Gesprek Alex Olieman

Gesprek 15-03

1. Advies en hergebruik
2. Weten wat er allemaal speelt rondom deze data

# Zijn onderzoek

Locatieherkenning in teksten, gecombineerd met zoekmachine en kaart.

Het binnenhalen van raadsinformatie via de API. Is die dan al bruikbaar voor je applicatie?

Dan ga je aan de slag met LocLinkVis. Klopt dat?

Je vult de data via de open street maps, hoe werkt dat?

Volgens mij had je het in het artikel ook nog over detectie van locaties die niet in OSM, hoe werkt dat?

Dan typeer je locaties, en dat is gewoon kijken naar woorden in de omgeving?

Hoe bouw je dan een hiërarchie in de entiteiten?

Hoe spot je homoniemen?

Hoe spot je synoniemen?

Vaak heb je een groot bestand, pak je gewoon alles wat een keer voor komt?

Hoe evalueer je je werk, op welke metric?

Dan zet je het automatisch op de kaart, is dat lastig?

Wat vond je het leukst aan je onderzoek?

Wat nam meer tijd in dan je dacht?

Wat zou je als eerste verbeteren?

# Mijn onderzoek

Classificatie van het onderwerp

Op 2 manieren: Naive Bayes en CNN’s Met Word2Vec

Evaluatie op micro- en macro-averages van accuracy, F1, recall, precision

Ook nog evaluatie op de relevante data

---- Paar interessante problemen:

In hoeverre komt de data van de overheid goed overeen met die van de gemeentes?

Hoe ga je om met grotere documenten in CNN?

Veel documenten zijn multi-topic, is dat erg?

Heb je nog meet advies?

Transfer learning

Vind het belang van bepaalde woorden die veel voorkomen in een dataset, en die downplayen in belang. Kan in model, of in preprocessing

Misschien beter om als baseline maximum entropy te gebruiken, of misschien SVM.

Veelal binaire classifiers beter dan multitopic classifiers, maar.. ConvNet kan het wel multitopic.

Altijd even unsupervised technieken toepassen met analyse, want verschillende soorten documenten er in je set zitten. Probeer snel te visualiseren, bijvooorbeeld met histogrammen, wordclouds, en metadata visualiseren. Zorg dat je weet wat de notulen zijn, bestemmingsplannen etc, Die uit elkaar halen is misschien goed voor je classificatie.

Door te kijken naar welke commissie waar aan werkt kan je een inschatting maken van de topics, op zijn minst de overkoepelende topics.

Een set van enteiten bij naam, ambigue naam, dus twee dingen met een andere naam. Kan je dat unsupervised uit elkaar gaan halen, bijvoorbeeld met clusters. Cold-start kan je er ook in zien, je weet namelijk dat een persoon aantreed op een bepaald moment. Kan je dat automatisch dan