TEXT MINING

L01. INTRODUCTION

SUZAN VERBERNE 2021

Universiteit Leiden

COURSE INFORMATION

- Brightspace page: https://brightspace.universiteitleiden.nl/d2l/home/91534
  - Once you registered for the course in uSis you are automatically subscribed to the course in Brightspace
- Course web page: http://tmr.liacs.nl/TM.html
- Lectures:
  - Wednesday, 9.15-11.00
  - Sitterzaal, Huygens/Oort & online on https://weblectures.leidenuniv.nl/Mediasite/Channel/textmining

CONTACT INFORMATION

- dr. Suzan Verberne
  http://liacs.leidenuniv.nl/~verbernes/
- Teaching assistants:
  - Michiel van der Meer (PhD student)
  - Cheyenne Heath (master student)
  - Hainan Yu (LIACS TA)
  - Juan Bascur Cifuentes (LIACS TA)
- Contact: tmcourse@liacs.leidenuniv.nl

WHO ARE YOU?

Quick round (raise hands): what is your master program?

- Computer Science
- Artificial Intelligence
- Data Science
- Bio-informatics
- Media Technology
- ICT in Business and the Public Sector
- Other

WHO ARE YOU?

Quick round (raise hands)

- Who has taken a course in Data Mining or Machine Learning?
- Who can program in Python?
- Who knows what the vector space model is?
- Who could explain the difference between supervised and unsupervised learning?
- Who has experience with deep neural networks?

TODAY'S LECTURE

- Course goals
- Why text mining
- What is text mining
- Challenges of text data
- The text mining process
- Structure of this course

COURSE GOALS

COURSE GOALS

- https://studiegids.universiteitleiden.nl/courses/105087/text-mining
- You will learn about:
  - fundamentals of models (conceptual understanding)
  - practical applications
  - data, experimentation, evaluation
  - challenges and limitations

COURSE LITERATURE

- The majority of the chapters comes from this book:
  - Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd ed), December 2020 https://web.stanford.edu/~jurafsky/slp3/
  - And a number of papers / chapters from other sources
- The literature will be distributed on Brightspace, as are the slides

RELATED COURSES (SPRING SEMESTER)

- Information Retrieval
  https://studiegids.universiteitleiden.nl/courses/105168/information-retrieval
- Advanced topics in Deep Learning for Natural Language Processing
  https://studiegids.universiteitleiden.nl/courses/107995/advanced-topics-in-deep-learning-for-natural-language-processing
- Advances in Deep Learning
  https://studiegids.universiteitleiden.nl/courses/105089/advances-in-deep-learning

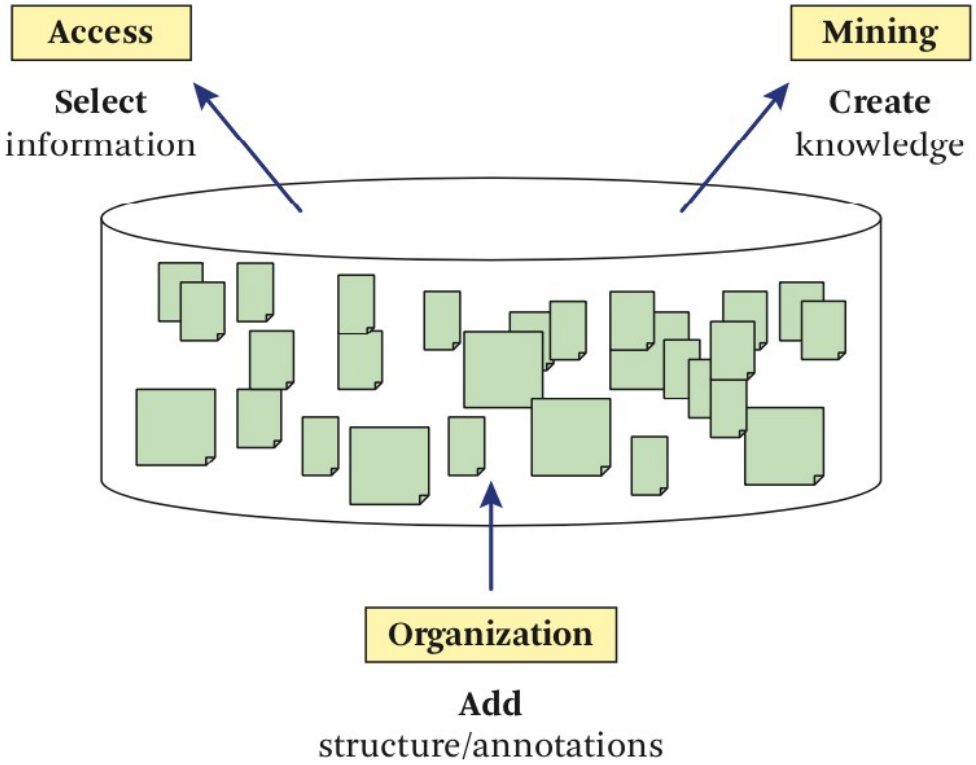
WHAT IS TEXT MINING

WHY TEXT MINING?

- A large portion of the world's knowledge is stored in text:
  - web pages
  - user-generated content on the web (social media)
  - electronic health records
  - scientific literature
  - patents
  - political/legal texts

WHAT IS TEXT MINING

- Text mining: Automatic extraction of knowledge from text
- Text = unstructured
- Knowledge = structured



Zhai & Massung (2016)

TEXT MINING AND DATA MINING

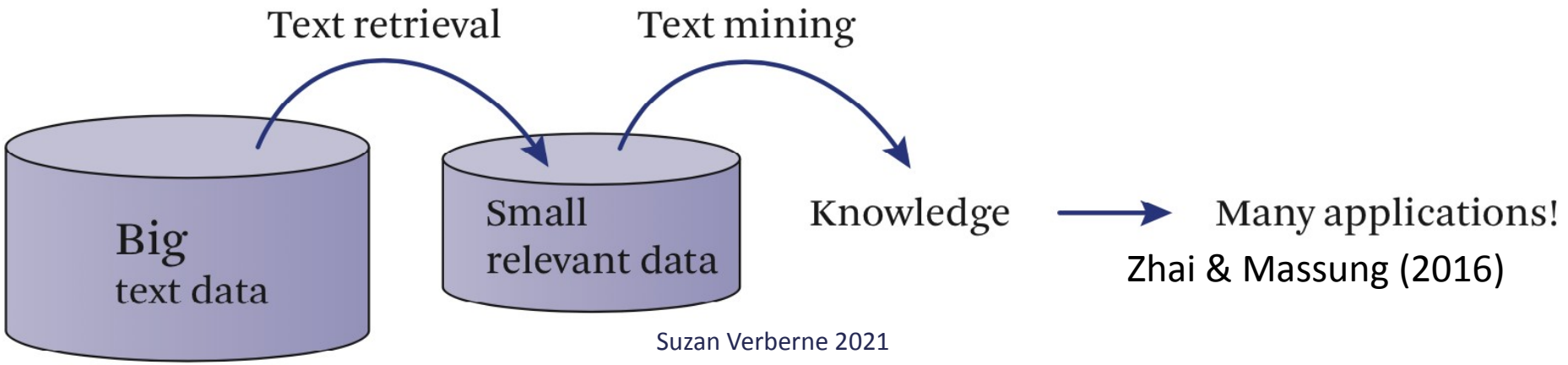
- Text mining is a form of data mining
- Many of the learning methods are similar
  - Classification
  - Clustering
- But text data is unstructured
- And requires text-specific processing
- We will see the specifics of text data later

TEXT MINING AND NLP

- NLP = Natural Language Processing
- Text Mining applications use NLP methods
- NLP is a large and active research field
- NLP has a fundamental component (computational linguistics)
- Current NLP methods heavily rely on deep neural networks
- Not all NLP tasks are TM tasks
  - e.g. Machine translation, Speech recognition
- Check http://nlpprogress.com/ for an overview of NLP tasks and the state-of-the-art methods for each task

TEXT MINING AND INFORMATION RETRIEVAL

- Text Mining (TM) and Information Retrieval (IR) are related disciplines
- In many applications, IR is the first step of the TM process
- First retrieve documents (IR), then extract and structure the relevant information



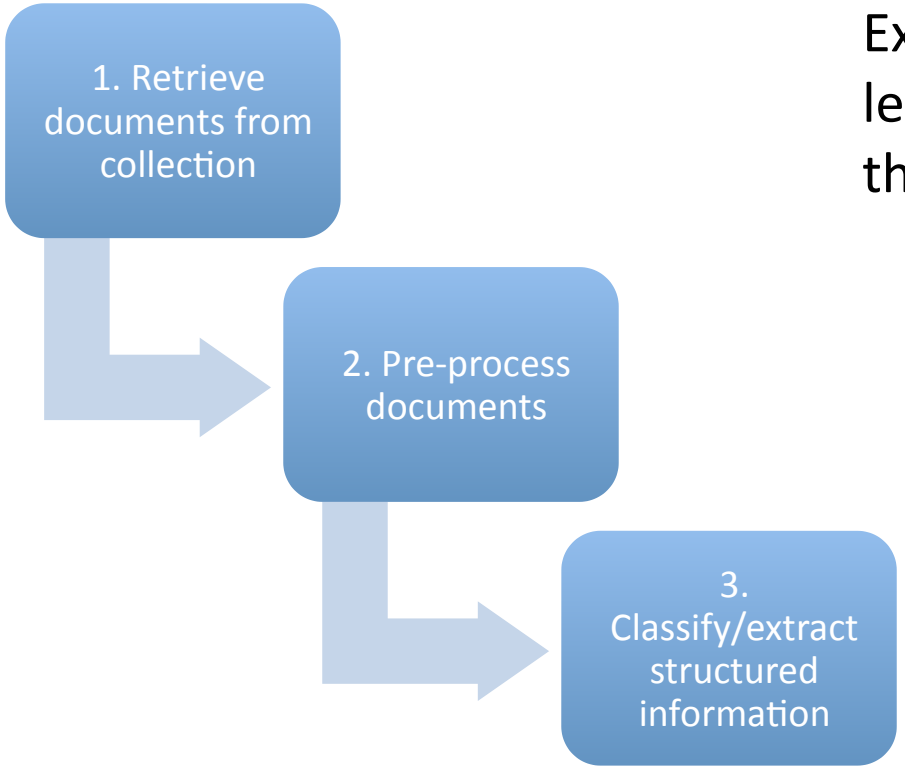
Zhai & Massung (2016)

THE TEXT MINING PIPELINE

Universiteit Leiden

Suzan Verberne 2021

THE TEXT MINING PIPELINE



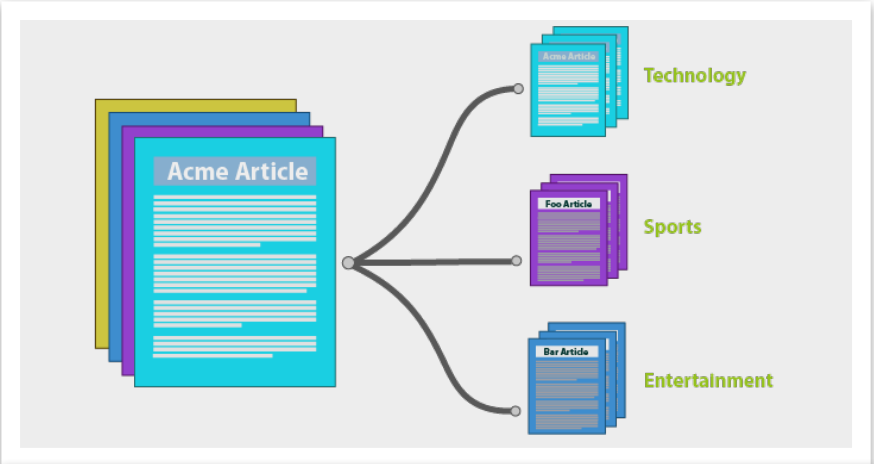
Example TM problem: estimate the level of support on social media for the use of mouth masks

1. IR: retrieve tweets that mention mouth masks
2. Pre-processing: Filter duplicates. Clean from noise. Anonymize if necessary
3. NLP: classify all messages in pro/against/neutral with respect to mouth masks

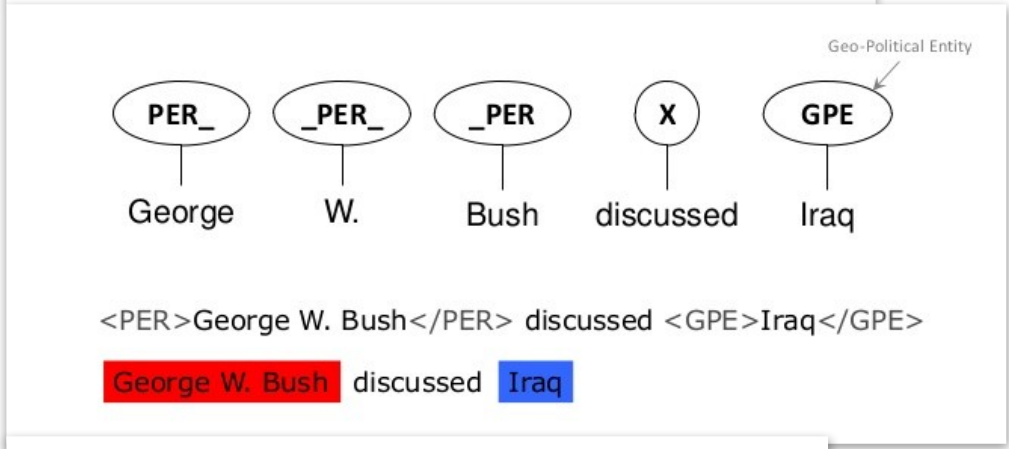
TYPES OF TEXT PROCESSING TASKS

- Roughly, we distinguish three types of text mining tasks:
  1. Text classification/clustering: assign a category or cluster per document
     - the 'document' can be any text type (newspaper article, tweet, e-mail, text message, patent, …)
     - the 'category' can be any type of label (topic, relevance indicator, author, sentiment, …)
  2. Sequence labelling: assign a category per word in a text
     - e.g. label the person names, dates and places in a text (named entity recognition)
  3. Text-to-text generation: input is text, output is text
     - summarization, translation

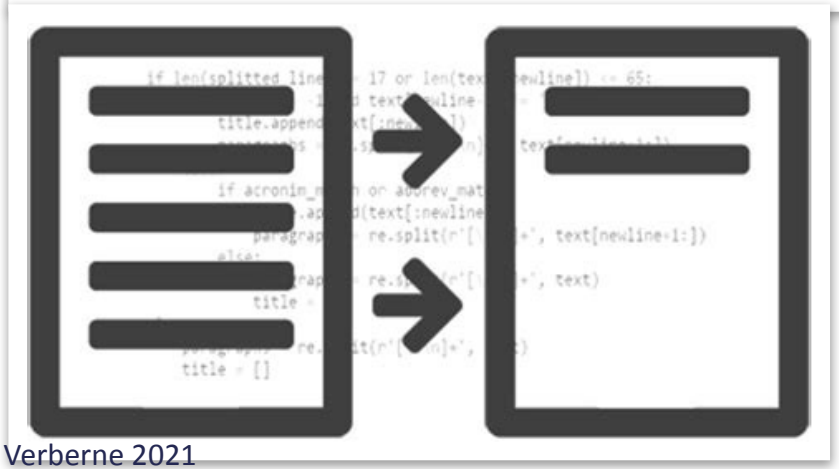
1. Text classification



2. Named entity recognition
(= sequence labelling)



3. Summarization
(= sequence-to-sequence)



CASE

- Goal: to discover side effects for hypertension medications
- Data:
  - 39,892 messages from a patient discussion forum on hypertension
- How would you address this problem? Discuss in small groups

worsening symptoms since starting medication

Follow

Posted 2 weeks ago, 5 users are following.

dean89033

Hi all. I started out at 184/100, and was put on 40mg Lisinopril in January. Almost immediately I started to suffer with vertigo and dizzy spells where my hearing would cut out, cold sweats and fatigue. I knew that the first two weeks there would be some expected side effects, so I waited to see if they would pass. About a month later, I went back to my doctor, with a BP reading of 122/78, and she swapped me to Losartan Potassium 25mg. The same side effects continued, and after another month or so I went back (BP 126/90) and was swapped to Amlodipine Besylate 5mg. I've been on that since March, as there aren't any other medications my doctor can switch me to without going to beta blockers. But my symptoms have worsened.

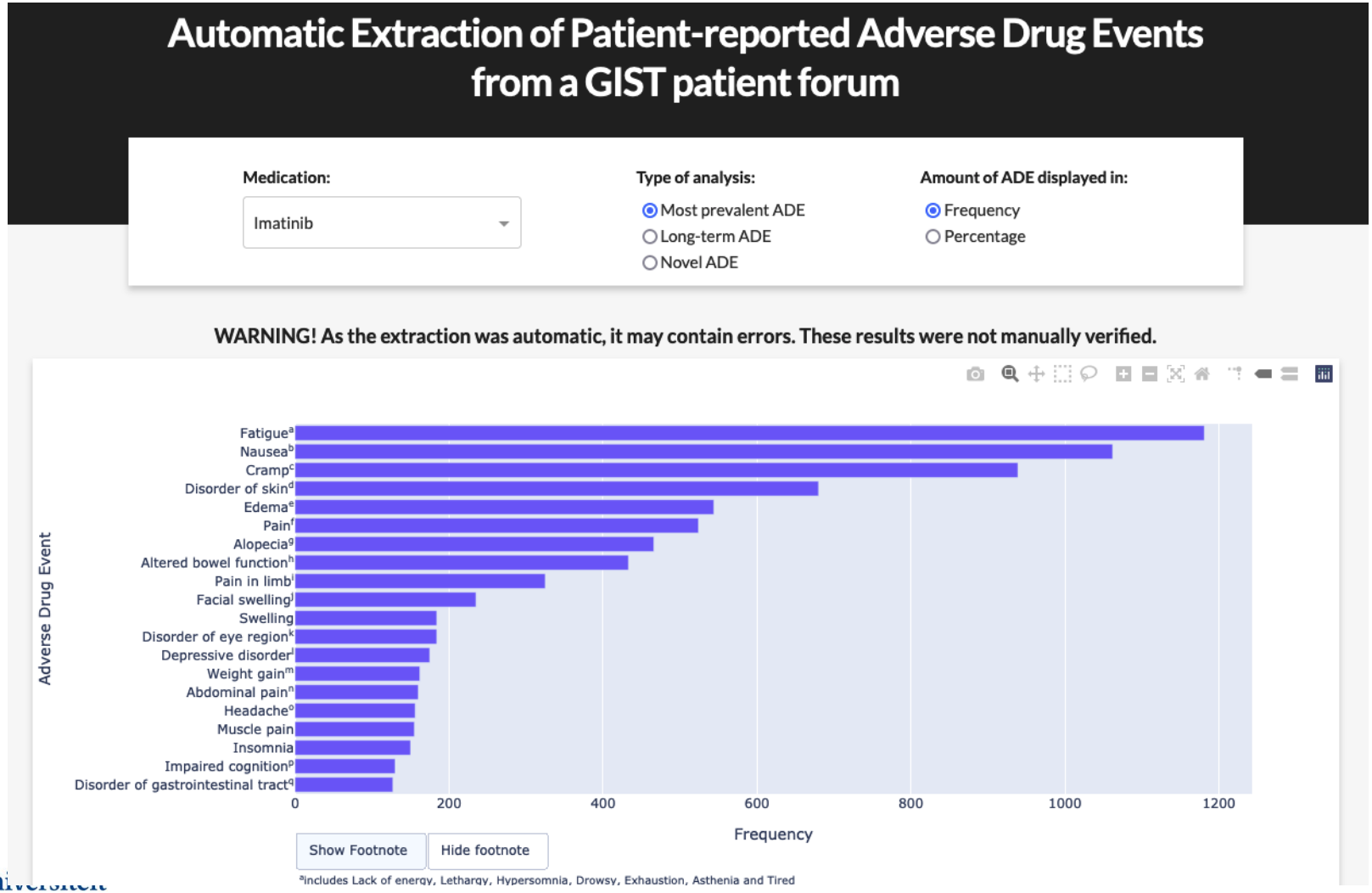
It's like spacing out but worse? But also not like dissociating. For 10 or 15 seconds, I'm not "there" but I have the after image of whatever I was looking at before. Sometimes my eyes cross and I can't un-cross them, or sometimes they close and I can't keep them open, what normally brings me "back" is that I'll sway too much to one side, or my head will jerk down. It used to be that parts of my body would jerk but not so much now. I wouldn't say I'm confused? I know who/what/where I am but I don't know what I was doing/am doing/should do next. It doesn't feel like falling asleep, it feels like my brain lagged out, or I'm behind a loading screen in a video game. I've been tracking them and can't find any triggers.

I've had an MRI and EEG, and I'm waiting for the follow-up appointment in September to go over those results. I've also got a consultation with a cardiologist in October. I got a CPAP machine a few months ago with a diagnose of sleep apnea, but with the meds and stress from my symptoms it's hard to say if that's helped my BP any. We're trying to be thorough and check all the bases, but I

WHAT DID YOU COME UP WITH?

- Filter the data? (Retrieve relevant messages)
- Process the data?
- Identify medication names? External knowledge needed?
- Identify side aspects? Relations between medications and side effects?
- What human input would you need?

EXTRACTING SIDE EFFECTS FROM PATIENT EXPERIENCES: RESULTS



https://dashboard-gist-adr.herokuapp.com/

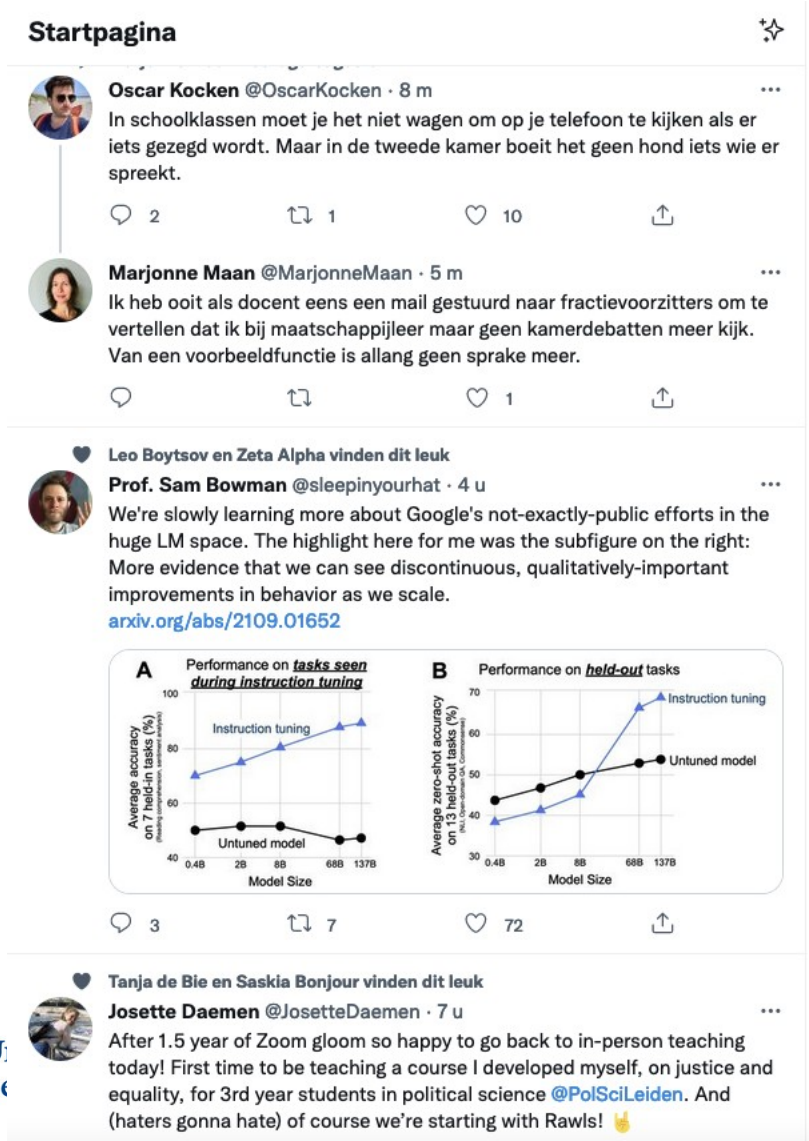
CHALLENGES OF TEXT DATA

1. TEXT DATA IS UNSTRUCTURED

- Or at best semi-structured:
  - PHYSICAL EXAMINATION: On physical examination, her blood pressure was 104/73, pulse 79. In general, she was a woman in no acute distress. HEENT: Nonicteric. Pupils are equal, round, and reactive to light. Extraocular movements are full. Pharynx is benign. Tongue midline. Neck is supple.

Parks and Recreation (also known as ***Parks and Rec***) is an American political satire mockumentary sitcom television series created by Greg Daniels and Michael Schur. The series aired on NBC from April 9, 2009, to February 24, 2015, for 125 episodes, over seven seasons. A special reunion episode aired on April 30, 2020.[1][2][3][4] The series stars Amy Poehler as Leslie Knope, a perky, mid-level bureaucrat in the Parks Department of Pawnee, a fictional town in Indiana. The ensemble and supporting cast features Rashida Jones as Ann Perkins, Paul Schneider as Mark Brendanawicz, Aziz Ansari as Tom Haverford, Nick Offerman as Ron Swanson, Aubrey Plaza as April Ludgate, Chris Pratt as Andy Dwyer, Adam Scott as Ben Wyatt, Rob Lowe as Chris Traeger, Jim O'Heir as Garry "Jerry" Gergich, Retta as Donna Meagle, and Billy Eichner as Craig Middlebrooks.

2. TEXT DATA CAN BE MULTI-LINGUAL



- (which means that we have to pre-filter it, especially when keywords have meanings in multiple languages)

3. TEXT DATA IS NOISY

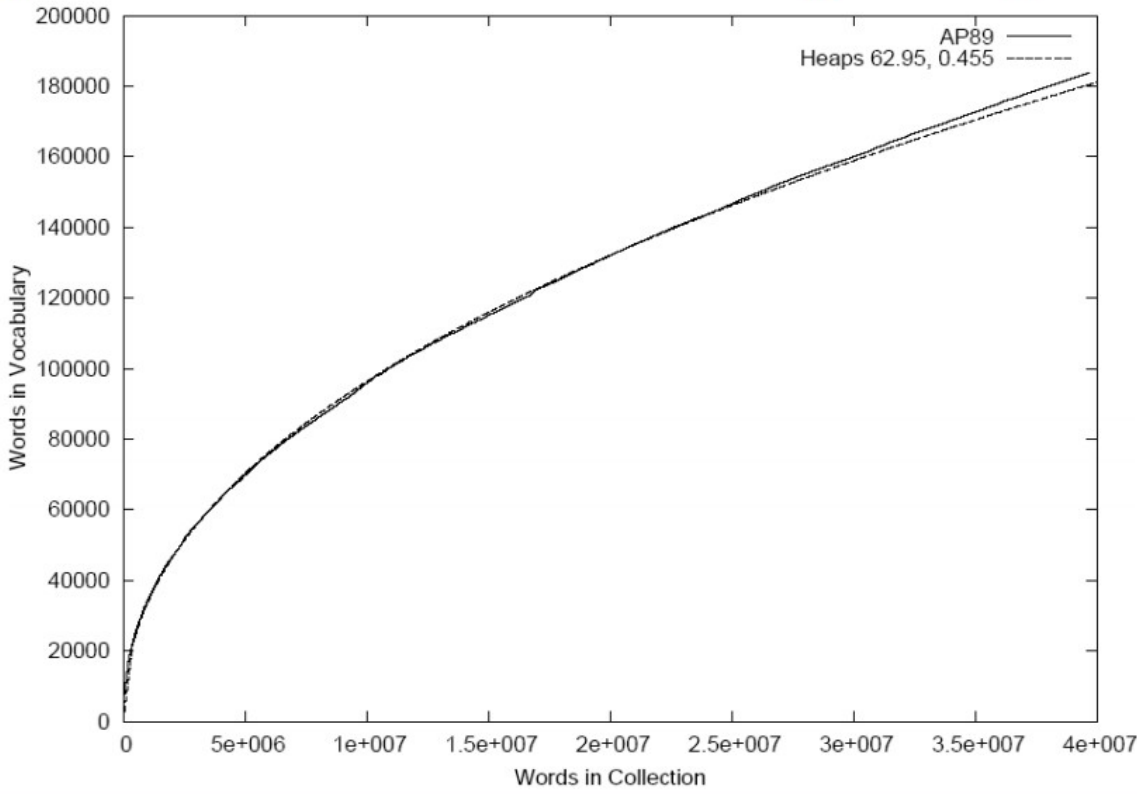
- Noisy encoding and typography might give challenges in processing
- Noisy attributes: spelling errors, OCR errors

```
  <p top="516" left="535" docno="test.VP_1977.0057.027">
   Invoering van het ' Pensioenwo
  </p>
  <p top="685" left="519" docno="test.VP_1977.0057.028">
   ningplan 0'66' (zie Sociale zekerheid) . Volgens dit plan krijgt iedereen het recht van de spaar- o
 verzekeringsinstelling die zijn pensioenbesparingen beheert, deze gespaarde gelden in de vorm van een
ypotheek voor de aan koop van een eigen huis terug te lenen . Hierbij zullen waardevaste hypotheeklenin
en worden verstrekt met een lage (maar reÃ«le) rente; dit leidt tot lage beginwoonlasten, waardoor zelf
beperking van algemene overheidssubsidies voor de nieuwbouw mogelijk wordt.
  </p>
  <p top="696" left="519" docno="test.VP_1977.0057.029">
   b
  </p>
  <p top="721" left="519" docno="test.VP_1977.0057.030">
   Evenals voor huurders: invoering van individuele woonsubsidies voor eigenaar-bewoners .
  </p>
  <p top="841" left="519" docno="test.VP_1977.0057.031">
   Opening van de mogelijkheid woningbouwstichtingen, woningbouwverenigingen en particuliere huurverho
dingen om te zetten in coÃ¶peratieve veren igingen van eigenaar-bewoners. Daartoe dienen groepen huurde
 een aankooprecht te krijgen . De bewoners worden bij deze vormen van bewoners-zelfbestuur eigenaar van
hun woning en beslissen in princi
  </p>
  <p top="853" left="520" docno="test.VP_1977.0057.032">
   pe zelf over indeling, afwerking , aan
  </p>
  <p class="footer" top="937" left="337" docno="test.VP_1977.0057.033">
   0'66 : 7-9
  </p>
 </page>
```

4. LANGUAGE IS INFINITE

- A new document in your collection is likely to add new terms ⇏ new dimensions

- The number of new words will increase very rapidly when the corpus is small and would continue to increase indefinitely, but at a slower rate for larger corpus (Heaps' Law)



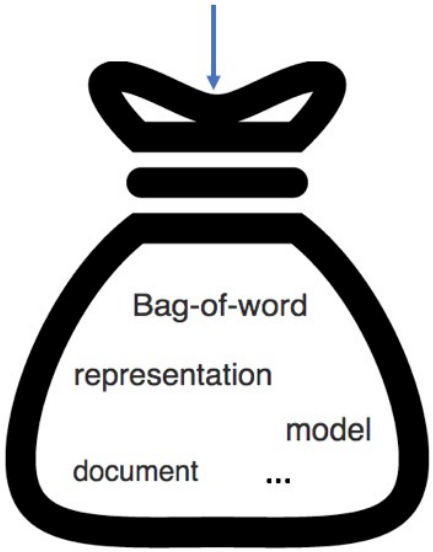
5. TEXT DATA IS SPARSE

- Important distinction here:
  Text as classification object vs. text as sequence

  - As classification object, text is sparse
  - Traditional text classification methods represent the text as a 'bag of words'
  - In the bag-of-words model, each word in the collection becomes a feature

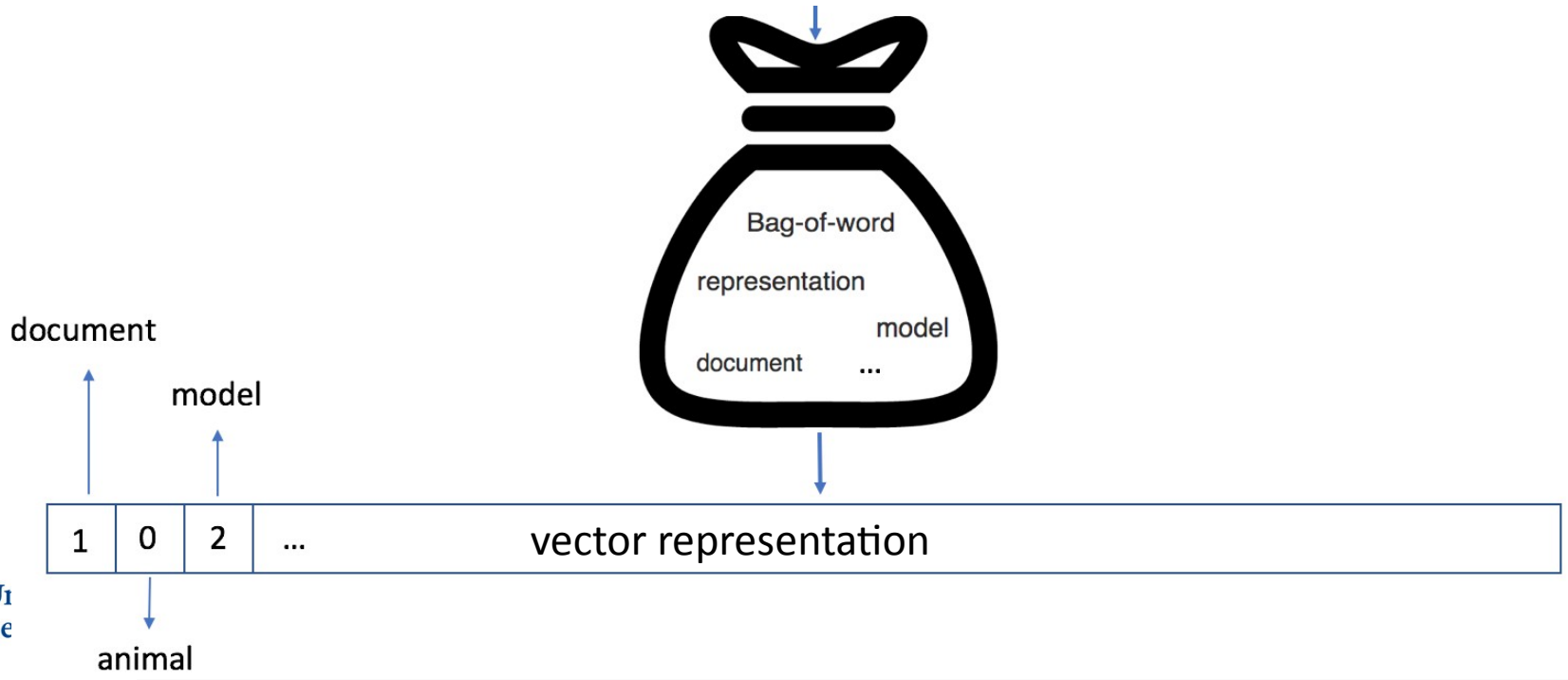
TEXT AS CLASSIFICATION OBJECT

Bag-of-word model is an orderless document representation
likes movies too", the bag-of-words representation will not re
gram model can be used to store this spatial information with
store the term frequency of each unit as before.



TEXT AS CLASSIFICATION OBJECT

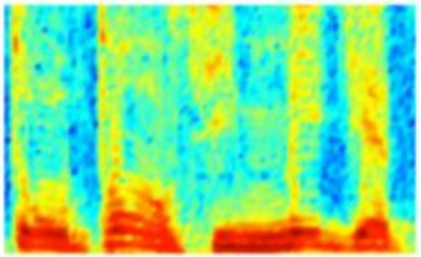
- Traditional Bag-of-words model:
  - Word order is not relevant
  - Punctuation is not relevant
  - Sentence and paragraph borders are not relevant



TEXT AS CLASSIFICATION OBJECT

- Typically, we use words as features in text classification
- Only a few of all words occur in a given document
- Hence, text vectors are sparse vectors

AUDIO



Audio Spectrogram

DENSE

IMAGES

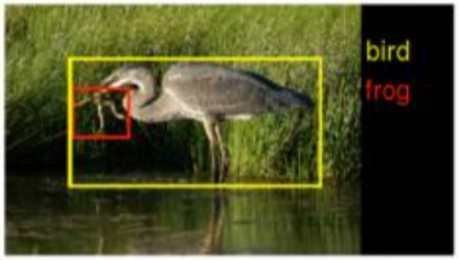


Image pixels

DENSE

TEXT

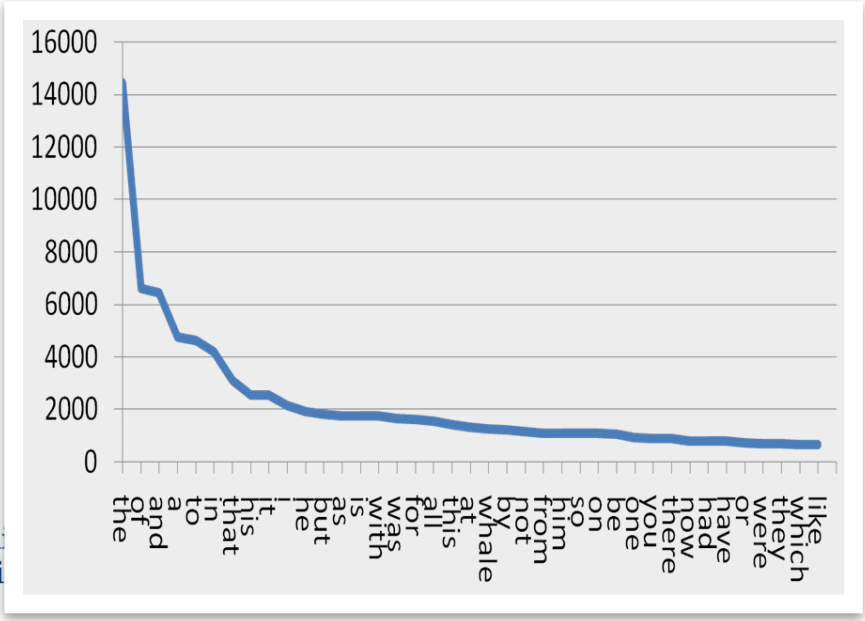
0	0	0	0.2	0	0.7	0	0	0	...	...

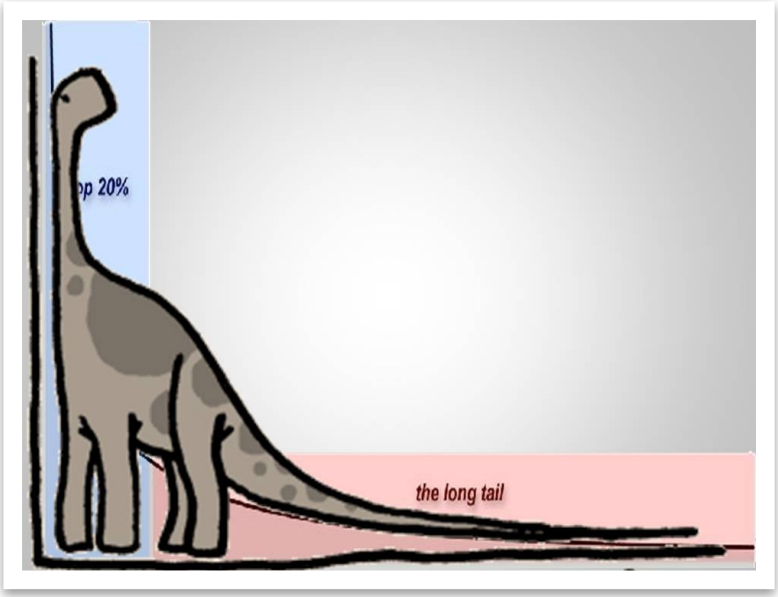
Word, context, or document vectors

SPARSE

ZIPF'S LAW

- Given a text collection, the frequency of any word is inversely proportional to its rank in the frequency table
- In English, the top four most frequent words are about 10-15% of all word occurrences. The top 50 words are 35-40% of word occurrences.





6. TEXT DATA IS SEQUENTIAL DATA

- Important distinction:
  Text as classification object vs. text as sequence
- If we want to extract knowledge from text, word order (sequence), punctuation and capitalisation do matter

6. TEXT DATA IS SEQUENTIAL DATA

- E.g. names, dates, and titles from biographical text:

Daisy Jazz Isobel Ridley (born 10 April 1992) is an English actress who rose to international prominence through playing the role of Rey in the Star Wars sequel trilogy: The Force Awakens (2015), The Last Jedi (2017), and The Rise of Skywalker (2019).

EVALUATION OF TEXT MINING METHODS

Universiteit Leiden

Suzan Verberne 2021

EVALUATION OF TEXT MINING

- Evaluation of complete application (extrinsic evaluation):
  - human vs. automatic
  - are humans helped/satisfied by the results?
- Evaluation of the components (intrinsic evaluation): ground truth labels needed
  - Existing labels in the data
  - Human-assigned labels in the data

EVALUATION OF TEXT MINING

- Evaluation metrics:
  - accuracy
  - precision
  - recall

- precision: proportion of the assigned labels that are correct
- recall: proportion of the relevant labels that were assigned

PRECISION AND RECALL

A = set of labels assigned by algorithm

T = set of true labels

$$\text{Precision} = \qquad \text{Recall} =$$

This will come back in many lectures (with specific definitions for each task)

PRECISION AND RECALL

A = set of labels assigned by algorithm

T = set of true labels

$$\text{Precision} = \frac{|A \cap T|}{|A|} \qquad \text{Recall} = \frac{|A \cap T|}{|T|}$$

This will come back in many lectures (with specific definitions for each task)

PRECISION AND RECALL EXAMPLE

- Think of spam classification as example task: messages are classified as either spam or no-spam
- We can measure accurracy: what proportion of messages is correctly labeled.
- But there are two ways the label can be wrong:

PRECISION AND RECALL EXAMPLE

- But there are two ways the label can be wrong:
  - a spam message ends up in the inbox
  - a non-spam message ends up in the spambox
- Precision and Recall measure these 2 evaluation aspects
  - precision of the 'spam class': what proportion of the messages in the spam box were indeed spam
  - recall of the 'spam class': what proportion of the true spam messages were correctly put in the spam box
  - (and you can also measure the precision and recall of the 'no spam' class)

COURSE STRUCTURE

COURSE OUTLINE

- Course website: http://tmr.liacs.nl/TM.html

Week	Lecture	Literature	Exercise / assignment
1 (8 Sept)	Introduction		
2 (15 Sept)	Text processing	J&M chapter 2. Regular Expressions, Text Normalization, Edit Distance	Exercise: Chapter 1 of "Advanced NLP with Spacy"
3 (22 Sept)	Vector Semantics	J&M chapter 6. Vector Semantics	Exercise: Word Embedding Tutorial: Word2vec with Gensim
4 (29 Sept)	Text categorization	J&M chapter 4.1-4.3. Naive Bayes Classification	Exercise: Text classification tutorial (sklearn)
5 (6 Oct)	Data collection and annotation	Finin (2010). Annotating Named Entities in Twitter Data with Crowdsourcing McHugh (2012). Interrater reliability: the kappa statistic	**Assignment 1. Text classification** (deadline 18 Oct)
(13 Oct)	No lecture		
6 (20 Oct)	Information Extraction	J&M chapter 18. Information Extraction	Exercise: Sequence labelling tutorial (crfsuite)
(26 Oct)	No lecture		
7 (3 Nov)	Neural NLP and transfer learning	J&M chapter 7. Neural Nets and Neural Language Models	Exercise: BERT Fine-Tuning with Huggingface
8 (10 Nov)	Text summarization	Kryściński et al (2019). Neural Text Summarization: A Critical Evaluation	**Assignment 2. Information Extraction** (deadline 15 Nov)
9 (17 Nov)	Sentiment analysis		Exercise: Sentiment analysis with BERT
10 (24 Nov)	Biomedical text mining	Lee et al. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining	
11 (1 Dec)	Industrial Text Mining: guest lecture	Paper reading for the final assignment	
12 (8 Dec)	Conclusions		**Final assignment**: multiple topics to choose from (deadline 16 Jan)
(13 Jan)	Exam		
(4 Feb)	Re-sit		

GENERAL STRUCTURE

- 12 lecture weeks
- Homework:
  - Literature after the lecture
  - In some weeks you work on a practical exercise (online tutorial)
  - In other weeks you work on an assignment that you need to submit (2 smaller assignments, and one large assignment)

EXAM AND GRADE

- The assessment of the course consists of
  - a written exam (50% of course grade)
  - practical assignments (50% of course grade)
    - Assignment 1 (10%): text classification
    - Assignment 2 (10%): information extraction
    - Assignment 4 (30%): multiple topics to choose from
- Groups: make teams of 2 students

DEADLINES

- All assignments will be submitted and graded through Brightspace. A TA will provide you with feedback
- Each assignment has a re-take opportunity,
  - but when submitted after the first deadline your maximum grade is 6

	Deadline	Re-sit deadline
Assignment 1	18 October	January 16 (maximum grade 6)
Assignment 2	15 November	January 16 (maximum grade 6)
Final assignment	16 January	6 February (maximum grade 6)
Written exam	13 January	4 February

EXAM AND GRADE

- Passing the course:
  - The grade for the written exam should be 5.5 or higher in order to complete the course.
  - The weighted average grade for the practical assignments should be 5.5 or higher in order to complete the course.
  - If a task is not submitted the grade for that task is 0.

CONCLUSIONS

SUZAN VERBERNE 2021

Universiteit Leiden

HOMEWORK

- Find a partner for the practical assignments
  - Enroll in a group on Brightspace (Groups -> Assignments) with your team mate
- (optional) If you want to improve your Python programming skills:
  - 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)

AFTER THIS LECTURE…

- You know what to expect from this course (both content and structure)
- You can explain the relation between text mining and data mining
- You can explain the relation between text mining and information retrieval
- You can explain the relation between text mining and natural language processing
- You can list and explain the most important challenges of text data
- You can describe the text mining process on a high level
- You can explain the difference between tasks that represent text as classification object and tasks that represent text as sequence