

# Natural Language Processing with Deep Learning

## Basics of NLP & Text Processing



Navid Rekab-Saz

[navid.rekabsaz@jku.at](mailto:navid.rekabsaz@jku.at)

# Course

- Exam and project, both possible!

## Exam or Project?

Visible groups All participants ▾

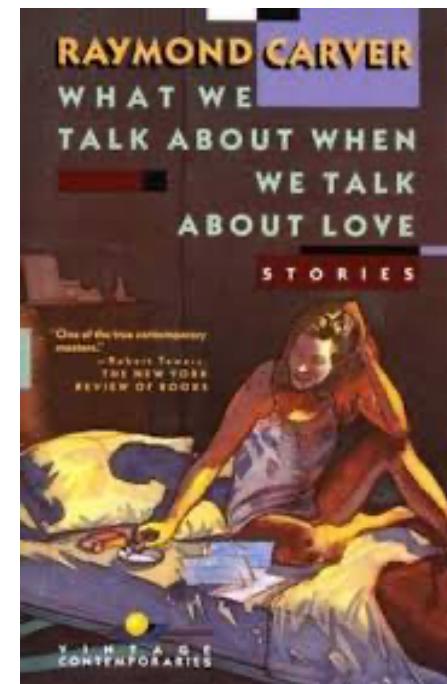
### Responses

Choice options	<b>I prefer to take an exam</b> <input type="checkbox"/>	<b>I prefer to do a project</b> <input type="checkbox"/>
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

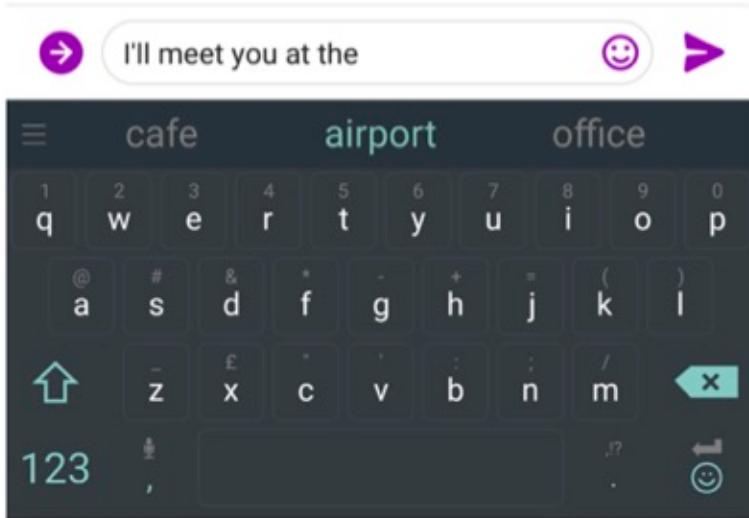
- Applications
- Low-level tasks



# **What We Talk About When We Talk About NLP**

- **Applications**
- Low-level tasks

# Auto Correction / Spell Checking / Predictive Text



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?



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

# Sentiment Analysis / Market Intelligence



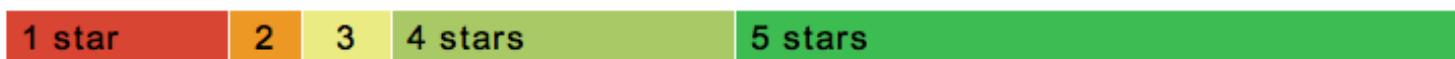
**HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner**

\$89 online, \$100 nearby    ★★★★☆ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

## Reviews

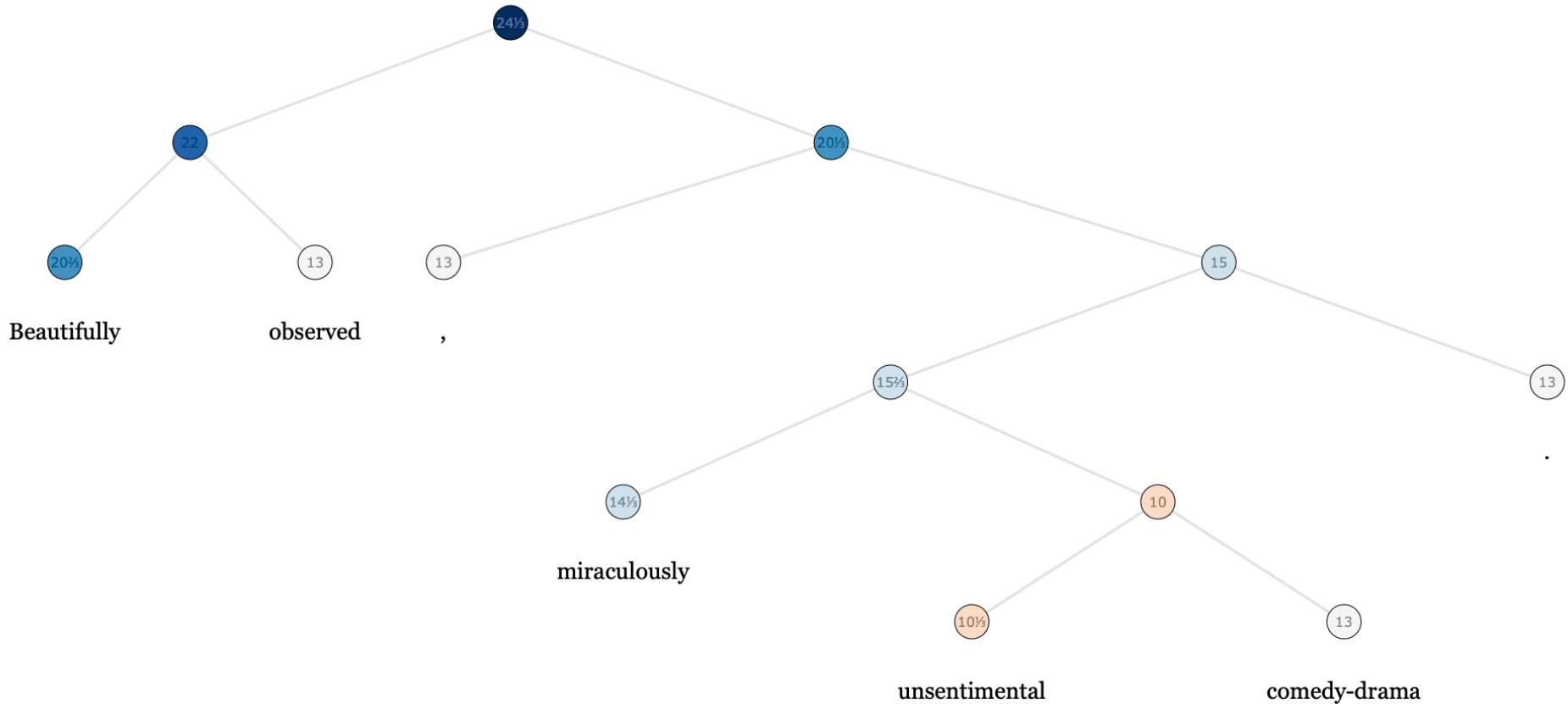
**Summary** - Based on 377 reviews



### What people are saying

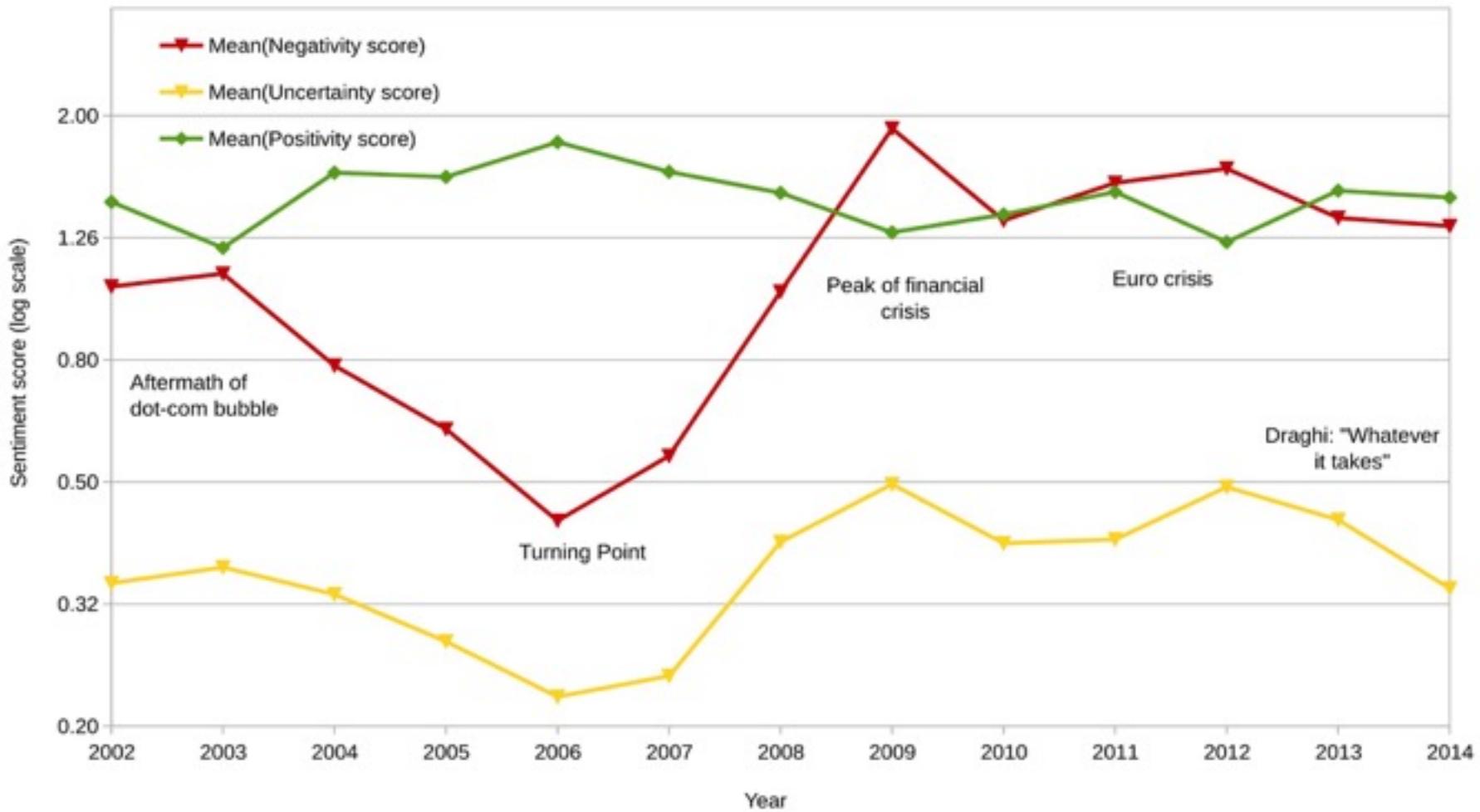
ease of use	1 star	"This was very easy to setup to four computers."
value	2 stars	"Appreciate good quality at a fair price."
setup	3 stars	"Overall pretty easy setup."
customer service	4 stars	"I DO like honest tech support people."
size	4 stars	"Pretty Paper weight."
mode	4 stars	"Photos were fair on the high quality mode."
colors	4 stars	"Full color prints came out with great quality."

# Sentiment Analysis / Market Intelligence

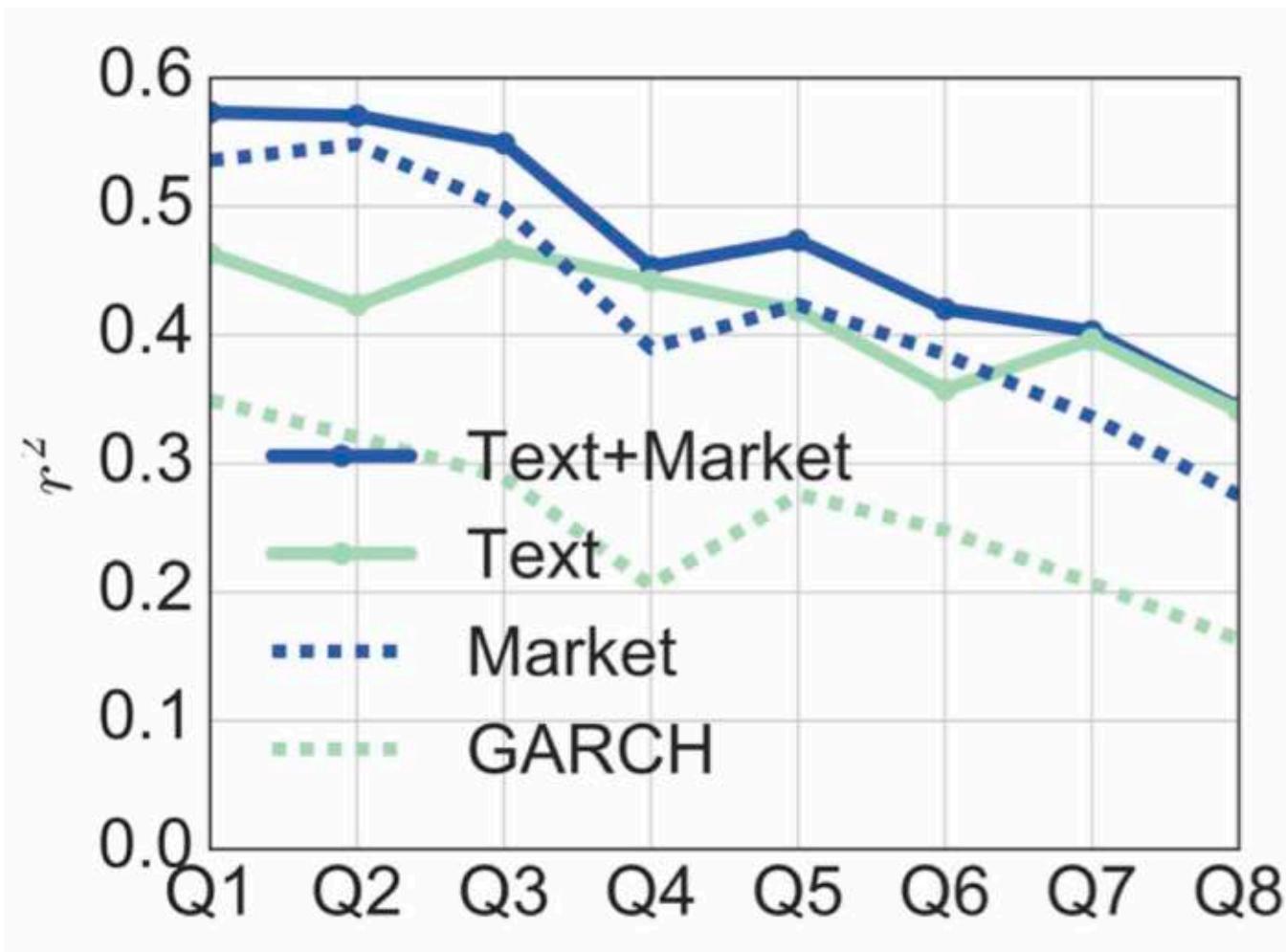


<https://nlp.stanford.edu/sentiment/treebank.html>

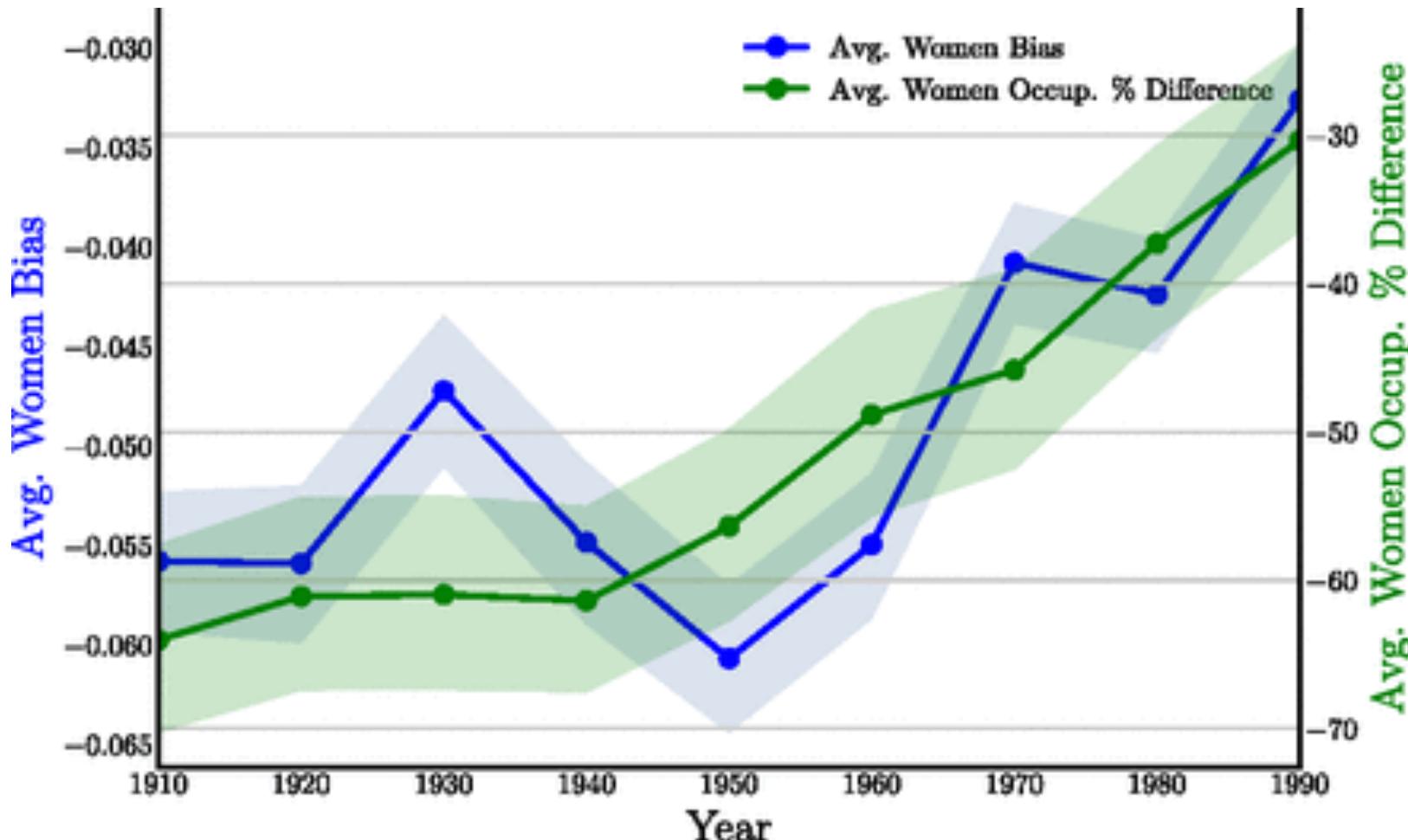
# Sentiment Analysis / Market Intelligence



# Sentiment Analysis / Market Intelligence



# Monitoring Societal Changes



# Text Classification

SIMPLIFIED ASSISTED ORIGINAL IMAGES ENTRIES

Famine may be unfolding 'right now' in Yemen, warns UN relief wing

17 November 2017 – The United Nations relief wing on Friday, warned of famine-like conditions unfolding in Yemen, as a blockade on aid and other essential goods by a Saudi-led coalition fighting Houthi rebels there enters its 12th day.

Jens Laerke, spokesperson for the UN Office for the Coordination of Humanitarian Affairs (OCHA), sounded the alarm during the regular bi-weekly news briefing in Geneva.

He was responding to a question from a journalist who asked him to clarify a warning yesterday from UN aid chiefs

that the closure of air, sea and land ports in Yemen threatened millions of vulnerable children and families.

"It means that these are the number of people in areas where there's an IPC4 – Integrated Phase Classification 4 – which is the last step before obviously 5, which is famine [...] But you are correct, there may be as we speak right now, famine happening, and we hear children are dying, I mean, there's excess mortality as a cause and consequence of undernourishment."

Yemen imports up to 90 per cent of its daily needs, including fuel, which has now reached crisis levels.

Reserves are in such short supply that three Yemeni cities have been unable to pump clean water to residents in recent days, according to UN partner the Red Cross.

This has left one million people at risk of a renewed cholera outbreak, just as the country emerges from the worst epidemic in modern times.

Other diseases are also a threat, including diphtheria, a serious infection of the nose and throat, that's easily prevented with a vaccine.

It's "spreading fast" and has already claimed 14 lives, according to the World Health Organization (WHO), which said that a vaccination campaign is planned in nine days' time.

In addition to water and sewage problems in Hodeida, Sa'ada and Taiz, the Red Cross warned that the capital Sana'a and other cities "will find themselves in the same situation" in two weeks – unless imports of essential goods resume immediately.

Also at the briefing, Alessandra Vellucci, for the UN Information Service (UNIS) recalled yesterday's statement in New York from Stéphane Dujarric, Spokesman for the UN Secretary-General regarding a letter the UN chief sent to the

SHOW SUGGESTIONS FOR  
6 selected

Food

Agriculture

Logistics

<https://www.thedep.io>

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

# Machine Translation

ENGLISH - DETECTED ENGLISH GERMAN SPANISH

"One of history's few iron laws is that luxuries tend to become necessities and to spawn new obligations." X "Eines der wenigen eisernen Gesetze der Geschichte ist, dass Luxus dazu neigt, zu Notwendigkeiten zu werden und neue Verpflichtungen zu schaffen." ☆

106/5000

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."

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

Translate document

12

# Machine Translation

The screenshot shows a machine translation interface with two main panels. The left panel is for English input, and the right panel is for German output. Both panels include a speaker icon for audio playback, a edit icon, and a more options icon.

**ENGLISH - DETECTED** ENGLISH GERMAN ENGLISH SPANISH

"Humans think in stories, and we try to make sense of the world by telling stories." X "Menschen denken in Geschichten, und wir versuchen, die Welt zu verstehen, indem wir Geschichten erzählen." ☆

84/5000

Translate from **English** (detected) Translate into **German**

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

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

Translate document

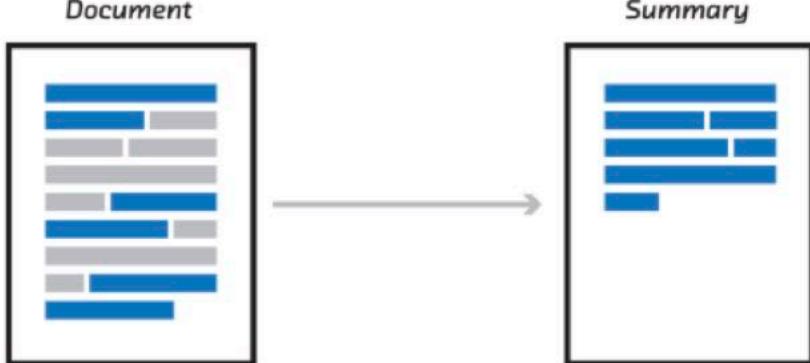
13

NEW YORK TIMES BESTSELLER  
Yuval Noah Harari  
New York Times Bestselling Author of *Sapiens*  
**Homo Deus**  
A Brief History of Tomorrow  
"Provocative... The kind of book of a lifetime."  
—NEW YORK TIMES

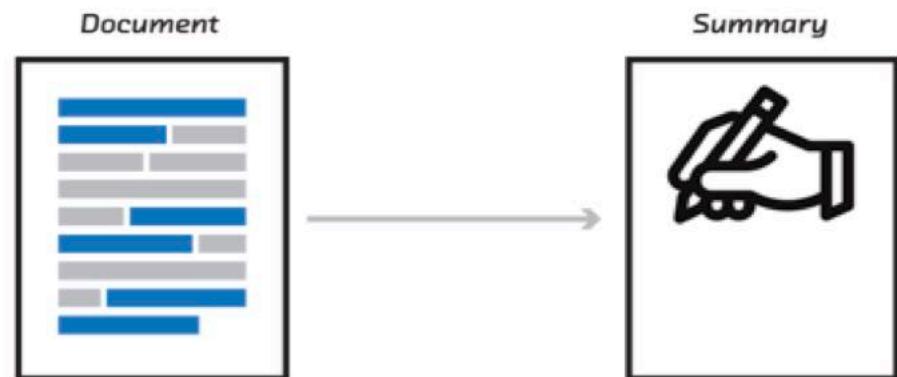
# Summarization

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

**Extractive Summarization**



**Abstractive Summarization**



# Question Answering



A screenshot of a Google search results page. The search query is "when was the last total eclipse in europe". The results show a large featured snippet box containing the date "August 11, 1999" and a brief description: "The last total eclipse in continental Europe occurred on August 11, 1999." Below the snippet is a link to "Solar eclipse of August 12, 2026 - Wikipedia". The page also includes standard Google navigation links like All, News, Images, Videos, Maps, More, Settings, and Tools, along with search and refresh icons.

Google

when was the last total eclipse in europe

All News Images Videos Maps More Settings Tools

About 42.000.000 results (0,92 seconds)

**August 11, 1999**

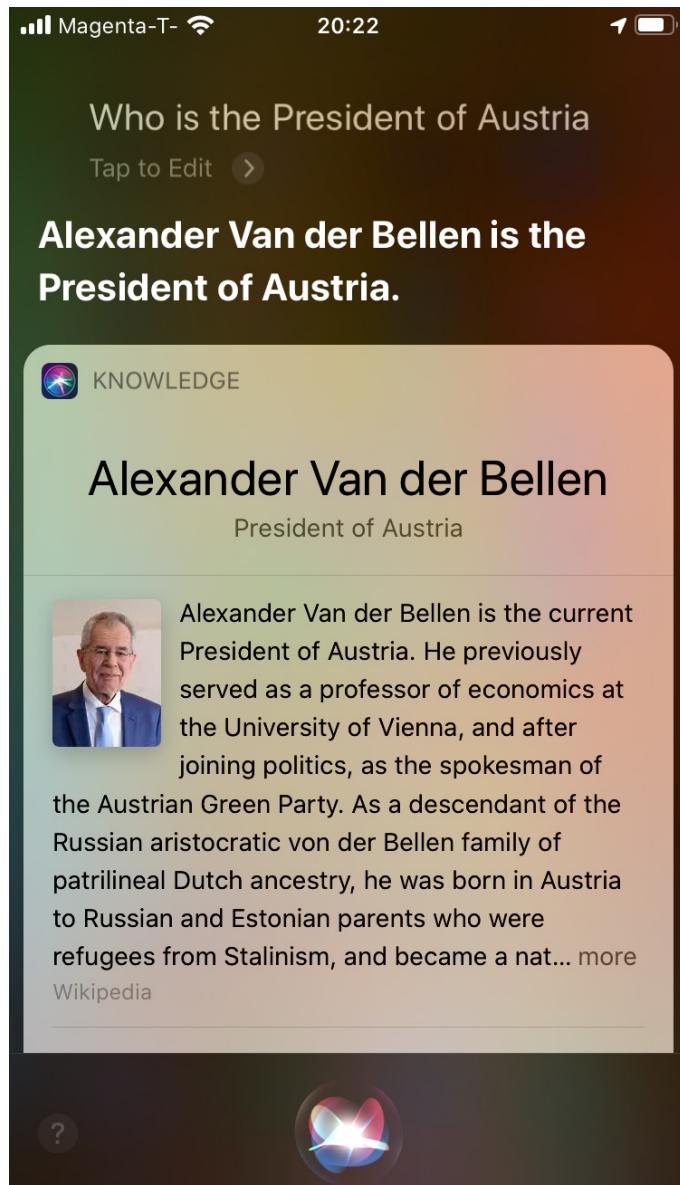
The last total eclipse in continental Europe occurred on August 11, 1999.

en.wikipedia.org › wiki › Solar\_eclipse\_of\_August\_12,\_2026

[Solar eclipse of August 12, 2026 - Wikipedia](#)

About Featured Snippets Feedback

# 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?

# Information Retrieval

The screenshot shows a search interface with a search bar containing "coronavirus vaccine". Below the search bar are filters for "Web", "Images", "Videos", "News", and "Maps", along with a "Settings" dropdown. A location filter for "Austria" is set to "Moderate" and "Any Time". The results section is titled "Recent News" and includes a thumbnail for a CNN article about a Chinese coronavirus vaccine, a thumbnail for a Fox News article by Dr. Marc Siegel, and a snippet from Trip's medical evidence database.

Recent News

Coronavirus Update: Vaccine Almost Complete Years Before Outbreak, But No On... International... | 5h

Dr. Marc Siegel on reported coronavirus mutation: 'I think we can get a vaccine to fit all' | Fox News | 7h

**Trip**  
Find evidence fast

SEARCH PICO ADVANCED PRO RECENT PRO

coronavirus vaccine

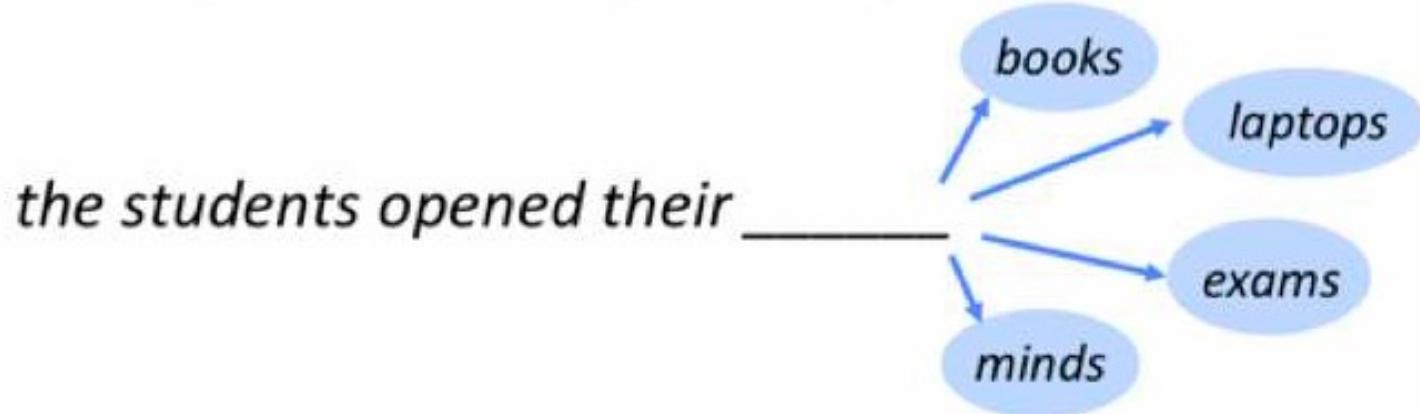
322 results for **coronavirus vaccine** by quality ▾ Latest & greatest Alerts Export Snippets

- 1. **Antibody response to equine coronavirus in horses inoculated with a bovine coronavirus vaccine**  
Full Text available with Trip Pro PRO  
2017 The Journal of Veterinary Medical Science  
Tweet this Star this Report broken link Primary Research
- 2. **Safety and immunogenicity of an anti-Middle East respiratory syndrome coronavirus DNA vaccine: a phase 1, open-label, single-arm, dose-escalation trial. (Abstract)**  
2019 Lancet infectious diseases Controlled trial quality: predicted high ⓘ  
Tweet this Star this Report broken link Primary Research
- 3. **Evaluation of a recombination-resistant coronavirus as a broadly applicable, rapidly implementable vaccine platform**  
Full Text available with Trip Pro PRO  
2018 Communications Biology  
Tweet this Star this Report broken link Primary Research

# 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 CCCCC	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 CCCCC	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.

# 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.”

President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

Annotations:

- President: Person
- of China: Loc
- first: ORDINAL
- United States: Location
- American: Misc
- Tuesday: Date
- night: Time

# Named Entity Recognition (NER)

Also at the briefing, Alessandra Vellucci, for the UN Information Service (UNIS) recalled yesterday's statement in New York from Stéphane Dujarric, Spokesman for the UN Secretary-General regarding a letter the UN chief sent to the Permanent Representative of Saudi Arabia to the United Nations.

In the letter, the Secretary-General said that the blockade imposed by the coalition since 6 November is already reversing the impact of humanitarian efforts. While he welcomed the reopening of Aden port, the Secretary-General noted that "this alone will not meet the needs of 28 million Yemenis."

As such, the Secretary-General called on the Saudi-led coalition to enable the resumption of UN Humanitarian Air Service (UNHAS) flights to Sana'a and Aden airports, and the reopening of Hodeida and Saleef ports so that fuel, food and medical supplies could enter Yemen.

NLP

Entities

Category Editor

SHOW SUGGESTIONS FOR :

2 selected

Locations

Organizations

# Part-of-Speech (POS) Tagging

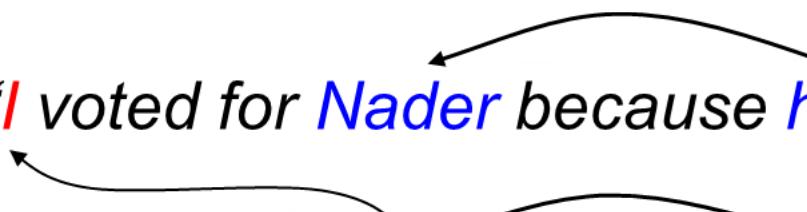
- “A Part-Of-Speech Tagger is a piece of software that reads text in some language and assigns parts of speech to each word, such as noun, verb, adjective”



# Coreference Resolution

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

*“I voted for Nader because he was most aligned with my values,” she said.*



Tell Me This 20 hours ago (edited)

Human: What do we want?

Computer: Natural language processing!

Human: When do we want it?

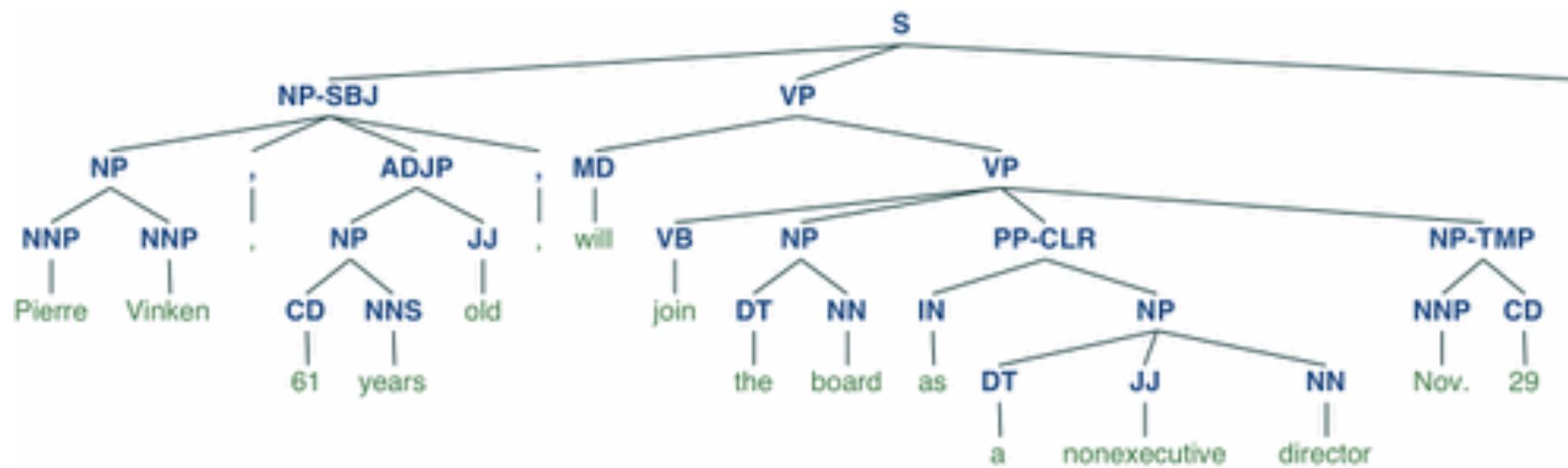
Computer: When do we want what?

Reply • 203

[View reply](#) ▾

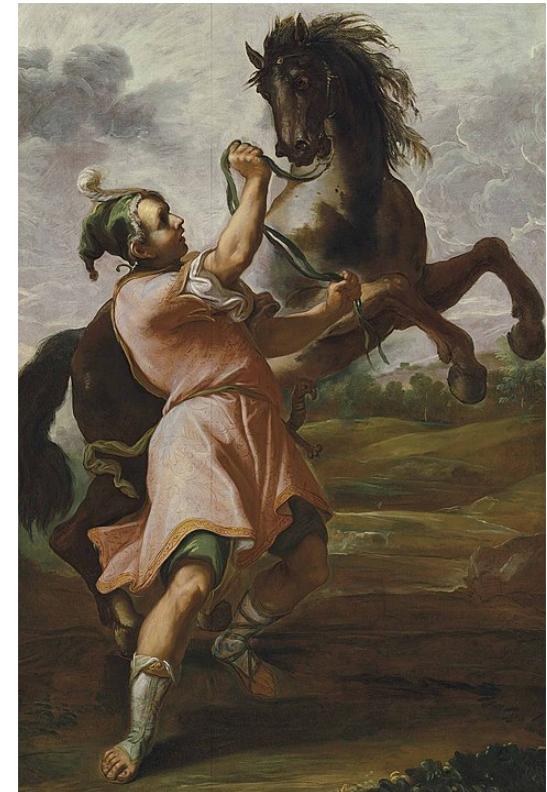
# Parsing

- 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

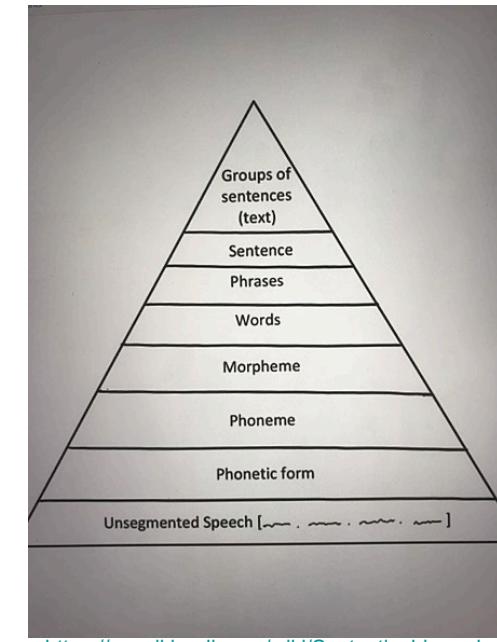
- Definitions
- Normalization
- Tokenization & Segmentation
- Stop words
- Stemming & Lemmatization
- Zipf's law & Dictionary



*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](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
  - Lemmatizers 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)

## Byte Pair Encoding – example

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

	<b>dictionary</b>	<b>vocabulary</b>
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**

5 l o w \_  
2 l o w e s t \_  
6 n e w e r \_  
3 w i d e r \_  
2 n e w \_

**vocabulary**

\_, d, e, i, l, n, o, r, s, t, w

## First merge

**dictionary**

5 l o w \_  
2 l o w e s t \_  
6 n e w e r \_  
3 w i d e r \_  
2 n e w \_

**vocabulary**

\_, d, e, i, l, n, o, r, s, t, w, r\_

## Next merge

**dictionary**

5 l o w \_  
2 l o w e s t \_  
6 n e w e r \_  
3 w i d e r \_  
2 n e w \_

**vocabulary**

\_, d, e, i, l, n, o, r, s, t, w, r\_, er\_

**dictionary**

5 l o w \_  
2 l o w e s t \_  
6 n e w er\_  
3 w i d er\_  
2 n e w \_

**vocabulary**

\_, d, e, i, l, n, o, r, s, t, w, r\_, er\_

Next merge

**dictionary**

5 l o w \_  
2 l o w e s t \_  
6 n ew er\_  
3 w i d er\_  
2 n ew \_

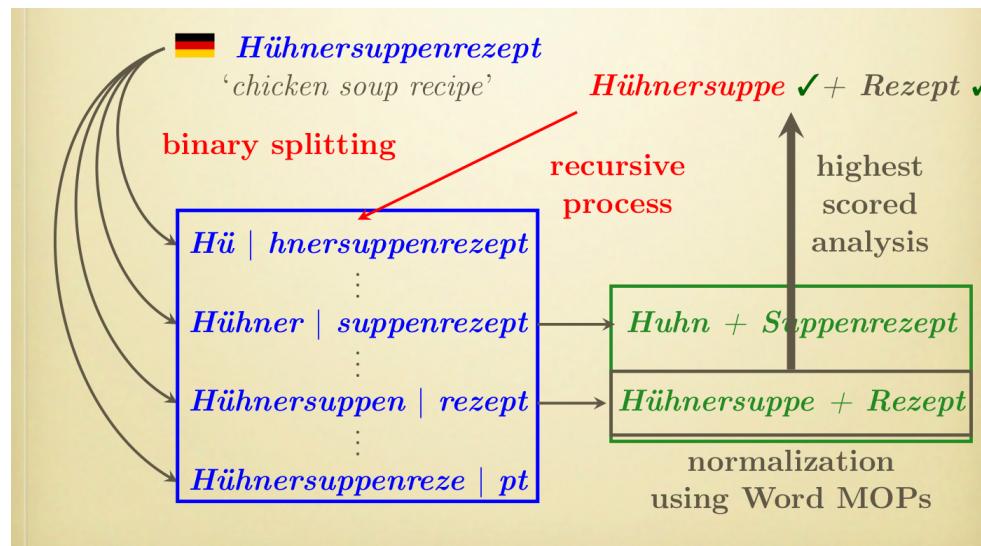
**vocabulary**

\_, d, e, i, l, n, o, r, s, t, w, r\_, er\_, 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**



# 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* → **automat**
  - *for example compressed and compression are both accepted as equivalent to compress* →  
**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* → *ss*
  - *ies* → *i*
  - *ational* → *ate*
  - *tional* → *tion*

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

## Zipf's Law

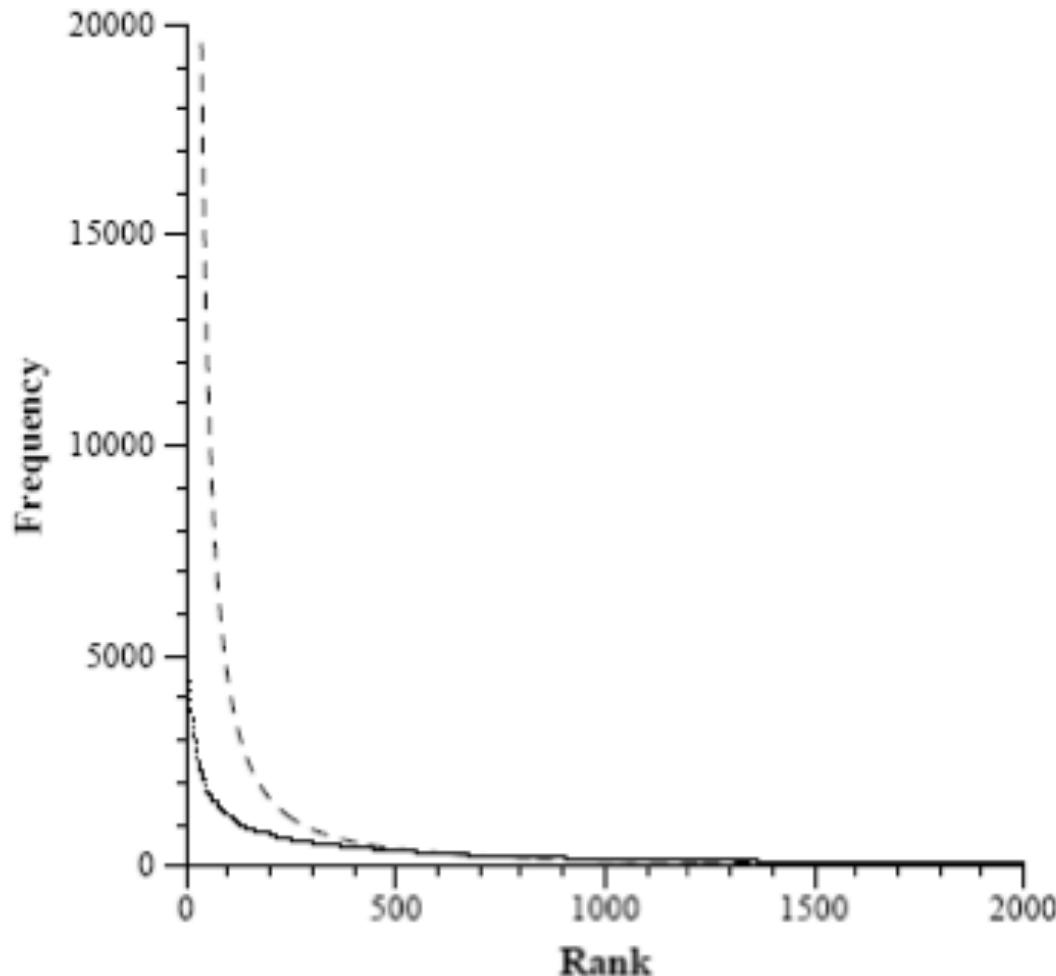
- Sort the vocabularies of a dictionary based on their corpus frequencies
- E.g. results for the [Brown Corpus](#)

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*

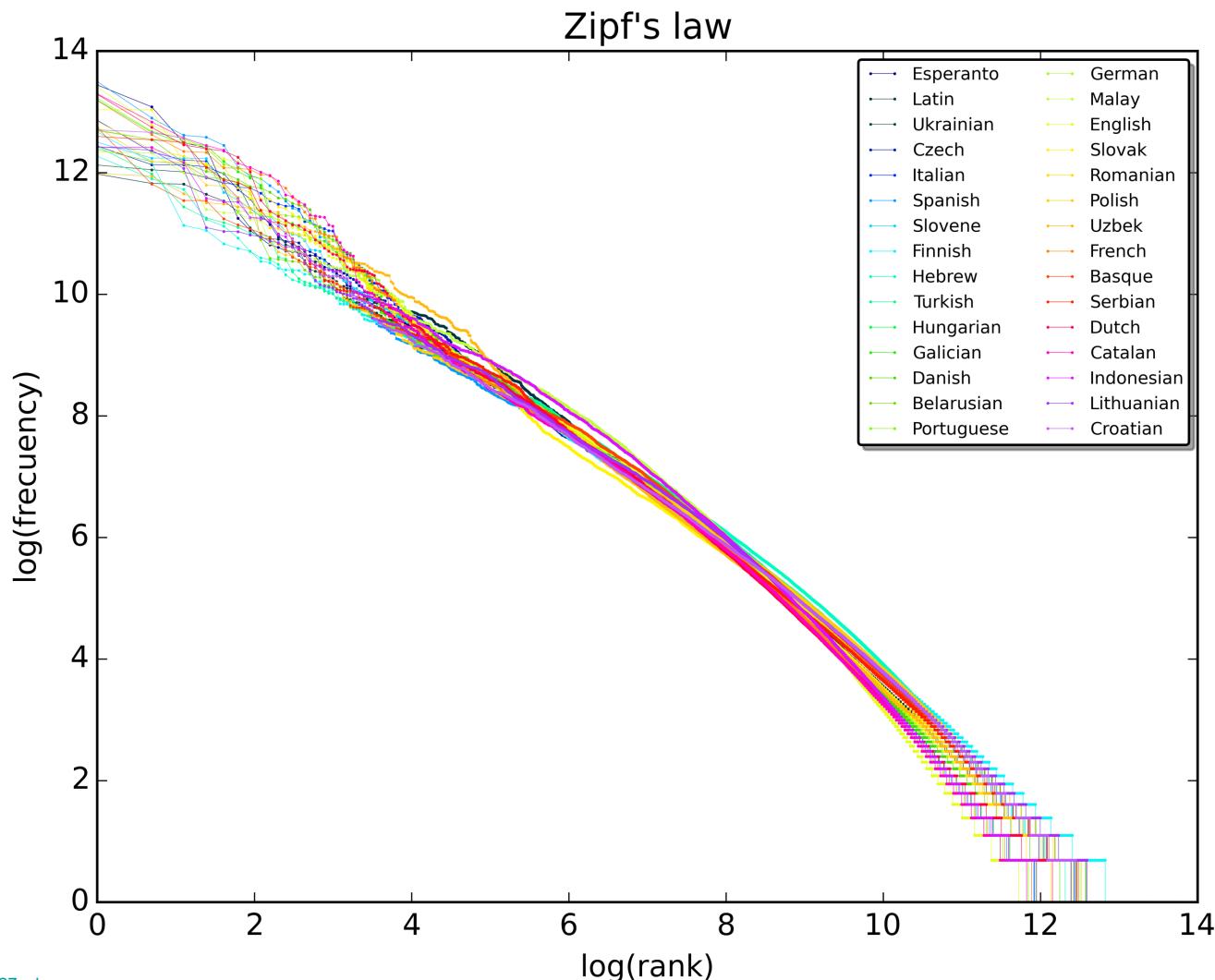
## Zipf's Law

- The plot of rank and frequency looks something like this



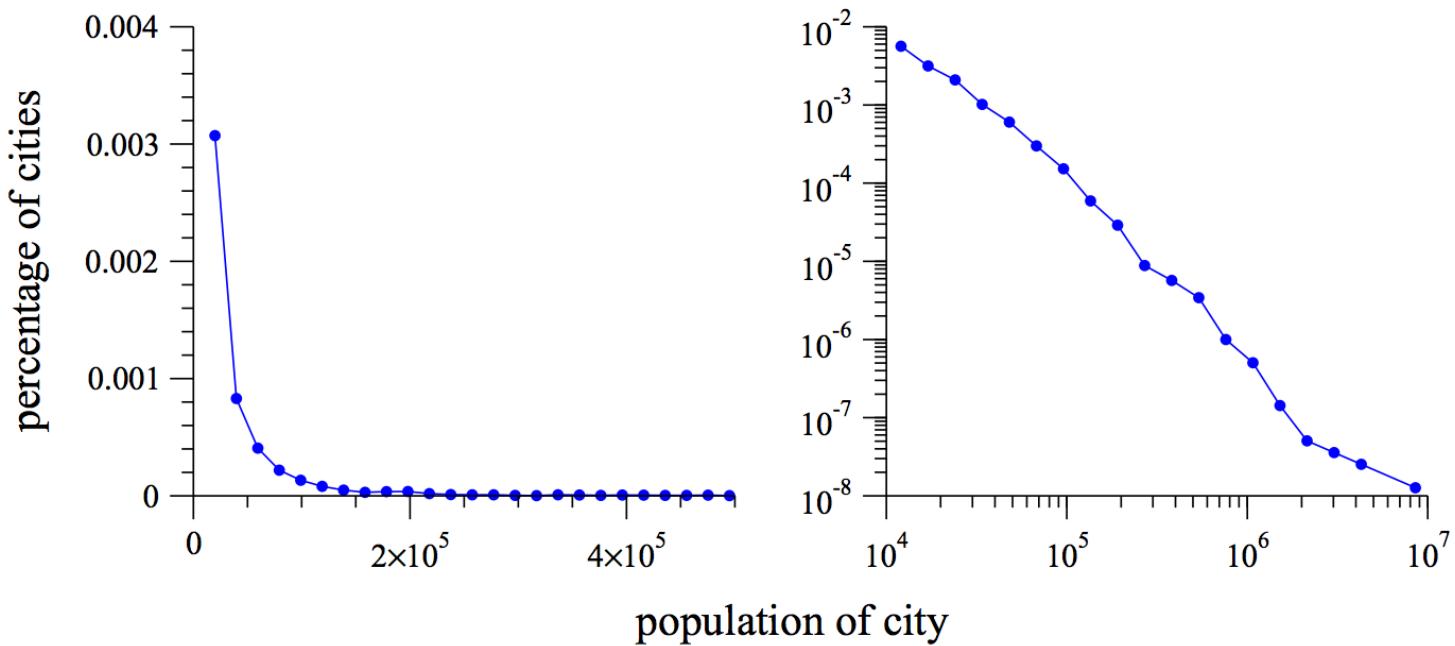
# Zipf's Law

- When transfer both axes with logarithms:



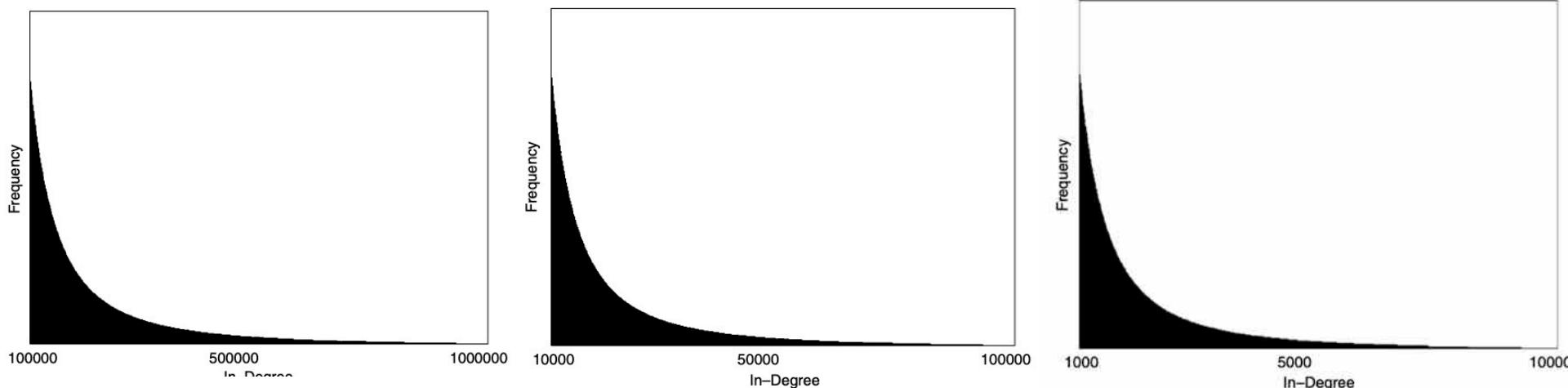
# Power Law in social science

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



# Power Law in web

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

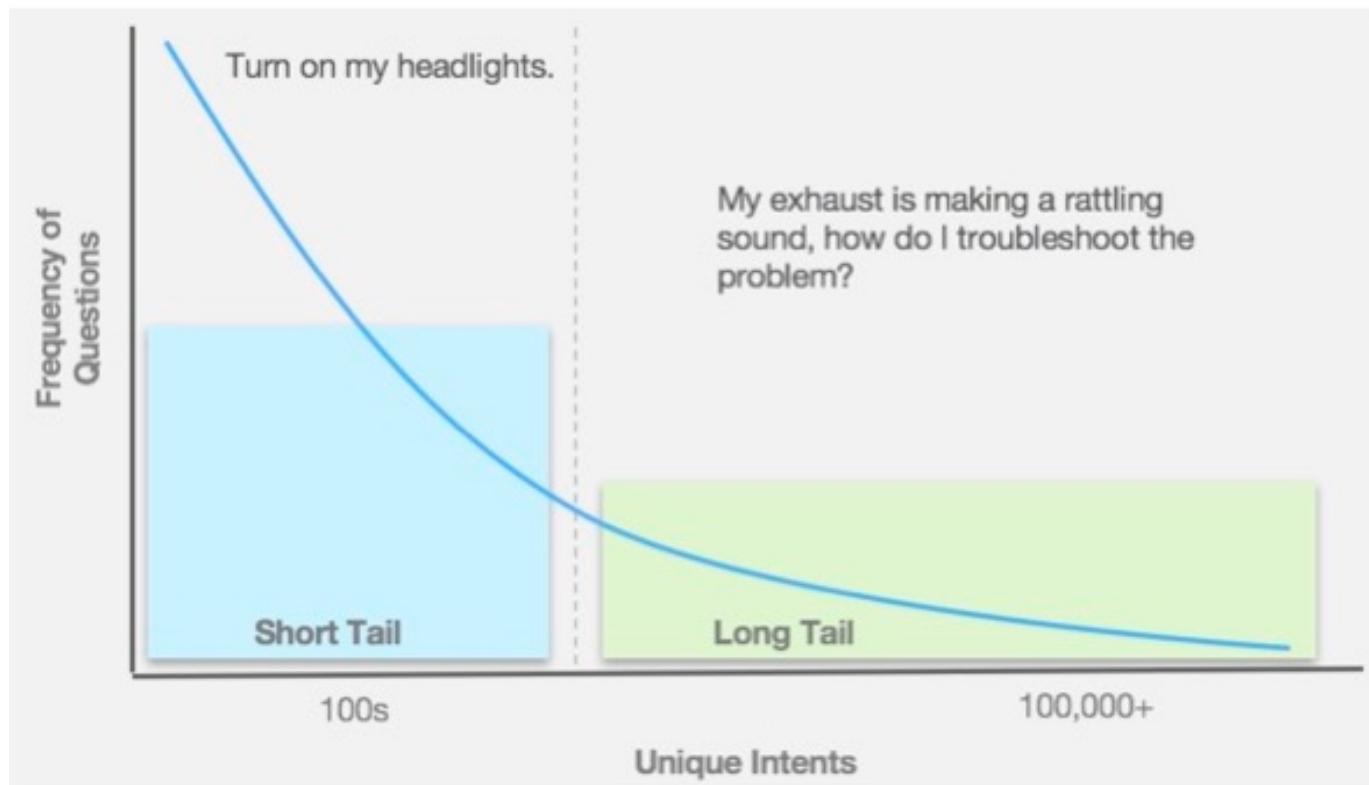


## What does Zipf' Law tell us when creating a dictionary?

- 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



## Libraries

AllenNLP

gensim

Stanford CoreNLP

spaCy



[huggingface.co](https://huggingface.co)

polyglot



PYTORCH

# Assignment 1