Onderzoeksvoorstel (KI): Inhoudelijke Clustering t.b.v. vindbaarheid

Kunnen we de clustering van overheidsdocumenten verbeteren d.m.v. Kunstmatige Intelligentie?

Er worden momenteel meer dan 2,5 miljoen[1] documenten gepubliceerd onder Bekendmakingen. De zoekfunctie stelt de burger in staat om te filteren op onderwerpen als “Economie” of “Douane” en sub-onderwerpen zoals “Economie | Markttoezicht”. In mijn werk wil ik onderzoeken hoe een ongereguleerd (*unsupervised*) algoritme deze set aan documenten clustert. De resultaten van dit onderzoek bieden een vergelijking tussen de “menselijke” groepering en een op basis van het model.

Een voordeel van ongereguleerd clusteren is dat het onderwerp van een tekst niet meer handmatig hoeft te worden bepaald. In plaats daarvan wordt er per document een verdeling zichtbaar over de set aan onderwerpen. Tegelijkertijd wordt er een verdeling getraind van onderliggende onderwerpen in de gehele dataset van overheidsdocumenten. Dit model heet een *Latent Dirichlet Allocation*[2].

Door te focussen op onderliggende onderwerpen neem je de volgende problemen weg bij de burger:

- Geen kennis nodig van politieke processen en namen van bestandstypes
- Relevante resultaten met algemene zoektermen, bijvoorbeeld “Onderwijs”

Mijn voorstel is om te onderzoeken of we een set tekstuele overheidsdocumenten kunnen clusteren op basis van overeenkomende inhoud met behulp van Kunstmatige Intelligentie. Computers kunnen veel grotere aantallen bits aan informatie verwerken en herkennen daardoor (wellicht) andere patronen in grote datasets. *Information Overload* weerhoudt ons ervan door de bomen het bos te zien en bijvoorbeeld ook vele documenten te lezen voordat we de juiste info vinden.

Clustering is een ongereguleerde methode waarbij de input data enkel bestaat uit een set tekstuele documenten en dus geen labels nodig heeft. Voor een voorspellend model zijn labels wel noodzakelijk, maar het gaat hier om het vinden van onderliggende patronen in de woordkeuze.

Per document wordt er geen label onderwerp gegeven zoals “Economie”. Het doel van het trainen met een puur tekstuele dataset is het herkennen van onderliggende onderwerpen zonder die vooraf mee te hoeven geven. Een voordeel hiervan is het verminderen van menselijk vooroordeel (*bias*) in de dataset. De onderzoeksvraag is of dit model, na training, goed presteert in het toekennen van onderwerpen. Een belangrijke vraag is, wat betekent hierbij “goed”?

Een *Latent Dirichlet Allocation (LDA)* model is gebaseerd op twee verdelingen. Tijdens het trainingsproces worden de parameters van deze verdelingen getraind op basis van de teksten in de dataset (overheidsdocumenten). De eerste verdeling is er een van de kans per woord dat het voorkomt in een bepaald thema. Een thema zoals “Economie” zal een grote kans hebben om het woord “euro” te bevatten. De tweede verdeling is er een van thema’s per document. Een document kan bijvoorbeeld voornamelijk gaan over “Economie” (80%), maar ook over “Bestuur” (20%). Deze verdelingen vormen samen een representatie van de set aan overheidsdocumenten.

---

[1] https://zoek.officielebekendmakingen.nl/resultaten?q=(c.product-area==%22officielepublicaties%22)and(((w.publicatienaam==%22Tractatenblad%22))or((w.publicatienaam==%22Staatsblad%22))or((w.publicatienaam==%22Staatscourant%22))or((w.publicatienaam==%22Gemeenteblad%22))or((w.publicatienaam==%22Provinciaal%20blad%22))or((w.publicatienaam==%22Waterschapsblad%22))or((w.publicatienaam==%22Blad%20gemeenschappelijke%20regeling%22)))&pg=10&svel=&svol=&zv=&col=AlleBekendmakingen

[2] https://jmlr.org/papers/volume3/blei03a/blei03a.pdf

Na training en validatie kan dit model dus een kansverdeling geven per overheidsdocument van de onderwerpen die in het document aangekaart worden. Naar mijn mening is het voordeel van een kansmodel, zoals LDA, dat we de onzekerheid kunnen aantonen. We kunnen bij een zoekopdracht voor “Economie” per document aangeven hoeveel procent de tekst in het document een economisch thema heeft. In de toekomst kunnen we onderzoeken of dit kan bijdragen aan een relevantie score.

Een ander voordeel van deze aanpak is dat we een verdeling vinden van woorden per thema. Dit betekent dat een thema gevonden kan worden door meerdere woorden in te geven. Momenteel vragen we burgers een thema te kiezen maar deze lijst is erg lang en uitgebreid. Kiezen uit een lange lijst onderwerpen vraagt denkenergie en maakt het zoeken vermoeiend en lastig. Het intikken van een zoekopdracht van een aantal woorden is dan eenvoudiger. We kunnen met de ingetikte woorden op zoek naar het thema van de zoekopdracht. Daarmee maken we de zoekfunctionaliteit beter en kan de burger het kiezen tussen alle onderwerpen overslaan.

Hoofdvraag: Werkt *unsupervised topic modeling* (KI) voor inhoudelijke clustering van overheidsdocumenten? Oftewel: Halen we met KI zinnige thema’s uit overheidsdocumenten?

- Welke onderwerpen zijn er te interpreteren uit de verdelingen over woorden?
- Hoeveel onderwerpen zijn er te vinden bij het maximaliseren van stabiliteit?[3]
- Zijn de onderwerpen vanuit LDA te vergelijken met de bestaande lijst onderwerpen?
- Is een *Latent Dirichlet Allocation* toepasbaar als methodiek?

Invulling nieuwe functie Kunstmatige Intelligentie (KI) naast onderzoek:

- Training geven en events organiseren rondom overheidsdata en data science
- Blogs schrijven over Kunstmatige Intelligentie bij het Rijk, Open Overheid, Open Algoritmes
- Presenteren van onderzoeksresultaten en verdere mogelijkheden met KI

Praktische invulling eerste periode (komende drie maanden):

1. Data verzamelen: documenten via https://www.rijksoverheid.nl/opendata/documenten
2. Uitleg video opnemen voor het gebruik van Open Data als promotie/informatief
3. Analyse uitvoeren: Latent Dirichlet Allocation toepassen op set documenten (bv nieuwsberichten)
4. Resultaten interpreteren: hoeveel thema’s zijn er gevonden? Welke woorden komen relatief veel voor per thema?
5. Komen deze thema’s overeen met de thema’s die we momenteel hanteren?
6. Voor een nieuw document, welk thema past erbij volgens de woord-document verdeling?
7. Werkt deze “unsupervised” methode naar behoren? Open discussie binnen KOOP? Bespreking van verschillende documenten en het gegeven thema + aantal thema’s
8. Presentatie van model en resultaten (ook in Jip en Janneke taal)

Ten slotte, ik vind het een hele mooie uitdaging om te onderzoeken hoe toepassingen van Kunstmatige Intelligentie de producten van KOOP kunnen verbeteren. Het hierboven beschreven onderzoek zal (hopelijk) leiden tot meer inzicht in de praktische mogelijkheden van deze methodes.

[3] Zie scriptie in bijlage voor meer informatie over het kiezen van het aantal onderwerpen