Introduction

Applied Text Mining

Ayoub Bagheri

Contents

Very useful

You can access the course materials quickly from

https://ayoubbagheri.nl/applied_tm/

Some guidelines

1- Please keep your microphone off

2- If you have a question, raise your hand or type your question in the chat

3- You may always interrupt me

4- We will introduce frequent question breaks

Lecturers and assistants

Dong Nguyen

Berit Janssen

Anastasia Giachanou

Nikos Bentis

Jelle Teijema

Program

```
## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.

## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.

## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.

## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.

## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.

## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.

## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
```

```
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.

## Warning in latex_new_row_builder(target_row, table_info, bold, italic,
## monospace, : Setting full_width = TRUE will turn the table into a tabu
## environment where colors are not really easily configable with this package.
## Please consider turn off full_width.
```

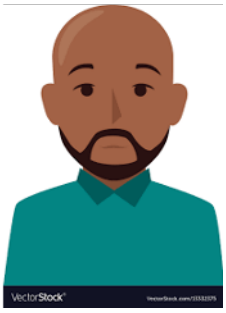
Time	Monday	Tuesday	Wednesday	Thursday
9:00 - 10:30	**Lecture 1**	**Lecture 3**	**Lecture 5**	**Lecture 7**
	Break	Break	Break	Break
10:45 – 11:45	Practical 1	Practical 3	Practical 5	Practical 7
11:45 – 12:15	Discussion 1	Discussion 3	Discussion 5	Discussion 7
	Lunch	Lunch	Lunch	Lunch
13:45 – 15:15	**Lecture 2**	**Lecture 4**	**Lecture 6**	**Lecture 8**
	Break	Break	Break	Break
15:30 – 16:30	Practical 2	Practical 4	Practical 6	Practical 8
16:30 – 17:00	Discussion 2	Discussion 4	Discussion 6	Discussion 8

Goal of the course

- Text data is everywhere!
- A lot of world's data is in unstructured text format
- The course teaches
  - text mining techniques
  - using Python
  - on a variety of applications
  - in many domains.

What is Text Mining?

Text mining in an example



- This is **Garry**!
- **Garry** works at Bol.com (a webshop in the Netherlands)
- He works in the dep of **Customer relationship management**.
- He uses Excel to read and search customers' reviews, extract aspects they wrote their reviews on, and identify their sentiments.
- Curious about his job? See two examples!

This is a nice book for both young and old. It gives beautiful life lessons in a fun way. Definitely worth the money!

+ Educational

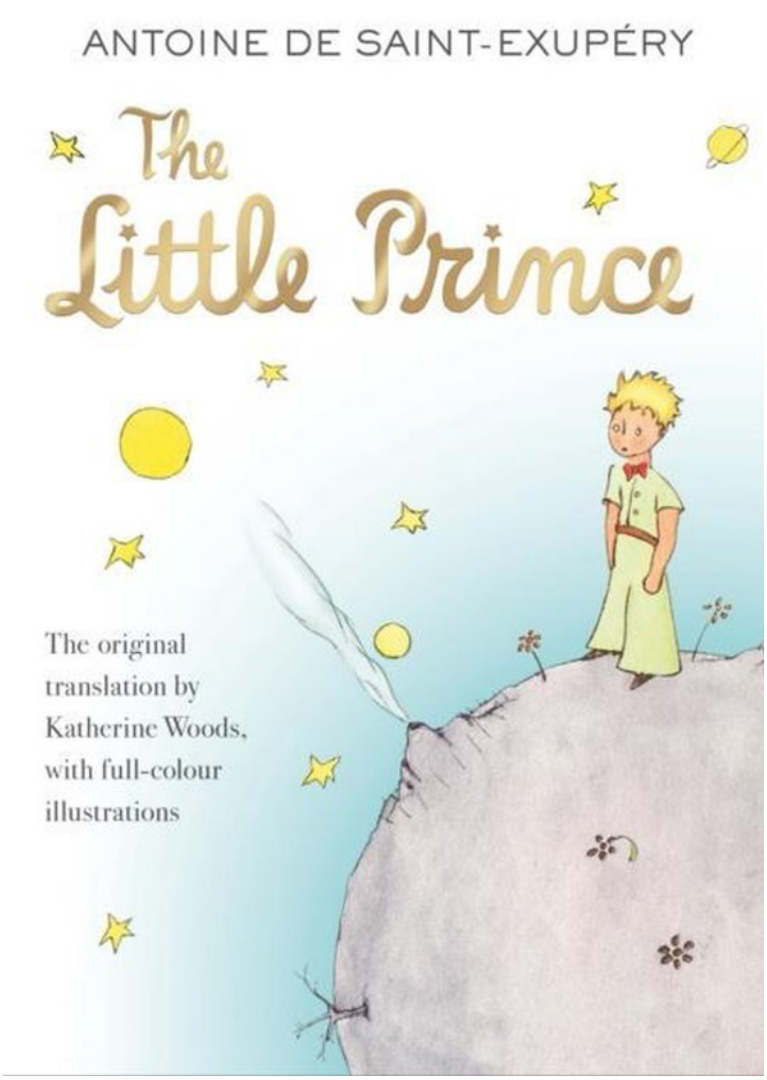
+ Funny

+ Price

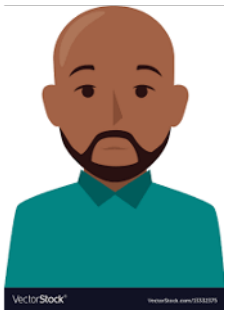
Nice story for older children.

+ Funny

- Readability

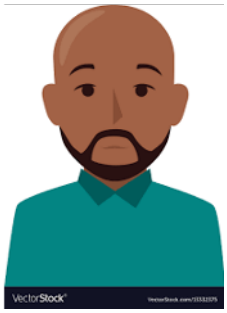


Example



- Garry likes his job a lot, but sometimes it is frustrating!
- This is mainly because their company is expanding quickly!
- Garry decides to hire **Larry** as his assistant.

Example



- Still, a lot to do for two people!
- Garry has some budget left to hire another assistant for couple of years!
- He decides to hire **Harry** too!
- Still, manual labeling using Excel is labor-intensive!

Challenges?

- Can you guess what are the challenges Garry, Larry, and Harry encounter in doing their job, when working with text data?
  - Go to www.menti.com and use the code 7338 2184

Challenges with Text Data

Challenges with text data

- Huge amount of data

- High dimensional but sparse
  - all possible word and phrase types in the language!!

Challenges with text data

- Ambiguity

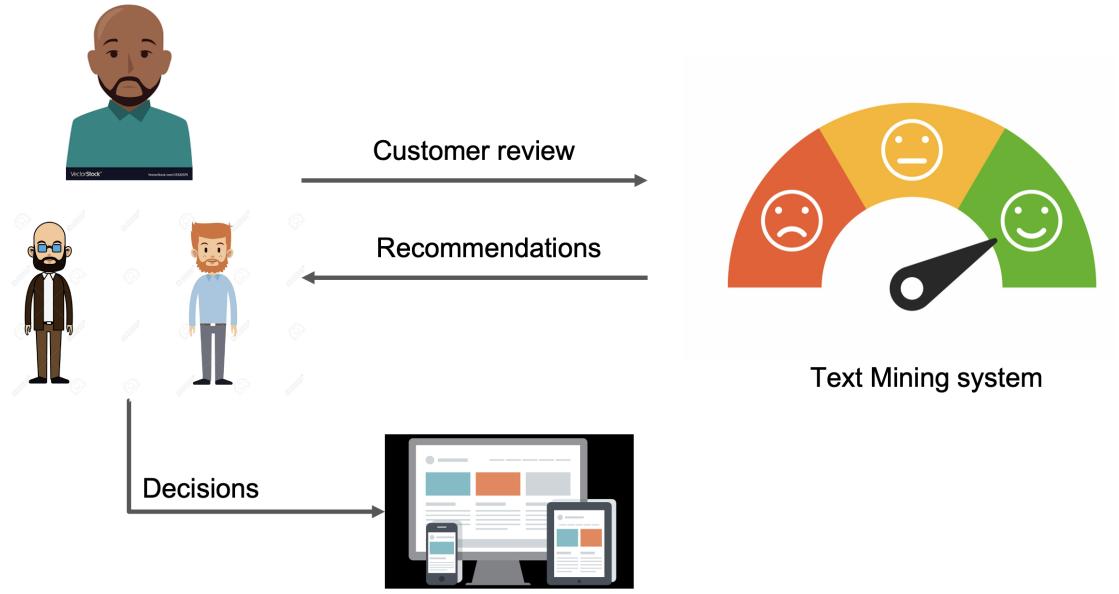
Challenges with text data

- Noisy data
  - Examples: Abbreviations, spelling errors, short text
- Complex relationships between words
  - "Hema merges with Intertoys"
  - "Intertoys is bought by Hema"

Example

- During one of the coffee moments at the company, **Garry** was talking about their situation at the dep of Customer relationship management.
- When **Carrie**, her colleague from the **Data Science department**, hears the situation, she offers Garry to use Text Mining!!
- She says: "Text mining is your friend; it can help you to make the process way faster than Excel by filtering words and recommending labels."
- She continues : "Text mining is a subfield of AI and NLP and is related to data science, data mining and machine learning."
- After consulting with Larry and Harry, they decide to give text mining a try!

Example



Language is hard

- Different things can mean more or less the same (“data science” vs. “statistics”)
- Context dependency (“You have very nice shoes”);
- Same words with different meanings (“to sanction”, “bank”);
- Lexical ambiguity (“we saw her duck”)
- Irony, sarcasm (“That’s just what I needed today!”, “Great!”, “Well, what a surprise.”)
- Figurative language (“He has a heart of stone”)
- Negation (“not good” vs. “good”), spelling variations, jargon, abbreviations
- All the above are different over languages, 99% of work is on English!

Language is hard

- We won’t solve linguistics . . .
- In spite of the problems, text mining can be quite effective!

Text mining definition?

- Which can be a part of Text Mining definition?
  - The discovery by computer of new, previously unknown information from textual data
  - Automatically extracting information from text
  - Text mining is about looking for patterns in text
  - Text mining describes a set of techniques that model and structure the information content of textual sources

(You can choose multiple answers)

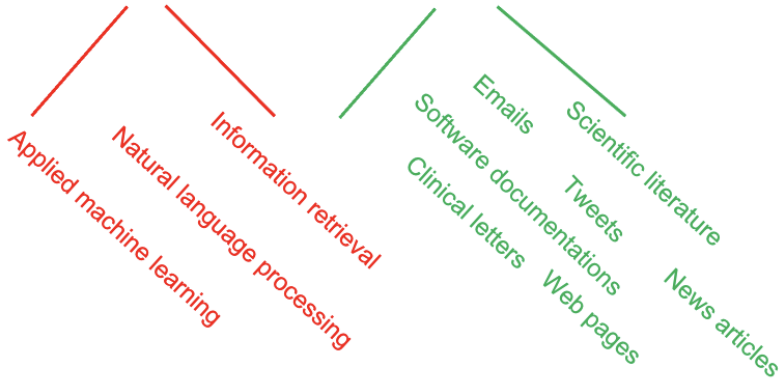
Go to www.menti.com and use the code 7338 2184

Text mining definition

- "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" Hearst (1999)
- Text mining is about looking for patterns in text, in a similar way that data mining can be loosely described as looking for patterns in data.
- Text mining describes a set of linguistic, statistical, and machine learning techniques that model and structure the information content of textual sources. (Wikipedia)

Another TM definition

- Text Mining = Data Mining + Text Data

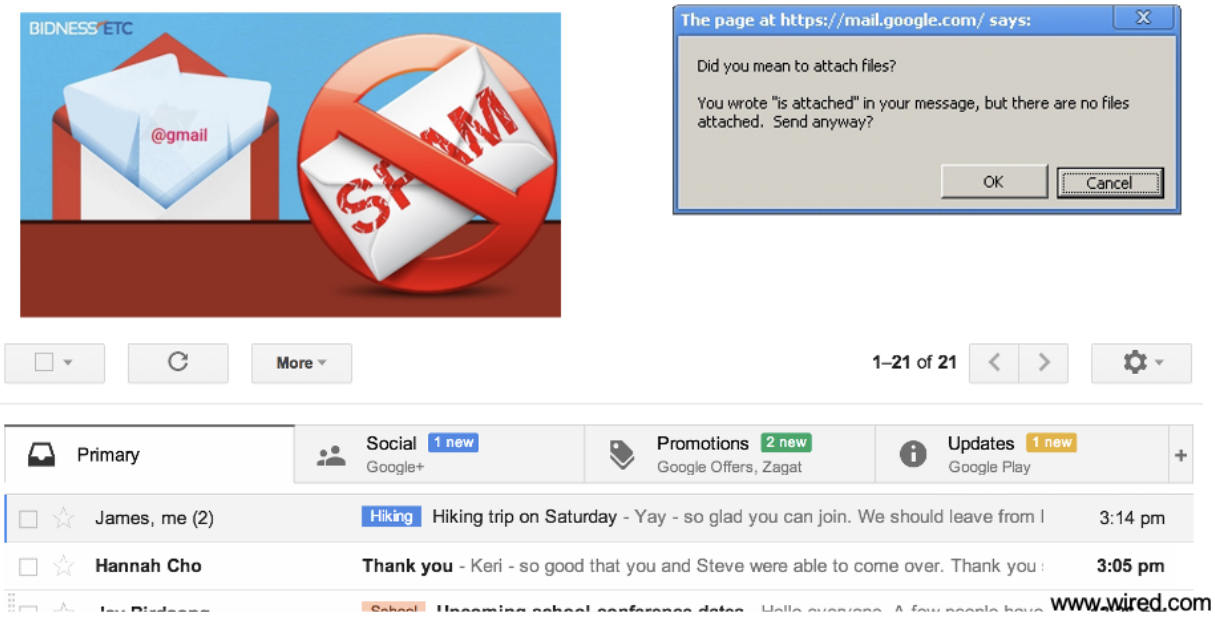


Examples & Applications

Who wrote the Wilhelmus?

https://dh2017.adho.org/abstracts/079/079.pdf

Text classification



Which ICD-10 codes should I give this doctor's note?

Bovengenoemde patiënt was opgenomen op op de voor het specialisme **Cardiologie.**

Cardiovasculaire risicofactoren: Roken(-) Diabetes(-) Hypertensie(?) Hypercholesterolemie (?)

Anamnese. Om 18.30 pijn op de borst met uitstraling naar de linkerarm, zweten, misselijk. Ambulance gebeld en bij aansluiten monitor beeld van acuut onderwandinfarct. AMBU overdracht:.500mg aspegic iv, ticagrelor 180mg oraal, heparine, zofran eenmalig, 3x NTG spray. HD stabiel gebleven. . .Medicatie bij presentatie.Geen..

Lichamelijk onderzoek. Grauw, vegetatief, Halsvenen niet gestuwd. Cor s1 s2 geen souffles.Pulm schoon. Extr warm en slank .

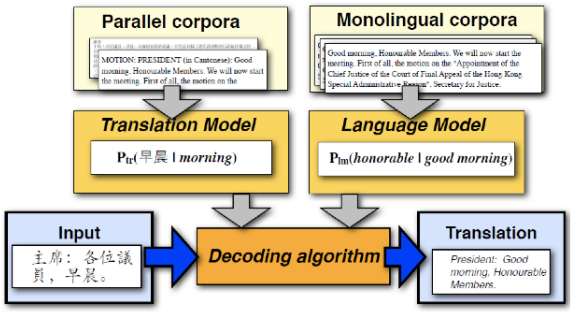
Aanvullend onderzoek. AMBU ECG: Sinusritme, STEMI inferior III)II C/vermoedelijk RCA. Coronair angiografie. (. . .) .Conclusie angio: 1-vatslijden..PCI

Conclusie en beleid Bovengenoemde jarige man, blanco cardiale voorgeschiedenis, werd gepresenteerd vanwege een STEMI inferior waarvoor een spoed PCI werd verricht van de mid-RCA. Er bestaan geen relevante nevenletsels. Hij kon na de procedure worden overgeplaatst naar de CCU van het . ..Dank voor de snelle overname. ..Medicatie bij overplaatsing. Acetylsalicylzuur dispertablet 80mg ; oraal; 1 x per dag 80 milligram ; .Ticagrelor tablet 90mg ; oraal; 2 x per dag 90 milligram ; .Metoprolol tablet 50mg ; oraal; 2 x per dag 25 milligram ; .Atorvastatine tablet 40mg (als ca-zout-3-water) ; oraal; 1 x per dag 40 milligram ; **Samenvatting** Hoofddiagnose: STEMI inferior wv PCI RCA. Geen nevenletsels. Nevendiagnoses: geen. Complicaties: geen Ontslag naar: CCU .

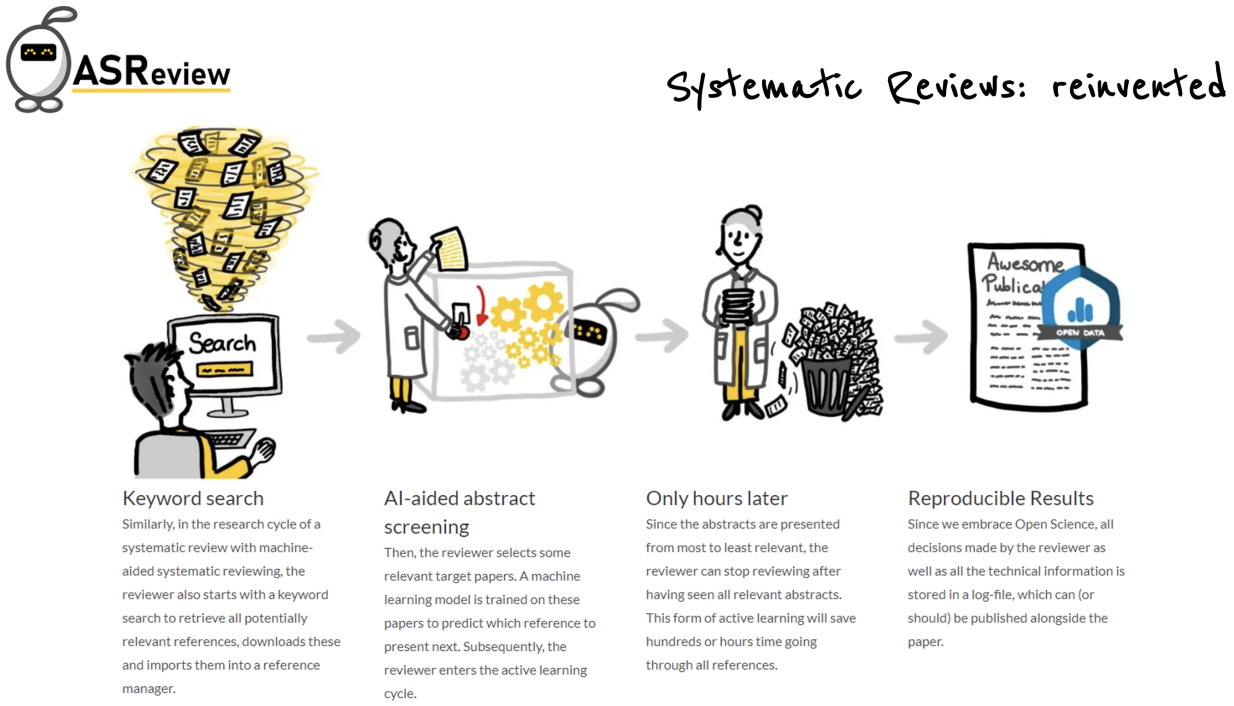
Which ICD-10 codes should I give this doctor's note?

Sentiment analysis / Opinion mining

Statistical machine translation



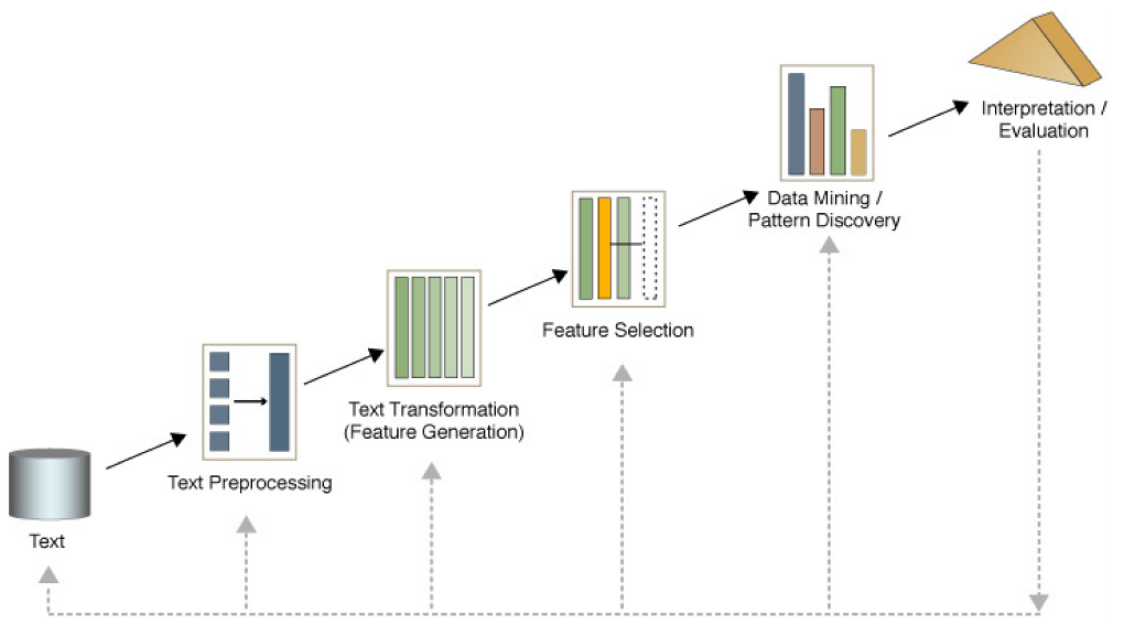
Which studies go in my systematic review?



https://asreview.nl/

Process & Tasks

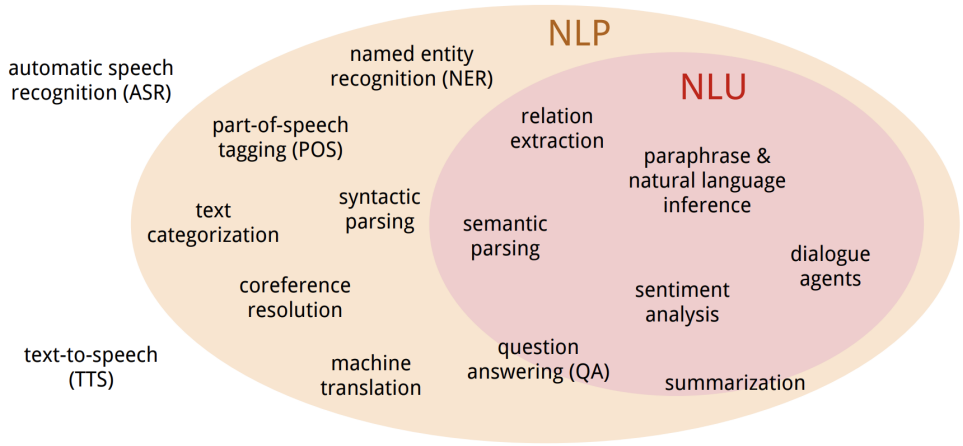
Text mining process



Text mining tasks

- Text classification
- Text clustering
- Sentiment analysis
- Feature selection
- Topic modelling
- Word embedding
- Deep learning models
- Responsible text mining
- Text summarization

And more in NLP



source: https://nlp.stanford.edu/~wcmac/papers/20140716-UNLU.pdf

Text Preprocessing

How to represent a document

- Represent by a string?
  - No semantic meaning
- Represent by a list of sentences?
  - Sentence is just like a short document (recursive definition)
- Represent by a vector?
  - A vector is an ordered finite list of numbers.

Vector space model

- A vector space is a collection of vectors
- Represent documents by concept vectors
  - Each concept defines one dimension
  - k concepts define a high-dimensional space
  - Element of vector corresponds to concept weight

Vector space model

- Distance between the vectors in this concept space
  - Relationship among documents
- The process of converting text into numbers is called Vectorization

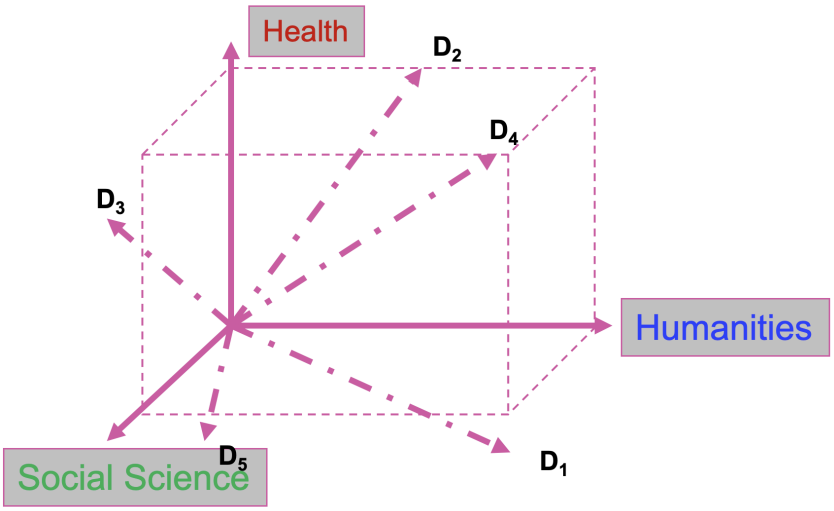
Vector space model

- Terms are generic features that can be extracted from text
- Typically, terms are single words, keywords, n-grams, or phrases
- Documents are represented as vectors of terms
- Each dimension (concept) corresponds to a separate term

$$d = (w_1, ..., w_n)$$

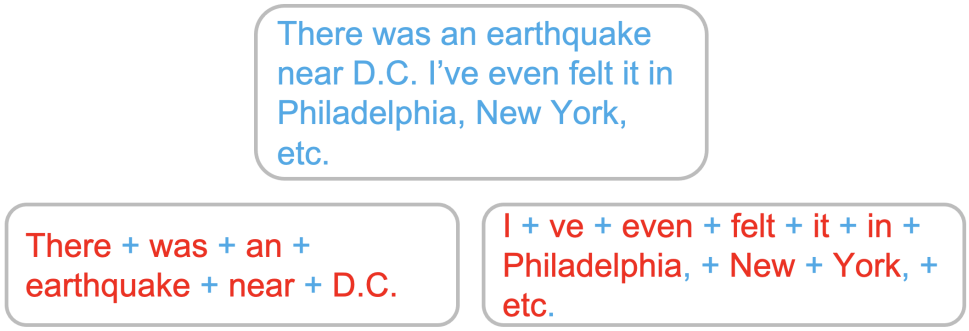
An illustration of VS model

- All documents are projected into this concept space



Tokenization/Segmentation

- Split text into words and sentences



N-grams

- N-grams: a contiguous sequence of N tokens from a given piece of text
  - E.g., *'Text mining is to identify useful information.'*
  - Bigrams: *'text_mining', 'mining_is', 'is_to', 'to_identify', 'identify_useful', 'useful_information', 'information_.'*
- Pros: capture local dependency and order
- Cons: a purely statistical view, increase the vocabulary size $O(V^N)$

Preprocessing

Typical steps:

- Stemming ("running"→"run") or Lemmatization ("were"→"is")
- Lowercasing ("And"→"and")
- Stopword removal ("evning morning is third day.")
- Punctuation removal ("evning morning is third day")
- Number removal ("day 3"→"day")
- Spell correction ("evning"→"evening")
- Tokenization ("evening", "morning", "is", "third", "day")

Not all of these are appropriate at all times!

Stemming

- Unifies variations in the text data:
  - e.g., 'walking', 'walks', 'walked' → walk
- Inflectional stemming:
  - Remove plurals

  - Normalize verb tenses
  - Remove other affixes
- Stemming to root:
  - Reduce word to most basic element
  - More aggressive than inflictional
  - e.g., 'denormalization' → norm;
  - e.g., 'Apply', 'applications', 'reapplied' → apply

Constructing a VSM representation

D1: 'Text mining is to identify useful information.'

Tokenization:

D1: 'Text', 'mining', 'is', 'to', 'identify', 'useful', 'information', '.'

Stemming/normalization:

D1: 'text', 'mine', 'is', 'to', 'identify', 'use', 'inform', '.'

N-gram construction:

D1: 'text-mine', 'mine-is', 'is-to', 'to-identify', 'identify-use', 'use-inform', 'inform-.'

Stopword/controlled vocabulary filtering::

D1: 'text-mine', 'to-identify', 'identify-use', 'use-inform'

VSM: How do we represent vectors?

After tokenization and pre-processing, we have three options:

- Bag of Words
- Topics
- Word Embeddings

Bag of Words (BOW)

- With Bag of Words (BOW), we refer to a Vector Space Model where:
  - Terms: words (more generally we may use n-grams, etc.)
  - Weights: number of occurrences of the terms in the document

BOW representation

- Term as the basis for vector space
  - Doc1: Text mining is to identify useful information.
  - Doc2: Useful information is mined from text.
  - Doc3: Apple is delicious.

BOW weights: Binary

- Binary
  - with 1 indicating that a term occurred in the document, and 0 indicating that it did not

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

BOW weights: Term frequency

- Idea: a term is more important if it occurs more frequently in a document
- TF Formulas
  - Let $t(c, d)$ be the frequency count of term t in doc d
  - Raw TF: $tf(t, d) = c(t, d)$

TF: Document - Term matrix (DTM)

Bag of words

- d1: "And God said, Let there be light: and there was light."
- d2: "And God saw the light, that it was good: and God divided the light from the darkness."
- d3: "And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day."

"Document - Term matrix" (DTM) (raw word counts)

	light	god	darkness	called	day	let	said	divided	good	saw	evening	first	morning	night
d1	2	1	0	0	0	1	1	0	0	0	0	0	0	0
d2	2	2	1	0	0	0	0	1	1	1	0	0	0	0
d3	1	1	1	2	2	0	0	0	0	0	1	1	1	1

BOW weights: TFiDF

- Idea: a term is more discriminative if it occurs a lot but only in fewer documents

Let $n_{d,t}$ denote the number of times the t-th term appears in the d-th document.

$$TF_{d,t} = \frac{n_{d,t}}{\sum_i n_{d,i}}$$

Let N denote the number of documents annd N_t denote the number of documents containing the t-th term.

$$IDF_t = log(\frac{N}{N_t})$$

TFiDF weight:

$$w_{d,t} = TF_{d,t} \cdot IDF_t$$

TFiDF: Document - Term matrix (DTM)

Bag of words

- d1: "And God said, Let there be light: and there was light."
- d2: "And God saw the light, that it was good: and God divided the light from the darkness."
- d3: "And God called the light Day, and the darkness he called Night. And the evening and the morning were the first day."

"Document - Term matrix" (DTM) (tf-idf)

	light	god	darkness	called	day	let	said	divided	good	saw	evening	first	morning	night
d1	0	0	0.000	0.0	0.0	1.1	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
d2	0	0	0.405	0.0	0.0	0.0	0.0	1.1	1.1	1.1	0.0	0.0	0.0	0.0
d3	0	0	0.405	2.2	2.2	0.0	0.0	0.0	0.0	0.0	1.1	1.1	1.1	1.1

Why document frequency

- How about total term frequency?
  - $ttf(t) = \sum_d c(t, d)$

Why document frequency

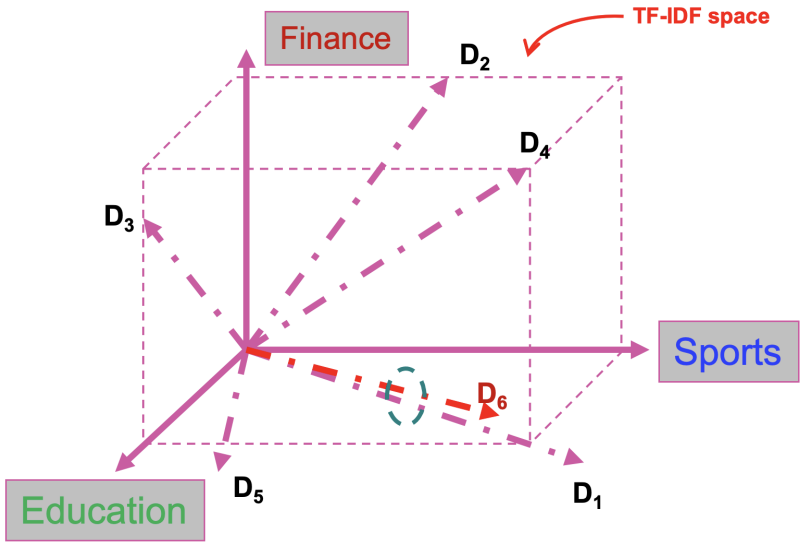
- How about total term frequency?
  - $ttf(t) = \sum_d c(t, d)$

Table 1. Example total term frequency v.s. document frequency in Reuters-RCV1 collection.

Word	ttf	df
try	10422	8760
insurance	10440	3997

- Cannot recognize words frequently occurring in a subset of documents

How to define a good similarity metric?



How to define a good similarity metric?

- Euclidean distance
  $dist(d_i, d_j) = \sqrt{\sum_{t \in V} [tf(t, d_i)idf(t) - tf(t, d_j)idf(t)]^2}$
  - Longer documents will be penalized by the extra words
  - We care more about how these two vectors are overlapped
- Cosine similarity
  - Angle between two vectors:
    $cosine(d_i, d_j) = \frac{V_{d_i}^T V_{d_j}}{|V_{d_i}|_2 \times |V_{d_j}|_2}$ ← TF-IDF vector
  - Documents are normalized by length

More pre-processing: Named entity recognition

- Determine text mapping to proper names

> Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

NER

- Determine text mapping to proper names

> Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

> Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe.

Organization, **Location**, **Person**

Part Of Speech (POS) tagging

- Annotate each word in a sentence with a part-of-speech.

I ate the spaghetti with meatballs.
Pro V Det N Prep N

John saw the saw and decided to take it to the table.
PN V Det N Con V Part V Pro Prep Det N

- Useful for subsequent syntactic parsing and word sense disambiguation.

Preprocessing demo

- CogComp: https://cogcomp.seas.upenn.edu/vsa8080/curator
- CCG Demos: https://cogcomp.seas.upenn.edu/page/demos/
- Stanford parser: http://nlp.stanford.edu:8080/parser/index.jsp

Python | before starting the first practical

How familiar are you with Python?

- What is your experience level with Python?
  - I don't know anything about Python.
  - I know a bit about Python and/or I have worked with Python years ago.
  - I am familiar with Python, but I use another programming language for my work.
  - I use Python for my daily work, but I don't know it very well.
  - I use Python for my daily work and I know it very well, but I am not an expert.
  - I am an expert with Python.
  - Go to www.menti.com and use the code 7338 2184

Python IDE?

- Which Python IDE do you mostly use? If you use more than one environment fill in the other text boxes.
  - Go to www.menti.com and use the code 7338 2184

Google Colab?

- From 0 (none) to 5 (expert), how familiar are you with Google Colab?
  - Go to www.menti.com and use the code 7338 2184

Python

- Latest: Python 3.9.1
- Follow the tutorial on Python in Google Colab for the Applied Text Mining course: link
- Python For Beginners
  - https://www.python.org/about/gettingstarted/
- The Python Language Reference
  - https://docs.python.org/3/reference/
- Python 3.9.1 documentation
  - https://docs.python.org/3/

Google Colab

- Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with
  - Zero configuration required
  - Free access to GPUs
  - Easy sharing
- [Intro](https://colab.research.google.com/notebooks/intro.ipynb)
- Cheat-sheet for Google Colab
- Keyboard shortcuts:

Actions	Colab	Jupyter
show keyboard shortcuts	Ctrl/Cmd M H	H
Insert code cell above	Ctrl/Cmd M A	A
Insert code cell below	Ctrl/Cmd M B	B
Delete cell/selection	Ctrl/Cmd M D	DD
Interrupt execution	Ctrl/Cmd M I	II
Convert to code cell	Ctrl/Cmd M Y	Y
Convert to text cell	Ctrl/Cmd M M	M
Split at cursor	Ctrl/Cmd M -	Ctrl Shift -

Summary

Summary

- Text data is everywhere!
- Language is hard!
- The basic problem of text mining is that text is not a neat data set
- Solution: text pre-processing & VSM

Practical 1

In a few moments:

- You will be automatically added to a practical session.
- There will be a practical instructor present.
- At the end of the practical, you will be automatically returned to the main meeting.