Winter semester 2022/23

344.175 VL: Natural Language Processing

Text Processing

Navid Rekab-saz

Email: navid.rekabsaz@jku.at
Office hours: https://navid-officehours.youcanbook.me

JOHANNES KEPLER
UNIVERSITY LINZ

Institute of
Computational
Perception

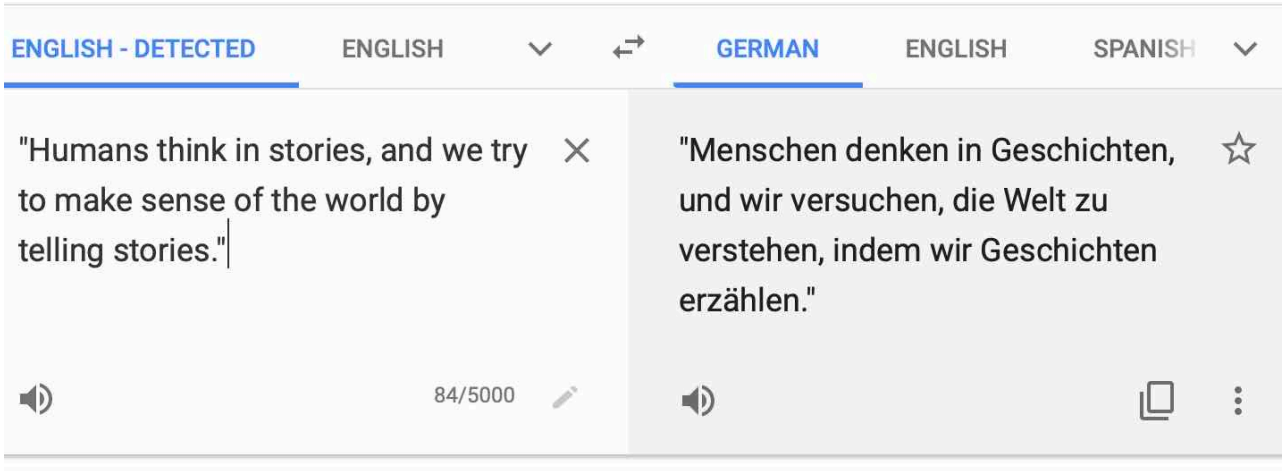
Agenda

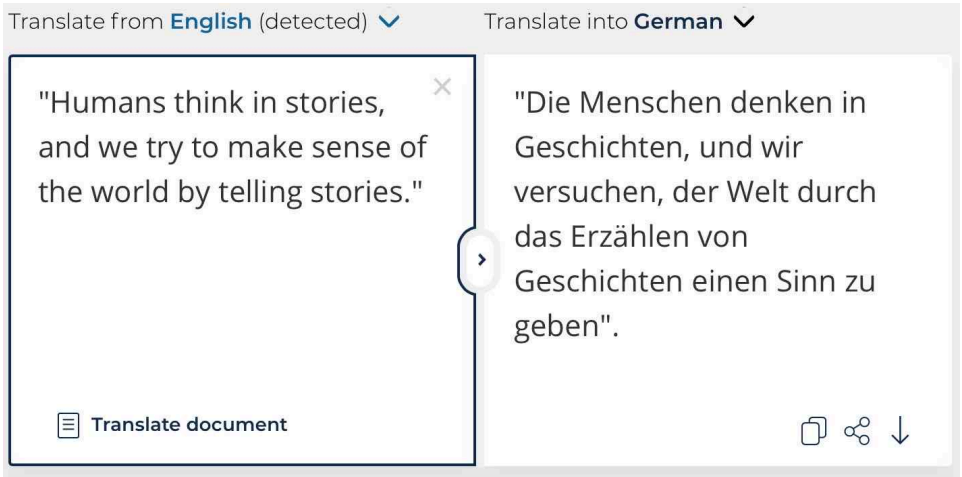
- NLP Applications
- Language as a complex system
- Text preprocessing

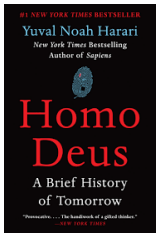
Agenda

- **NLP Applications**
- Language as a complex system
- Text preprocessing

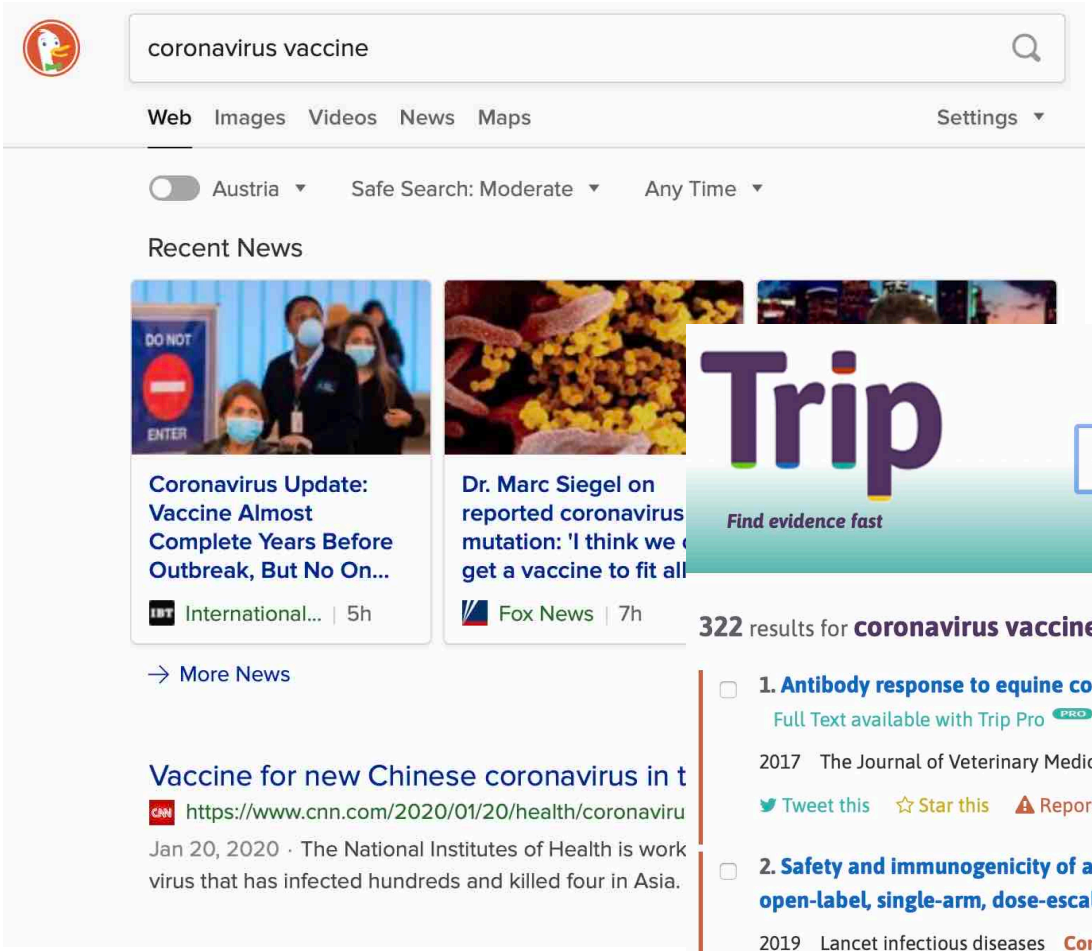
Machine Translation

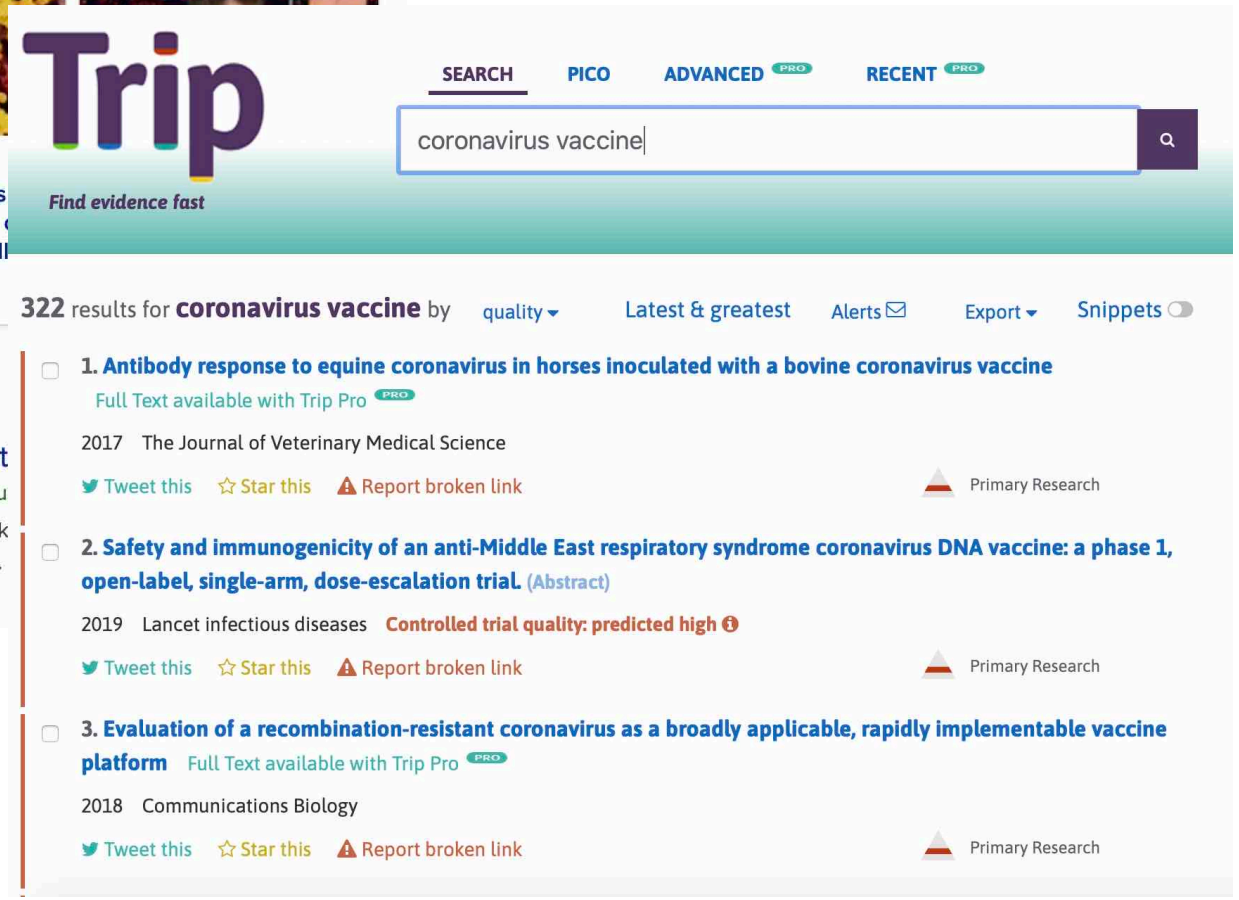






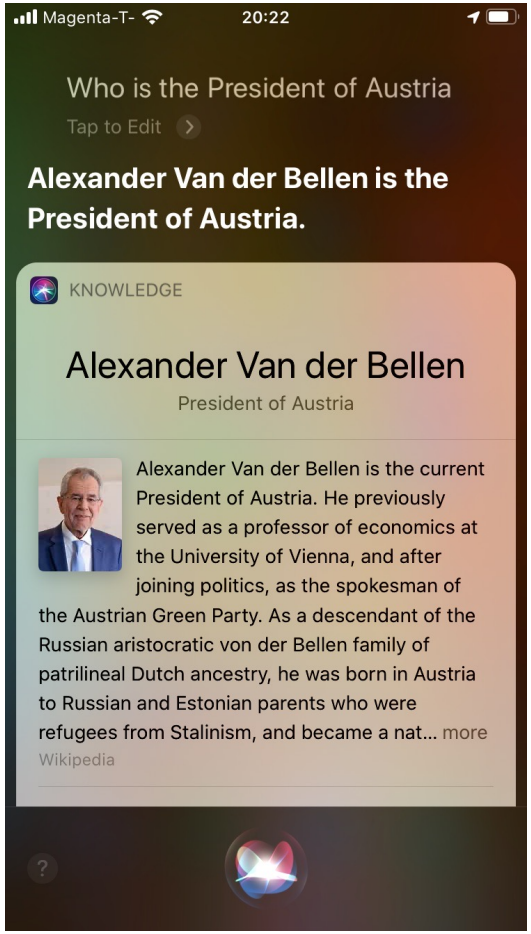
Information Retrieval (IR)





Factoid Q&A


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


Magenta-T-
20:22
Who is the President of Austria
Tap to Edit
Alexander Van der Bellen is the President of Austria.
KNOWLEDGE
Alexander Van der Bellen
President of Austria
Alexander Van der Bellen is the current President of Austria. He previously served as a professor of economics at the University of Vienna, and after joining politics, as the spokesman of the Austrian Green Party. As a descendant of the Russian aristocratic von der Bellen family of patrilineal Dutch ancestry, he was born in Austria to Russian and Estonian parents who were refugees from Stalinism, and became a nat... more
Wikipedia

Factoid 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/

Non-factoid Question Answering

Query 2402:
Question: What is the structure of Australia's members of parliament?
Document ID: 400
Document Name: Member_of_parliament.html
Answer Passages:
Passage 1 A Member of Parliament is the representative of the voters to a parliament. In many countries with bicameral parliaments, this category includes specifically members of the lower house, as upper houses often have a different title. Members of parliament tend to form parliamentary groups with members of the same political party. The Westminster system is a democratic parliamentary system of government modelled after the politics of the United Kingdom. This term comes from the Palace of Westminster, the seat of the Parliament of the United Kingdom. A member of parliament is a member of the House of Representatives, the lower house of the Commonwealth parliament. Members may use "MP" after their names; "MHR" is not used, although it was used as a post-nominal in the past.
Passage 2 A member of the upper house of the Commonwealth parliament, the Senate, is known as a "Senator". In the Australian states of New South Wales, Victoria and South Australia, a Member of the Legislative Assembly or "lower house," may also use the post-nominal "MP." Members of the Legislative Council use the post-nominal "MLC." Members of the Jatiyo Sangshad, or National Assembly, are elected every five years and are referred to in English as members of Parliament. The assembly has directly elected 300 seats, and further 50 reserved selected seats for women. The Parliament of Canada consists of the monarch, the Senate , and the House of Commons.

Question: How are people coping in the lockdown?

Headline: China coronavirus: Death toll rises as more cities restrict travel

Document: China has widened its travel restrictions in Hubei province - the centre of the coronavirus outbreak - as the death toll climbed to 26. The restrictions will affect at least 20 million people across 10 cities, including the capital, Wuhan, where the virus emerged. On Thursday, a coronavirus patient died in northern Hebei province - making it the first death outside Hubei. [...] We now know this is not a virus that will burn out on its own and disappear. [...] And we still don't know when people are contagious. Is it before symptoms appear, or only after severe symptoms emerge? One is significantly harder to stop spreading than the other. [...] **One doctor, who requested anonymity, describes the conditions at a hospital in Wuhan. [...] "I was planning to stay in my apartment because I'm scared to go to the gym, and I'm scared to go to out in public, and not many people are willing to go out." (141 words).** Vietnam and Singapore were on Thursday added to the nations recording confirmed cases, joining Thailand, the US, Taiwan and South Korea. [...] Taiwan has banned people arriving from Wuhan and the US state department warned American travellers to exercise increased caution in China. (document length: 921 words)

https://aclanthology.org/2021.eacl-main.106.pdf
https://arxiv.org/pdf/1805.03797.pdf

Conversational IR / Chat Bots

- A task-oriented dialog system

DB State

Attraction

Restaurant

1	**H:** I am looking for a museum in the center of town do you have any suggestion?
2	**M:** there are *6* museums in the *centre* of town . how about the *Broughton house gallery* at *98 king street* ?
3	**H:** nice, can you tell me what is the phone number and address ?
4	**M:** the phone number for the *Broughton house gallery* is *01223314960* and they are located at *98 king street* .
5	**H:** ok thanks, can you recommend a restaurant in the same area serving Indian food
6	**M:** *curry garden* serves *Indian* food and is located in the *centre* . would you like me to make a reservation ?
7	**H:** sure, please book a table for two people on thursday at 18:00
8	**M:** ok , i have a table for *2* at *18:00* on *Thursday* . your reference number is *1234567890* . is there anything else I can help you with ?
9	**H:** that's it, bye.
10	**M:** have a great time and please contact us if you need any further assistance . goodbye .

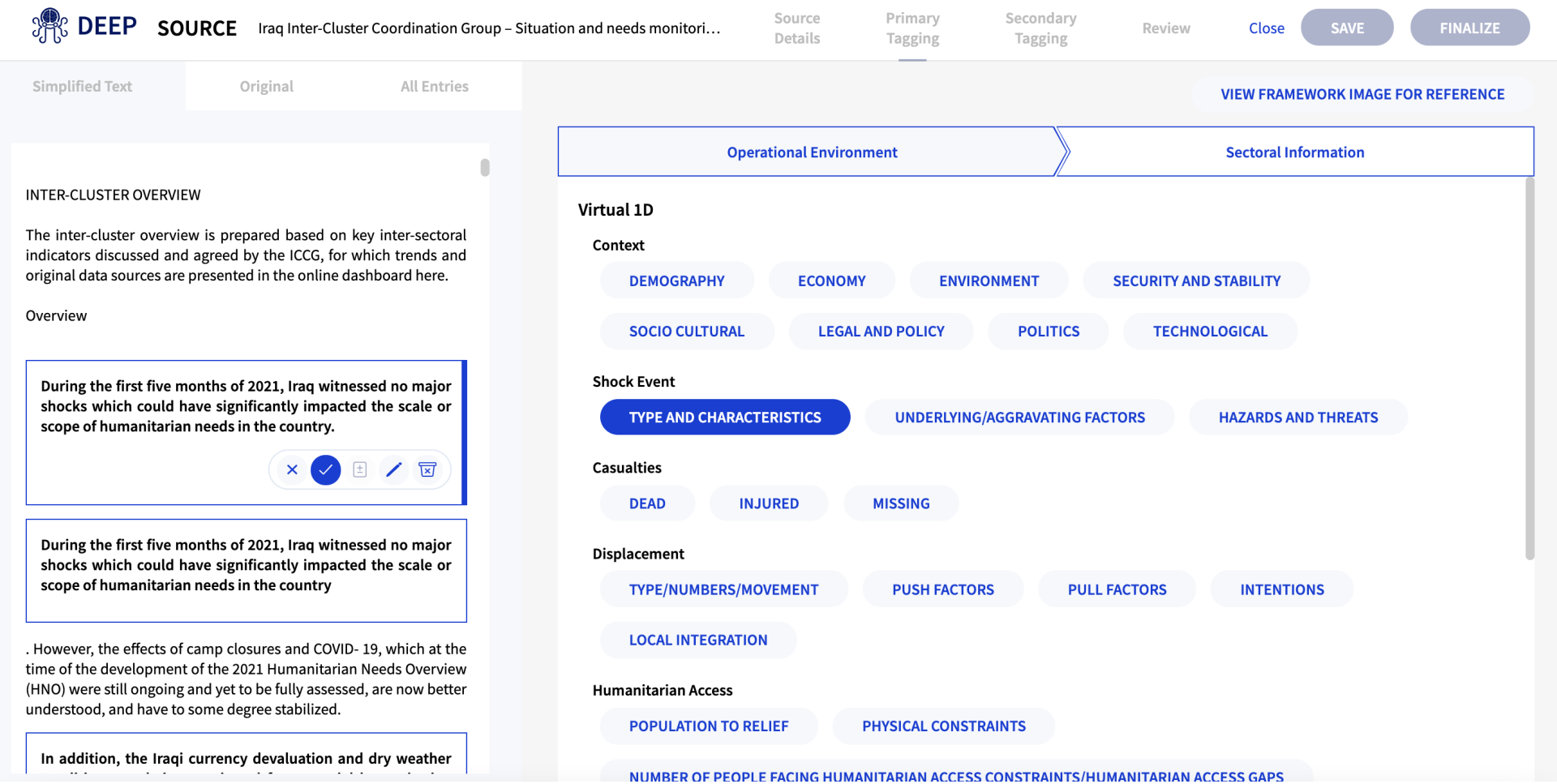
Belief State

Belief state:
{
'attraction':{'type': 'museum',
'area': 'centre'}
}

Belief state:
{
'restaurant': {'food': 'indian', 'area': 'centre'},
'booking': {'day': 'Thursday',
'people': '2', 'time': '18:00'},
'attraction':{'type': 'museum',
'area': 'centre'}
}

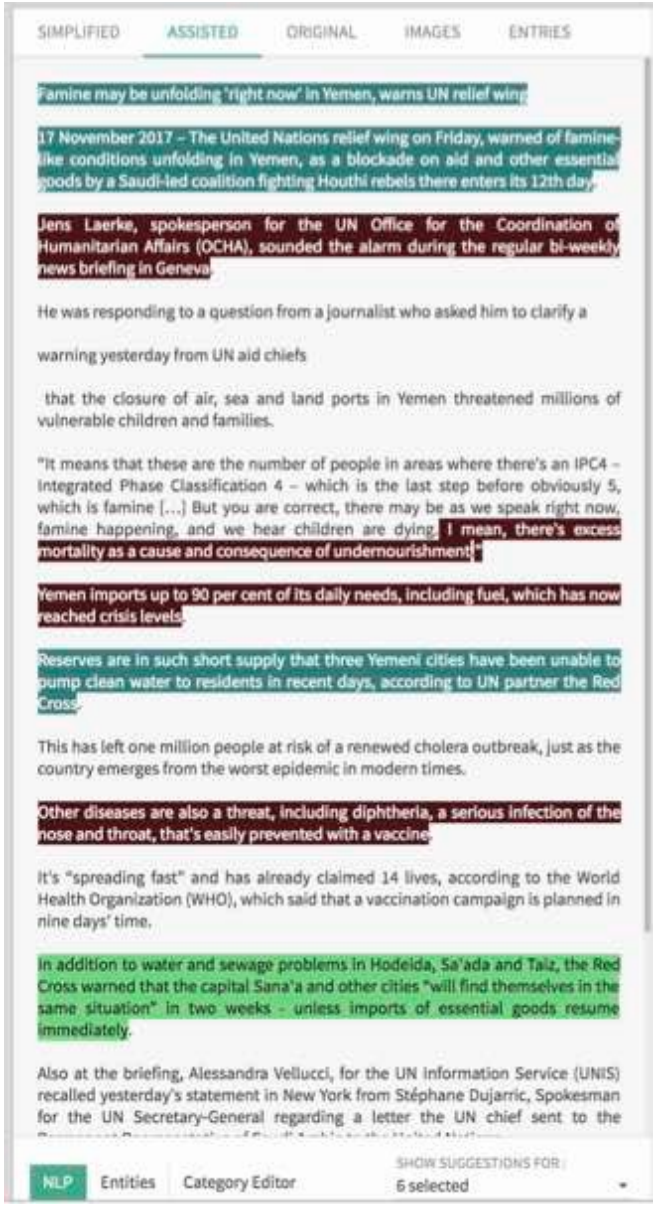
https://www.microsoft.com/en-us/research/uploads/prod/2020/07/SIGIR-2020-Tutorial-Slides-recent-advances-in-conversational-information-retrieval.pdf

Text/Document analysis



https://www.thedeep.io
https://primer.ais

Excerpt extraction and classification



Food

Agriculture

Logistics

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

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

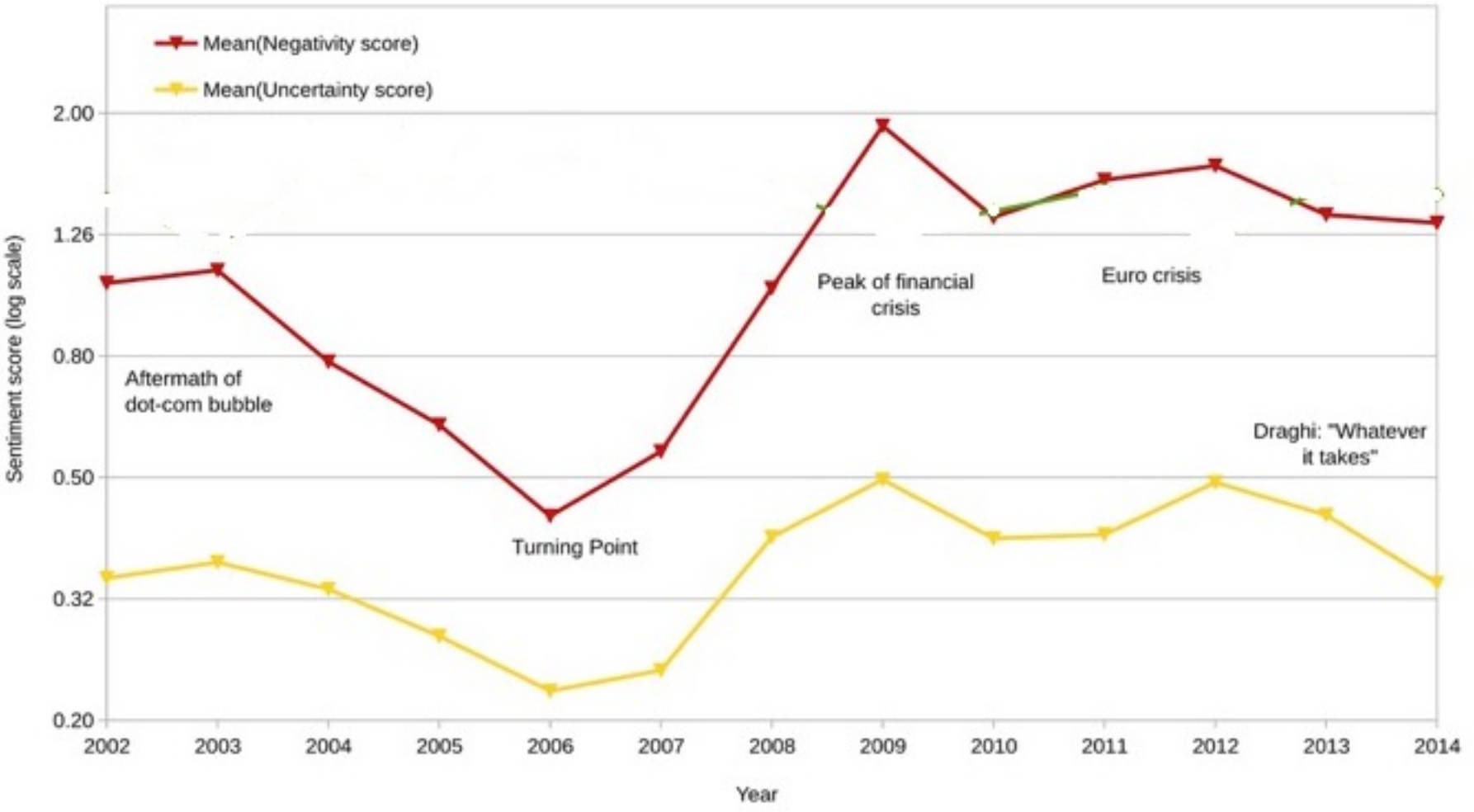
Summary - Based on 377 reviews

1 star	2	3	4 stars	5 stars

What people are saying

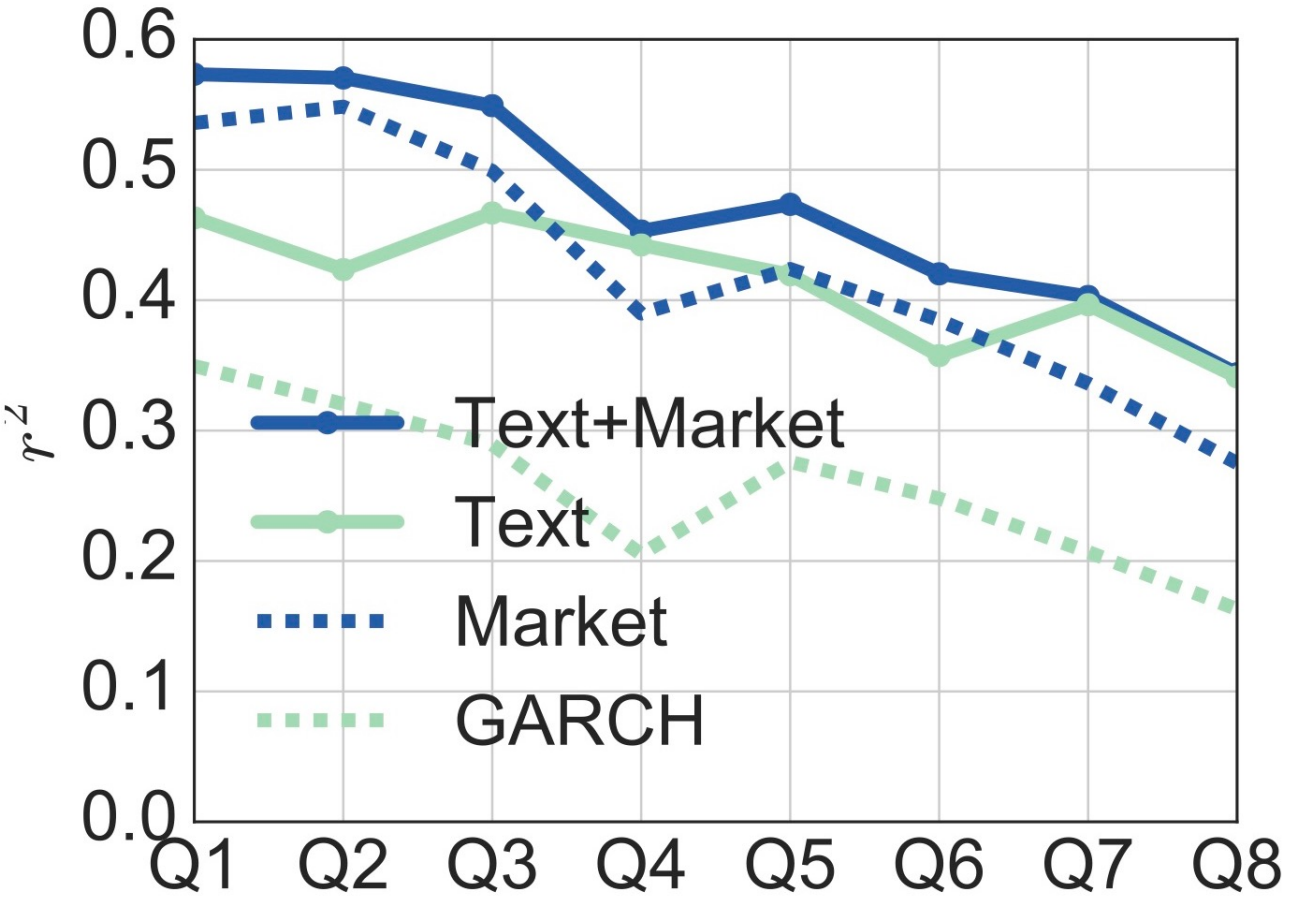
ease of use	"This was very easy to setup to four computers."
value	"Appreciate good quality at a fair price."
setup	"Overall pretty easy setup."
customer service	"I DO like honest tech support people."
size	"Pretty Paper weight."
mode	"Photos were fair on the high quality mode."
colors	"Full color prints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

Monitoring market change from text data



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

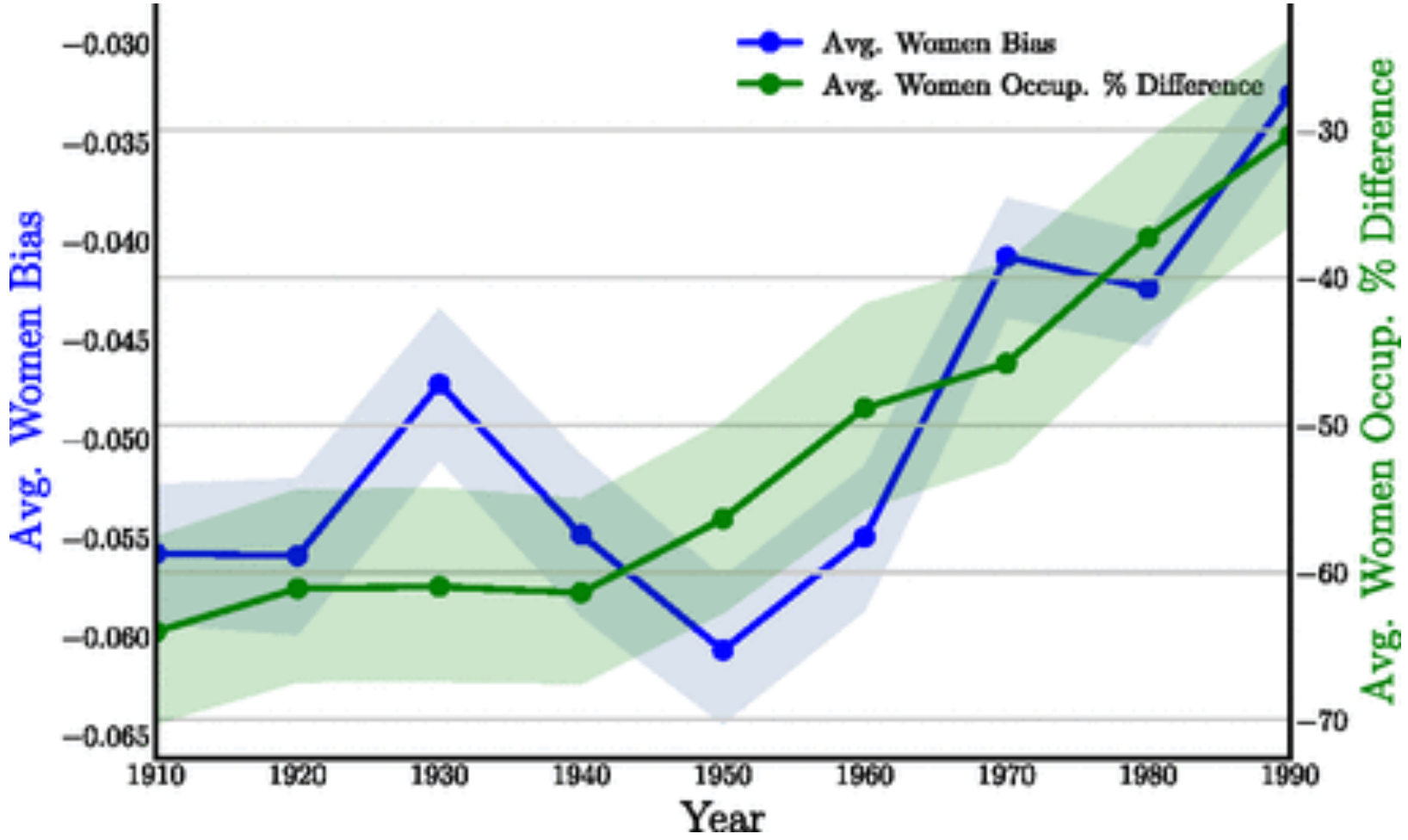
Market/Volatility prediction



https://aclanthology.org/P17-1157.pdf

[5]

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*

Information Extraction (IE)

Involvement of Tumor Necrosis Factor Receptor-associated Protein 1 (TRAP1) in Apoptosis Induced by β-Hydroxyisovalerylshikonin*

IE

Subject	Relation	Object
p53	is_a	protein
Bax	is_a	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

textual abstract:
summary for human

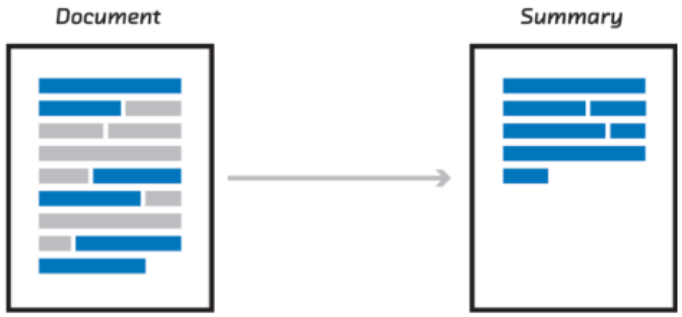
structured knowledge extraction:
summary for machine

https://courses.cs.washington.edu/courses/cse517/13wi/slides/cse517wi13-RelationExtraction.pdf

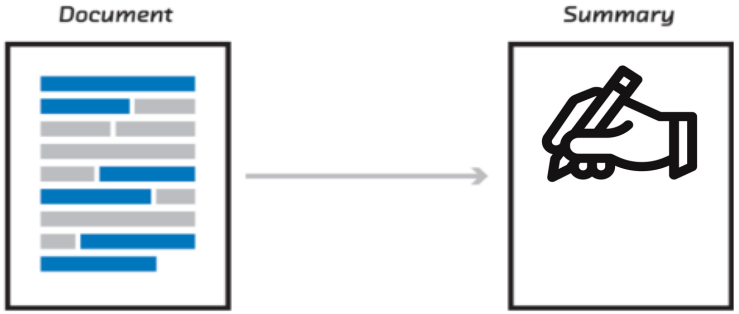
Summarization

- Aims to provide a concise and comprehensive summary of a text

Extractive Summarization



Abstractive Summarization



Abstractive summarization

Text:

We can't settle iPhone vs. Android or "Star Wars" vs. "Star Trek" for you. But another long-running geek debate was put to rest Tuesday night. Those short, animated loops that have captivated the Web for decades? They're pronounced like a brand of peanut butter. Steve Wilhite created the Graphics Interchange Format, or GIF, while working for Compuserve in 1987. On Tuesday, he received a Webby Award for it and delivered his five-word acceptance speech (that's all the Webbys allow) by flashing a GIF on the big screens at the Cipriani Wall Street in New York. And, in a flash, it all became clear: . "It's pronounced JIF, not GIF." Of course, in the grand tradition of heated debate, a flat statement of fact by the creator wasn't enough to sway some partisans. On Twitter, "GIF" became a trending topic as some folks pushed back. "Graphics Interchange Format. Graphics. Not Jraphics. #GIF #hardg," wrote Web designer Dan Cederholm. "So instead of GIF, we've got to say JIF? YEAH RIGHT," chimed in October Jones, creator of the "Texts From Dog" Tumblr and book. "And I suppose those animals with long necks are called 'JIRAFFES.'" And, of course, the peanut butter brand was getting lots of free publicity along the way. The always amusing HAL 9000 account (yes, somebody tweets as the robot from "2001") posted an "animated JIF" -- which is to say, a swirling, animated jar of the tasty, high-protein spread. So, it's perhaps no surprise that the company got into the act itself. Wednesday afternoon, the company took to Twitter with a post reading, "It's pronounced Jif® ." The tweet linked to, what else, a multi-colored GIF flashing the same phrase. Animated GIFs were a staple of the early Internet. Remember The Dancing Baby? That's a GIF. They fell out of favor as more advanced graphics technology emerged. But in the past couple of years, the Web has remembered how much fun it is to watch ridiculous things happen over and over again. Appropriately, Wilhite received his Lifetime Achievement Award from David Karp, the founder of Tumblr, one prominent place where GIFs found a new fanbase. In less publicized interviews, Wilhite had argued for the soft-G pronunciation for years. So, will a widely covered "speech" in front of some of the Web's most influential folks finally be the turning point? Maybe not. Last month, no less an authority than the White House posted an image on its new Tumblr feed advocating for the hard-G. And the Oxford English Dictionary says both pronunciations are acceptab le. So, here's wishing Mr. Wilhite "Jood Luck."

Summary:

GIF creator: It's pronounced "JIF" Steve Wilhite created the Graphics Interchange Format in 1987 at Compuserve . He pronounced the issue closed at the Webby Awards . And yet, some partisans remain unswayed .

https://huggingface.co/datasets/viewer/?dataset=cnn_dailymail&config=3.0.0

Abstractive summarization

The bottleneck is no longer access to information; now it's our ability to keep up.
AI can be trained on a variety of different types of texts and summary lengths.
A model that can generate long, coherent, and meaningful summaries remains an open research problem.

The last few decades have witnessed a fundamental change in the challenge of taking in new information. The bottleneck is no longer access to information; now it's our ability to keep up. We all have to read more and more to keep up-to-date with our jobs, the news, and social media. We've looked at how AI can improve people's work by helping with this information deluge and one potential answer is to have algorithms automatically summarize longer texts. Training a model that can generate long, coherent, and meaningful summaries remains an open research problem. In fact, generating any kind of longer text is hard for even the most advanced deep learning algorithms. In order to make summarization successful, we introduce two separate improvements: a more contextual word generation model and a new way of training summarization models via reinforcement learning (RL). The combination of the two training methods enables the system to create relevant and highly readable multi-sentence summaries of long text, such as news articles, significantly improving on previous results. Our algorithm can be trained on a variety of different types of texts and summary lengths. In this blog post, we present the main contributions of our model and an overview of the natural language challenges specific to text summarization.

https://blog.einstein.ai/your-tldr-by-an-ai-a-deep-reinforced-model-for-abstractive-summarization/

Extractive summarization

A section of DNA that contains instructions to make a protein is called a gene. Each gene has the sequence for at least one polypeptide. Proteins form structures, and also form enzymes. The enzymes do most of the work in cells. Proteins are made out of smaller polypeptides, which are formed of amino acids. To make a protein to do a particular job, the correct amino acids have to be joined up in the correct order.

Proteins are made by tiny machines in the cell called ribosomes. Ribosomes are in the main body of the cell, but DNA is only in the nucleus of the cell. The codon is part of the DNA, but DNA never leaves the nucleus. Because DNA cannot leave the nucleus, the cell makes a copy of the DNA sequence in RNA. This is smaller and can get through the holes – pores – in the membrane of the nucleus and out into the cell.

Genes encoded in DNA are transcribed into messenger RNA (mRNA) by proteins such as RNA polymerase. Mature mRNA is then used as a template for protein synthesis by the ribosome. Ribosomes read codons, 'words' made of three base pairs that tell the ribosome which amino acid to add. The ribosome scans along an mRNA, reading the code while it makes protein. Another RNA called tRNA helps match the right amino acid to each codon.

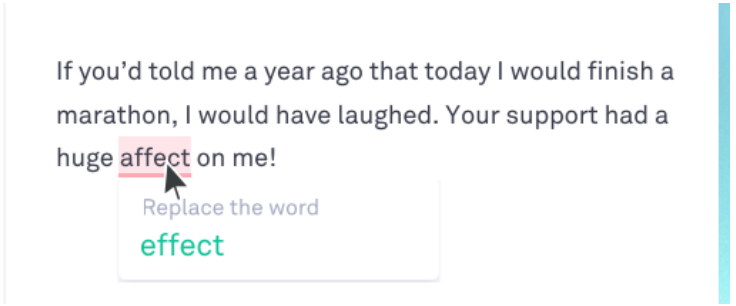
A section of DNA that contains instructions to make a protein is called a gene. Each gene has the sequence for at least one polypeptide. Proteins form structures, and also form enzymes. The enzymes do most of the work in cells. Proteins are made out of smaller polypeptides, which are formed of amino acids. To make a protein to do a particular job, the correct amino acids have to be joined up in the correct order.

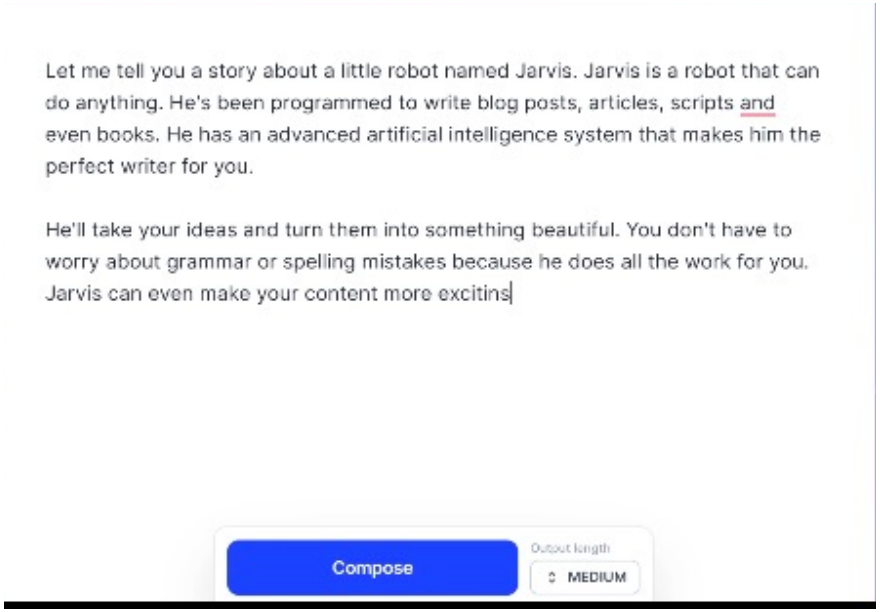
Proteins are made by tiny machines in the cell called ribosomes. Ribosomes are in the main body of the cell, but DNA is only in the nucleus of the cell. The codon is part of the DNA, but DNA never leaves the nucleus. Because DNA cannot leave the nucleus, the cell makes a copy of the DNA sequence in RNA. This is smaller and can get through the holes – pores – in the membrane of the nucleus and out into the cell.

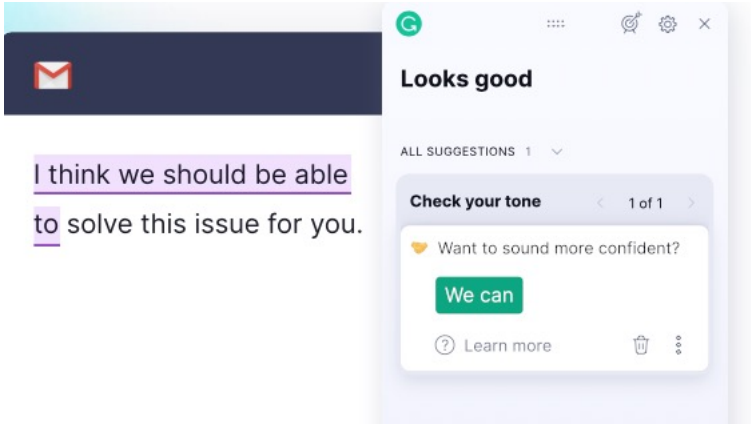
Genes encoded in DNA are transcribed into messenger RNA (mRNA) by proteins such as RNA polymerase. Mature mRNA is then used as a template for protein synthesis by the ribosome. Ribosomes read codons, 'words' made of three base pairs that tell the ribosome which amino acid to add. The ribosome scans along an mRNA, reading the code while it makes protein. Another RNA called tRNA helps match the right amino acid to each codon.

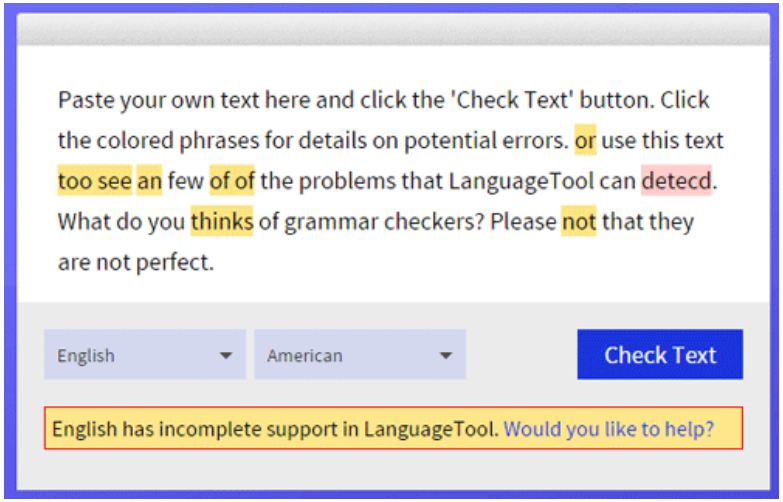
https://fastforwardlabs.github.io/luhn/

Content generator / Writing assistant / Spell checker



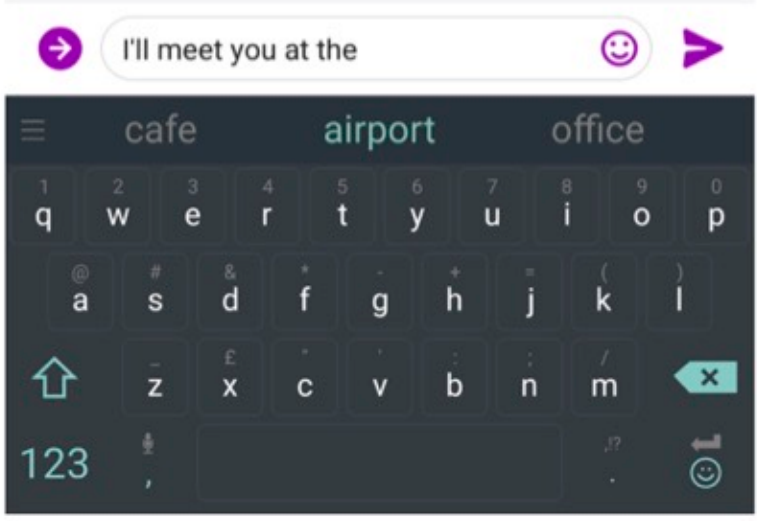


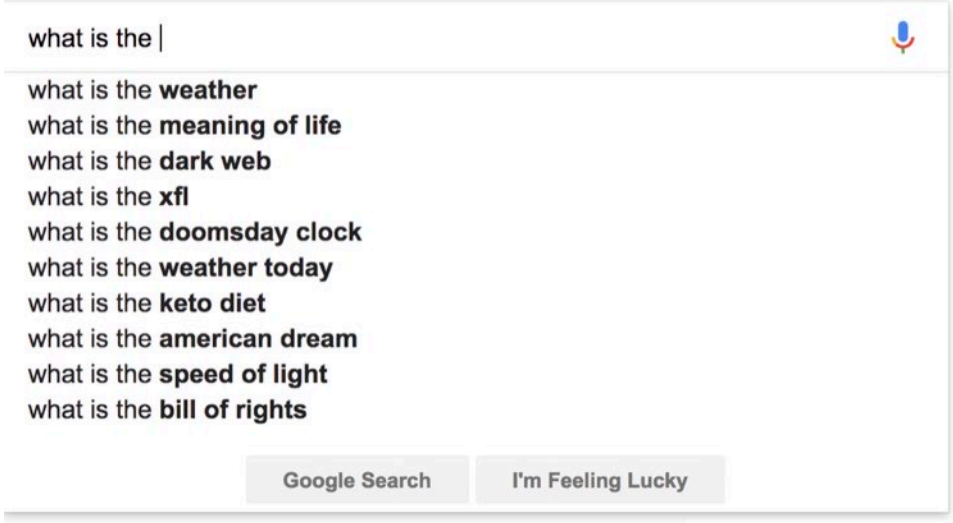




https://www.jarvis.ai/free-trial?fpr=prowriter
https://www.grammarly.com

Language Modeling





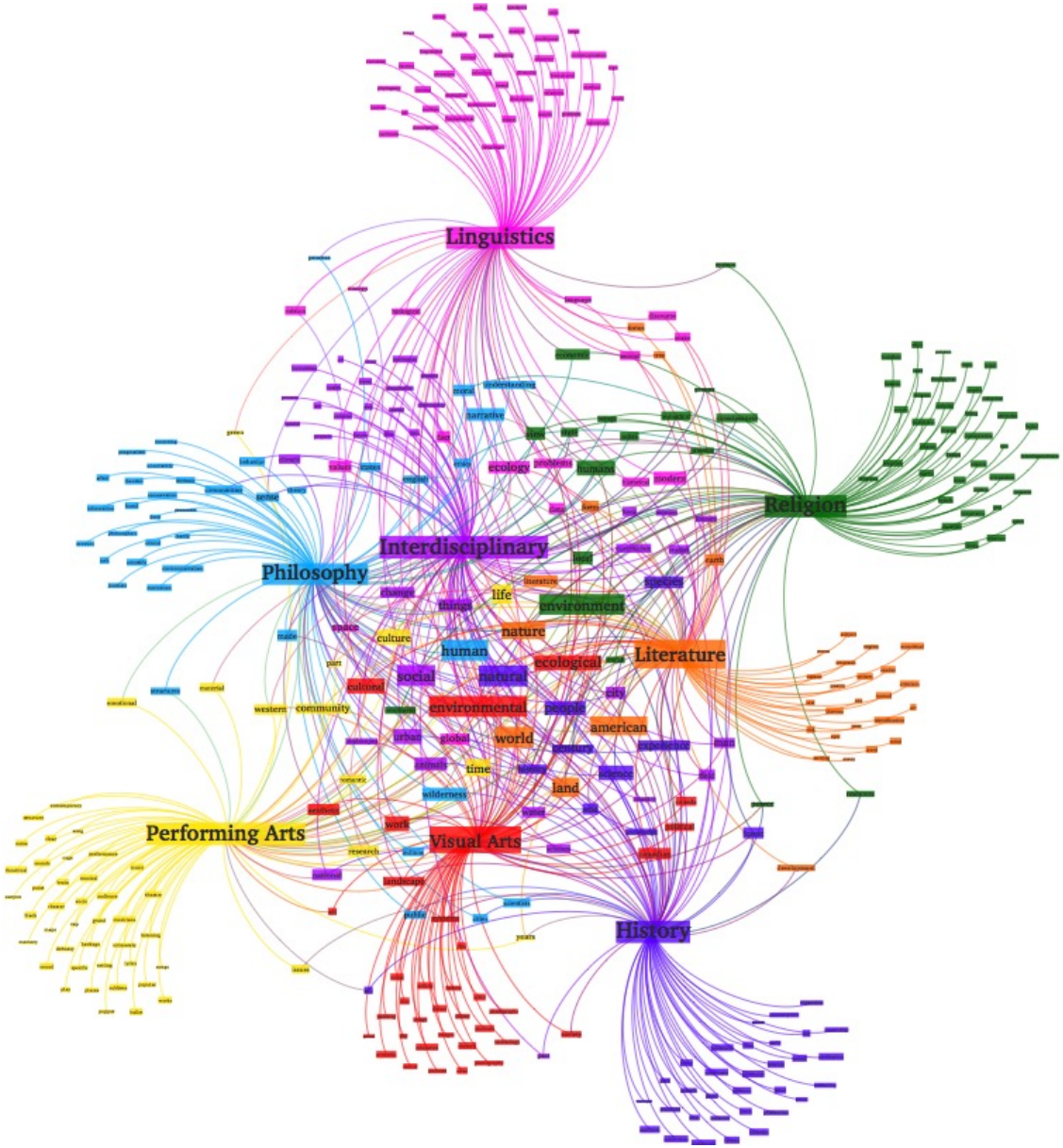
Look more:
https://talktotransformer.com
https://www.youtube.com/watch?v=gcHkxP9adiM

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/

Topic modeling / Clustering



https://knitdata.org/post/r-markdown/

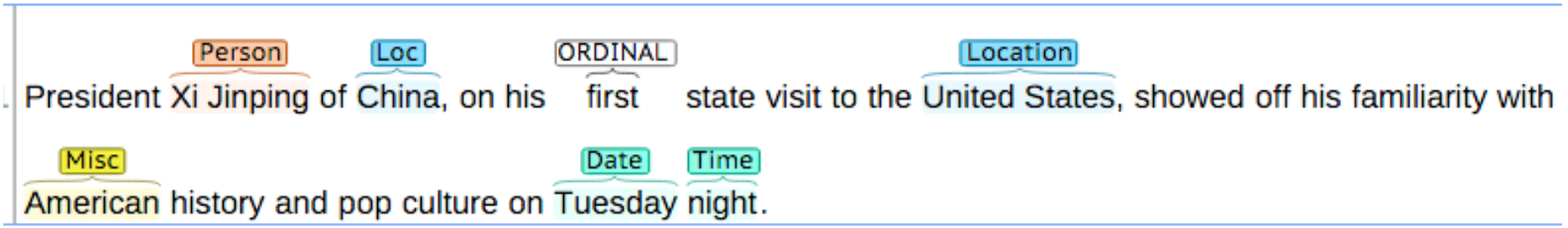
Word Sense Disambiguation

- 2 senses of ***bank***
  - "The ***bank*** will not be accepting cash on Saturdays. "
  - "The river overflowed the ***bank.***"

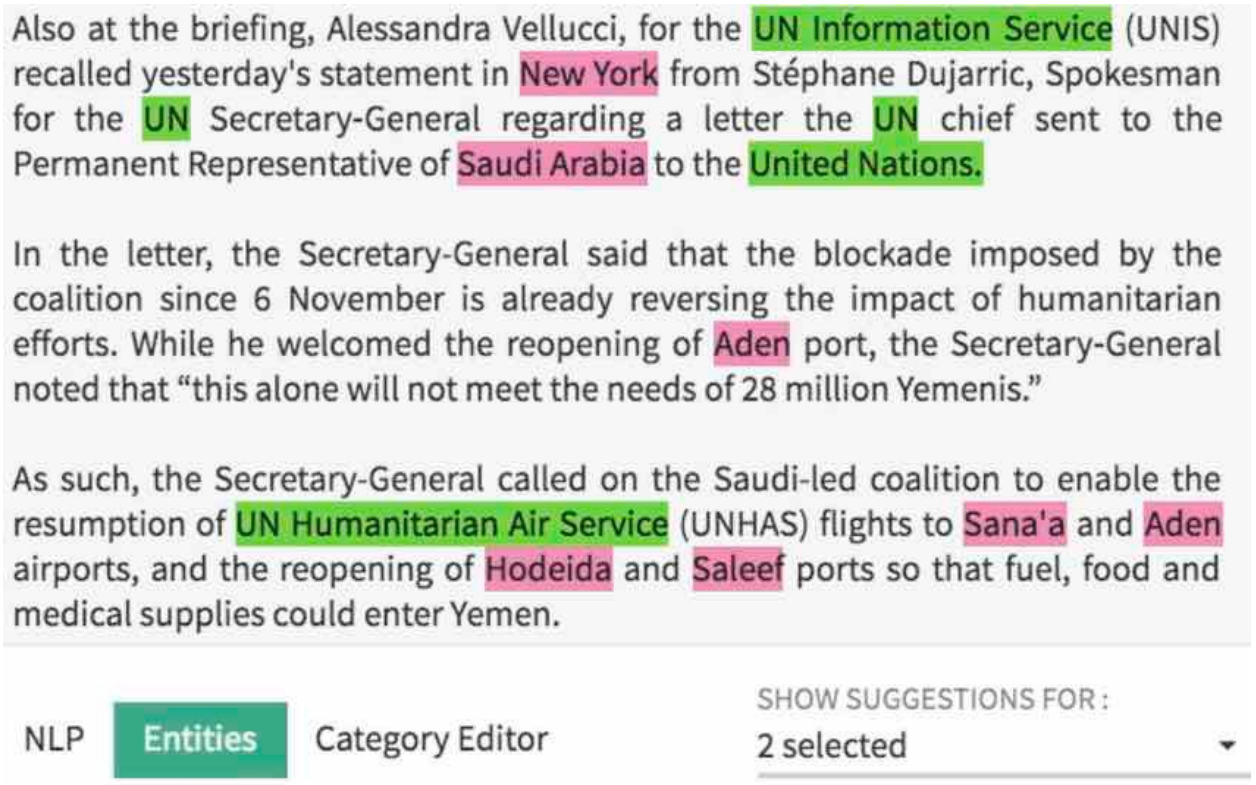
- 8 senses of ***bass***, as defined in WordNet
  - bass - (the lowest part of the musical range)
  - bass, bass part - (the lowest part in polyphonic music)
  - bass, basso - (an adult male singer with the lowest voice)
  - sea bass, bass - (flesh saltwater fish of the family Serranidae)
  - freshwater 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



Locations

Organizations

Relation Extraction

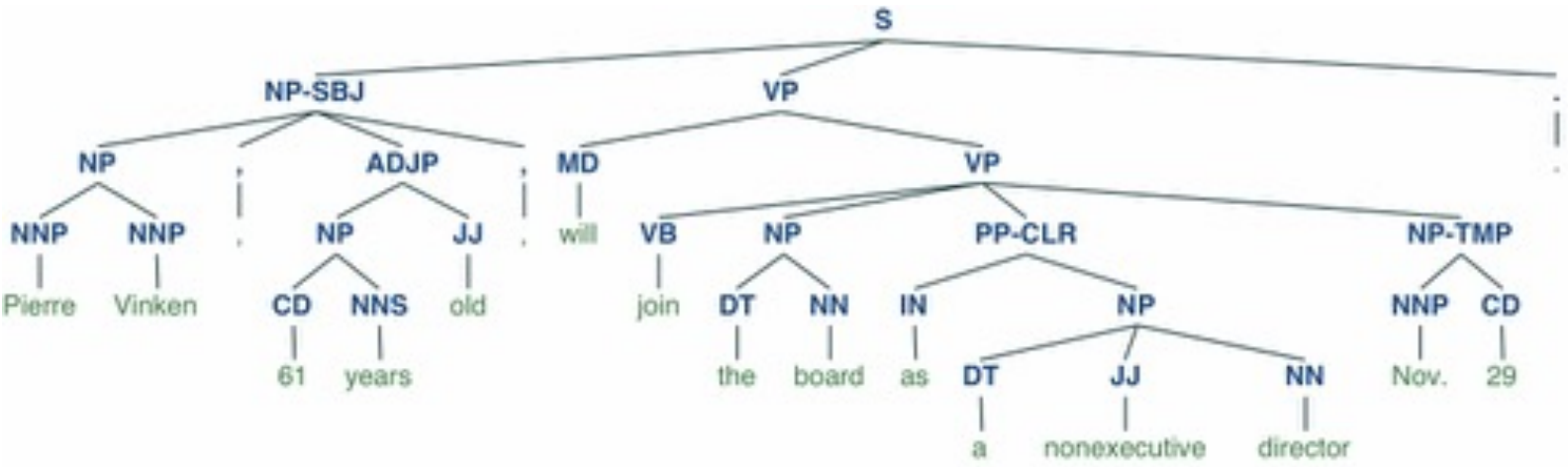
CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR, immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Subject	Relation	Object
American Airlines	subsidiary	AMR
Tim Wagner	employee	American Airlines
United Airlines	subsidiary	UAL

example from Jim Martin

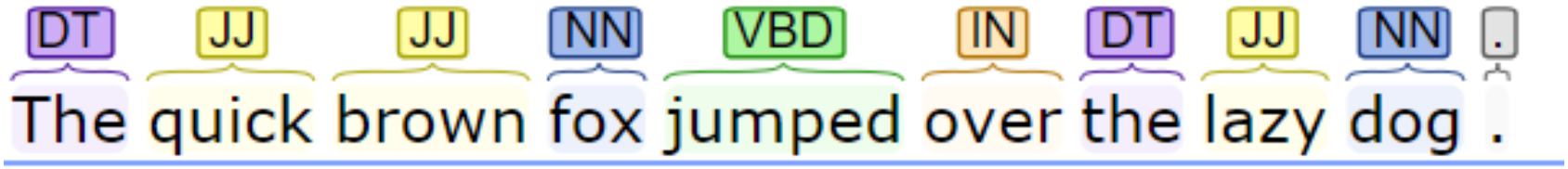
Parsing

- 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



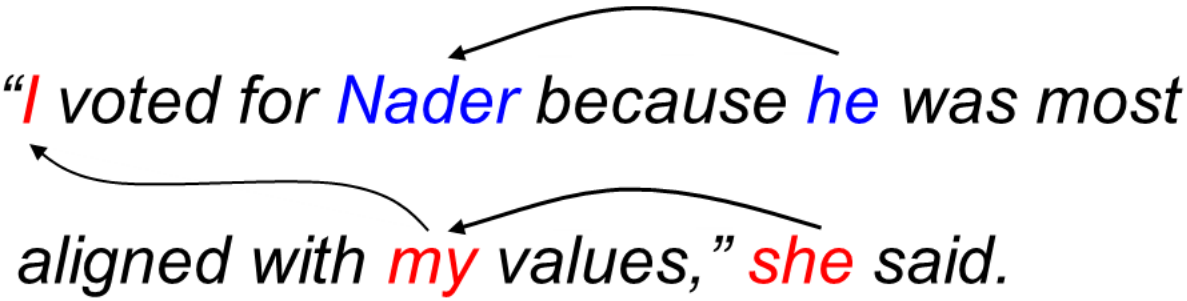
Part-of-Speech (POS) Tagging

- “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



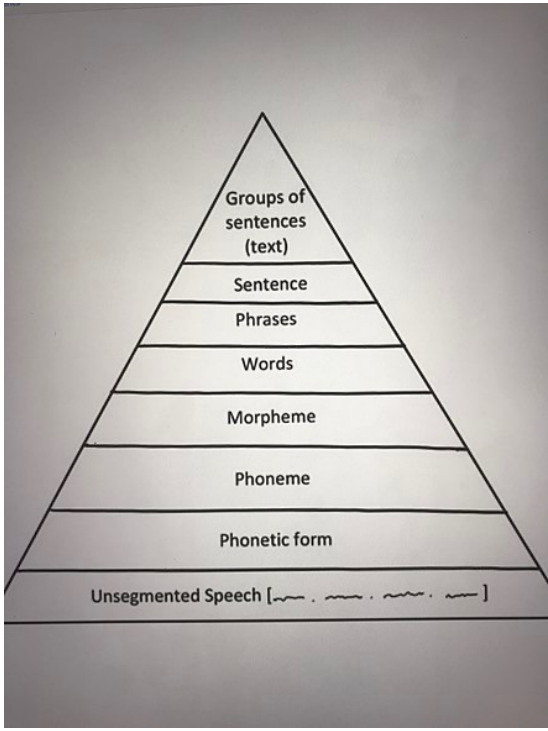
Agenda

- NLP Applications
- **Language as a complex system**
- Text preprocessing

Language hierarchy – linguistic perspective

Sample sentence: "*a fox jumped over the lazy dog.*"

- 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: ***a fox***
  - Verb phrase: "***jumped over the lazy dog***"
- Sentence



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

Language hierarchy – computational perspective

Sample sentence: "*a fox jumped over the lazy dog.*"

- Character
  - like "***a***", "***f***", "***x***", etc.
- N-gram character
  - E.g. tri-gram characters like "***a f***", " ***fo***", "***fox***", and "***ox*** "
- Word
- N-gram word
  - E.g. tri-grams like "***a fox jumped***", "***fox jumped over***" , and "***jumped over the***"
- 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 perspective (cont.)

Sample sentence: "*a fox jumped over the lazy dog.*"

- Token
  - used as the unit of processing
  - can be any of the previously mentioned units and any sequence of characters
  - E.g. when tokenized by words:
    - ["***a***", "***fox***", " ***jumped***", "***over***", "***the***", "***lazy***", "***dog***"]
- Dictionary or vocabulary list or lexicon
  - List of unique tokens in the given text
- Sentence
- Paragraph
- Document
- Corpus
  - a collection of text documents

Natural Language

- Learned from experience
- A symbolic/discrete system …

 kangaroo

 volcano

- Encodings in brains and models are continuous

What makes language hard?

- Language is a complex social process
- Tremendous ambiguity at every level of representation
- Variability
- Grounding
- Sparsity

Ambiguity in word level

“One morning I shot an elephant in my pajamas”

https://ell.stackexchange.com/questions/77370/what-is-the-meaning-of-this-famous-groucho-marx-joke

Ambiguity in sentence level

“Finally, a computer that understands you like your mother”

(Ad,1985)

1. The computer understands you as well as your mother understands you.
2. The computer understands that you like your mother.
3. The computer understands you as well as it understands your mother.

Variability – Semantic change

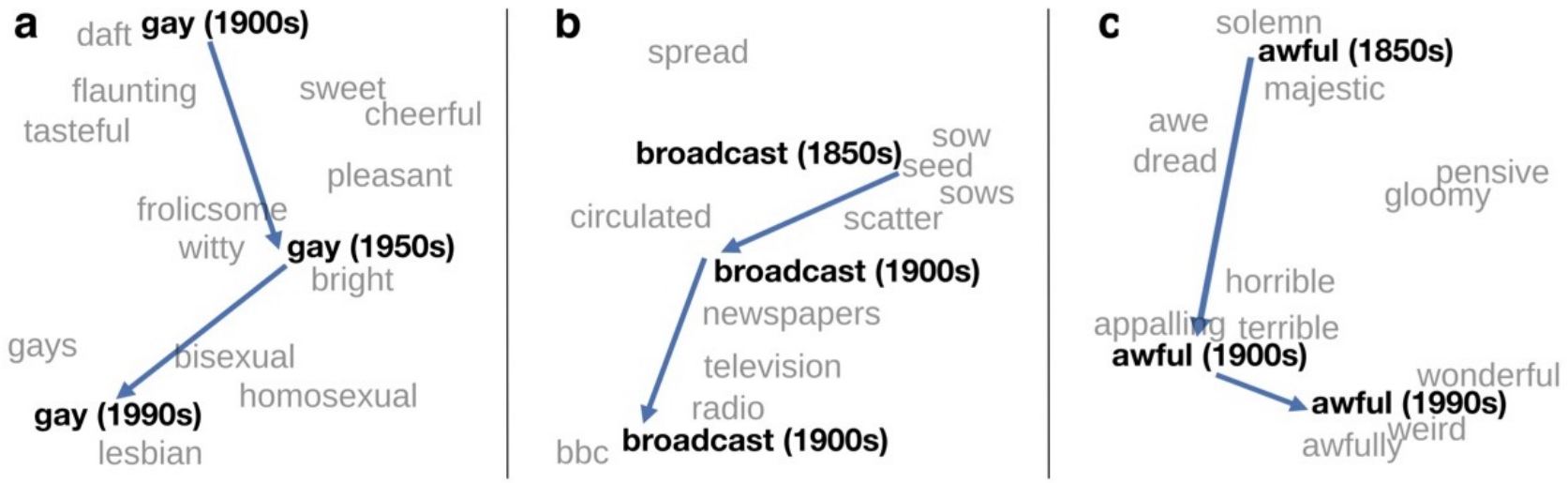


Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

Grounding

- Humans do not learn language by observing an endless stream of text

- Common sense

- An explanation of common sense in infant kids

https://youtu.be/7ROelYvo8f0?t=2411

Challenges in language processing

- Language is a complex social process
- Tremendous ambiguity at every level of representation
- Variability
- Grounding
- **Sparsity**

Zipf's Law

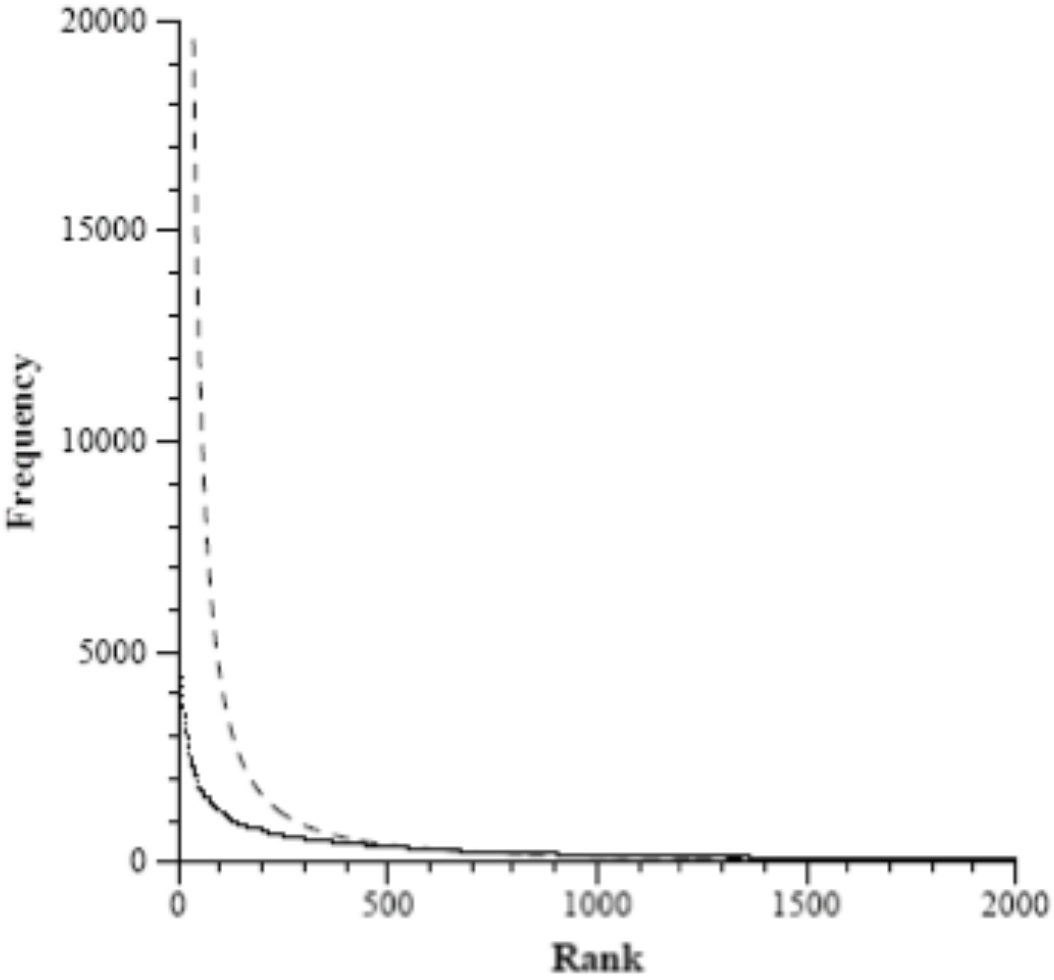
- Sort the vocabularies of a dictionary based on their frequencies in a text corpus
  - E.g. the results from the Brown Corpus:

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

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

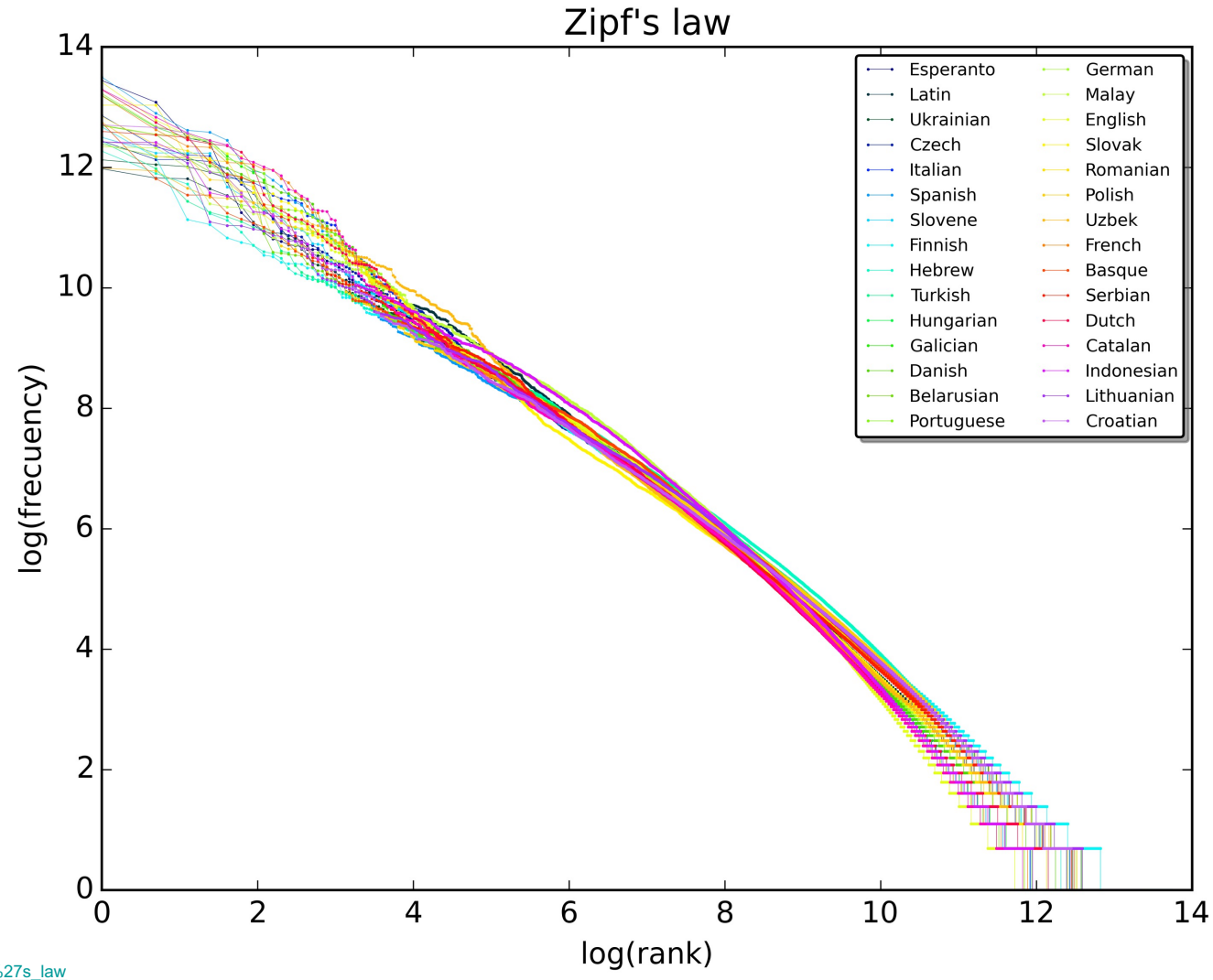
Zipf's Law

- The plot of rank versus frequency looks like this:



Zipf's Law

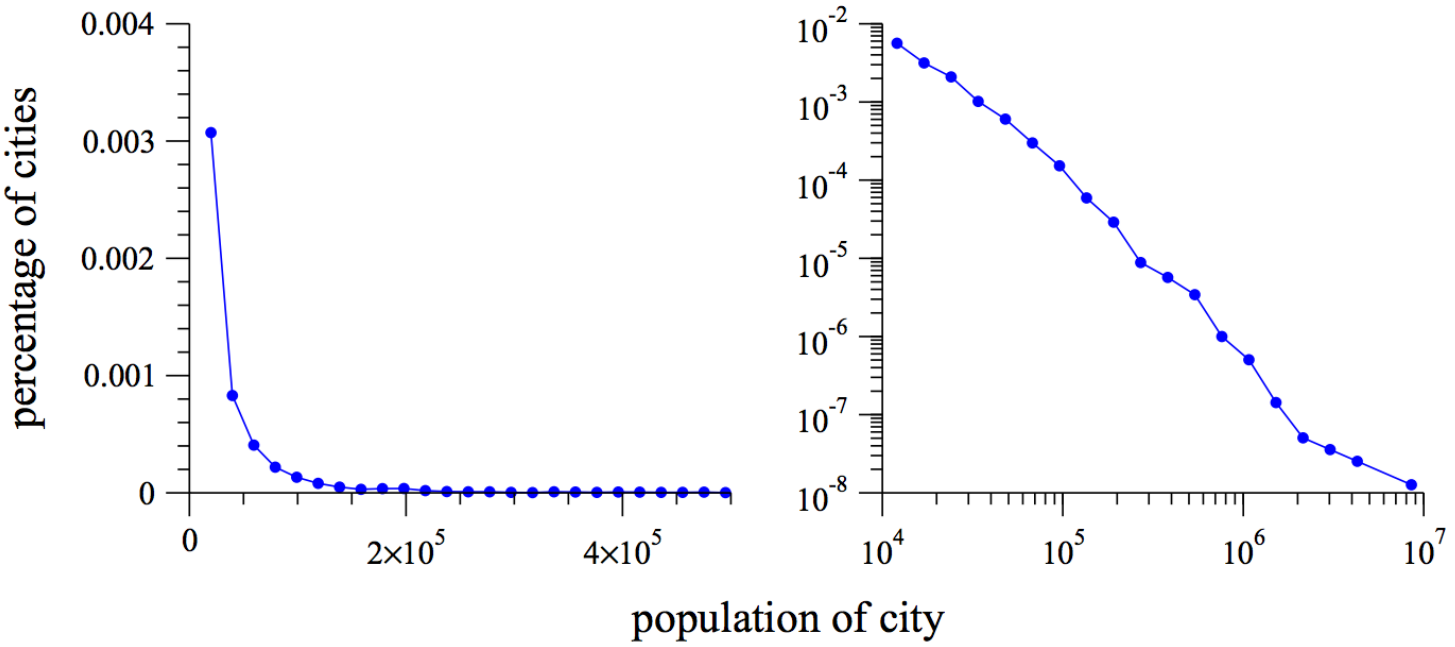
- Applying logarithm to both axes:



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

Power Law in social science

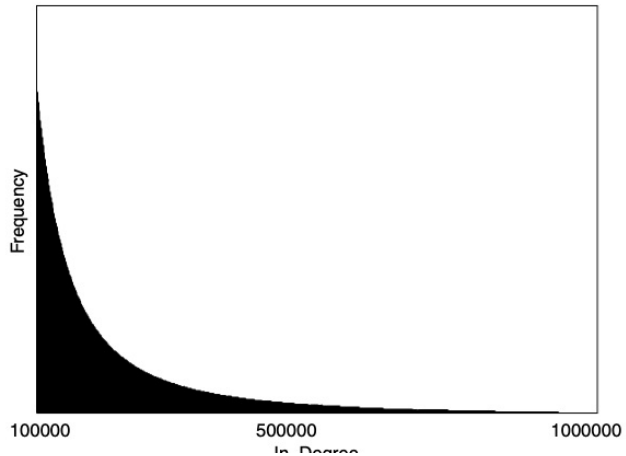
- Observed in other human-related phenomena/systems
  - Population rank of cities
  - Pagerank scores of websites
  - Income distribution
  - Corporation sizes

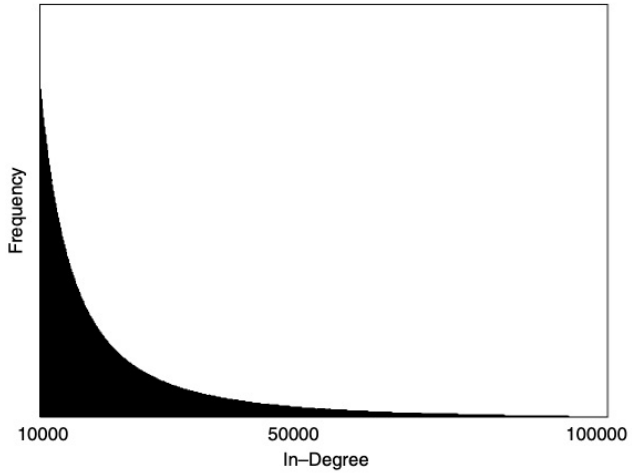


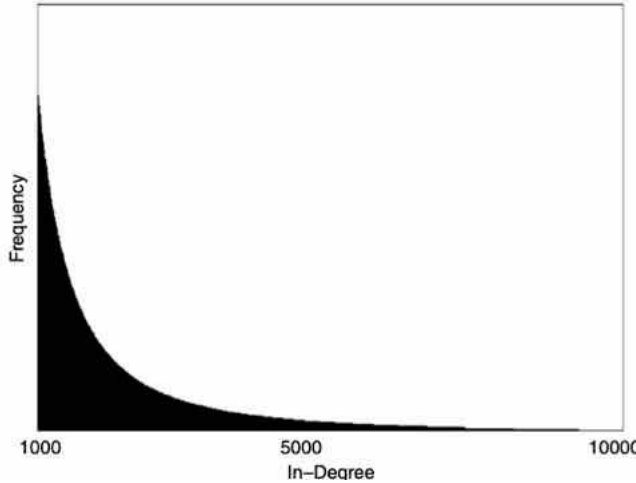
https://statweb.stanford.edu/~owen/courses/306a/ZipfByHera.pdf

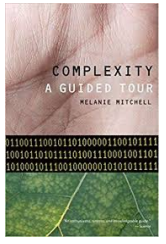
Power Law in web

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





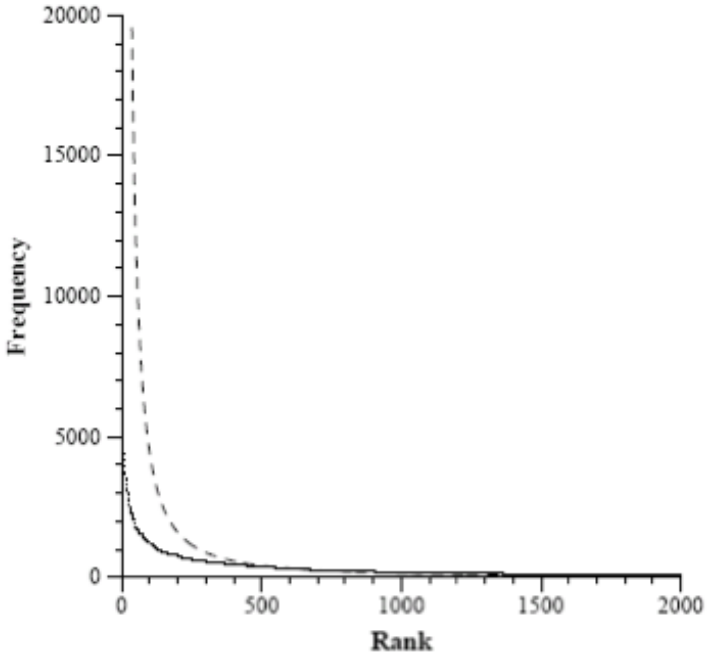




What does Zipf' Law tell us?

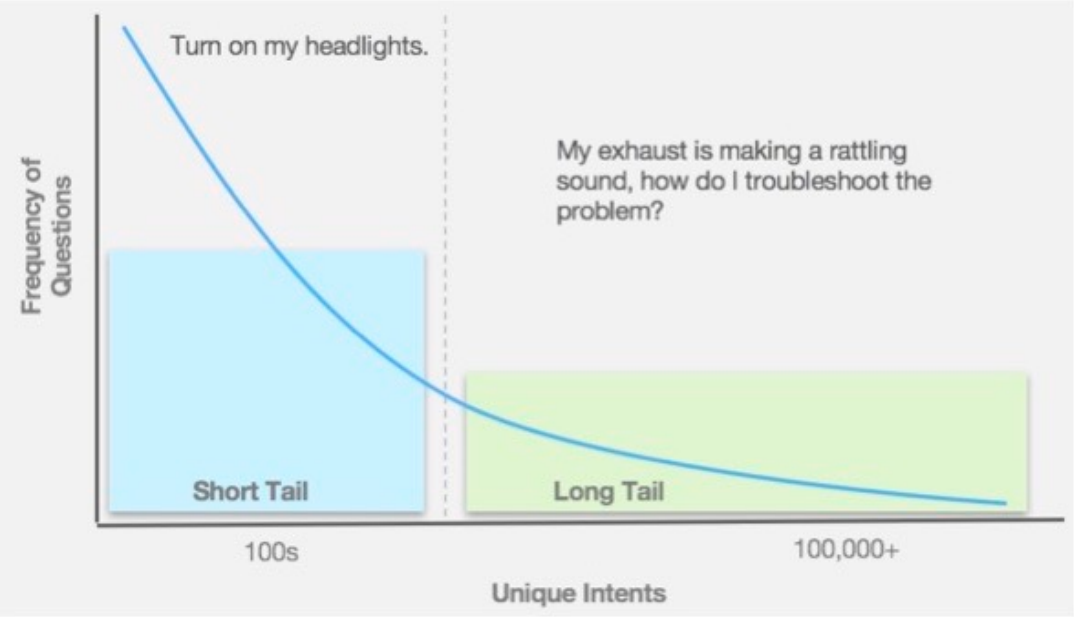
- Highly frequent tokens cover a large portion of a corpus
  - Only 135 most frequent words cover half of the Brown Corpus

- 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!



What does Zipf' Law tell us?

- The challenge of long tail
  - Phenomena with low frequency (resided on the tail of distribution) are the challenging parts of language processing
  - Example of the long tail problem in a question answering system:



- Why could it be challenging for statistical models?

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

Agenda

- NLP Applications
- Language as a complex system
- **Text preprocessing**

Text preprocessing

- Preprocessing is the first and a crucial step of NLP task
  - The objective of preprocessing is to clean/harmonize the text, reduce language fluctuations if necessary, and prepare the tokens for being processed in the next steps
  - Preprocessing partially addresses the issue with sparsity, why?

We will learn:

- Text cleaning/harmonization/reduction of fluctuations
  - Text normalization
  - Segmentation
  - Stop words
  - Stemming & Lemmatization

Text Normalization

- Normalization harmonizes the written forms of the words with same meanings

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

Text Normalization

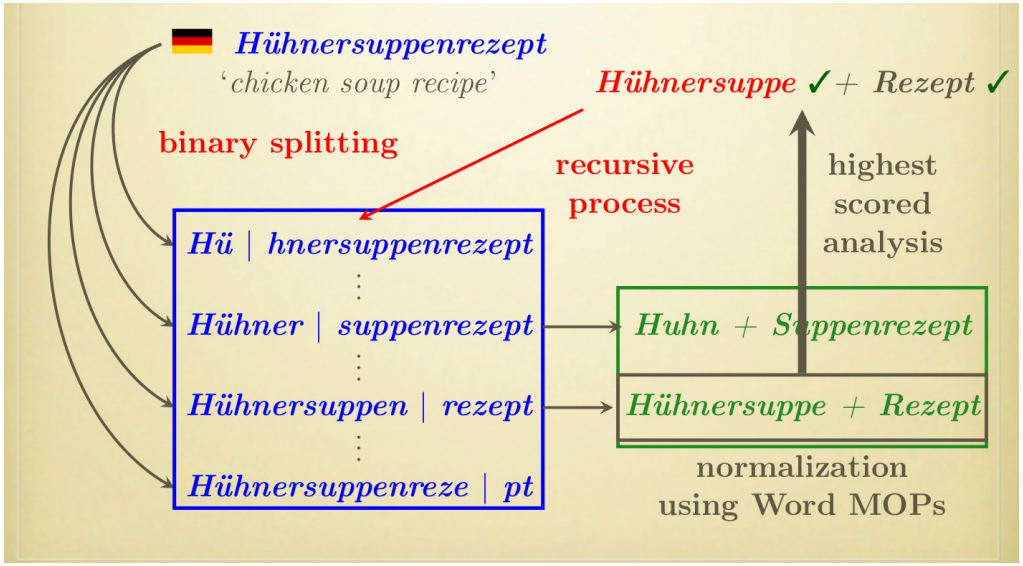
- Case folding: reduce all letters to lower case
  - It may cause ambiguity but typically helpful
    - **General Motors** vs. **general motors**
    - **Fed** vs. **fed**
    - **CAT** (City Airport Train) vs. **cat**

- Longstanding Google example:
  - Search **C.A.T.**

- Do the numbers, dates, etc. bring information?
  - If included, the dictionary size may explode!
  - Numbers and dates are commonly replaced by special tokens, e.g.
    - Numbers with <num>
    - Dates with <dates>

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**
  - A Possible algorithm (look in the link for more details):



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 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**"

- Stop words are sometimes excluded from the corpus
  - Commonly in bag-of-words approaches

Stemming

- Morphemes in English 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. Examples:
  - ***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

Porter's stemmer

- 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/

Lemmatization

- A lemmatizer uses a knowledge resource (like WordNet) to find and replace base forms

- Lemmatzer reduces inflectional/variant forms to base forms. Examples:
  - ***am***, ***are***, ***is*** $\rightarrow$ ***be***
  - ***car***, ***cars***, ***car's***, ***cars'*** $\rightarrow$ ***car***
  - ***the boy's cars are different colors*** $\rightarrow$ ***the boy car be different color***

- Lemmatization versus Stemming:
  - Both reduce variation
  - Stemming is typically faster
  - Stemming may harm precision and increase ambiguity
  - If a given word does not exist in the knowledge resource, lemmatization may not be able to process it

NLP Libraries

AllenNLP

gensim

Stanford CoreNLP

spaCy

huggingface.co

polyglot

PYTORCH