

Automatisk Textsammanfattning av Patent

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 patent.

En stor del av arbetet för ett patentombud består av att läsa patent. Detta är en tidskrävande uppgift då ett patent 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 patentet så att läsaren kan avgöra om det fulla patentet 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 il-

lustrera har jag kört denna populärvetenskapliga sammanfattning genom mitt verktyg. Följande textruta är den genererade sammanfattningen. Resultaten från studien visar att den valda meto-

Automatisk Textsammanfattning av Patent.

Mitt verktyg genererar sammanfattningar baserat på patentbeskrivningar för att minska tiden ett patentombud behöver lägga på att läsa hela patent. 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.

den fungerar bra. Sju av tio sammanfattningar lyckades förmedla vad patentet 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.