Natural Language Processing with Deep Learning
Basics of NLP & Text Processing

Navid Rekab-Saz
navid.rekabsaz@jku.at

JOHANNES KEPLER
UNIVERSITY LINZ

Institute of
Computational
Perception

Course

- Exam and project, both possible!

Exam or Project?

Visible groups: All participants

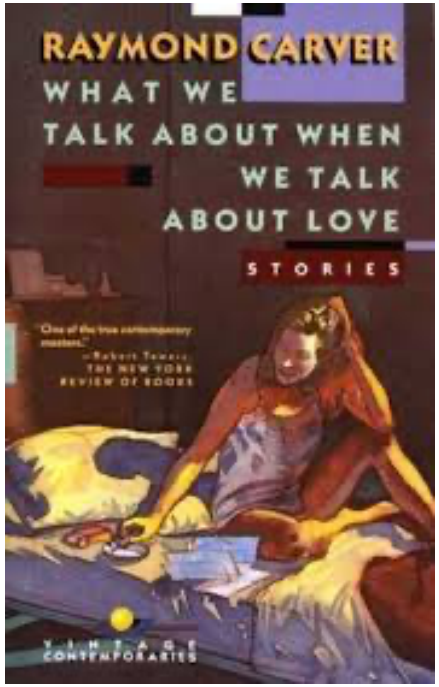
Responses

Choice options	I prefer to take an exam	I prefer to do a project
Number of responses	9	9

- Updates on projects
  - http://recsys-twitter.com
- Group registrations for assignments
- Release of assignment 1

What We Talk About When We Talk About NLP

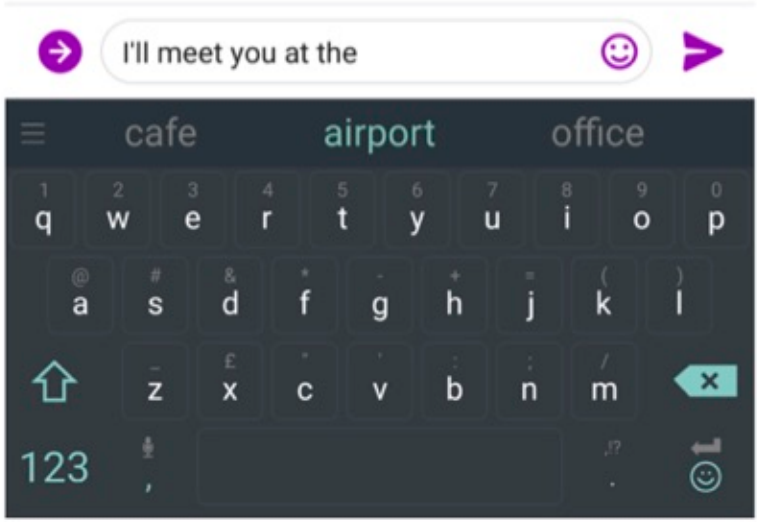
- Applications
- Low-level tasks



What We Talk About When We Talk About NLP

- **Applications**
- Low-level tasks

Auto Correction / Spell Checking / Predictive Text



Google

what is the

what is the **weather**
what is the **meaning of life**
what is the **dark web**
what is the **xfl**
what is the **doomsday clock**
what is the **weather today**
what is the **keto diet**
what is the **american dream**
what is the **speed of light**
what is the **bill of rights**

Google Search I'm Feeling Lucky

Paste your own text here and click the 'Check Text' button. Click the colored phrases for details on potential errors. or use this text too see an few of of the problems that LanguageTool can detecd. What do you thinks of grammar checkers? Please not that they are not perfect.

English American Check Text

English has incomplete support in LanguageTool. Would you like to help?

Sentiment Analysis / Market Intelligence

HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner
$89 online, **$100** nearby ★★★★☆ 377 reviews
September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

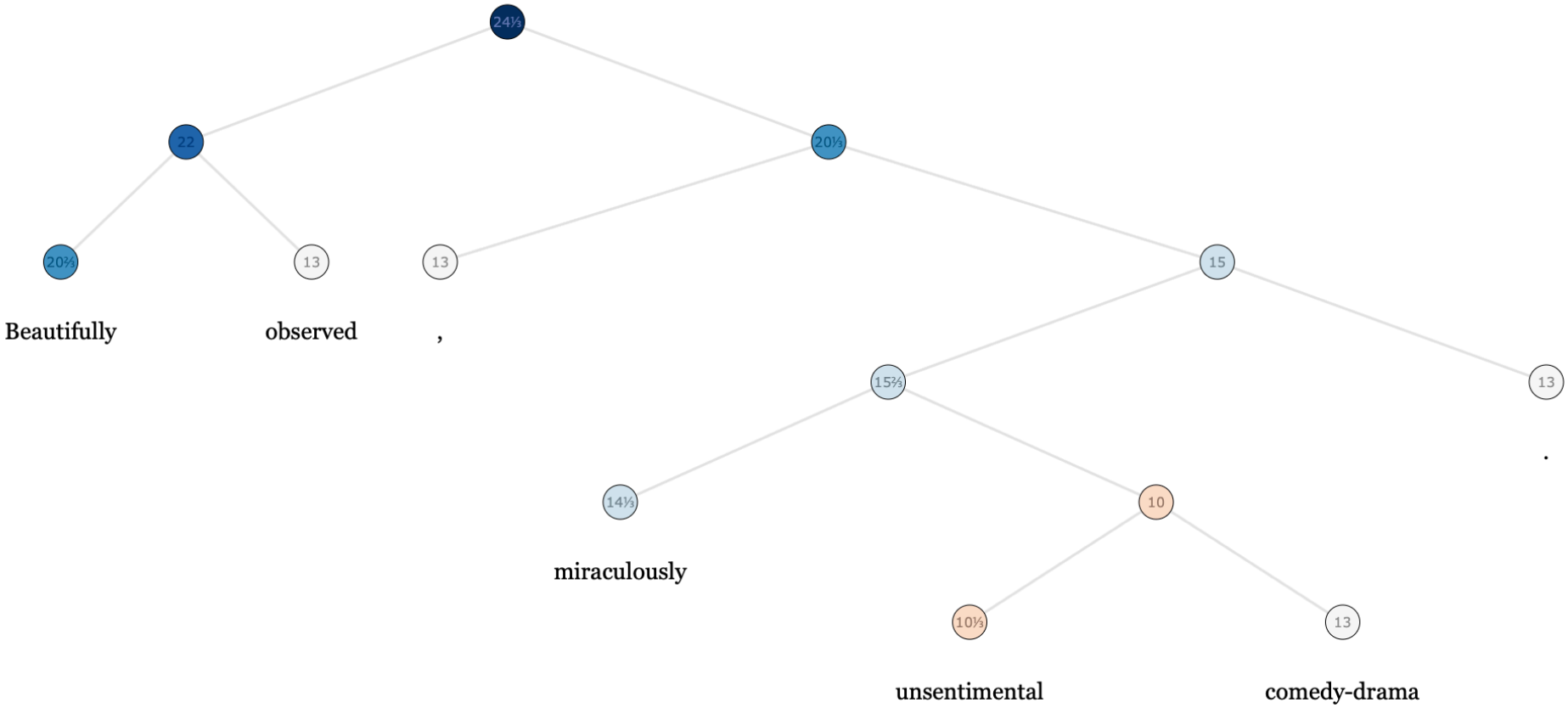
Summary - Based on 377 reviews

1 star	2	3	4 stars	5 stars

What people are saying

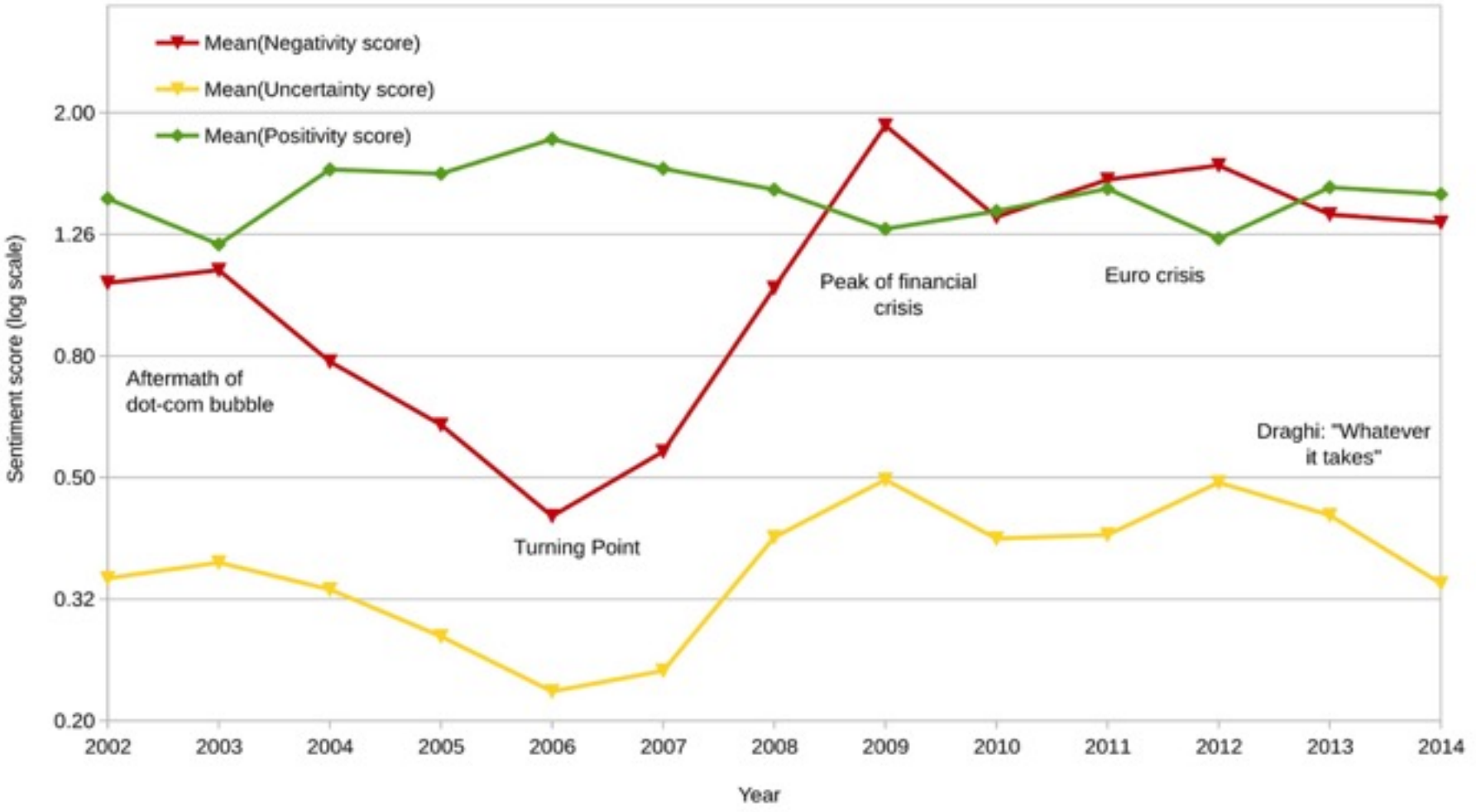
ease of use	"This was very easy to setup to four computers."
value	"Appreciate good quality at a fair price."
setup	"Overall pretty easy setup."
customer service	"I DO like honest tech support people."
size	"Pretty Paper weight."
mode	"Photos were fair on the high quality mode."
colors	"Full color prints came out with great quality."

Sentiment Analysis / Market Intelligence



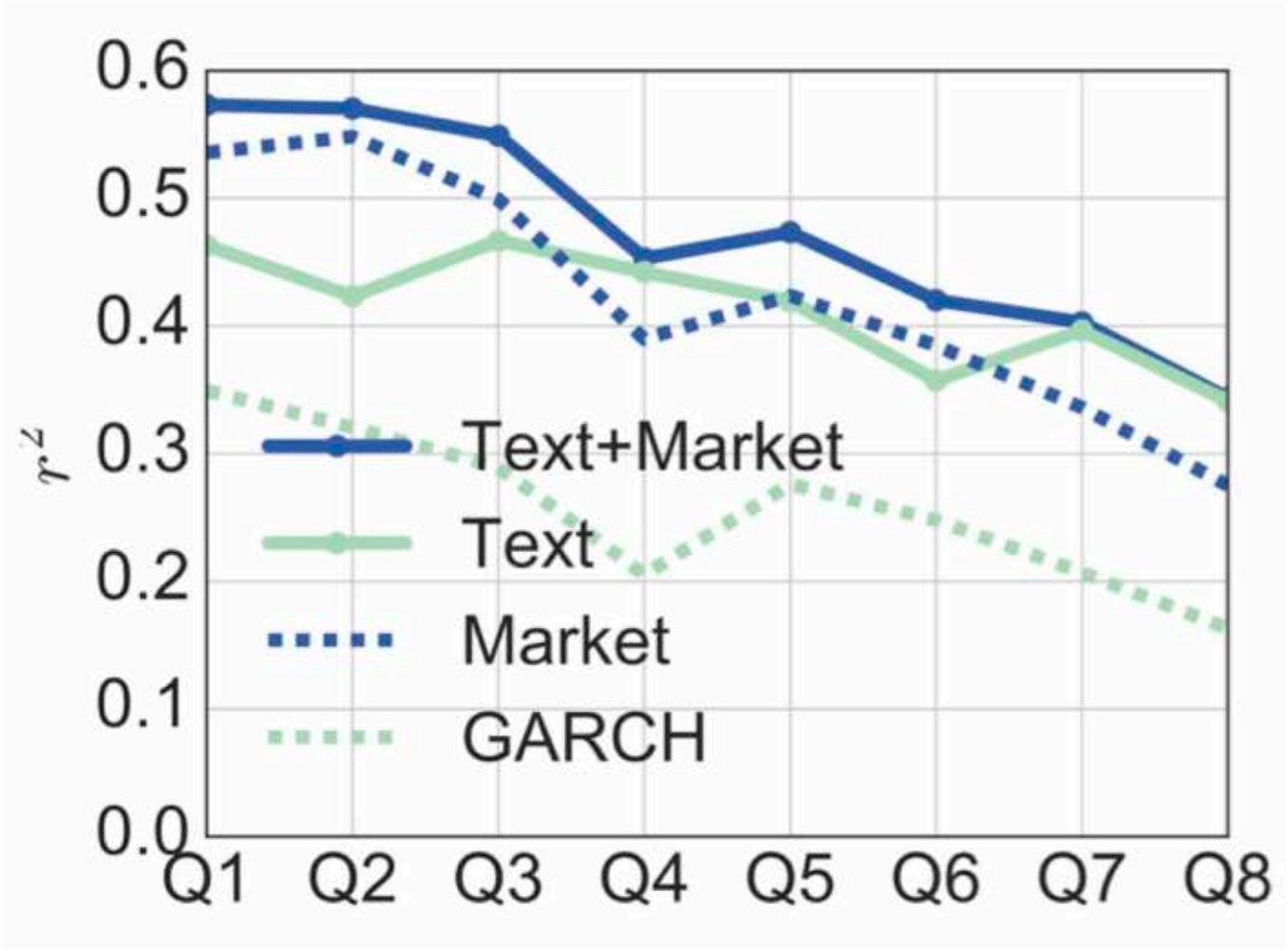
https://nlp.stanford.edu/sentiment/treebank.html

Sentiment Analysis / Market Intelligence



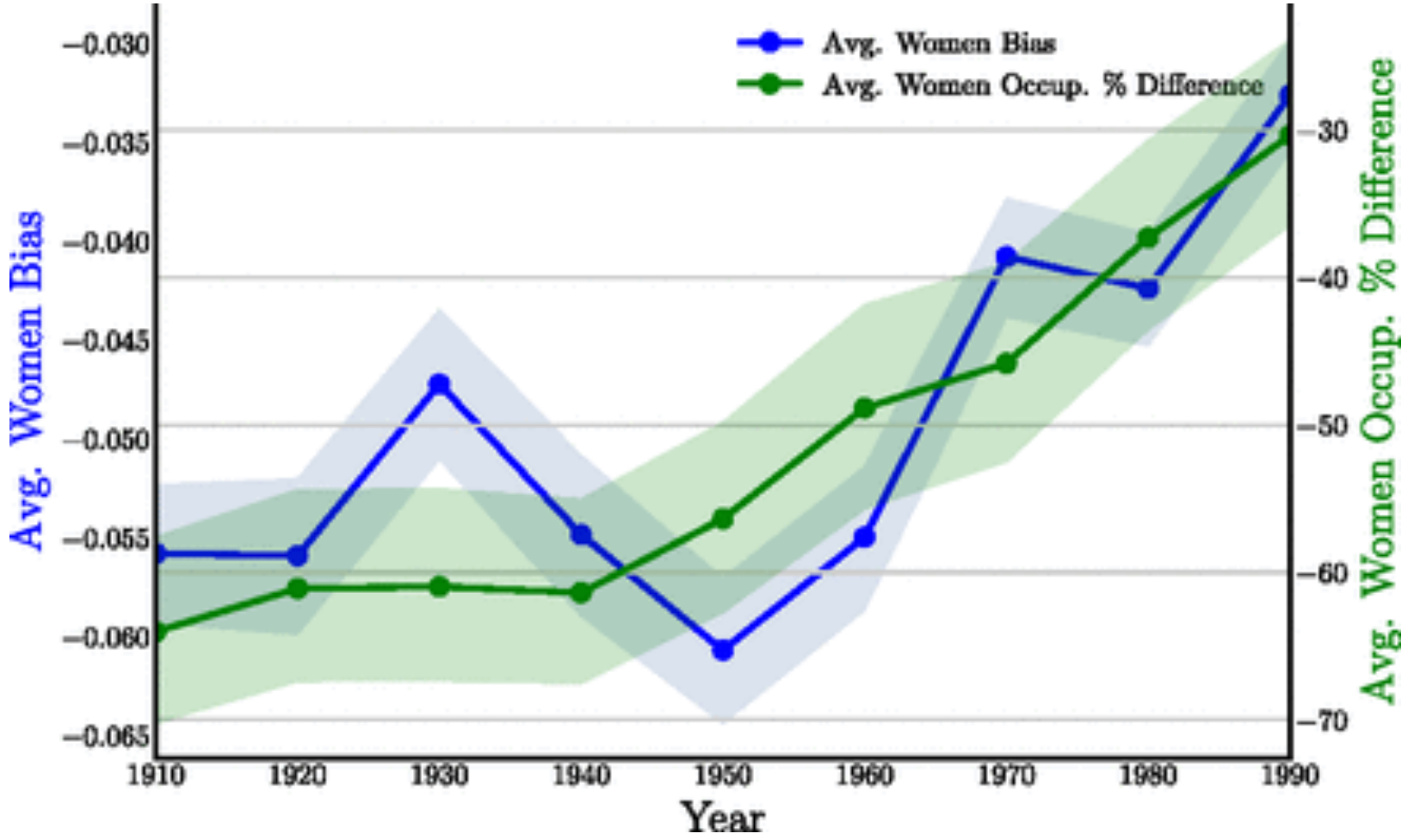
Nopp, Clemens, and Allan Hanbury. *"Detecting Risks in the Banking System by Sentiment Analysis."* Proceedings of the EMNLP 2015

Sentiment Analysis / Market Intelligence



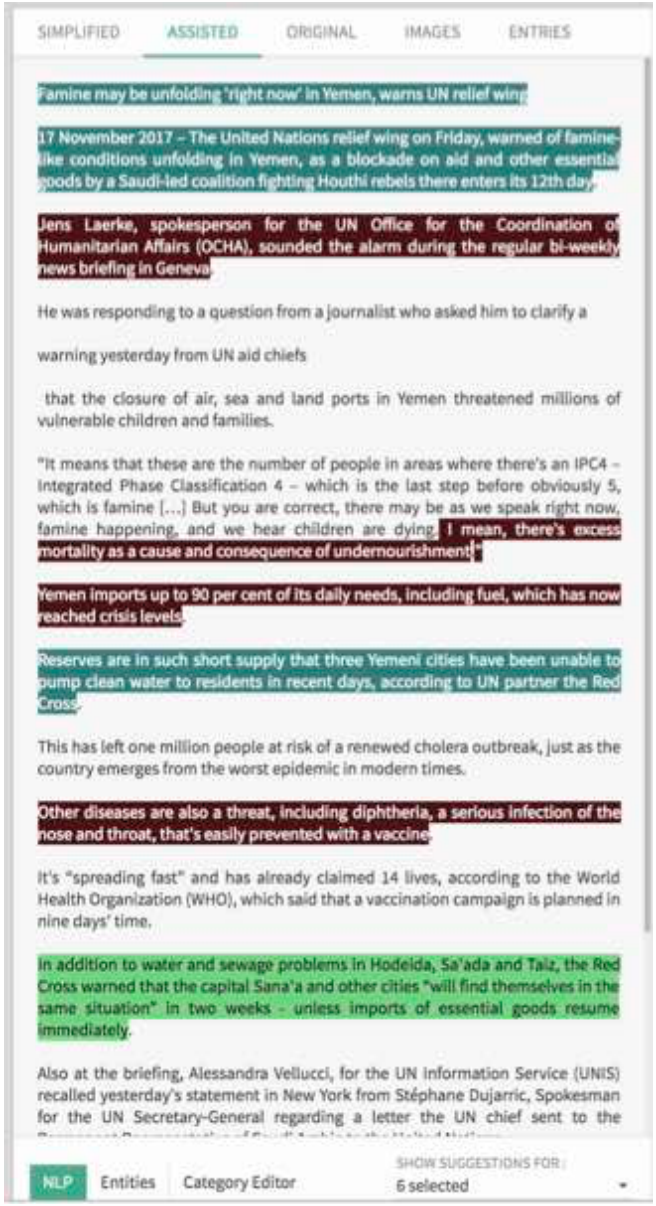
Volatility Prediction using Financial Disclosures Sentiments with Word Embedding-based IR Models
Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Duer, Linda Anderson In Proc. ACL 2017

Monitoring Societal Changes



Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*

Text Classification



Food

Agriculture

Logistics

https://www.thedeep.io

https://www.aicrowd.com/challenges/amld-2020-transfer-learning-for-international-crisis-response

Machine Translation

ENGLISH - DETECTED | ENGLISH | ENGLISH | GERMAN | SPANISH

"One of history's few iron laws is that luxuries tend to become necessities and to spawn new obligations."

106/5000

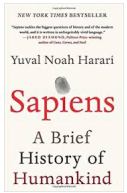
"Eines der wenigen eisernen Gesetze der Geschichte ist, dass Luxus dazu neigt, zu Notwendigkeiten zu werden und neue Verpflichtungen zu schaffen."

Translate from English (detected) | Translate into German

"One of history's few iron laws is that luxuries tend to become necessities and to spawn new obligations."

Translate document

"Eines der wenigen eisernen Gesetze der Geschichte ist, dass Luxus zur Notwendigkeit wird und neue Verpflichtungen hervorbringt".



Machine Translation

ENGLISH - DETECTED | ENGLISH | GERMAN | ENGLISH | SPANISH

"Humans think in stories, and we try to make sense of the world by telling stories."

84/5000

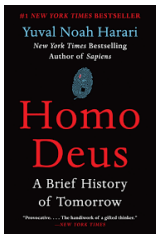
"Menschen denken in Geschichten, und wir versuchen, die Welt zu verstehen, indem wir Geschichten erzählen."

Translate from English (detected) | Translate into German

"Humans think in stories, and we try to make sense of the world by telling stories."

Translate document

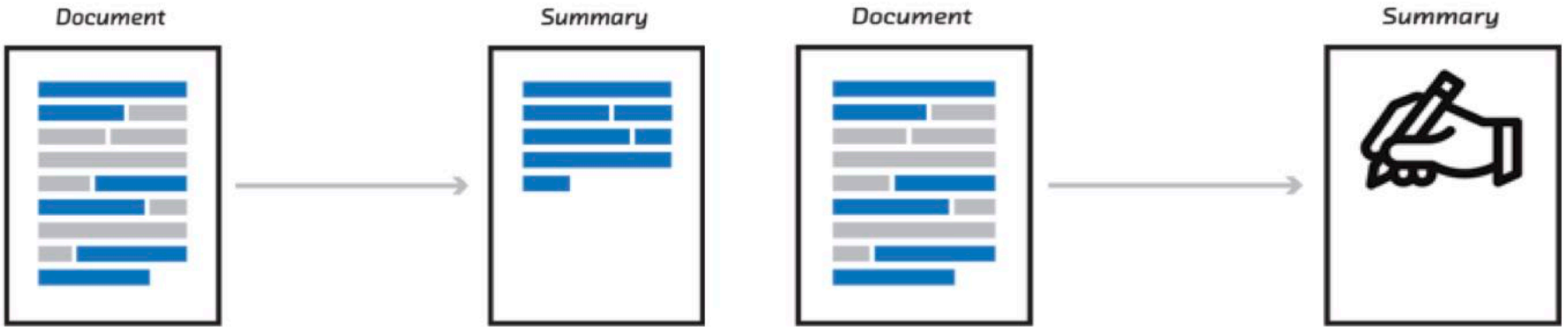
"Die Menschen denken in Geschichten, und wir versuchen, der Welt durch das Erzählen von Geschichten einen Sinn zu geben".



Summarization

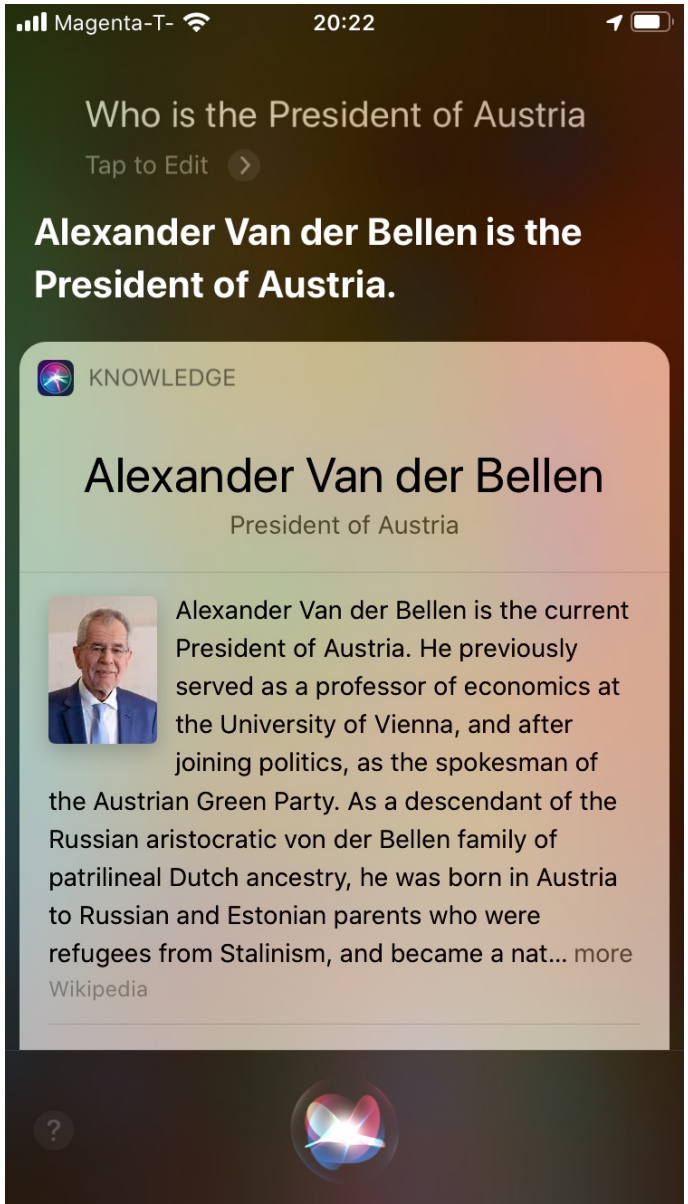
- Extractive Summarization
  - Extracting informative sections of a document
- Abstractive Summarization
  - *Reading* the document and *writing* a summary

Extractive Summarization **Abstractive Summarization**



Question Answering

Conversational Question Answering



Question Answering

Passage Segment

...The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process...

Question

Which governing bodies have veto power?

Passage Segment

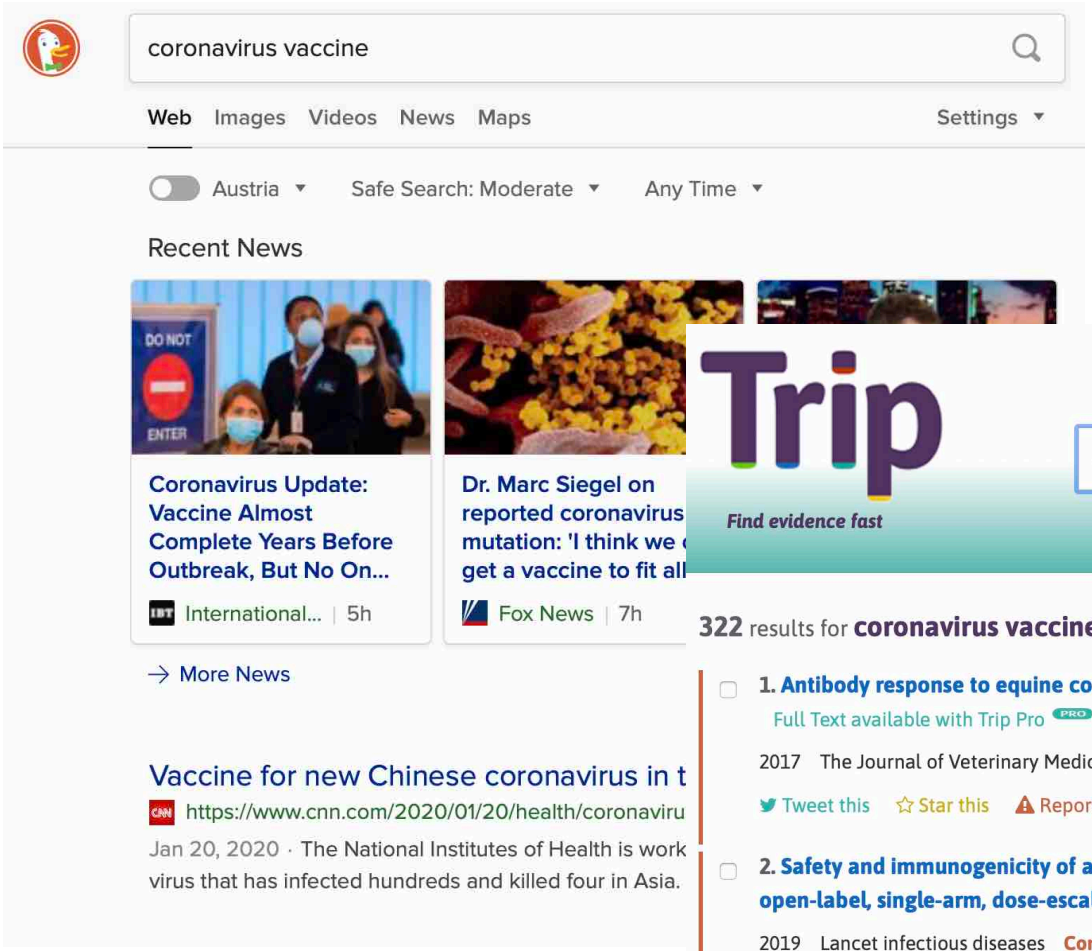
...The Rankine cycle is sometimes referred to as a practical Carnot cycle...

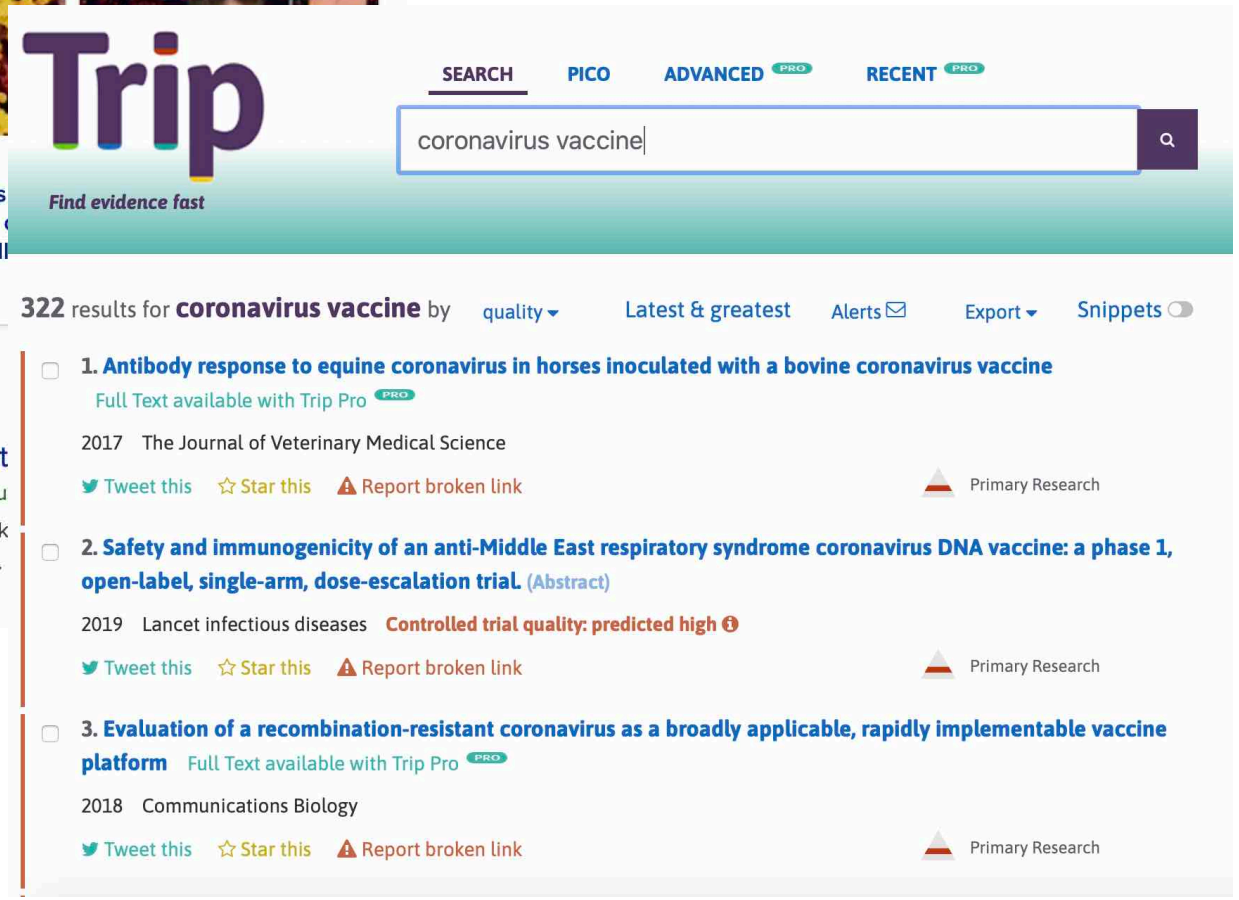
Question

What is the Rankine cycle sometimes called?

https://rajpurkar.github.io/mlx/qa-and-squad/

Information Retrieval

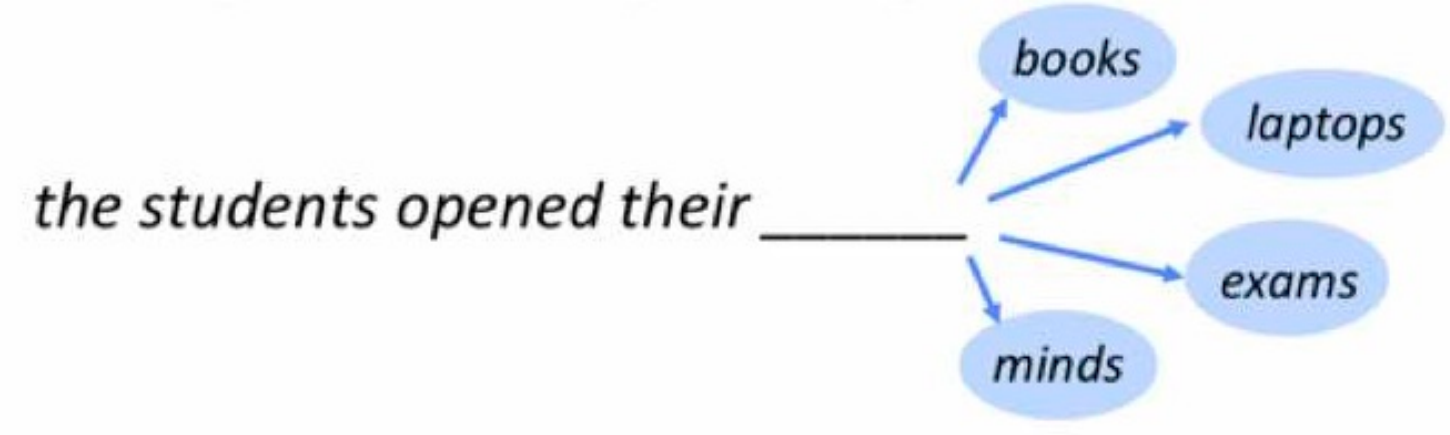




What We Talk About When We Talk About NLP

- Applications
- **Low-level tasks**

Language Modeling



https://talktotransformer.com

https://www.youtube.com/watch?v=gcHkxP9adiM

Natural Language Inference

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

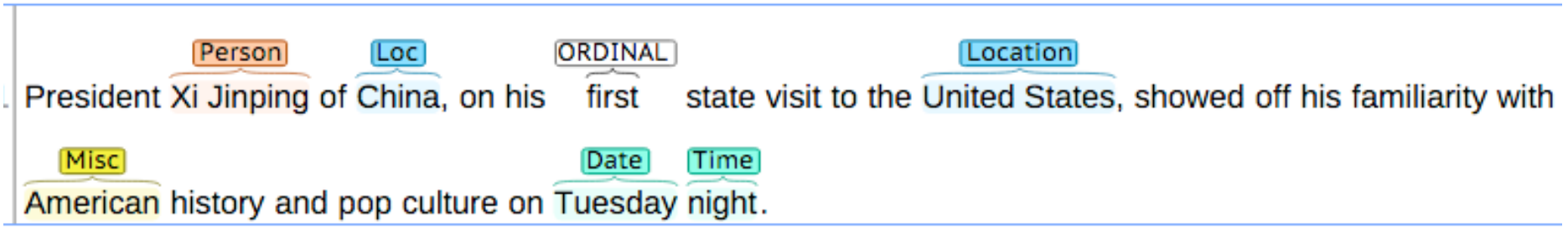
https://nlp.stanford.edu/projects/snli/

Word Sense Disambiguation

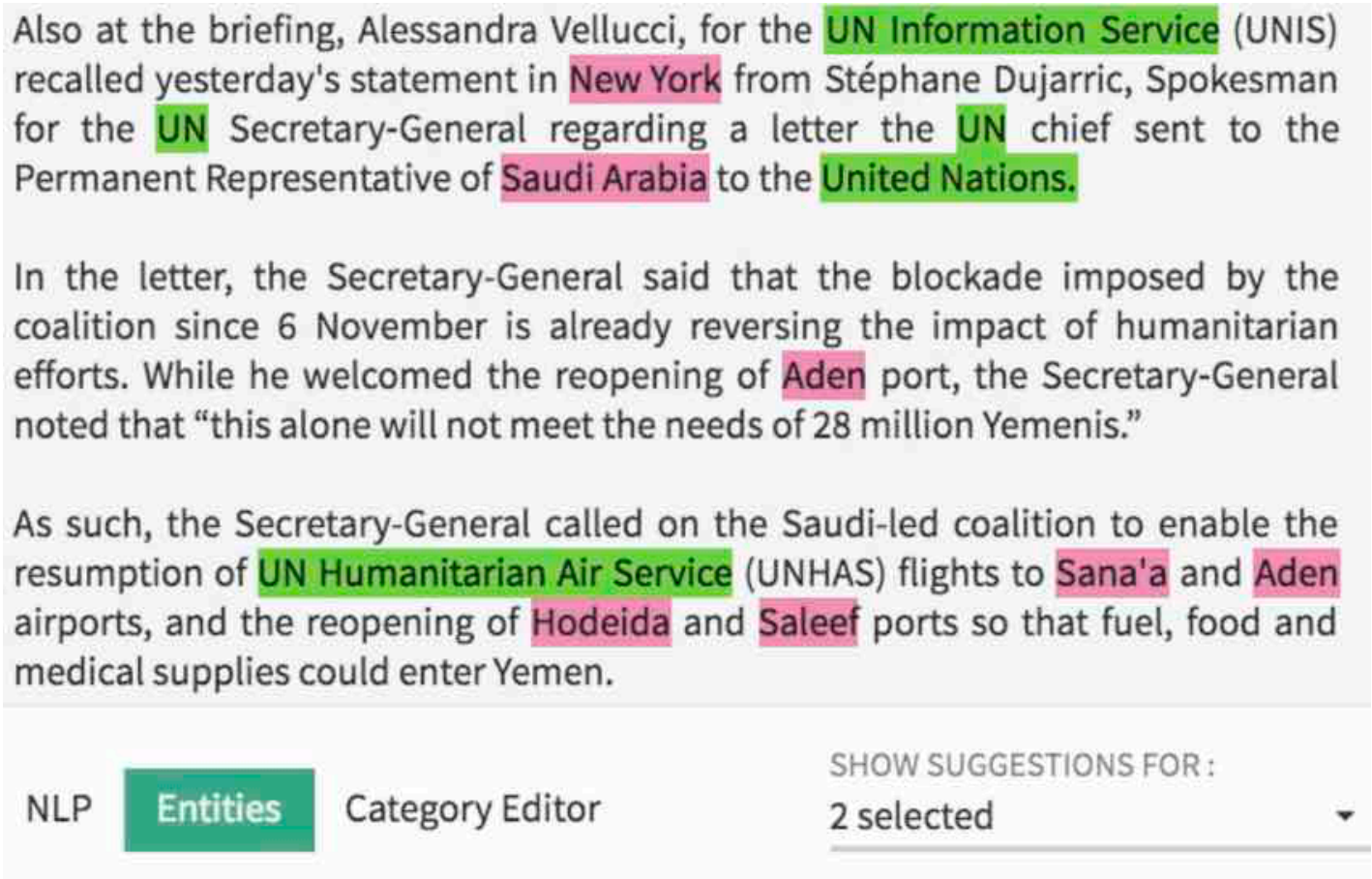
- 2 senses of ***bank***
  - "The ***bank*** will not be accepting cash on Saturdays. "
  - "The river overflowed the ***bank.***"
- 8 senses of ***bass***, as defined in WordNet
  - bass - (the lowest part of the musical range)
  - bass, bass part - (the lowest part in polyphonic music)
  - bass, basso - (an adult male singer with the lowest voice)
  - sea bass, bass - (flesh saltwater fish of the family Serranidae)
  - freshwater bass, bass - (any of various North American lean-fleshed freshwater fishes especially of the genus Micropterus)
  - bass, bass voice, basso - (the lowest adult male singing voice)
  - bass - (member with the lowest range of a family of musical instruments)
  - bass - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Named Entity Recognition (NER)

- "Named Entity Recognition labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names."



Named Entity Recognition (NER)

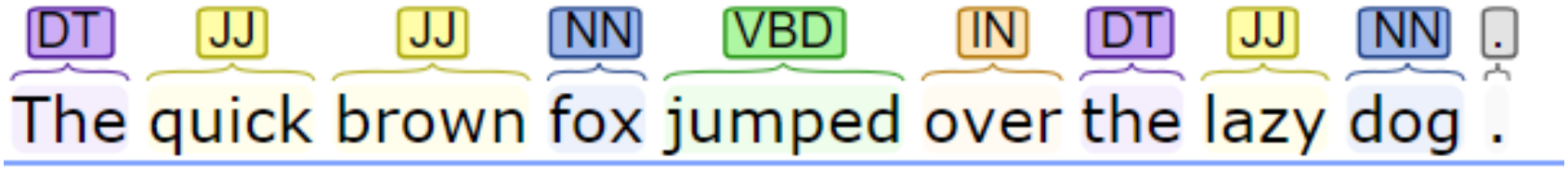


Locations

Organizations

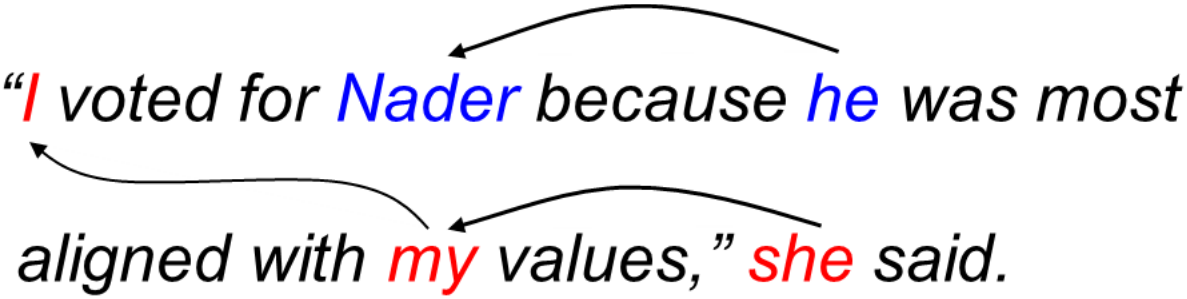
Part-of-Speech (POS) Tagging

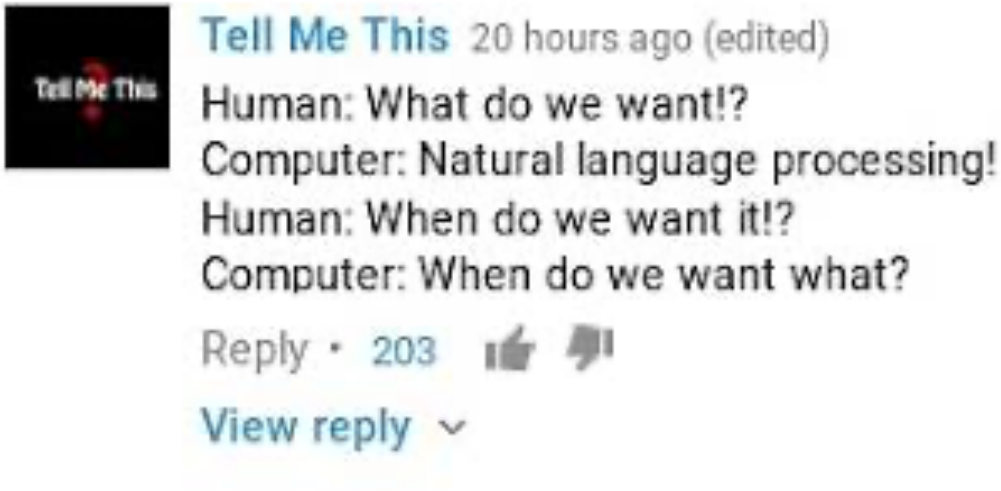
- “A Part-Of-Speech Tagger is a piece of software that reads text in some language and assigns parts of speech to each word, such as noun, verb, adjective”



Coreference Resolution

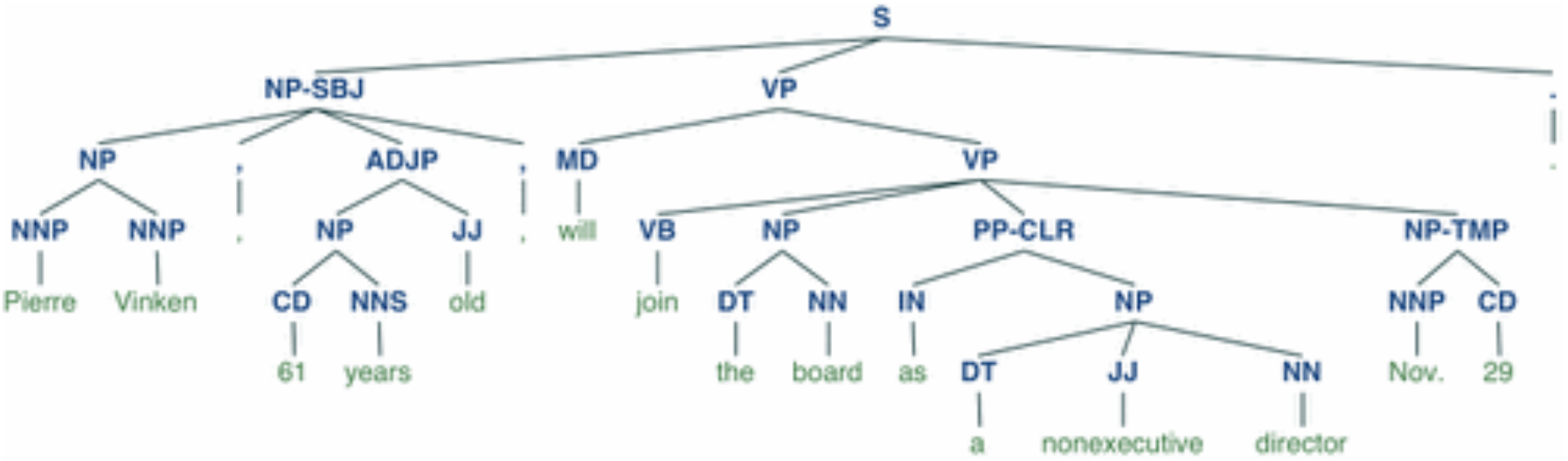
- Coreference resolution is the task of finding all expressions that refer to the same entity in a text





Parsing

- A natural language parser is a program that works out the grammatical **structure of sentences**, for instance, which groups of words go together (as "phrases") and which words are the **subject** or **object** of a verb



Taming Text Data, aka Preprocessing

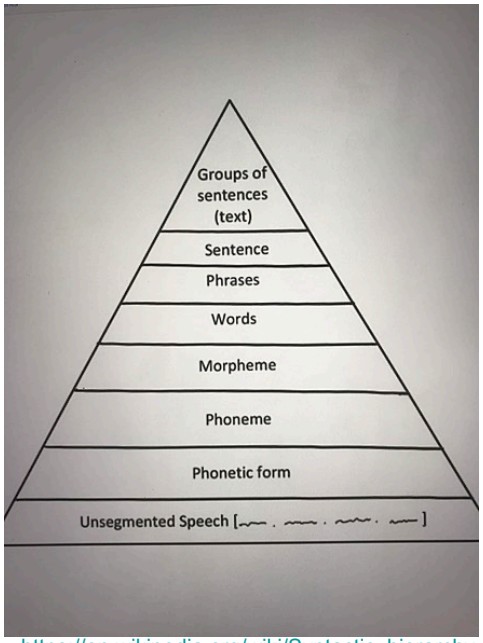
- Definitions
- Normalization
- Tokenization & Segmentation
- Stop words
- Stemming & Lemmatization
- Zipf's law & Dictionary

Alexander and Bucephalus

Language Syntactic Hierarchy – Linguistic View

- Phonemes
  - A unit of sound, e.g. /p/, /t/ and /æ/
- Morphemes
  - Smallest meaningful unit in a word, e.g. the word *national* has two morphemes: ***nation*** a noun, and ***-al*** a suffix
- Words
- Phrase
  - Noun phrase: ***the cat***
  - Verb phrase: "***jumped over the lazy dog***"

  in the sentence "*a fox jumped over the lazy dog*"
- Sentence



https://en.wikipedia.org/wiki/Syntactic_hierarchy

Language Terminology – Computational View

- Character
- N-gram character
  - E.g. tri-gram characters like "***a f***", " ***fo***", "***fox***", and "***ox*** " in the sentence "*a fox jumped over the lazy dog*"
- Word
- N-gram word
  - E.g. tri-grams like "***a fox jumped***", "***fox jumped over***" , and "***jumped over the***" in "*a fox jumped over the lazy dog*"
- Compound noun
  - is made up of at least two nouns like ***post office***, ***San Francisco***
- Multiword Expression
  - Made up of at least two words like
    - ***hang up the gloves*** (idiom)
    - ***in short***

Language Hierarchy – Computational View (cont.)

- Token
  - A token is an instance of a sequence of characters
  - used as the unit of processing
  - can be any of the previously mentioned units
  - E.g. when tokenized by words: "***a***", "***fox***", " ***jumped***", "***over***", "***the***", "***lazy***" and "***dog***" in "*a fox jumped over the lazy dog*"
- Dictionary or Vocabulary list
  - List of unique tokens in the text data
- Sentence
- Paragraph
- Document
- Corpus
  - a collection of text data

Text Normalization

- "Normalization" harmonizes the written forms of the words with same meanings
  - ***U.S.A.*** and ***USA***

- Some examples:
  - deleting periods
    - ***U.S.A.*** → ***USA***
  - deleting hyphens
    - ***anti-discriminatory*** → ***antidiscriminatory***
  - Accents
    - French ***résumé*** → ***resume***
  - Umlauts
    - German: ***Tuebingen*** → ***Tübingen***

Text Normalization

- Case folding: reduce all letters to lower case
  - It may cause ambiguity but typically helpful
    - **General Motors** vs. **general motors**
    - **Fed** vs. **fed**
    - **CAT** (City Airport Train) vs. **cat**
- Longstanding Google example:
  - Search **C.A.T.**

Numbers and dates

- Do the numbers, dates, etc. bring information?
  - If included, the dictionary size may explode!

- Usually replaced by special tokens, e.g.
  - Numbers with <num>
  - Dates with <dates>

Tokenization

- Tokenization
  - Splitting a running text into tokens

- Sample questions when tokenizing:
  - ***Finland's capital*** → ***Finland*** AND ***s***? ***Finlands***? ***Finland's***?
  - ***Hewlett-Packard*** → ***Hewlett*** and ***Packard*** as two tokens?
  - ***state-of-the-art*** → break up hyphenated sequence?
  - ***San Francisco*** → one token or two? ***San_Francisco*** or "***San***" and "***Francisco***"?
  - ***lowercase*** → ***lower-case***, ***lowercase***, or ***lower case***?
  - what's, don't → ***what is***, ***do not***?

Tokenization approaches

- Two general tokenization approaches
  - Rule-based tokenization
  - Subword tokenization

- Rule-based tokenization
  - Tokenization using a set of rules
    - done in libraries like spaCy or Moses, or by using Regular Expressions
  - needs language-specific knowledge
  - can become problematic in morphologically rich languages

Subword tokenization

- Tokenization using the vocabulary list, extracted from the statistics of subwords in a training corpus
  - E.g. "*structurally*" appears rarely, however, its meaning can be inferred from "*structure*" which may appear much more often in a corpus
  - Lemmetizers and stemmers turn "*structurally*" and "*structure*" to the same stem (like "*structur*"), however
    1. They also required language knowledge
    2. They typically remove the differences between these two words

- Core approach: create a vocabulary list of more frequent subwords, such that you can decompose less frequent words into the defined subwords

- Subword tokenization radically decrease Out-Of-Vocabulary tokens

Byte Pair Encoding (BPE)

- The core idea of BPE comes from information theory and compression

Sketch of tokenization with BPE:

1. First, pre-tokenize the training corpus to a dictionary of words and counts
   - Pre-tokenization is done e.g. by white space splitting
2. Add special character "_" to the end of each word in the dictionary
3. Start from a vocabulary list with all single characters
4. Find the most frequent pair of characters, merge the characters, and add them to the vocabulary list
5. Repeat step 4 till some limits on vocabulary size are reached

- Learn more at "*Speech and Language Processing (3rd ed.) D. Jurafsky and J. H. Martin*", section 2.4.3. https://web.stanford.edu/~jurafsky/slp3/2.pdf (the resource of the example)

Sennrich, R., Haddow, B., & Birch, A.. Neural Machine Translation of Rare Words with Subword Units. In *Proc. Of ACL (2016)*

Byte Pair Encoding – example

- Consider a tiny training corpus that leads to the following dictionary and vocabulary

	dictionary	**vocabulary**
5	l o w _	_, d, e, i, l, n, o, r, s, t, w
2	l o w e s t _	
6	n e w e r _	
3	w i d e r _	
2	n e w _	

	dictionary	vocabulary
5	l o w _	_, d, e, i, l, n, o, r, s, t, w
2	l o w e s t _	
6	n e w e r _	
3	w i d e r _	
2	n e w _	

First merge

	dictionary	vocabulary
5	l o w _	_, d, e, i, l, n, o, r, s, t, w, r_
2	l o w e s t _	
6	n e w e r_	
3	w i d e r_	
2	n e w _	

Next merge

	dictionary	vocabulary
5	l o w _	_, d, e, i, l, n, o, r, s, t, w, r_, er_
2	l o w e s t _	
6	n e w er_	
3	w i d er_	
2	n e w _	

	dictionary	vocabulary
5	l o w _	_, d, e, i, l, n, o, r, s, t, w, r_, er_
2	l o w e s t _	
6	n e w er_	
3	w i d er_	
2	n e w _	

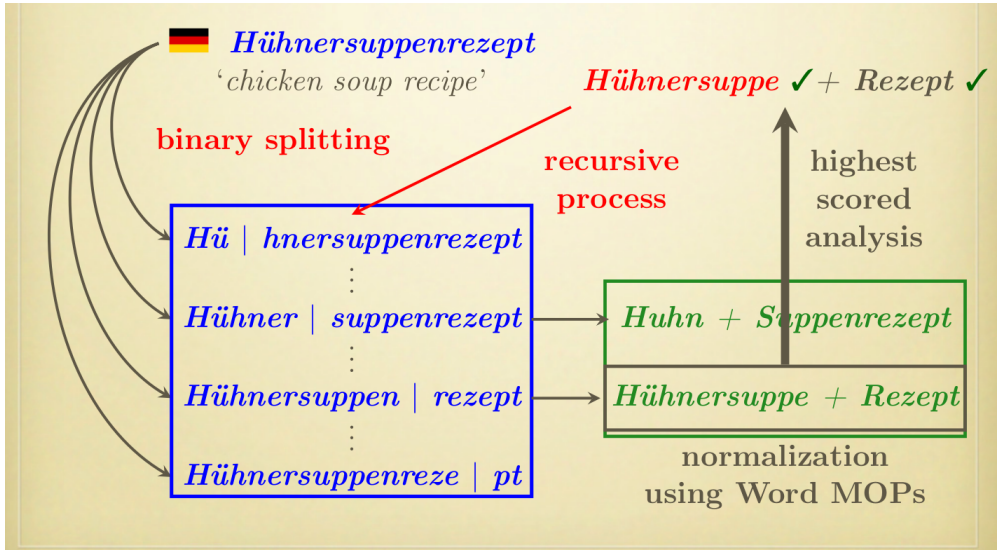
Next merge

	dictionary	vocabulary
5	l o w _	_, d, e, i, l, n, o, r, s, t, w, r_, er_, ew
2	l o w e s t _	
6	n ew er_	
3	w i d er_	
2	n ew _	

…

Segmentation

- Segmentation
  - Splitting a compound word into tokens
- French
  - ***L'ensemble*** → one token or two? ***L*** ? ***L'*** ? ***Le*** ?
- German compound nouns
  - ***Halsschlagader*** → ***Hals Schlag Ader***?
  - Compound words in German usually require **compound splitter**



https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/mcs/

Stop words

- Stop words
  - The commonest words, like ***the***, ***a***, ***and***, ***to***, ***be***
  - They carry little or no semantic information

- Stop words are typically excluded from the corpus

- can also be important, especially in combination with other words, e.g.:
  - Phrases: ***"King of Denmark"***, ***"To be or not to be"***
  - Titles, etc.: ***"Let it be"***
  - Definitional purposes: "**flights to London**"

Lemmatization

- reduces inflectional/variant forms to base forms, e.g.
  - ***am***, ***are***, ***is*** → ***be***
  - ***car***, ***cars***, ***car's***, ***cars'*** → ***car***
  - ***the boy's cars are different colors*** → ***the boy car be different color***

- A lemmatizer uses the resources like WordNet to find and replace base forms

Stemming

- Morphemes consists of
  - **Stem**: the core meaning-bearing units
  - **Affixes**: pieces that adhere to stems
- A stemmer reduces words to their "stems" by crude affix chopping, e.g.
  - ***automate, automates, automatic, automation*** $\rightarrow$ ***automat***
  - ***for example compressed and compression are both accepted as equivalent to compress*** $\rightarrow$

    for exampl compress and compress ar both accept as equival to compress

- Stemming
  - Pros: reduces variation, is typically faster than Lemmatization
  - Cons: may harm precision, and increase ambiguity

Porter's algorithm

- Commonest algorithm for stemming English
  - consists of a set of grammatical commands

- Typical rules:
  - *sses* $\rightarrow$ *ss*
  - *ies* $\rightarrow$ *i*
  - *ational* $\rightarrow$ *ate*
  - *tional* $\rightarrow$ *tion*

Give it a try: https://text-processing.com/demo/stem/

Zipf's Law

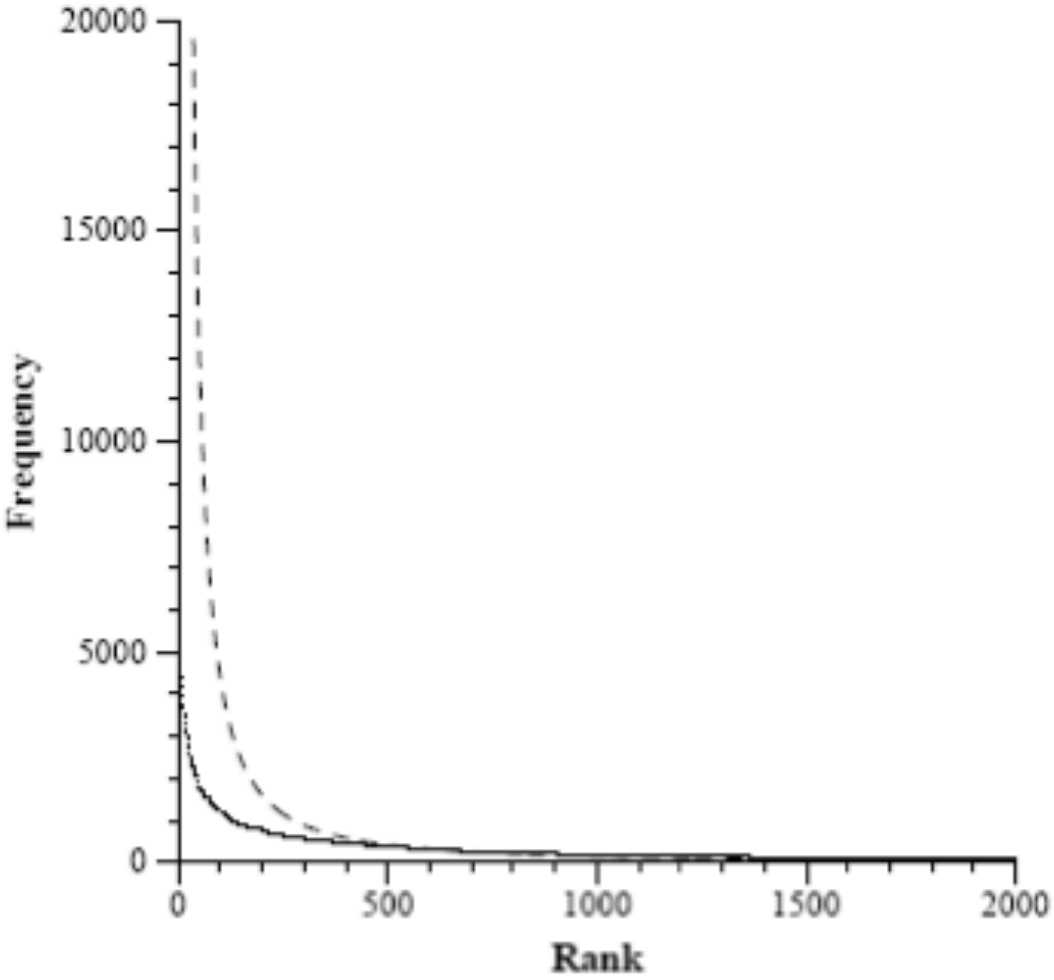
- Sort the vocabularies of a dictionary based on their corpus frequencies
- E.g. results for the Brown Corpus

Rank	Token	Frequency
1	the	22038615
2	be	12545825
3	and	10741073
4	of	10343885
5	A	10144200
…	…	…

- Top words are stop words
- The ones at the bottom are *rare words*

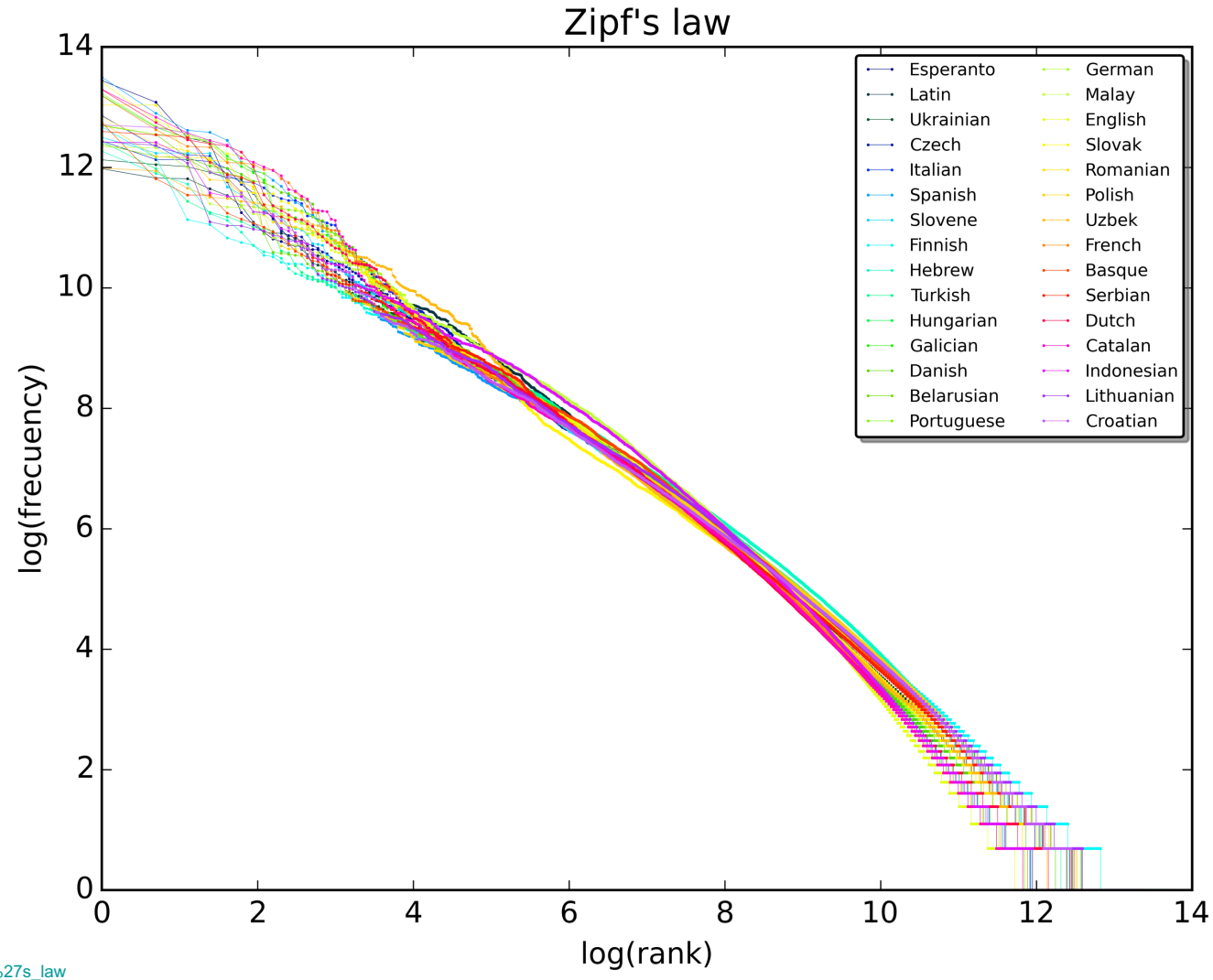
Zipf's Law

- The plot of rank and frequency looks something like this



Zipf's Law

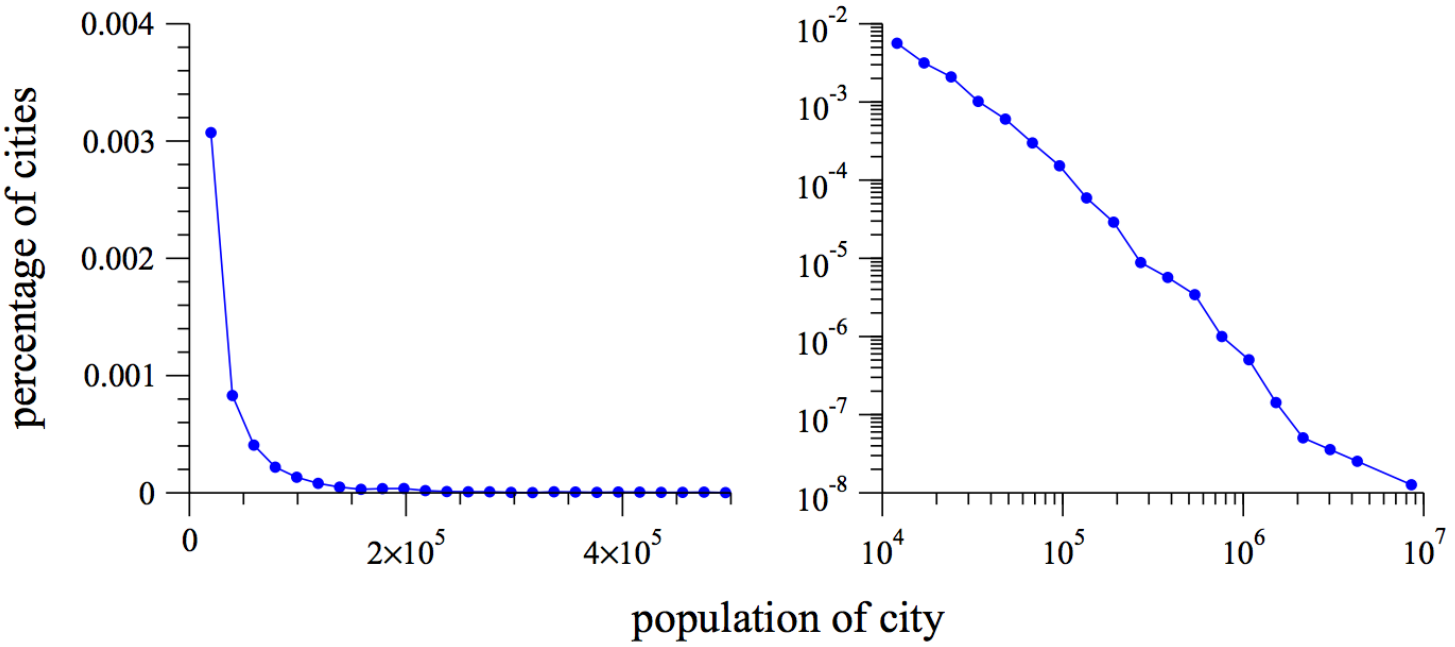
- When transfer both axes with logarithms:



https://en.wikipedia.org/wiki/Zipf%27s_law

Power Law in social science

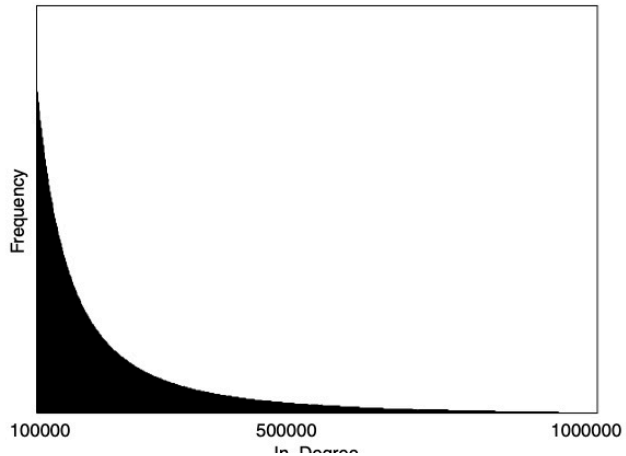
- Observed in other ranking statistics
  - Population rank of cities
  - Pagerank scores of websites
  - Income distribution
  - Corporation sizes

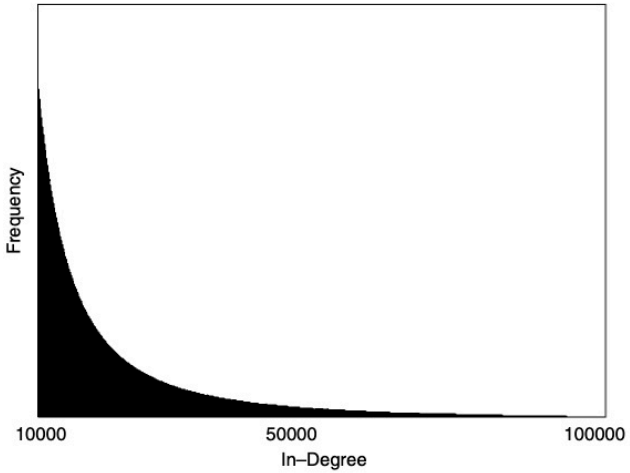


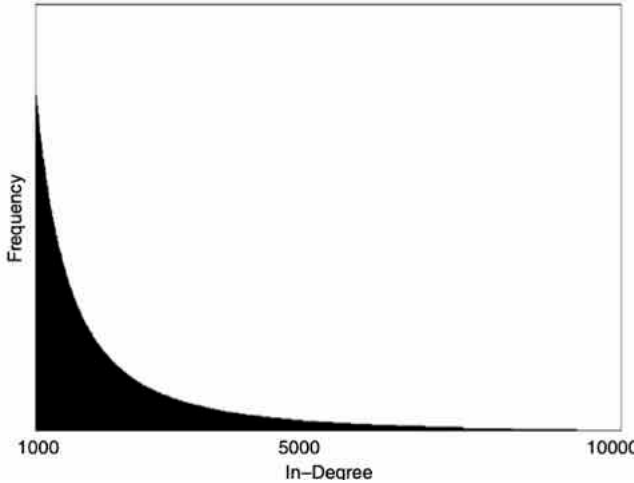
https://statweb.stanford.edu/~owen/courses/306a/ZipfByHera.pdf

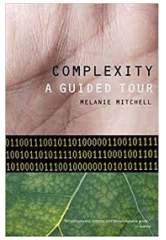
Power Law in web

- Approximation of the shape of web based on the number of incoming links to each website







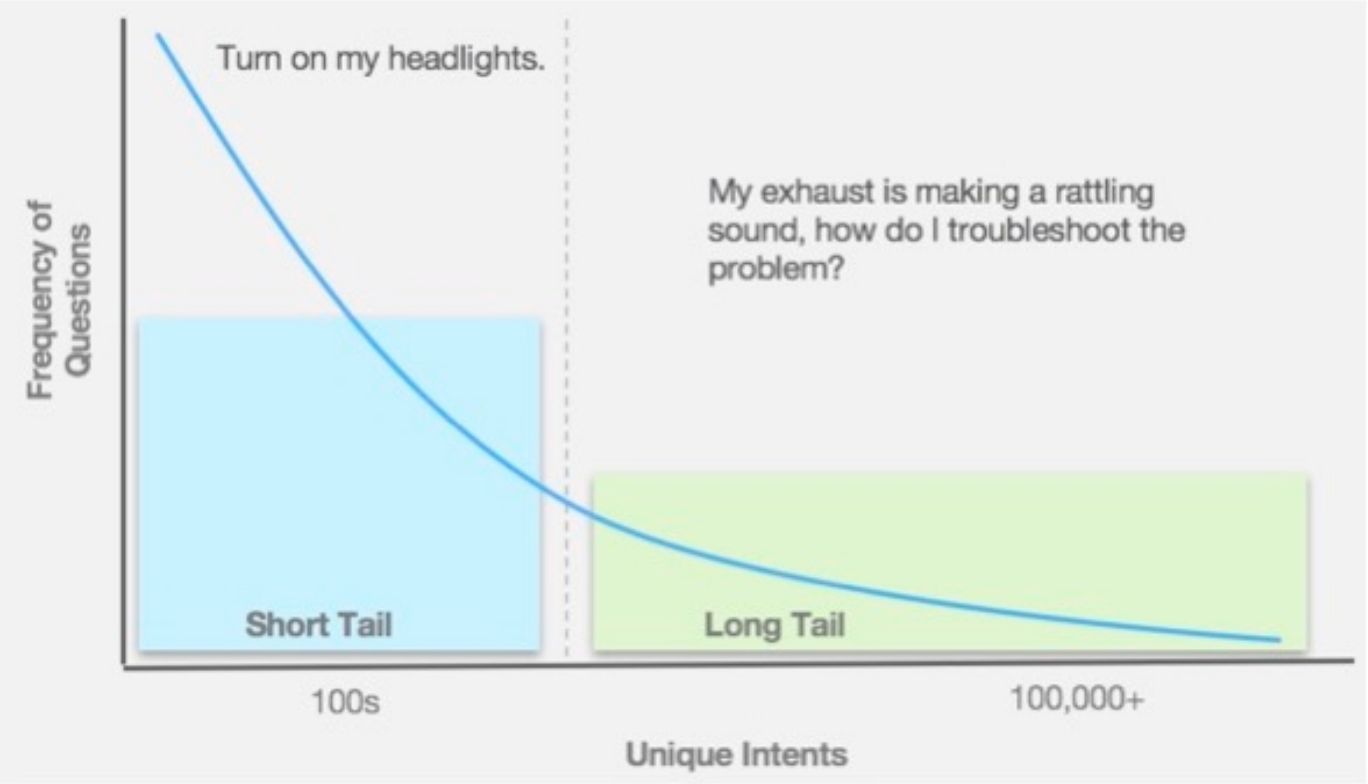


What does Zipf' Law tell us when creating a dictionary?

- Highly frequent words cover a large portion of tokens in a corpus
  - Only 135 most frequent words cover half of the Brown Corpus
- A large portion of the dictionary consists of words with very low frequencies
  - Typically removing words with frequencies of lower than 3 halves the size of dictionary
  - Selecting a *proper* threshold to remove low-frequent words from dictionary is necessary

What does Zipf' Law tell us when creating a dictionary?

- The challenge of *long tail*
  - Phenomena (here words) with low frequency are challenging for statistical model
  - Very typical in NLP



https://www.syntravlaanderen.be/sites/default/files/atoms/files/Conversational%20Bots%20IBM%20Watson.pdf

Libraries

AllenNLP

gensim

Stanford CoreNLP

spaCy

huggingface.co

polyglot

scikit learn

PYTORCH

Assignment 1