VOGIN-IP-lezing 2019
Eric Sieverts
Workshop Automatisch metadateren en categoriseren

Oefening 1: Entity recognition

1. Automatische "entity recognition" en onderwerpsclassificatie met "OpenCalais"

- Ga naar de website van de Open Calais service: **http://www.opencalais.com/opencalais-demo/**
- Open een ander browservenster en zoek daar een Engelstalige tekst, bijvoorbeeld een krantenartikel (NY Times, Guardian, ...) of een persbericht.
- Knip de tekst en plak hem in het tekst-venster van Open Calais.
- Klik op [TAG IT] en kijk welke (soorten) entiteiten in de tekst zijn herkend (en met gekleurde onderstreping gemarkeerd).
- Door met de muis over gehighlighte woorden te gaan krijg je daar meer informatie over te zien.
- In de linker kolom kun je de verschillende soorten entiteiten ook openklikken om ze allemaal opgesomd te krijgen.
- Je ziet daar onder “Topics” ook welke onderwerpscategorieën aan de tekst zijn toegekend.
- Herhaal dit desgewenst nog met een andere tekst.

2. Eenvoudig "entity recognition" tool "Annie".

Ga naar de website: **http://services.gate.ac.uk/annie/**
Hier kun je of een URL opgegeven van een te analyseren webpagina, of een ergens vandaan uitgeknipt stuk tekst inplakken. Klik dan op "Process Text".
Het systeem beperkt zich tot het herkennen of iets een mens, een locatie of een organisatie is.
Beperk je tot Engelstalige teksten.

3. Voorbeeld toepassing in KB onderzoeksportal

Ga naar de website: **http://www.kbresearch.nl/xportal**
Kies onder "Select collection" bijvoorbeeld "Newspapers" (dat is Delpher)

- Tik een zoekwoord in en klik één van de resultaten aan. Boven de gevonden tekst zie je blauwe blokjes met de in die tekst als entiteiten herkende namen of begrippen. Door met de muis over zo'n blokje te gaan of het aan te klikken, krijg je meer informatie over die entiteit te zien.
- Zie eventueel nog **https://youtu.be/EvWPsXplWnk** voor uitleg hoe je fout herkende entiteiten kunt corrigeren.
- Tik een generiek zoekwoord tussen rechte haken in, bijv. [romeinse keizer] (geen meervoud) of [rolling stones]. Het zoekresultaat bevat nu ook artikelen waarin niet noodzakelijkerwijze die zoekterm voorkomt, maar wel de naam van een specifieke keizer of van een lid van die groep waarvan in Wikidata bekend was dat ze tot die categorie behoren.

Oefening 2: Automatische classificatie

1. Automatische toekenning van Dewey Decimale Classificatie

- Ga naar de website van ACT-DL: **http://act-dl.base-search.net/**
  Je kunt daar Engelse of Duitse teksten laten categoriseren op basis van de Dewey Decimale Classificatie
- Kies één van de opties om tekst te laten voorzien van een Dewey-code:
  - hetzij een ergens "uitgeknipt" stuk tekst (via "text categorizer"),
  - hetzij de inhoud van een webpagina (URL via "web categorizer"),
  - hetzij een PDF-document (upload via "PDF categorizer").

2. Automatische trefwoordtoekenning voor een catalogus

(zie ook: https://www.youtube.com/watch?time_continue=3&v=lSrFP3D-uTg)

- Ga naar de website van het Finse Annif: **http://annif.org/**
  Je kunt daar een stuk Engelse (of Finse ….) tekst inplakken en er trefwoorden aan laten toekennen.
- Als je de tekst heb ingeplakt, kies dan bij " Project" (recht naast het venster) voor Engelse analyse, bijvoorbeeld YSO of Wikidata, en klik op "Analyze".

3. Automatische toekenning van thesaurustermen met Climate Tagger

- Ga naar de website **http://api.climatetagger.net/demo/**
- Plak een stuk tekst over een klimaat-gerelateerd onderwerp in het tekstvenster (in één van de daaronder te kiezen talen) en klik op [Extract]
- De "extracted concepts" zijn afkomstig uit hun thesaurus. De zwarte driehoekjes geven meer informatie over de die thesaurustermen. De score geeft een maat voor de waarschijnlijkheid dat dit een voor deze tekst correcte thesaurusterm is.
- Verder omlaag zie je ook nog geografische aanduidingen en niet gestandaardiseerde termen die uit de tekst zijn afgeleid.
- Onder "Format"kun je desgewenst ook nog kiezen om andere, meer gestructureerde uitvoer van de resultaten te krijgen

4. Automatische genrebepaling met genre-classifier van de KB

Ga naar de genre-classifier van de KB: **http://www.kbresearch.nl/genre/**
Vervang de tekst in het venster links op het scherm door een zelfgekozen (uitgeknipte) tekst en klik op [Toon genre].
Bekijk de mate waarin het programma die tekst tot allerlei genres vindt behoren.
NB: Werkte recent niet meer

5. Automatische bepaling van het "leesbaarheidsniveau" van teksten

Wizenose is een organisatie die eenvoudig leesbare informatieve teksten verzamelt ten behoeve van een zoekmachine voor kinderen en laaggeletterden (en daarvoor tevens eenvoudig taalgebruik propageert). Wize-scan is hun test-tool om automatisch het leesbaarheidsniveau van teksten te bepalen.
Maak op **https://wizescan.com/** een account aan.
Je kunt nu Nederlandse of Engelse teksten laten analyseren door die in te plakken, het URL ervan op te geven of een tekstbestand te uploaden.
Kijk welke "niveaus" aan jouw teksten worden toegekend (en wat daarin de moeilijke woorden waren).

6. Nog een aantal producten waarvan je de demo's kunt uitproberen

Aylien	https://developer.aylien.com/text-api-demo
Data Harmony	http://demo.newsindexer.com/
IBM Watson	https://natural-language-understanding-demo.ng.bluemix.net/
Intellexer	http://demo.intellexer.com/
Lexalytics	https://www.lexalytics.com/demo
Meaning Cloud	https://www.meaningcloud.com/demo
PoolParty Power Tagging	https://drupal.poolparty.biz/powertagging
Text Razor	https://www.textrazor.com/demo

Oefening 3: Sentiment analysis

1. Voorbeelden van systemen voor "sentiment analysis"

Probeer enkele van deze systemen met je eigen tekst of zoekwoorden

- Python NLTK Text Classification: **http://text-processing.com/demo/sentiment/**
  Analyse van (in te plakken) stukken Engelse, Nederlandse of Franse tekst.
- Opinion Crawl: **http://www.opinioncrawl.com/**
  Nieuwsanalyse op basis van een zoekterm
- Social Searcher: **https://www.social-searcher.com/**
  Zoekt in diverse sociale media en toont onder "Detailed statistics" het "sentiment" van gevonden berichten voor elk van de doorzochte bronnen.
- Twinword: **https://www.twinword.com/api/sentiment-analysis.php**
  Analyse van ingeplakte (Engelse) tekst; toont scores op woord niveau.

2. Nog wat meer voorbeelden voor "sentiment analysis".

Overzichten van meer voorbeelden van systemen voor sentiment analysis o.a. in:
https://www.talkwalker.com/blog/best-sentiment-analysis-tools
https://www.softwareadvice.com/resources/free-twitter-sentiment-analysis-tools/
http://barnraisersllc.com/2017/01/best-tools-sentiment-analysis-free-fee/
https://www.brandwatch.com/blog/top-free-social-media-monitoring-tools/

- Verder heeft ook Google nog een tool, waarvoor je sinds kort echter eerst ingewikkeld een account moet aanmaken. Niettemin voor de diehards:
  Google's Cloud Natural Language API: **https://cloud.google.com/natural-language/**
  Analyse van stukken tekst, in te plakken in het venster onder "Try the API".
  Geeft niet alleen sentiment per alinea weer, maar haalt ook entities uit de tekst, doet aan zinsontleding en kent een onderwerpscategorie toe.

Oefening 4: Automatische beeldherkenning

1. Voorbeeld van automatische beeldherkenning

Probeer de Wolfram Image Identification: **https://www.imageidentify.com/**
Sleep daar met de muis een foto naar het witte vierkante venster. Dat kan zijn:

- een foto die je al op je computer (of een USB-stick) hebt staan en die je er uit de Windows Verkenner heen sleept,
- een foto die je met bijvoorbeeld Google Image gevonden hebt; daarvoor moet je dan Google Image in een apart venster openen, zodat je een foto daaruit naar het witte Wolfram vierkant kunt slepen.

2. Voorbeeld van automatisch aan foto's toegekende tags in Flickr

Ga naar de site van Flickr: **https://www.flickr.com/**
[NB: Soms blijk je geen toegang tot Flickr te krijgen als je niet bent ingelogd.]

- Doe daar een zoekactie en klik een van de gevonden foto's aan. Scroll vervolgens omlaag om de daaraan toegekende tags te zien te krijgen. De tags in de grijze blokjes zijn door de maker van de foto toegevoegd. Die met witte achtergrond zijn automatisch door Flickr toegevoegd.
- Beoordeel voor een paar foto's deze automatische tags.
- Je kunt dit bijvoorbeeld ook doen met foto's van VOGIN-IP-2018: **https://www.flickr.com/photos/voginacademie/sets/72157692116442312**

3. Nog twee andere voorbeelden van automatische beeldherkenning

- Google Cloud Vision demo: **https://cloud.google.com/vision/**
  Sleep een plaatje of foto in het venster dat verschijnt als je even naar beneden doorscrollt en klik aan dat je geen robot bent.
  [Je kunt ook direct naar **https://cloud.google.com/vision/docs/drag-and-drop***]*
  Kijk hoe uitgebreid (en correct?) elementen uit de afbeelding worden herkend.
- Imagga: **https://imagga.com/auto-tagging-demo**
  Sleep een plaatje of foto in het venster, klik aan dat je geen robot bent en klik op [Analyze]
  Beoordeel ook weer het resultaat.

Oefening 5: Automatisch termen ontlenen en automatisch samenvatten

1. Automatische term extractie

Ga naar **http://fivefilters.org/term-extraction/**
Zoek een stuk Engelse tekst of hergebruik een stuk dat je al bij voorgaande systemen hebt gebruikt. Plak dat in het daarvoor bedoelde venster van Fivefilters Term Extraction en kijk of de woorden en begrippen (termen) die uit de tekst "geëxtraheerd" worden, inderdaad de essentie daarvan representeren.

2. Automatische tekst summarizer

Ga naar **http://freesummarizer.com/** en kies "Summarize Text".
Zoek een stuk Engelse tekst of hergebruik een stuk dat je al bij voorgaande systemen hebt gebruikt. Plak dat in het daarvoor bedoelde venster van de summarizer en kijk wat die daarvan maakt.

Oefening 6: Unsupervised clustering

1. **Yippy metasearch: https://yippy.com**
   - Tik als zoekterm in *stones* en kijk welke clusters (clouds) gevormd worden uit de vele honderden resultaten die uit diverse zoekmachines zijn opgehaald.
   - Wat gebeurt er als je op bijv. de categorie (cloud) “rolling stones” klikt?
   - En als je op het +-teken voor bijvoorbeeld “kidney stones” klikt?
   - Doe ook nog eens zoekacties op bijvoorbeeld *bse*, *sinterklaas*, *jsf*, *climate* of *ajax* en kijk weer welke clusters uit de resultaten worden afgeleid.

2. **Carrot2 metasearch: http://search.carrot2.org/**
   - Tik hier ook een paar van de hierboven gebruikte zoektermen in en kijk welke clusters hier uit de resultaten gevormd worden.
   - Welke presentatievorm vind je overzichtelijker: Folders, Circles of Foam Tree?

Zelf wat doen?

Python leren:

- https://www.codecademy.com/learn/python
- https://www.coursera.org/learn/python (Python for everybody)
- https://www.coursera.org/learn/python-machine-learning (applied machine learning in Python)
- http://www.karsdorp.io/python-course/ (Python for the Humanities)

Machine learning in Python met sklearn:

- http://scikit-learn.org/stable/tutorial/basic/tutorial.html

Text processing and categorization in Python:

- http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
- https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a

Toolbox/library for text classification, natural language understanding, machine translation, image classification etc. to train and test deep learning models - Ludwig

- https://uber.github.io/ludwig/

Meer lezen over de techniek?

[Een ongeordende, niet geëvalueerde lijst]

- Face Recognition for Beginners (2019) - https://towardsdatascience.com/face-recognition-for-beginners-a7a9bd5eb5c2
- Automate Twitter Sentiment Analysis using Zapier and Watson (no coding required) (2019) - https://medium.com/ibm-watson/automate-twitter-sentiment-analysis-using-zapier-and-watson-no-coding-reqd-406aabd8ee66
- Simply deep learning: an effortless introduction (2019) - How to join the deep learning conversation in mere minutes: conquering neural networks for newbies, novices, and neophytes - https://towardsdatascience.com/intro-to-deep-learning-c025efd92535
- The Most Intuitive and Easiest Guide for Artificial Neural Network (2019) - Demystifying neural networks for complete starters - https://towardsdatascience.com/the-most-intuitive-and-easiest-guide-for-artificial-neural-network-6a3f2bc0eecb
- Everything you need to know about Neural Networks and Backpropagation (2019) - Machine Learning Easy and Fun - Neural Network explanation from the ground including understanding the math behind it - https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a
- The Naïve Bayes Classifier (2018) - https://towardsdatascience.com/the-naive-bayes-classifier-e92ea9f47523
- Everything you need to know about Neural Networks and Backpropagation (2019) - Machine Learning Easy and Fun - Neural Network explanation from the ground including understanding the math behind it - https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a
- Deep Learning (MIT Technology Review, 2018) - With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart - https://www.technologyreview.com/s/513696/deep-learning/
- EXAM—State-of-The-Art Method for Text Classification (2018) - https://neurohive.io/en/state-of-the-art/exam-text-classification/
- Multi-Class Text Classification Model Comparison and Selection (2018) - Natural Language Processing, word2vec, Support Vector Machine, bag-of-words, deep learning - https://towardsdatascience.com/multi-class-text-classification-model-comparison-and-selection-5eb066197568
- Bayes' Theorem: The Holy Grail of Data Science (2018) - Intuitive derivation of the Bayes' Theorem - https://towardsdatascience.com/bayes-theorem-the-holy-grail-of-data-science-55d93315defb
- Sentiment Analysis of Tweets using Multinomial Naive Bayes (2018) - https://towardsdatascience.com/sentiment-analysis-of-tweets-using-multinomial-naive-bayes-1009ed24276b
- Text Classification with State of the Art NLP Library (2018) - Flair - A new version of Flair - simple Python NLP library has just been released by Zalando Research! - https://towardsdatascience.com/text-classification-with-state-of-the-art-nlp-library-flair-b541d7add21f
- Document Classification Using Machine Learning (2018) - https://medium.com/mlrecipies/document-classification-using-machine-learning-f1dfb1171935

- Building a text classification model with TensorFlow Hub and Estimators (2018) - https://medium.com/tensorflow/building-a-text-classification-model-with-tensorflow-hub-and-estimators-3169e7aa568
- Automated Keyword Extraction from Articles using NLP (2018) - https://medium.com/analytics-vidhya/automated-keyword-extraction-from-articles-using-nlp-bfd864f41b34
- Sentiment Analysis with Text Mining (2018) - Learn how to prepare text data and run two different classifiers to predict the sentiment of tweets. - https://towardsdatascience.com/sentiment-analysis-with-text-mining-13dd2b33de27
- Text Classification: Applications and Use Cases (2018) - https://towardsdatascience.com/text-classification-applications-and-use-cases-beab4bfe2e62
- Named Entity Recognition (NER) with keras and tensorflow (2018) - Meeting Industry's Requirement by Applying state-of-the-art Deep Learning Methods - https://towardsdatascience.com/named-entity-recognition-ner-meeting-industrys-requirement-by-applying-state-of-the-art-deep-698d2b3b4ede
- Multi-Class Text Classification with SKlearn and NLTK in python (2018) - A Software Engineering Use Case - https://towardsdatascience.com/multi-class-text-classification-with-sklearn-and-nltk-in-python-a-software-engineering-use-case-779d4a28ba5
- The Seductive Diversion of 'Solving' Bias in Artificial Intelligence (2018) - https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53
- Tutorial on Text Classification (NLP) using ULMFiT and fastai Library in Python (2018) - https://www.analyticsvidhya.com/blog/2018/11/tutorial-text-classification-ulmfit-fastai-library/
- Named Entity Recognition and Classification with Scikit-Learn (2018) - How to train machine learning models for NER using Scikit-Learn's libraries - https://towardsdatascience.com/named-entity-recognition-and-classification-with-scikit-learn-f05372f07ba2
- Machine Learning, NLP: Text Classification using scikit-learn, python and NLTK (2017) - https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a
- How Do Machine Learning Programs "Learn"? (2017) - In this article, we look at two machine learning (ML) techniques, Naive Bayes classifier and neural networks, and demystify how they work. - https://medium.com/iotforall/how-do-machine-learning-programs-learn-215303338d7 / https://www.leverege.com/blogpost/machine-learning-naive-bayes-neural-networks