Figure 4.1. A total maximum length of around one page is the consolidated view, although evaluators within the patent profession (patent attorneys at AWA) seem to have acceptance for something slightly longer.

The second question in the evaluation form (see Appendix B) related to how important the overall aspects of the content of a summary are, or more specifically how important grammar, structure, coherence, and length are. The result is presented in Figure 4.2, Figure 4.3, Figure 4.4, and Figure 4.5. The spread of

opinions is large, but one can see that coherence stands out as an aspect of high importance.

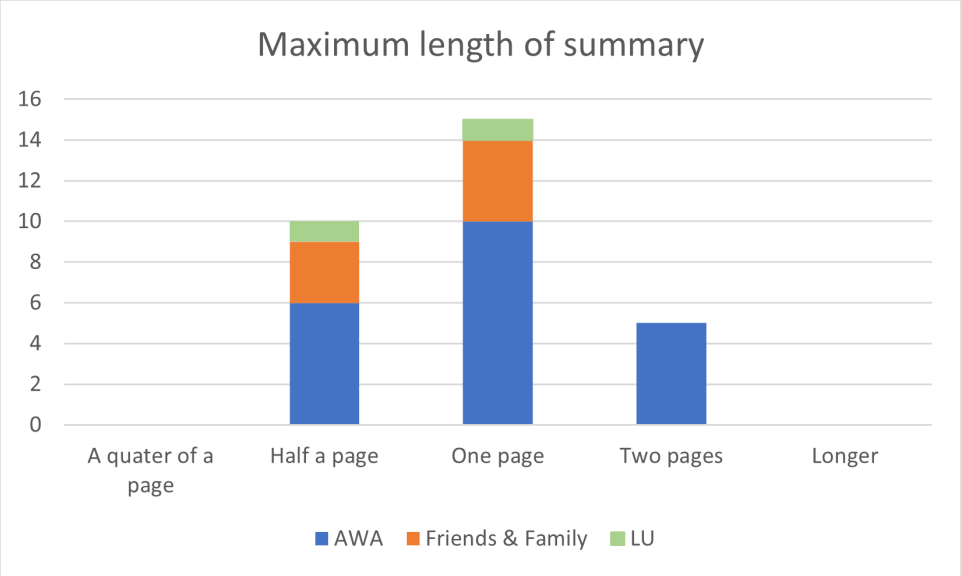


Figure 4.1: Result from Question 1 i evaluation form.

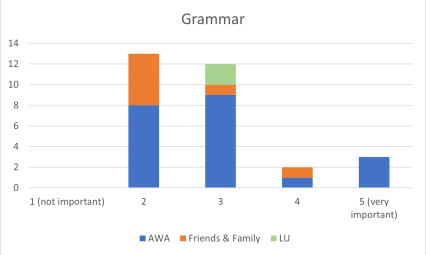


Figure 4.2: Importance of grammar in a summary.

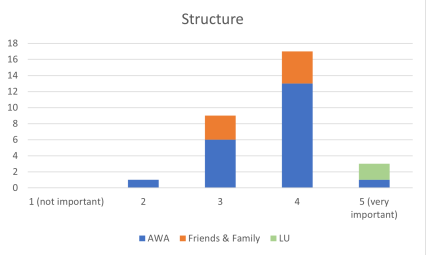


Figure 4.3: Importance of structure in a summary.

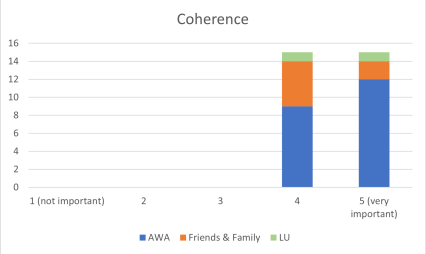


Figure 4.4: Importance of coherence in a summary.

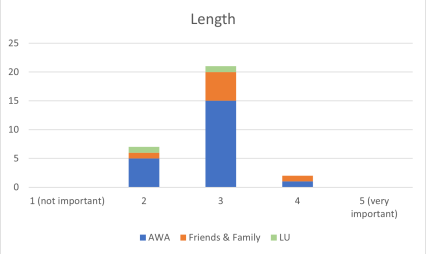


Figure 4.5: Importance of length of a summary.

4.3 Summary results

The score of each summary is calculated as explained in Section 3.4.3 where a summary receives a point if the evaluator can determine the subject of the patent description solely based on the summary conditioned on the evaluator also determining the correct subject based on the full patent description. The score of each summary is presented in Table 4.1. Six summaries received a full score (WO9200640A1, EP3439458A1, EP3035664A1, US2016094765A1, EP0205073A2, WO2020120499A1). Two summaries received scores from some evaluators (EP3696022A1, US6216772B1). One summary received a score of zero (WO2020098963A1). One summary (EP3627321A1) could not be scored as none of the evaluators could determine the subject based on the full patent description. Further discussion of the results can be found in Chapter 5.

	Score of the summary	Could determine the subject based on the full description	Number of evaluators
EP3627321A1	0	0	3
WO9200640A1	3	3	3
WO2020098963A1	0	2	2
EP3439458A1	2	2	2
EP3696022A1	1	3	3
US6216772B1	2	3	3
EP3035664A1	3	3	3
US2016094765A1	3	3	3
EP0205073A2	4	4	4
WO2020120499A1	3	3	4

Table 4.1: Score for all summaries calculated as explained in Section 3.4.3, the number of evaluators who could correctly determine the subject based on the full description, and the total number of evaluators for each summary.

The results from Question 5 and 7 from the evaluation form are presented for each summary in Tables 4.2-4.11. Question 5 asked the evaluators to score the aspects of grammar, structure, coherence, and length of the generated summary from 1-5 (where 1 is the lowest and 5 the highest). Question 7 asked the evaluators to score how well the summary captured the information from the full patent description, and the alternatives were Extremely well, Somewhat well, Neutral, Somewhat not well, Extremely not well. The results will be discussed in Chapter 5.

EP3627321A1

Table 4.2 presents the result from Question 5 and 7 for the generated summary of patent EP3627321A1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Patent attorney)	4	3	2	4	Somewhat well
Evaluator 2 (Patent attorney)	5	3	3	5	Neutral
Evaluator 3 (Patent attorney)	3	2	3	1	Neutral

Table 4.2: Answers from the evaluation of the generated summary of patent EP3627321A1.

WO9200640A1

Table 4.3 presents the result from Question 5 and 7 for the generated summary of patent WO99200640A1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Patent attorney)	3	2	2	3	Neutral
Evaluator 2 (Know a bit about patents)	4	3	2	5	Somewhat well
Evaluator 3 (Patent attorney)	3	3	3	2	Somewhat well

Table 4.3: Answers from the evaluation of the generated summary of patent WO9200640A1.

WO2020098963A1

Table 4.4 presents the result from Question 5 and 7 for the generated summary of patent WO2020098963A1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Patent attorney)	4	4	2	4	Somewhat well
Evaluator 2 (Have worked with patents)	4	3	3	4	Neutral

Table 4.4: Answers from the evaluation of the generated summary of patent WO2020098963A1.

EP3439458A1

Table 4.5 presents the result from Question 5 and 7 for the generated summary of patent EP3439458A1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Know a bit about patents)	3	2	2	4	Somewhat not well
Evaluator 2 (Patent attorney)	5	3	4	5	Somewhat well

Table 4.5: Answers from the evaluation of the generated summary of patent EP3439458A1.

EP3696022A1

Table 4.6 presents the result from Question 5 and 7 for the generated summary of patent EP3696022A1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Patent attorney)	4	4	3	5	Neutral
Evaluator 2 (Know very little about patents)	3	2	3	3	Somewhat not well
Evaluator 3 (Patent attorney)	4	3	3	5	Somewhat well

Table 4.6: Answers from the evaluation of the generated summary of patent EP3696022A1.

US6216772B1

Table 4.7presents the result from Question 5 and 7 for the generated summary of patent US6216772B1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Patent attorney)	3	3	3	4	Somewhat well
Evaluator 2 (Patent attorney)	4	3	3	5	Neutral
Evaluator 3 (Have worked with patents)	3	4	4	4	Extremely well

Table 4.7: Answers from the evaluation of the generated summary of patent US6216772B1.

EP3035664A1

Table 4.8 presents the result from Question 5 and 7 for the generated summary of patent EP3035664A1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Patent attorney)	5	4	4	3	Somewhat well
Evaluator 2 (Patent attorney)	4	4	5	4	Somewhat well
Evaluator 3 (Have worked with patents)	4	4	5	3	Extremely well

Table 4.8: Answers from the evaluation of the generated summary of patent EP3035664A1.

US2016094765A1

Table 4.9 presents the result from Question 5 and 7 for the generated summary of patent US2016094765A1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Patent attorney)	4	3	5	2	Somewhat well
Evaluator 2 (Patent attorney)	5	4	4	2	Somewhat well
Evaluator 3 (Patent attorney)	3	3	4	2	Somewhat well

Table 4.9: Answers from the evaluation of the generated summary of patent US2016094765A1.

EP0205073A2

Table 4.10 presents the result from Question 5 and 7 for the generated summary of patent EP0205073A2.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Know very little about patents)	3	2	2	4	Somewhat well
Evaluator 2 (Patent attorney)	4	3	3	2	Somewhat not well
Evaluator 3 (Patent attorney)	5	4	4	4	Somewhat well
Evaluator 3 (Patent attorney)	4	4	5	5	Somewhat well

Table 4.10: Answers from the evaluation of the generated summary of patent EP0205073A2.

WO2020120499A1

Table 4.11 presents the result from Question 5 and 7 for the generated summary of patent WO2020120499A1.

	Grammar (1-5)	Structure (1-5)	Coherence (1-5)	Length (1-5)	Capture information
Evaluator 1 (Have worked with patents)	2	3	2	4	Neutral
Evaluator 2 (know a bit about patents)	2	4	4	2	Somewhat well
Evaluator 3 (Patent attorney)	4	3	4	5	Somewhat not well
Evaluator 3 (Patent attorney)	4	3	3	2	Somewhat well

Table 4.11: Answers from the evaluation of the generated summary of patent WO2020120499A1.

Chapter 5
Discussion

5.1 Generating summaries

The generated summaries (see Appendix C) demonstrate that the implemented program could do what it was designed and intended for. It produced summaries of each patent description, with the correct length, fully extracted sentences, and reasonable run times. This is expected as a consequence of the implemented algorithms. More important aspects are the quality of the generated summaries and how well they manage to convey information about the patent's subject.

5.2 Convey information

The purpose of the summaries is to convey information about the patent document's subject. The reader should after reading the summary be able to tell the subject of the full patent document in order to decide whether the patent is relevant and should be read in full or can be discarded. The majority of the summaries were successful in this task, some were unsuccessful, and one failed due to other reasons.

5.2.1 Successful summaries

A successful result is if the evaluator was able to understand the patent description and chose the correct patent subject solely based on the summary. This is the case for the majority of the summaries (WO9200640A1, EP3439458A1, EP3035664A1, US2016094765A1, EP0205073A2, and WO2020120499A1) where all evaluators could determine the subject based on the summary giving these summaries a perfect score. For the summary of patent US6216772B1 the majority of the evaluators (two out of three) could determine the subject from the summary and it does therefore also count as a successful summary.

5.2.2 Unsuccessful summaries

An unsuccessful summary is the case where the evaluator could determine the subject based on the full description but not based on the summary, i.e. information is missing in the summary. This was the case for two summaries. The summary for

EP3696022A1 where only one out of three evaluators could determine the subject based on the summary, and the summary for WO2020098963A1 where none of the evaluators could determine the subject based on the summary.

A closer look into these two summaries shows some shortages with the method, i.e. where there is room for improvement.

WO2020098963A1 is a patent about a cutting tool such as a hedge clipper. The words hedge and clipper(s) are mentioned 9 times each in the patent description but zero times in the summary. The word cutting is however mentioned 99 times in the patent description and 12 times in the summary. This example clearly shows that the method favors sentences with frequently used words and that the extracted sentences are very likely to be similar due to this. With the current summary the reader will understand that it is about a cutting tool, but not that this cutting tool is a hedge clipper. To improve the method a feature can be added to make sure that the extracted sentences are not too similar, making room for information that still appears quite frequently, but not as frequent as the highest scored.

EP3696022A1 is about a vehicle interior lightning system for motorized vehicles. Only once in the full description is it mentioned that this is specifically for motorized vehicles. All other text is mentioning vehicles in general. This is a common practice when writing patents. The patent will reach more coverage the more general it can be formulated, which is something the inventor aims for. In this case the mentioning of motorized vehicles appears in the section ”Technical field”. In WO2020098963A1 information that the cutting tool was a hedge clipper could also be located twice in the section ”Technical field”. Another added feature that could improve the summarization method is therefore to add the whole section ”Technical field” to the summary. The section is usually just a couple of sentences long and would therefore not increase the summary length significantly.

5.2.3 Other

For EP3627321A1 none of the evaluators could determine the subject from neither the summary nor the full description.

The choice to only score the summary if the evaluator also could determine the correct subject based on the full description was based on the limitations with extraction-based summarization. There are two possible reasons why the reader cannot determine the subject based on the full description.

- The full description is written in a way where the subject is not clear.
- The reader has either not understood or misunderstood the subject.

As the generated summaries are extracted sentences from the full description the summary will retain features from the full description. If the full description cannot clearly explain the subject then the summary, which essentially is a shorter version of the description, will probably not either. Subsequently, if one misunderstands the subject having all facts from the full description, it is likely that one will do so based on the summary as well.

5.3 Quality of the generated summaries

The choice of extraction-based summarization sets boundaries on the quality of the generated summaries. The quality of the summary reflects the quality of the input text. If something, for example the grammar, language, or sentences, is bad in the input text they will be bad in the summary as well since extraction-based summarization extracts text without modifying it. The summary can also be perceived as incoherent due to the fact that the sentences which in the input text ties the text together are not among the extracted sentences. Despite the summary being perceived as incoherent, the subject may still be clear in the summary so the incoherence may not affect the ability to understand the subject based on the summary.

The results for each patent vary a lot, clearly demonstrating the evaluator's personal opinion of how important some features are in a summary. We can, however, compare the results for the summaries with the general opinions of summaries derived from Questions 1 and 2 in the evaluation (see Figures 4.1-4.5).

5.3.1 Grammar

Based on the general opinion, grammar is of low or medium importance in a summary. For the generated summaries grammar received mixed scores, but generally favorable (3-5 out of 5), and the conclusion is that the grammar was adequately good. This is, however, not the merit of the summarization technique, but rather due to the fact that the grammar in the full descriptions is good.

5.3.2 Structure

Based on the general opinion, structure is an important feature for a summary. The structure on the generated summaries received scores from 2-4 (out of 5) but mostly 3s and 4s. The added features to improve the structure in the method was to include the title of the patent in bold text, and to make sure the extracted sentences were sorted in the order they appear in the full description. The acceptable score tells us that these were good additions to the structure of the summary.

5.3.3 Coherence

Based on the general opinion, coherence is a very important feature for a summary. Coherence is the measure of the quality of being logical and consistent. Coherence in the generated summaries has received very mixed scores from 2-5 (out of 5), where some summaries have high scores (4 or 5) and some quite low (2 or 3).

As the summaries are extracted sentences combined to a text the coherence is dependent on whether the extracted sentences feel logical to combine in the order they are extracted and whether the text is readable and has flow. When studying the summaries it is clear that the logical flow in the summaries is closely related to the coherence-result.

5.3.4 Length

Based on the general opinion, length is of medium importance in a summary. From the general questions we know that a majority prefers summaries of 1/2-1 page long.

Looking at the results the longer summaries have received a lower score on length than the shorter ones, concluding that a summary should rather be on the shorter side around a half-page, up to one page, rather than longer.

5.4 Future work

This thesis shows that the chosen summarization method is successful in conveying information from a patent description to an extracted summary. It does, however, have room for improvement as we can see in the patent about the hedge clipper (WO2020098963A) and the patent about interior lightning system for motorized vehicles (EP3696022A).

Important information in a patent can be found in the structural parts such as the title, subheadings, and full passages of text such as ”Techincal field”. An interesting extension to the method would be to combine the extraction-based method tested in this thesis with adding more structural parts from the patent description. These structural parts could be subheadings to understand from which passages the extracted sentences are extracted and/or direct incorporation of the Technical field passage into the summary. This would be a way to improve both the structure of the summary as well as the ability to convey information.

The current method has a drawback in favoring sentences with frequently occurring words so that the summary is dominated by sentences containing those words. In this way the summary can be perceived as repetitive and other important information can be missed. A clear example is the patent about the hedge clipper (WO2020098963A). The method can be improved by adding a feature to limit the number of similar sentences or set a boundary for how similar two sentences in the summary are allowed to be. If two sentences are too similar another high ranked sentence should be extracted instead. This way more information can be extracted and the repetitive pattern will be reduced.

All patent attorneys skim through patents in different ways to assess if they are relevant or not. One way is to look for certain words that can point to the subject of the patent. These words are for example ”relates”, ”invention”, ”advantage”, and ”specifically”. If one can add semantic analysis to the summarization method to identify sentences containing these words and score them higher then the method will work closer to how a patent attorney looks for information in a patent description thus improving its output.

Other things many patent attorneys look at are the first part of the claims and the figures that are included. It would therefore be interesting to know if patent attorneys would like these parts to be included in the summary as well.

The frequency-driven and graph-based methods are limited by a structural property of the English language. The methods do a word for word count or

comparison not taking into account that in English there are many nouns, called compounds nouns, that consists of two separate words. Examples of this are bus stop, washing machine, and water tank. In the current method these are treated as two separate words even though they together form a noun with a different meaning than the two words separately. As the current method also norms the similarity measure between two sentences based on the total number of words in the sentences it also affects the method if the compound noun should be counted as one or two words. To improve in this area one needs domain knowledge of the patents one wants to summarize. For each patent classification area one can create a list of compound nouns relevant to the area and when summarizing add to the method that these combinations of words should be counted as one.

Lastly, it is practice to include an abstract for patents. It would be very interesting if the summarizer created here could be used to create the abstract for the patent attorney.

Chapter 6
Conclusion

To, finally, summarize this thesis about Automatic Text Summarization I conclude that the chosen method can successfully be used on patent descriptions creating good summaries to be used by patent attorneys in their daily work. The method is very promising and could with some improvements work perfectly.

The results show that a majority (7) of the summaries were able to convey information about the patent's subject successfully. The three which did not succeed gave clear leads of the drawbacks of the method and pointed out areas of improvement.

To further improve the method, a number of suggestions for improvements and future work were suggested. The main suggestions relate to capturing more information from the patent description and preventing repetitive sentences in the summary to make room for sentences with additional information. In order to make the method resemble the way patent attorneys work today, a suggestion to implement semantic analysis searching for words a patent attorney would look for in a patent description, was also suggested.

References

[1] C. Aggarwal and C. e. Zhai. *Mining Text Data, A survey of text summarization techniques*. Springer, Boston, Massachusetts, 2012. Chapter authors: A. Nenkova and K. McKeown.

[2] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. Trippe, J. B. Gutierrez, and K. Kochut. Text summarization techniques: A brief survey. *International Journal of Advanced Computer Science and Applications*, 8(10), Jul 2017. http://dx.doi.org/10.14569/IJACSA.2017.081052.

[3] A. Aries, D. eddine Zegour, and W. K. Hidouci. Automatic text summarization: What has been done and what has to be done, 2019.

[4] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *International Journal of Computer Applications*, 2016.

[5] G. Como and F. Fagnani. Lecture notes on network dynamics. May 2019. Lecture notes recieved as course material in Network DynamicsFRTN30, Lund University.

[6] D. Das and A. F. Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics*, 3(3):1–12, 2007.

[7] T. Earnsy Liu.

[8] M. J. Garbade. A quick introduction to text summarization in machine learning. *Medium*, Sept 2018.

[9] E. Lloret and M. Palomar. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37:1–41, 2012.

[10] N. Munot and S. S. Govilkar. Comparative study of text summarization methods. *International Journal of Computer Applications*, Sept 2014.

[11] N. Nazari and M. A. Mahdavi. A survey on automatic text summarization. *Journal of AI and Data Mining*, 7(1):121–125, 2019.

[12] M. Nekic. Automatic text summarization. https://youtu.be/_d0OXm0dRZ4, Jun 2019. Lecture from NDC Conferences 17-21 June 2019, accessed on 2020-05-12.

[13] P. och Registreringsverket. Utformning och innehåll. https://www.prv.se/sv/patent/lagar-och-regler/riktlinjer-for-patentarenden/del-b---nationell-patentansokans-innehall/b1-utformning-och-innehall/. Accessed on 2020-09-23.

[14] A. J. C. Trappey and C. V. Trappey. An R&D knowledge management method for patent document summarization. *Industrial Management & Data Systems*, Oct 2007.

[15] A. J. C. Trappey, C. V. Trappey, and C.-Y. Wu. Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering*, Mar 2009.

[16] Wikipedia. Automatic summarization. https://en.wikipedia.org/wiki/Automatic_summarization, Apr 2020. Accessed on 2020-05-04.

[17] Wikipedia. Legal technology. https://en.wikipedia.org/wiki/Legal_technology, Oct 2020. Accessed on 2020-11-03.

[18] Wikipedia. Lemmatisation. https://en.wikipedia.org/wiki/Lemmatisation, Oct 2020. Accessed on 2020-11-12.

[19] Wikipedia. Stemming. https://en.wikipedia.org/wiki/Stemming, May 2020. Accessed on 2020-09-16.

Appendix A

Text from Harry Potter Wikipedia-page used as example in chapter 3.3

Harry Potter is a series of fantasy novels written by British author J K Rowling. The novels chronicle the lives of a young wizard, Harry Potter, and his friends Hermione Granger and Ron Weasley, all of whom are students at Hogwarts School of Witchcraft and Wizardry. The main story arc concerns Harry's struggle against Lord Voldemort, a dark wizard who intends to become immortal, overthrow the wizard governing body known as the Ministry of Magic and subjugate all wizards and Muggles (non-magical people). Since the release of the first novel, Harry Potter and the Philosopher's Stone, on 26 June 1997, the books have found immense popularity, critical acclaim and commercial success worldwide. They have attracted a wide adult audience as well as younger readers and are often considered cornerstones of modern young adult literature.[2] As of February 2018, the books have sold more than 500 million copies worldwide, making them the best-selling book series in history, and have been translated into eighty languages.[3] The last four books consecutively set records as the fastest-selling books in history, with the final installment selling roughly eleven million copies in the United States within twenty-four hours of its release. The series was originally published in English by two major publishers, Bloomsbury in the United Kingdom and Scholastic Press in the United States. A play, Harry Potter and the Cursed Child, based on a story co-written by Rowling, premiered in London on 30 July 2016 at the Palace Theatre, and its script was published by Little, Brown. The original seven books were adapted into an eight-part namesake film series by Warner Bros. Pictures, which is the third highest-grossing film series of all time as of February 2020. In 2016, the total value of the Harry Potter franchise was estimated at $25 billion,[4] making Harry Potter one of the highest-grossing media franchises of all time. A series of many genres, including fantasy, drama, coming of age, and the British school story (which includes elements of mystery, thriller, adventure, horror, and romance), the world of Harry Potter explores numerous themes and includes many cultural meanings and references.[5] According to Rowling, the main theme is death.[6] Other major themes in the series include prejudice, corruption, and madness.[7] The success of the books and films has allowed the Harry Potter franchise to expand with numerous derivative works, a travelling exhibition that premiered in Chicago in 2009, a studio tour in London that opened in 2012, a digital platform on which J.K.

Rowling updates the series with new information and insight, and a pentalogy of spin-off films premiering in November 2016 with Fantastic Beasts and Where to Find Them, among many other developments. Most recently, themed attractions, collectively known as The Wizarding World of Harry Potter, have been built at several Universal Parks Resorts amusement parks around the world. The central character in the series is Harry Potter, a boy who lives in the fictional town of Little Whinging, Surrey with his aunt, uncle, and cousin – the Dursleys – and discovers at the age of eleven that he is a wizard, though he lives in the ordinary world of non-magical people known as Muggles.[8] The wizarding world exists parallel to the Muggle world, albeit hidden and in secrecy. His magical ability is inborn, and children with such abilities are invited to attend exclusive magic schools that teach the necessary skills to succeed in the wizarding world.[9] Harry becomes a student at Hogwarts School of Witchcraft and Wizardry, a wizarding academy in Scotland, and it is here where most of the events in the series take place. As Harry develops through his adolescence, he learns to overcome the problems that face him: magical, social, and emotional, including ordinary teenage challenges such as friendships, infatuation, romantic relationships, schoolwork and exams, anxiety, depression, stress, and the greater test of preparing himself for the confrontation that lies ahead in wizarding Britain's increasingly-violent second wizarding war.[10] Each novel chronicles one year in Harry's life[11] during the period from 1991 to 1998.[12] The books also contain many flashbacks, which are frequently experienced by Harry viewing the memories of other characters in a device called a Pensieve. The environment Rowling created is intimately connected to reality. The British magical community of the Harry Potter books is inspired by 1990s British culture, European folklore, classical mythology and alchemy, incorporating objects and wildlife such as magic wands, magic plants, potions, spells, flying broomsticks, centaurs and other magical creatures, and the Philosopher's Stone, beside others invented by Rowling. While the fantasy land of Narnia is an alternate universe and the Lord of the Rings' Middle-earth a mythic past, the wizarding world of Harry Potter exists parallel to the real world and contains magical versions of the ordinary elements of everyday life, with the action mostly set in Scotland (Hogwarts), the West Country, Devon, London, and Surrey in southeast England.[13] The world only accessible to wizards and magical beings comprises a fragmented collection of overlooked hidden streets, ancient pubs, lonely country manors, and secluded castles invisible to the Muggle population.[9] When the first novel of the series, Harry Potter and the Philosopher's Stone, opens, it is apparent that some significant event has taken place in the wizarding world – an event so very remarkable that even Muggles (non-magical people) notice signs of it. The full background to this event and Harry Potter's past is revealed gradually throughout the series. After the introductory chapter, the book leaps forward to a time shortly before Harry Potter's eleventh birthday, and it is at this point that his magical background begins to be revealed. Despite Harry's aunt and uncle's desperate prevention of Harry learning about his abilities,[14] their efforts are in vain. Harry meets a half-giant, Rubeus Hagrid, who is also his first contact with the wizarding world. Hagrid reveals himself to be the Keeper of Keys and Grounds at Hogwarts as well as some of Harry's history.[14] Harry learns that, as a baby, he witnessed his parents' murder by the power-obsessed dark wizard Lord Voldemort, who sub-

sequently attempted to kill him as well.[14] Instead, the unexpected happened: Harry survived with only a lightning-shaped scar on his forehead as a memento of the attack, and Voldemort disappeared soon afterwards, gravely weakened by his own rebounding curse. As its inadvertent saviour from Voldemort's reign of terror, Harry has become a living legend in the wizarding world. However, at the orders of the venerable and well-known wizard Albus Dumbledore, the orphaned Harry had been placed in the home of his unpleasant Muggle relatives, the Dursleys, who have kept him safe but treated him poorly, including confining him to a cupboard without meals and treating him as their servant. Hagrid then officially invites Harry to attend Hogwarts School of Witchcraft and Wizardry, a famous magic school in Scotland that educates young teenagers on their magical development for seven years, from age eleven to seventeen. With Hagrid's help, Harry prepares for and undertakes his first year of study at Hogwarts. As Harry begins to explore the magical world, the reader is introduced to many of the primary locations used throughout the series. Harry meets most of the main characters and gains his two closest friends: Ron Weasley, a fun-loving member of an ancient, large, happy, but poor wizarding family, and Hermione Granger, a gifted, bright, and hardworking witch of non-magical parentage.[14][15] Harry also encounters the school's potions master, Severus Snape, who displays a conspicuously deep and abiding dislike for him, the rich brat Draco Malfoy whom he quickly makes enemies with, and the Defence Against the Dark Arts teacher, Quirinus Quirrell, who later turns out to be allied with Lord Voldemort. He also discovers a talent of flying on broomsticks and is recruited for his house's Quidditch team, a sport in the wizarding world where players fly on broomsticks. The first book concludes with Harry's second confrontation with Lord Voldemort, who, in his quest to regain a body, yearns to gain the power of the Philosopher's Stone, a substance that bestows everlasting life and turns any metal into pure gold.[14] The series continues with Harry Potter and the Chamber of Secrets, describing Harry's second year at Hogwarts. He and his friends investigate a 50-year-old mystery that appears uncannily related to recent sinister events at the school. Ron's younger sister, Ginny Weasley, enrolls in her first year at Hogwarts, and finds an old notebook in her belongings which turns out to be the diary of a previous student, Tom Marvolo Riddle, later revealed to be Voldemort's younger self, who is bent on ridding the school of "mudbloods", a derogatory term describing wizards and witches of non-magical parentage. The memory of Tom Riddle resides inside of the diary and when Ginny begins to confide in the diary, Voldemort is able to possess her. Through the diary, Ginny acts on Voldemort's orders and unconsciously opens the "Chamber of Secrets", unleashing an ancient monster, later revealed to be a basilisk, which begins attacking students at Hogwarts. It kills those who make direct eye contact with it and petrifies those who look at it indirectly. The book also introduces a new Defence Against the Dark Arts teacher, Gilderoy Lockhart, a highly cheerful, self-conceited wizard with a pretentious facade, later turning out to be a fraud. Harry discovers that prejudice exists in the Wizarding World through delving into the school's history, and learns that Voldemort's reign of terror was often directed at wizards and witches who were descended from Muggles. Harry also learns that his ability to speak the snake language Parseltongue is rare and often associated with the Dark Arts. When Hermione is attacked and pet-

rified, Harry and Ron finally piece together the puzzles and unlock the Chamber of Secrets, with Harry destroying the diary for good and saving Ginny, and, as they learn later, also destroying a part of Voldemort's soul. The end of the book reveals Lucius Malfoy, Draco's father and rival of Ron and Ginny's father, to be the culprit who slipped the book into Ginny's belongings. The third novel, Harry Potter and the Prisoner of Azkaban, follows Harry in his third year of magical education. It is the only book in the series which does not feature Lord Voldemort in any form, only being mentioned. Instead, Harry must deal with the knowledge that he has been targeted by Sirius Black, his father's best friend, and, according to the Wizarding World, an escaped mass murderer who assisted in the murder of Harry's parents. As Harry struggles with his reaction to the dementors – dark creatures with the power to devour a human soul and feed on despair – which are ostensibly protecting the school, he reaches out to Remus Lupin, a Defence Against the Dark Arts teacher who is eventually revealed to be a werewolf. Lupin teaches Harry defensive measures which are well above the level of magic generally executed by people his age. Harry comes to know that both Lupin and Black were best friends of his father and that Black was framed by their fourth friend, Peter Pettigrew, who had been hiding as Ron's pet rat, Scabbers.[16] In this book, a recurring theme throughout the series is emphasised – in every book there is a new Defence Against the Dark Arts teacher, none of whom lasts more than one school year. During Harry's fourth year of school (detailed in Harry Potter and the Goblet of Fire), Harry is unwillingly entered as a participant in the Triwizard Tournament, a dangerous yet exciting contest where three "champions", one from each participating school, must compete with each other in three tasks in order to win the Triwizard Cup. This year, Harry must compete against a witch and a wizard "champion" from overseas schools Beauxbatons and Durmstrang, as well as another Hogwarts student, causing Harry's friends to distance themselves from him.[17] Harry is guided through the tournament by their new Defence Against the Dark Arts professor, Alastor "Mad-Eye" Moody, who turns out to be an impostor – one of Voldemort's supporters named Barty Crouch, Jr. in disguise. The point at which the mystery is unravelled marks the series' shift from foreboding and uncertainty into open conflict. Voldemort's plan to have Crouch use the tournament to bring Harry to Voldemort succeeds. Although Harry manages to escape, Cedric Diggory, the other Hogwarts champion in the tournament, is killed by Peter Pettigrew and Voldemort re-enters the Wizarding World with a physical body. In the fifth book, Harry Potter and the Order of the Phoenix, Harry must confront the newly resurfaced Voldemort. In response to Voldemort's reappearance, Dumbledore re-activates the Order of the Phoenix, a secret society which works from Sirius Black's dark family home to defeat Voldemort's minions and protect Voldemort's targets, especially Harry. Despite Harry's description of Voldemort's recent activities, the Ministry of Magic and many others in the magical world refuse to believe that Voldemort has returned. In an attempt to counter and eventually discredit Dumbledore, who along with Harry is the most prominent voice in the Wizarding World attempting to warn of Voldemort's return, the Ministry appoints Dolores Umbridge as the High Inquisitor of Hogwarts and the new Defence Against the Dark Arts teacher. She transforms the school into a dictatorial regime and refuses to allow the students to learn ways to defend themselves against dark magic.[18]

Hermione and Ron form "Dumbledore's Army", a secret study group in which Harry agrees to teach his classmates the higher-level skills of Defence Against the Dark Arts that he has learned from his previous encounters with Dark wizards. Through those lessons, Harry begins to develop a crush on the popular and attractive Cho Chang. Juggling schoolwork, Umbridge's incessant and persistent efforts to land him in trouble and the defensive lessons, Harry begins to lose sleep as he constantly receives disturbing dreams about a dark corridor in the Ministry of Magic, followed by a burning desire to learn more. An important prophecy concerning Harry and Lord Voldemort is then revealed,[19] and Harry discovers that he and Voldemort have a painful connection, allowing Harry to view some of Voldemort's actions telepathically. In the novel's climax, Harry is tricked into seeing Sirius tortured and races to the Ministry of Magic. He and his friends face off against Voldemort's followers (nicknamed Death Eaters) at the Ministry of Magic. Although the timely arrival of members of the Order of the Phoenix saves the teenagers' lives, Sirius Black is killed in the conflict. In the sixth book, Harry Potter and the Half-Blood Prince, Voldemort begins waging open warfare. Harry and his friends are relatively protected from that danger at Hogwarts. They are subject to all the difficulties of adolescence – Harry eventually begins dating Ginny, Ron establishes a strong infatuation with fellow Hogwarts student Lavender Brown, and Hermione starts to develop romantic feelings towards Ron. Near the beginning of the novel, lacking his own book, Harry is given an old potions textbook filled with many annotations and recommendations signed by a mysterious writer titled; "the Half-Blood Prince." This book is a source of scholastic success and great recognition from their new potions master, Horace Slughorn, but because of the potency of the spells that are written in it, becomes a source of concern. With war drawing near, Harry takes private lessons with Dumbledore, who shows him various memories concerning the early life of Voldemort in a device called a Pensieve. These reveal that in order to preserve his life, Voldemort has split his soul into pieces, used to create a series of Horcruxes – evil enchanted items hidden in various locations, one of which was the diary destroyed in the second book.[20] Draco, who has joined with the Death Eaters, attempts to attack Dumbledore upon his return from collecting a Horcrux, and the book culminates in the killing of Dumbledore by Professor Snape, the titular Half-Blood Prince. Harry Potter and the Deathly Hallows, the last original novel in the series, begins directly after the events of the sixth book. Lord Voldemort has completed his ascension to power and gained control of the Ministry of Magic. Harry, Ron and Hermione drop out of school so that they can find and destroy Voldemort's remaining Horcruxes. To ensure their own safety as well as that of their family and friends, they are forced to isolate themselves. A ghoul pretends to be Ron ill with a contagious disease, Harry and the Dursleys separate, and Hermione wipes her parents' memories and sends them abroad. As the trio searches for the Horcruxes, they learn details about an ancient prophecy of the Deathly Hallows, three legendary items that when united under one Keeper, would supposedly allow that person to be the Master of Death. Harry discovers his handy Invisibility Cloak to be one of those items, and Voldemort to be searching for another: the Elder Wand, the most powerful wand in history. At the end of the book, Harry and his friends learn about Dumbledore's past, as well as Snape's true motives – he had worked on Dumbledore's behalf since the mur-

der of Harry's mother. Eventually, Snape is killed by Voldemort out of paranoia. The book culminates in the Battle of Hogwarts. Harry, Ron and Hermione, in conjunction with members of the Order of the Phoenix and many of the teachers and students, defend Hogwarts from Voldemort, his Death Eaters, and various dangerous magical creatures. Several major characters are killed in the first wave of the battle, including Remus Lupin and Fred Weasley, Ron's older brother. After learning that he himself is a Horcrux, Harry surrenders himself to Voldemort in the Forbidden Forest, who casts a killing curse (Avada Kedavra) at him. The defenders of Hogwarts do not surrender after learning of Harry's presumed death and continue to fight on. Harry awakens and faces Voldemort, whose Horcruxes have all been destroyed. In the final battle, Voldemort's killing curse rebounds off Harry's defensive spell (Expelliarmus), killing Voldemort. An epilogue "Nineteen Years Later" (set on 1 September 2017)[21] describes the lives of the surviving characters and the effects of Voldemort's death on the Wizarding World. In the epilogue, Harry and Ginny are married with three children, and Ron and Hermione are married with two children.[22]

Appendix B

An example of the form used in the evaluation process

Evaluation of automatically generated summary EP3627321A1

The purpose of this evaluation is to score an automatically generated summary of a patent description based on how well it can convey information. The evaluation is part of my master thesis project and your answers much appreciated.

1. The first page will ask you questions of how important some aspects are to you regarding a summary. In the next page you will read the summary and score that.
2. In the second page you will answer questions having only read the summary.
3. In the third page you will answer a question after having read the full description from which the summary was derived.

***** Obligatoriskt

1. How long do you think a summary should be? (maximum length) *

- ◯ A quarter of a page
- ◯ Half a page
- ◯ One page
- ◯ Two pages
- ◯ Longer

2. Please score how important you find the following parts of the summary are (1 is "not important", and 5 is "very important") *

	1 (not important)	2	3	4	5 (very important)
Grammar	◯	◯	◯	◯	◯
Structure	◯	◯	◯	◯	◯
Coherence (coherence is the measurement of the quality of being logical and consistent)	◯	◯	◯	◯	◯
Length	◯	◯	◯	◯	◯

3. As this evaluation will be sent to a range of people it would help to understand your knowledge of patents. Please define how much experience you have with reading/working with patents? *

- ◯ I work as a patent attorney or similar
- ◯ I don't work as a patent attorney but have experience of reading/working with patents
- ◯ I know a bit about patents but have never worked with them
- ◯ I know very little about patents

Summary of EP3627321A1

Please read the generated summary marked "Summary of EP3627321A1" in the attached files and answer the questions regarding it.

4. Solely based on the summary please choose the one subject you consider the patent to be about. If you can't choose one of the subjects, please choose "None of the above/Can't tell". *

- ◯ Permission of access to the native code of the mobile terminal platform for 3G systems
- ◯ Permission of access to the native code of the mobile terminal platform for computers
- ◯ Permission of access to the native code of the mobile terminal platform for mobile applications
- ◯ Permission of access to the native code of the mobile terminal platform for base stations
- ◯ None of the above/Can't tell

5. Please score the summary based on the following aspects (1 is the lowest, and 5 is the highest) *

	1 (lowest)	2	3	4	5 (highest)
Grammar	◯	◯	◯	◯	◯
Structure	◯	◯	◯	◯	◯
Coherence (coherence is the measurement of the quality of being logical and consistent)	◯	◯	◯	◯	◯
Length	◯	◯	◯	◯	◯

Full description of EP3627321A1

Please read the full description, marked " Description of EP3627321A1 in the attached files and answer the questions after.

6. Based on the full description please choose the one subject you consider the patent to be about. If you can't choose one of the subjects, please choose "None of the above/Can't tell". *

- ◯ Permission of access to the native code of the mobile terminal platform for 3G systems
- ◯ Permission of access to the native code of the mobile terminal platform for computers
- ◯ Permission of access to the native code of the mobile terminal platform for mobile applications
- ◯ Permission of access to the native code of the mobile terminal platform for base stations
- ◯ None of the above/Can't tell

7. How well you think the summary captured the information from the full text? *

- ◯ Extremely well
- ◯ Somewhat well
- ◯ Neutral
- ◯ Somewhat not well
- ◯ Extremely not well

8. Is there anything you would like to add to your answers?

Appendix C

Generated summaries

EP3627321A1 MOBILE TERMINAL WITH MIDDLEWARE SECURITY ACCESS MANAGER.

[0008] A platform system such as described above, wherein mobile terminal platform assembly software and application software are developed separately and then later combined by installing, loading, and running the application software in the mobile terminal platform assembly, may require a non-native application such as a Java midlet to run on a virtual machine.

The requesting application domain software is granted access to the software services component via the at least one interface when the request is granted.

The platform system is generally designated by reference number 10 and includes a mobile terminal platform assembly 12 and one or more applications (ie, application software) 14 that have been installed, loaded, and run in the mobile terminal platform assembly 12.

[0017] The interface component 26 includes a middleware services layer that includes at least one application programming interface (API) for installing, loading, and running one or more applications 14 in mobile terminal platform assembly 12, that isolates the mobile terminal platform assembly 12 from the applications 14 using the assembly 12 via the interfaces, and that provides various other services for the applications 14.

In addition, software services component 22 includes basic system services layers 94 that provide general services that are needed by the platform assembly.

The middleware services layer functions to provide a well-defined interface between the software in the mobile terminal platform assembly 12 and the application software 14 to be installed, loaded, and run in the platform assembly, and encapsulates the mobile terminal platform assembly 12 and isolates the assembly 12 from applications via the middleware services layer, and provides various other services for the applications.

[0036] With reference to FIGURES 6A and 7 , a non-native application 250 requests a service that requires access to the native platform services at step 280.

At step 282, the IM 223 intercepts the service request, which includes an ID tag of the requesting non-native application 250.

If the permission request is granted, then, at step 288, the service request is forwarded to the native platform service or services requested by the non-native application 250.

According to FIGURE 6B , a non-native application 250 requests a service.

As shown at steps 280 and 282, the non-native application 250 invokes a service request and the service request is intercepted, along with an ID tag, at the JJVI 223.

Each access record includes the ID tags of specific applications that have permission to access the requested native platform service.

At step 301, the IM 223 searches the access record of the requested native platform service to determine, at step 303, if the ID tag of the requesting non-native application 250 is associated therewith and the request should thus be granted.

If the ID tag of the requesting non-native application 250 is found in the access record, then at step 303, permission is granted for the non-native application 250 to access the requested native platform service.

If the ID tag of the requesting non-native application 250 is not found in the access record for the requested native platform service, the request is rejected at step 292, aborted and returned to the client that issued the request at step 296.

[0044] The first time the non-native application 250 makes a service request, the SAM 518 accesses an Access Control List (ACL) 312 to determine if permission should be granted to the requested native platform service.

The next time the service request from the same non-native application 250 is intercepted by the JJVI 223 and forwarded to the SAM 518, the decision cache 310 is searched for the permission request.

Each access record 318 includes the JJD tags 320 of the non-native applications 250 that are allowed

access to the particular native platform service (or group of services) associated with the access record 318.
The update requests include the ID tag 320 of the non-native application 250 associated with the update and an identification of the requested native platform service or services where the permissions must be changed.
If the ID tag 320 of the requesting non-native application 250 matches one of the LD tags 320 included in the located access record 318, then permission is granted to the requesting non-native application 250 and the service request is forwarded to the native platform service handler.

WO9200640A1 A HANDS-FREE MODULE.

[0001] A HANDS-FREE MODULE

[0002] BACKGROUND OF THE INVENTION

[0003] The present invention relates to a hands-free module * 5 for a mobile telephone of the kind which includes a housing which houses electronic circuits of which one circuit is intended for an internal loudspeaker and one circuit is intended for an internal microphone.

[0007] DISCLOSURE OF THE INVENTION 25 The present invention relates to a hands-free module which leaves both hands of the subscriber free during a telephone call, wherein the mobile telephone is kept in the holster or in some other place and the subscriber has a head-set or some corresponding device which is 30 connected to the mobile telephone either directly or

[0008] * through the intermediary of a separate external housing connected mechanically and electrically to the mobile k telephone.

The switch 23 has a second fixed connector 22B which is connected to the internal microphone 8 via a microphone amplifier 24.

[0019] The second wire 21 of the cord 13 connects the output of the speaker amplifier 18 to the external phone unit 14.

The input of the microphone amplifier is connected to the output of the limiter 19, the input of which is connected directly to the internal loudspeaker 7 of the mobile telephone, via connector pins (not shown) in the connectors 12 and 10.

The cord 13 of the external phone unit 14 and the external microphone 15 are connected directly to the housing 3 by means of the connector 27.

WO2020098963A1 CUTTING TOOL.

The cutting tool includes a first lever having a first handle portion and a first cutting blade.
The cutting tool includes a second lever having a second handle portion.
The cutting tool includes an adjustment mechanism to selectively allow rotation of the gear lever relative to the first gear portion to change mechanical advantage between the first lever and the second lever.
The gear lever includes a second cut-out portion.
[0014] According to an embodiment of the present invention, the cutting tool further includes a spring buffer integrated on any of the first lever and the second lever.
The cutting tool 100 includes a first lever 110 having a first handle portion 112 and a first cutting blade 114.
The cutting tool 100 includes a second lever 120 having a second handle portion 122.
The cutting tool 100 includes an adjustment mechanism 140 to selectively allow rotation of the gear lever 130 relative to the first gear portion 116 to change mechanical advantage between the first lever 110 and the second lever 120.
As illustrated, the second lever 120 includes a first cut-out portion 202.
The gear lever 130 includes a second cut-out portion 204.
Further, the adjustment mechanism (here the screw mechanism 140) includes at least one protrusion 210 to selectively abut with at least one of the first cut-out portion 202 and the second cut-out portion 204 to restrict further rotational movement of the first cut-out portion 202 and the second cut-out portion 204 respectively.
[0031] As used herein, the present disclosure refers to protrusion or spacer which are used in different embodiments of the adjustment mechanism 140, 500 to selectively engage with cut-out portions (ie the first cut-out portion 202 and the second cut-out portion 204), whenever a change between a dynamic cutting mode and a power cutting mode is required.
Now after turning to ON state of the toggle switch mechanism 500 by actuation of the toggle button 502, the protrusion 506 abuts any or both of the first cut-out portion 202 and the second cut-out portion 204 leading to a higher mechanical advantage between the first lever 110 and the second lever 120.

EP3439458A1 MULTI-PURPOSE CAN.

Said edge comprises a edge portion forming a shovel edge.
In a possible embodiment that edge portion is formed as a straight edge portion to ease its use as a shovel.
In a possible embodiment the edge portion is formed as a straight edge portion .
According to embodiments, the liguid dispensing portion is arranged adjacent to said edge portion.
[0012] According to embodiments, the liguid dispensing portion is arranged in a corner region between said edge portion and a lateral edge portion.
In other words, the liguid dispensing portions are arranged at the front edge portion and spaced from each other by said edge portion.
[0027] At the upper side, the main body 2 comprises an edge 2.2.
The edge 2.2 comprises a straight edge portion 4 and lateral edge portions 2.2a, 2.2b being arranged between said straight edge portion 4 and the handle 3.
Said liguid dispensing portions 5, 5' are provided at the edge 2.2 of the main body 2.
The liquid dispensing portions 5, 5' may be arranged between the straight edge portion 4 and the lateral edge portion 2.2a, 2.2b.

EP3696022A1 VEHICLE INTERIOR LIGHTING SYSTEM.

The vehicle lighting system a light guide and a light dispensing device configured to transmit or dispense light from a light source into the interior of the vehicle.

[0006] In one example, a vehicle lighting system for a vehicle includes a light guide and a light dispensing device.

The light dispensing device is configured to receive light from the light source via the light guide and dispense the light from the light guide to an interior of the vehicle.

The light dispensing device includes a textile configured to dispense the light from the light guide to the interior of the vehicle and a reflector configured to reflect the light within the light dispensing device that is received from the light guide towards the textile, and a lens located between the textile and the reflector.

[0007] In another example, a vehicle lighting system for a vehicle includes a light guide and a light dispensing device.

The light dispensing device is configured to receive light from the light source via the light guide and dispense the light from the light guide to an interior of the vehicle.

[0008] In another example, a vehicle lighting system for a vehicle includes a light guide and a light dispensing device.

The light dispensing device is configured to receive light from the light source via the light guide and dispense the light from the light guide to an interior of the vehicle.

- a light dispensing device coupled to the second end of the light guide and configured to receive light from the light source via the light guide and dispense the light received from the light guide to an interior of the vehicle, the light dispensing device comprising:a textile configured to dispense the light from the light guide to the interior of the vehicle;a reflector configured to reflect the light within the light dispensing device that is received from the light guide towards the textile; anda diffusor lens located between the textile and the reflector, the diffusor lens including a plurality of scattering objects disposed between a front surface of the diffusor lens and a back surface of the diffusor lens to scatter the light within the lens in a plurality of directions.

In the example of Fig 2A , vehicle lighting system 202 includes a light dispensing device 210, light guide 211, and light source 212.

Light dispensing device 210, light guide 211, and light source 212 may be examples of light dispensing devices 110, light guides 111, and light sources 112 of Fig 1 , respectively.

[0020] Light guide 211 is configured to dispense light emitted by light source 212 to light dispensing device 210, such that light dispensing device 210 dispenses the light to an interior of a vehicle.

[0022] Textile layer 220 is configured to dispense light from light source 212 (via light guide 211) to an interior of the vehicle.

In the example of Fig 3A , vehicle lighting system 302 includes a light dispensing device 310, light guide 311, and light source 312.

Light dispensing device 310, light guide 311, and light source 312 may be examples of light dispensing devices 110, light guides 111, and light sources 112 of Fig 1 , respectively.

[0030] Light guide 311 is configured to dispense light emitted by light source 312 to light dispensing device 310, such that light dispensing device 310 dispenses the light to an interior of a vehicle.

US6216772B1 Device for filtering and cooling.

According to known techniques, cooling of the oil is in this case carried out by means of an oil cooler being mounted between the gearbox pump and the oil filter.

The oil that is utilized in the gearbox will thus be fed through the oil cooler, in which the oil is cooled down and is filtered in the oil filter.

According to one embodiment of the present invention, the filtering component is arranged as a detachable part in a protector for the filter; ie, the filtering component constitutes an insertion filter.

An oil pump 2 is arranged in the interior of the gearbox, by means of which oil can be supplied from the oil pump to the device according to the present invention.

The device according to the present invention contains two main components, namely a filter unit in the form of an oil filter 9 and an oil cooler 10.

The filter unit 9 is designed having an essentially tubular filter protector 11 which supports a similarly tubular filtering component 12 which is constructed of a filtering material, preferably filtering paper which in a known manner is designed for the separation of particles that are present in percolating, contaminated oil.

The oil cooler 10 according to the present invention is arranged at the end of the filter protector 11 which is not facing towards the gearbox 1.

The filtered oil, which has thus passed through the oil-filtering component 12, is further fed in the direction of the end of the cavity 26 that is farthest from the gearbox.

By means of the present invention, the filtered oil that is present in the cavity 26 will be forced against the oil cooler 10 through the holes 27 and 28, in which oil cooler the oil is cooled down.

Due to the fact that the oil cooler 10 is arranged downstream of the filtering component 12, the entire amount of oil will first be filtered, and then cooled down.

Furthermore, the present invention may be constructed so that the cavity, through which the oil is fed before it reaches the oil cooler 10, can be arranged inside as well as outside the filtering component 12, provided that the oil from the cooling device is fed to the outlet 31 through the cavity by means of a connection element.

EP3035664A1 Method for processing a video stream.

[0001] The present invention relates to processing a stream of images in a video application, in particular a stream in which the images are affected by varying zoom settings for a video camera in a situation where the images are affected by geometric distortions.

Most image post-processing software includes a barrel distortion correction function in which a user may alter various parameters for reducing the effects of barrel distortion in images acquired.

The method comprises acquiring a continuous flow of barrel distorted images in a video camera and processing the image data in an image processing unit within the camera and adding the processed image data as an output image to an image stream, and the processing includes applying a barrel-distortion correction so as enable formation of a corrected image having a minimum width and a minimum height.

[0013] In one or more embodiments the barrel distortion correction may be based on input of a current zoom setting of the camera, such that a particular zoom setting refer to a particular correction function to be applied in the barrel distortion correction.

[0014] In any embodiment a functionality for a user to select a view comprising an output image corresponding to an image displayed without barrel-distortion correction or an output image displayed with barrel-distortion correction, while the same user-defined aspect ratio is used.

[0015] In any of the above embodiments the barrel-distortion correction may be based on a current zoom setting for the camera.

[0016] The method may also, in one or more embodiments, enable the receipt of a client request including a selection between an output image displayed without barrel distortion correction or an output image corrected for barrel distortion, wherein the same client-defined aspect ratio is used for any output image stream.

Barrel distortion will depend on the zoom settings and therefore a database or transformation table for barrel-distortion correction may use a current value of the zoom settings as input.

If there are no distortions, or if a user prefers not applying any distortion correction the processed video stream may be forwarded as is, or at least without barrel distortion correction.

The transformation table uses as input data regarding the characteristics of the imaging optics used, ie present focal length settings and more particularly data regarding distortions etc, and its purpose is to move image information (eg pixel intensity information) from the imaged position on the image sensor to the position in which it should have been without distortions.

Fig 3 illustrates the corrected image data in the form of a corrected image 54.

It may not be evident from the drawings alone, yet due to the properties of barrel distortion and its dependence on distance from the optical axis the application of a barrel-distortion correction may alter the aspect ratio of the corrected image as compared to the aspect ratio prior to barrel-distortion correction.

[0028] A common solution used is to crop the corrected image to the particular aspect ratio used such that the video stream may continue uninterrupted when a user switches between viewing an uncorrected view and a view in which distortion correction is applied, ie adapting the distortion corrected view to the selected capture mode having the user-defined aspect ratio.

[0030] Since barrel distortion has a dependence on the distance from the optical axis, the effects of barrel distortion would be reduced if the user were to zoom in on a detail in the scene, and the discrepancy between a corrected view and a distorted view would decrease.

[0035] The mere application of a mask to an image may be more or less known, yet there is a benefit in performing the distortion correction, the cropping and the masking on a live video stream, preferably in the image processing unit of a camera before transmitting the information.

US2016094765A1 METHOD AND IMAGE PROCESSING DEVICE FOR IMAGE STABILIZATION OF A VIDEO STREAM.

In particular it relates to a method for image stabilization of a video stream and an image processing device arranged to perform image stabilization of a video stream.

This electronic manipulation is based on determining an image shift between image frames of the video stream captured by the video camera.

[0008] In Fig 1(B) the image frames of the video stream are displayed after determining and applying image shift between the image frames.

[0011] According to a first aspect of the invention, the above object is achieved by a method for image stabilization of a video stream comprising image frames captured by a video camera, the video stream depicting a scene.

The method comprises: performing electronic image stabilization to a first sub-set of image frames of the image frames of the video stream to compensate for a first oscillating movement of the video camera such that at least a portion of each image frame of the image frames of the first sub-set of image frames are building up a stable first view of the scene; applying a first edge mask to the first stable view of the scene, the first edge mask having a first width based on a first camera oscillation amplitude; comparing the first camera oscillation amplitude with a second camera oscillation amplitude, if the second camera oscillation amplitude differs from the first camera oscillation amplitude by a predetermined amount, the method further comprises: performing electronic image stabilization to a second sub-set of image frames of the image frames of the video stream to compensate for a second oscillating movement of the video camera such that at least a portion of each image frame of the image frames of the second sub-set of image frames are building up a stable second view of the scene; applying a second edge mask to the second stable view of the scene, the second edge mask having a second width based on the second camera oscillation amplitude; wherein the second sub-set of image frames is a set of image frames being captured later in time than the first sub-set of image frames.

[0013] By applying electronic image stabilization and applying edge masks being dependent on the camera oscillation amplitude for any given set of image frames the full resolution of the image sensor may be used at all times and still having a stable view.

[0020] The performing of electronic image stabilization may comprise measuring movement of the video camera using a motion sensor arranged in the video camera.

[0024] According to a second aspect of the present invention an image processing device arranged for image stabilization of a video stream comprising image frames captured by a video camera is provided.

The image processing device comprises: an electronic image stabilization module arranged to perform electronic image stabilization to sub-sets of image frames of the image frames of the video stream to compensate for a oscillating movement of the video camera; and a masking module arranged to apply an edge mask to each sub-set of image frames, wherein each edge mask is having a fixed width, wherein the fixed width is based on a camera oscillation amplitude being specific for the sub-set of image frames to which the edge mask is applied.

[0027] The electronic image stabilization module may comprise: an image shift determination module arranged to determine image shifts between image frames of the video stream, the image shifts being caused by an oscillating movement of the video camera; and an image shift module arranged to produce a stable view of a scene depicted by the video stream by applying shifts on image content of the image frames in accordance with the determined image shifts to compensate for the oscillating movement of the video camera.

[0028] The image processing device may be a video camera comprising a motion sensor arranged to sense oscillating movement of the video camera, wherein the masking module is arranged to determine the camera oscillation amplitudes based on data from the motion sensor.

The image stabilizing system 10 comprises a video camera 20 , an image processing device 30 and a display 40 .

EP0205073A2 An opening arrangement for packing containers.

[0001] The present invention relates to an opening arrangement for packing containers comprising a packing material with a pouring opening and a tear-off strip covering the same.

[0006] These and other objects have been achieved in accordance with the invention in that an opening arrangement for packing containers, comprising a packing material with a pouring opening and a tear-off strip covering the same, has been given the characteristic that the pouring opening has a projection extending against the direction of tearing of the strip whose area is considerably smaller than the total area of the pouring opening.

On the top side 3 of the packing container is an opening arrangement in the form of a pouring opening provided with a tear-off cover strip 6.

[0012] The form of the pouring opening 8 and the seal between the cover strip 6 and the different material layers of the packing container 1 are illustrated in greater detail in Figure 3 and Figure 4.It is evident from Figure 3 how the pouring opening 8 is placed close to one corner of the upper surface 3 of the packing container and is covered by the cover strip 6, indicated only in Figure 3, whose one end serves as a grip part 6' and extends out over the edge line 9 of the packing container.

Figure 3 shows further how the pouring opening 8 is provided with a projection 8'extending against the direction of tearing of the cover strip, whose area is considerably smaller than the total area of the pouring opening 8.

[0018] In the embodiment of the opening arrangement shown the projection 8' of the pouring opening is facing towards the edge line 9 of the packing container over which the product is intended to be poured, but it is also possible, of course, to place the projection 8' on the opposite end of the pouring opening if the tearing of the cover strip 6 too is in the opposite direction.

WO2020120499A1 A CENTRIFUGAL SEPARATOR.

[0009] In another aspect of the invention, this is achieved by a method for separating liquid food into a light phase, a heavy phase and an ejection phase that comprises solid impurities in a separator, the method comprises distributing a flow of the liquid food through a first set of discs in a disc stack arranged in the separator so that the liquid food is limited to flow between a periphery and a center portion of the disc stack, and distributing a flow of the liquid food through a second set of discs in the disc stack so that the light phase flows from distribution openings, located between the periphery and the center portion, towards a center channel at the center portion, and the heavy phase flows from the distribution openings towards the periphery.

Fig 1 is a cross-sectional side view of a centrifugal separator for separating liquid food into a light phase, a heavy phase, and an ejection phase that comprises solid impurities;

[0014] Fig 2 is a diagram showing a time interval for the ejection of an ejection phase in relation to the timing of changes in flow through an inlet valve and a first outlet valve; and

[0015] Figs 3a-c are flowcharts of methods for separating liquid food into a light phase, a heavy phase and an ejection phase.

[0018] Fig 1 is a schematic illustration of a centrifugal separator 100 for separating liquid food (RP) into a light phase (LP), a heavy phase (HP), and an ejection phase (SI) that comprises solid impurities (SI). Reducing the flow of liquid food (RF) through the inlet 102 provides for reducing turbulence in the separator 100 as the ejection phase (SI) is ejected through the ejection port 106.

The separator 100 may comprise a first outlet valve 118 that is arranged at the outlet 104´ for the light phase (LP) to increase the flow of the light phase (LP) through the outlet 104´ for the light phase (LP) when the ejection phase (SI) is ejected through the ejection port 106.

[0031] In one example, the first outlet valve 118 increases the flow of the light phase (LP) through the outlet 104´ for the light phase (LP) while the inlet valve 116 reduces the flow of liquid food (RF) through the inlet 102 when the ejection phase (SI) is ejected.

[0034] The separator 100 may comprise a second outlet valve 119 that is arranged at the outlet 104 for the heavy phase (HP) to reduce the flow of the heavy phase (HP) through the outlet 104 for the heavy phase (HP) when the ejection phase (SI) is ejected through the ejection port 106.

The flow of the heavy phase (HP) through the second outlet valve 119 may be reduced or completely stopped when the ejection phase (SI) is ejected through the ejection port 106.

[0035] Fig 3a illustrates a flow chart of a method 200 for separating liquid food (RP) into a light phase (LP), a heavy phase (HP) and an ejection phase (SI) that comprises solid impurities (SI) in a separator 100.

The method 200 may comprise reducing 203 the flow of the liquid food (RP) into the separator 100 while ejecting 204 the ejection phase (SI) from the separator 100.

The method 200 may comprise increasing 203´ the flow of the light phase (LP) through an outlet 104´ for the light phase (LP) while ejecting 204 the ejection phase (SI) from the separator 100.

The method 200 may comprise reducing 203 the flow of the heavy phase (HP) through an outlet 104 for the heavy phase (HP) while ejecting 204 the ejection phase (SI) from the separator 100.

Appendix D

Links to patents used for summarization

- EP0205073A2 An opening arrangement for packing containers (Tetra Pak)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DEP0205073A2` (2020-10-27)
- EP3035664A1 Method for processing a video stream (Axis)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DEP3035664A1` (2020-10-27)
- EP3439458A1 Multi-purpose can (Husqvarna)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DEP3439458A1` (2020-10-27)
- EP3627321A1 Mobile terminal with middleware security access manager (Ericsson)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DEP3627321A1` (2020-10-27)
- EP3696022A1 Vehicle interior lightning system (Volvo)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DEP3696022A1` (2020-10-27)
- US6216772B1 Device for filtering and cooling (Volvo)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DUS6216772B1` (2020-10-27)
- US2016094765A1 Method and image processing device for image stabilization of a video stream (Axis)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DUS2016094765A1` (2020-10-27)
- WO9200640A1 A hands-free module (Ericsson)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DWO9200640A1` (2020-10-27)
- WO2020098963A1 Cutting tool (Husqvarna)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DWO2020098963A1` (2020-10-27)

- WO2020120499A1 A Centrifugal separator (Tetra Pak)
  `https://worldwide.espacenet.com/patent/search?q=pn%3DWO2020120499A1`
  (2020-10-27)

Printed by Tryckeriet i E-huset, Lund 2020

LUND
UNIVERSITY

Series of Master's theses
Department of Electrical and Information Technology
LU/LTH-EIT 2020-798
http://www.eit.lth.se