Kan en dator sortera dokument lika bra som en människa?

Med en växande mängd digital information hos företag uppkommer behovet av verktyg för att organisera och konsumera data. Vi valde att undersöka om maskininlärning skulle kunna vara ett sådant verktyg. Två olika metoder undersöktes: en som behöver tränas av människor och en som inte behöver det. Den tränade metoden visar sig som väntat fungera bättre men båda är användbara. En viktig slutsats är att det är minst lika viktigt att lära känna sin data som det är att lära känna sina maskininlärningsmetoder.

Stora företag producerar stora mängder dokument. De flesta av dagens system för att organisera och hitta den information som används är byggda för mycket mindre datamängder och kräver manuellt arbete från människor. Detta blir en begränsning när informationen inte längre är överblickbar. Tänk om datorer skulle kunna lösa det här problemet, så att vi människor kan fokusera på att skapa nytt och intressant innehåll istället?

Vi har använt maskininlärning för att organisera en stor samling av textdokument. Maskininlärning handlar om att få datorer att själva lära sig olika uppgifter med hjälp av data. Detta utan att en människa aktivt behöver programmera exakt hur de ska bära sig åt.

Vi testade två olika approacher. Den ena var en ”blind” metod. Där får datorn helt enkelt tillgång till alla dokument, och får själv räkna ut vilka som påminner om varandra och hur de skulle kunna kategoriseras. Detta gör den genom att ta reda på vilka ord som är vanliga i de olika dokument. Resultaten från den här metoden blev överskande bra. Det gick klart och tydligt att se att datorn hade lyckats hitta samband mellan dokumenten som var meningsfulla även för en människa.

Den andra approachen var en ”seende” metod. Här användes samma information om vilka ord som är vanliga i de olika dokumenten, men dessutom hur en människa hade valt att kategorisera datan. Utifrån detta tränades den på två tredjedelar av dokumenten och utvärderades sedan på resterande del. Förhoppningen var att kategoriseringarna skull vara bättre. Resultaten visar att den seende metoden vida överträffar den blinda metoden. Dessutom har den seende metoden en mycket större utvecklingspotential: ju mer data den får att lära av, desto bättre kan resultatet bli.

Vi upptäckte under projektets gång hur viktigt det är att verkligen förstå dokumenten man arbetar med. Om man använder en seende metod är det dessutom viktigt att välja bra ”inlärningsdata” att träna datorn med. Även om det fanns mycket man kunde arbetat vidare med inom matematiken och metoderna, så är vi övertygade om att mer arbete med själva datan skulle vara det som har störst potential att göra lösningen ännu bättre.

Sammanfattningsvis ser resultaten från detta projekt mycket lovande ut. Vi tror att vi i framtiden kommer se en mer utbredd användning av maskininlärning som hjälper företag organisera sin data. Därmed kommer de anställda i gengäld kunna fokusera på att ta fram ny spännande information.