

344.175 VL: Natural Language Processing

Text Processing



Navid Rekab-saz

Email: navid.rekabsaz@jku.at

Office hours: <https://navid-officehours.youcanbook.me>

Agenda

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

Agenda

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

Machine Translation

The screenshot shows a machine translation interface with the following settings at the top:

- ENGLISH - DETECTED
- ENGLISH
- GERMAN
- ENGLISH
- SPANISH

The English input text is:

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

The German output text is:

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

Below the text are various interaction icons: a speaker icon, a progress bar (84/5000), a pencil icon, a speaker icon, a copy icon, and a more options icon.

The screenshot shows a machine translation interface with the following settings:

Translate from English (detected) ▾

Translate into German ▾

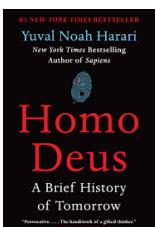
The English input text is:

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

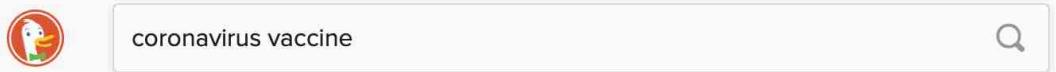
The German output text is:

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

Below the text are various interaction icons: a speaker icon, a copy icon, a link icon, and a download icon.



Information Retrieval (IR)



coronavirus vaccine

Web Images Videos News Maps Settings ▾

Austria ▾ Safe Search: Moderate ▾ Any Time ▾

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



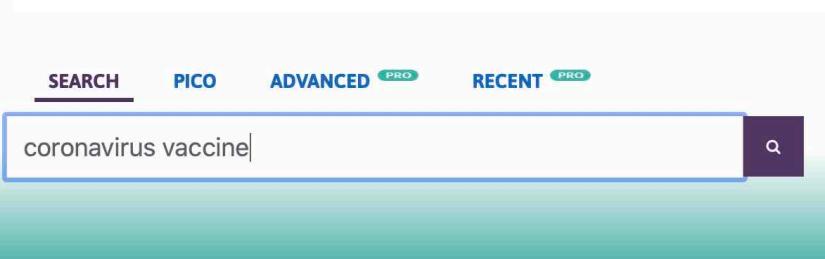
322 results for coronavirus vaccine by quality ▾ Latest & greatest Alerts Export Snippets

→ More News

Vaccine for new Chinese coronavirus in t

CNN <https://www.cnn.com/2020/01/20/health/coronavirus-china/index.html>

Jan 20, 2020 · The National Institutes of Health is working on a vaccine for the new coronavirus that has infected hundreds and killed four in Asia.



SEARCH PICO ADVANCED PRO RECENT PRO

coronavirus vaccine

322 results for coronavirus vaccine by quality ▾ Latest & greatest Alerts Primary Research

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

5

Factoid Q&A

A screenshot of a search results page. The search bar at the top contains the query "when was the last total eclipse in europe". Below the search bar, there are navigation links for "All", "News", "Images", "Videos", "Maps", "More", "Settings", and "Tools". A message indicates "About 42.000.000 results (0,92 seconds)". The main result is a featured snippet box with a large title "August 11, 1999". Below the title is the text: "The last total eclipse in continental Europe occurred on August 11, 1999." At the bottom of the snippet box, there is a link to "en.wikipedia.org › wiki › Solar_eclipse_of_August_12,_2026" and the text "Solar eclipse of August 12, 2026 - Wikipedia". At the very bottom of the page, there are links for "About Featured Snippets" and "Feedback".

A screenshot of an iPhone screen displaying a knowledge card. The top status bar shows signal strength, "Magenta-T-", Wi-Fi, and the time "20:22". The main content area asks "Who is the President of Austria" and provides the answer "Alexander Van der Bellen is the President of Austria." Below this, a "KNOWLEDGE" section is shown for "Alexander Van der Bellen", identified as "President of Austria". It includes a small profile picture of him, his name, his title, and a detailed biography. The biography mentions he is the current President of Austria, previously served as a professor of economics at the University of Vienna, and joined politics as the spokesman of the Austrian Green Party. It notes his Russian and Estonian ancestry, and his birth in Austria to refugees from Stalinism. A "more" link and a "Wiki" link are at the end of the bio. At the bottom of the screen is the iOS home screen dock.

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?

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)

Conversational IR / Chat Bots

- A task-oriented dialog system

DB State

Attraction

```
"address": "98 king street",
"area": "centre",
"entrance fee": "free",
"id": "5",
"location": [
  52.207409,
  0.126738
],
"name": "broughton house gallery",
"openhours": "",
"phone": "01223314960",
"postcode": "cb11ln",
"pricerange": "free",
"type": "museum"
```

Restaurant

```
"address": "106 Regent Street City Centre",
"area": "centre",
"food": "indian",
"id": "19214",
"introduction": "curry garden serves tradition bangladeshi cuisine cooked with fresh produce",
"location": [
  52.200187,
  0.126407
],
"name": "curry garden",
"phone": "01223302330",
"postcode": "cb21dp",
"pricerange": "expensive",
"type": "restaurant"
```

Belief State

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

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

Text/Document analysis

DEEP SOURCE Iraq Inter-Cluster Coordination Group – Situation and needs monitor...

Simplified Text Original All Entries Source Details Primary Tagging Secondary Tagging Review Close SAVE FINALIZE

VIEW FRAMEWORK IMAGE FOR REFERENCE

INTER-CLUSTER OVERVIEW

The inter-cluster overview is prepared based on key inter-sectoral indicators discussed and agreed by the ICCG, for which trends and original data sources are presented in the online dashboard here.

Overview

During the first five months of 2021, Iraq witnessed no major shocks which could have significantly impacted the scale or scope of humanitarian needs in the country.

During the first five months of 2021, Iraq witnessed no major shocks which could have significantly impacted the scale or scope of humanitarian needs in the country

. However, the effects of camp closures and COVID-19, which at the time of the development of the 2021 Humanitarian Needs Overview (HNO) were still ongoing and yet to be fully assessed, are now better understood, and have to some degree stabilized.

In addition, the Iraqi currency devaluation and dry weather

Operational Environment Sectoral Information

Virtual 1D

Context

DEMOGRAPHY ECONOMY ENVIRONMENT SECURITY AND STABILITY

SOCIO CULTURAL LEGAL AND POLICY POLITICS TECHNOLOGICAL

Shock Event

TYPE AND CHARACTERISTICS UNDERLYING/AGGRAVATING FACTORS HAZARDS AND THREATS

Casualties

DEAD INJURED MISSING

Displacement

TYPE/NUMBERS/MOVEMENT PUSH FACTORS PULL FACTORS INTENTIONS

LOCAL INTEGRATION

Humanitarian Access

POPULATION TO RELIEF PHYSICAL CONSTRAINTS

NUMBER OF PEOPLE FACING HUMANITARIAN ACCESS CONSTRAINTS/HUMANITARIAN ACCESS GAPS

Excerpt extraction and 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

NLP Entities Category Editor SHOW SUGGESTIONS FOR 6 selected

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

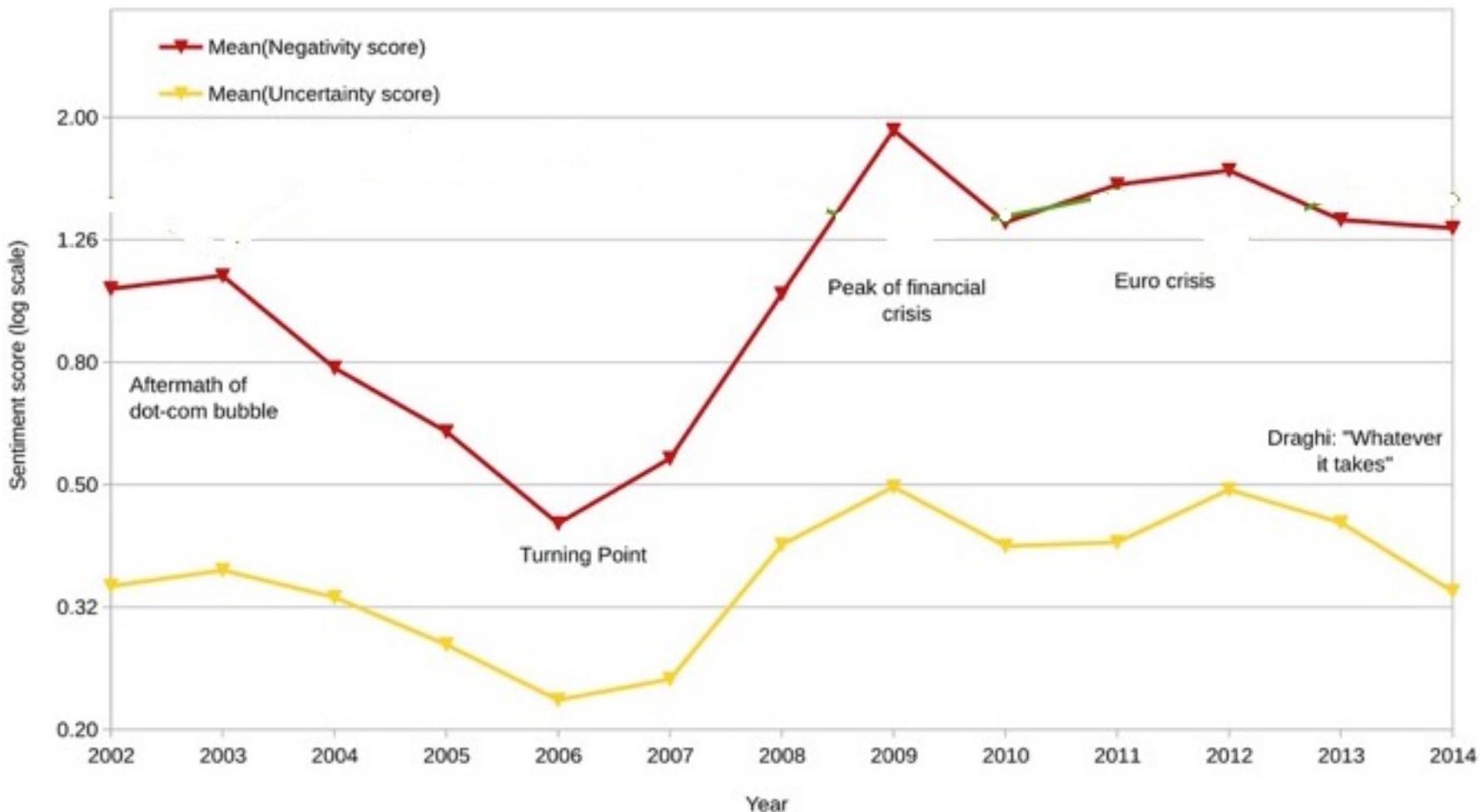
Summary - Based on 377 reviews



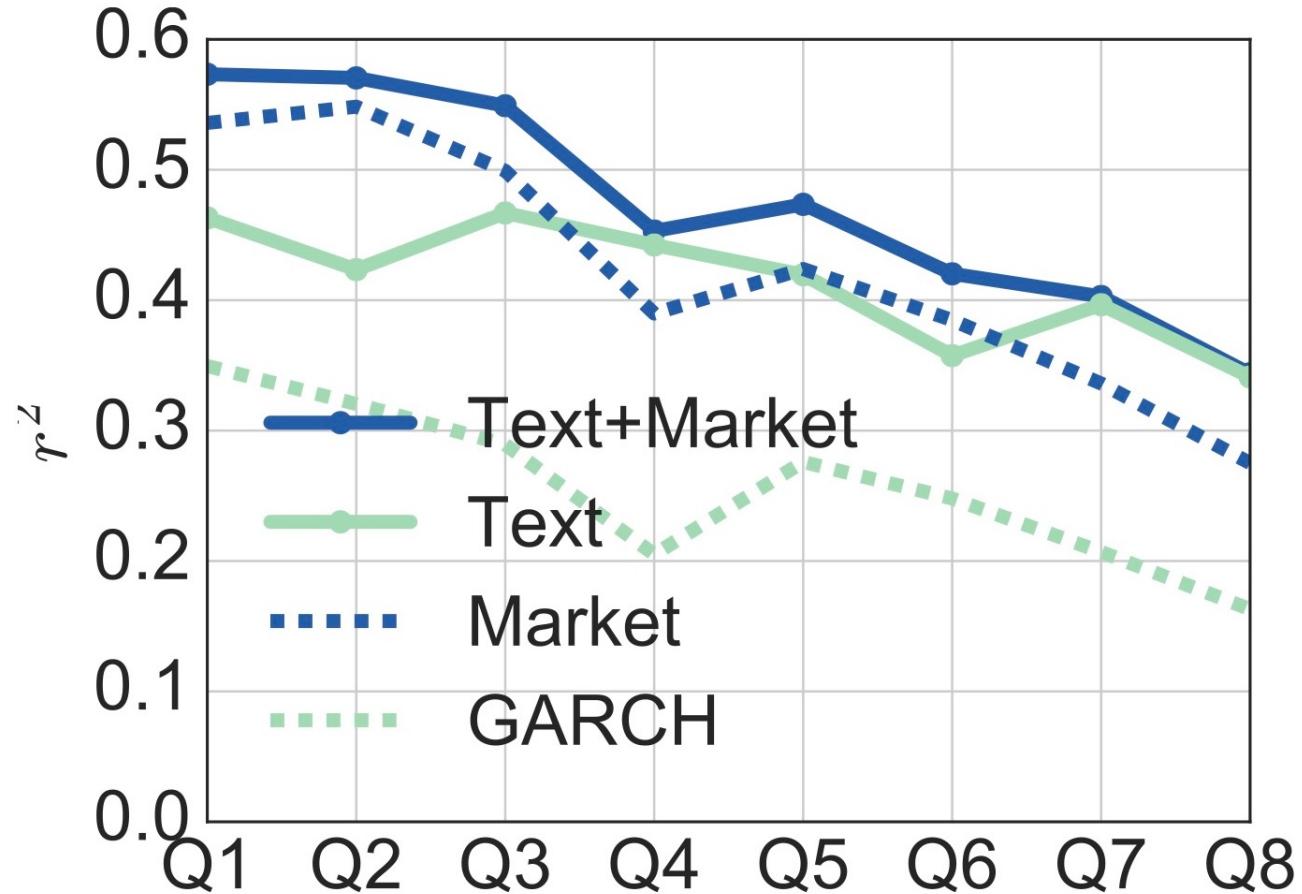
What people are saying

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."

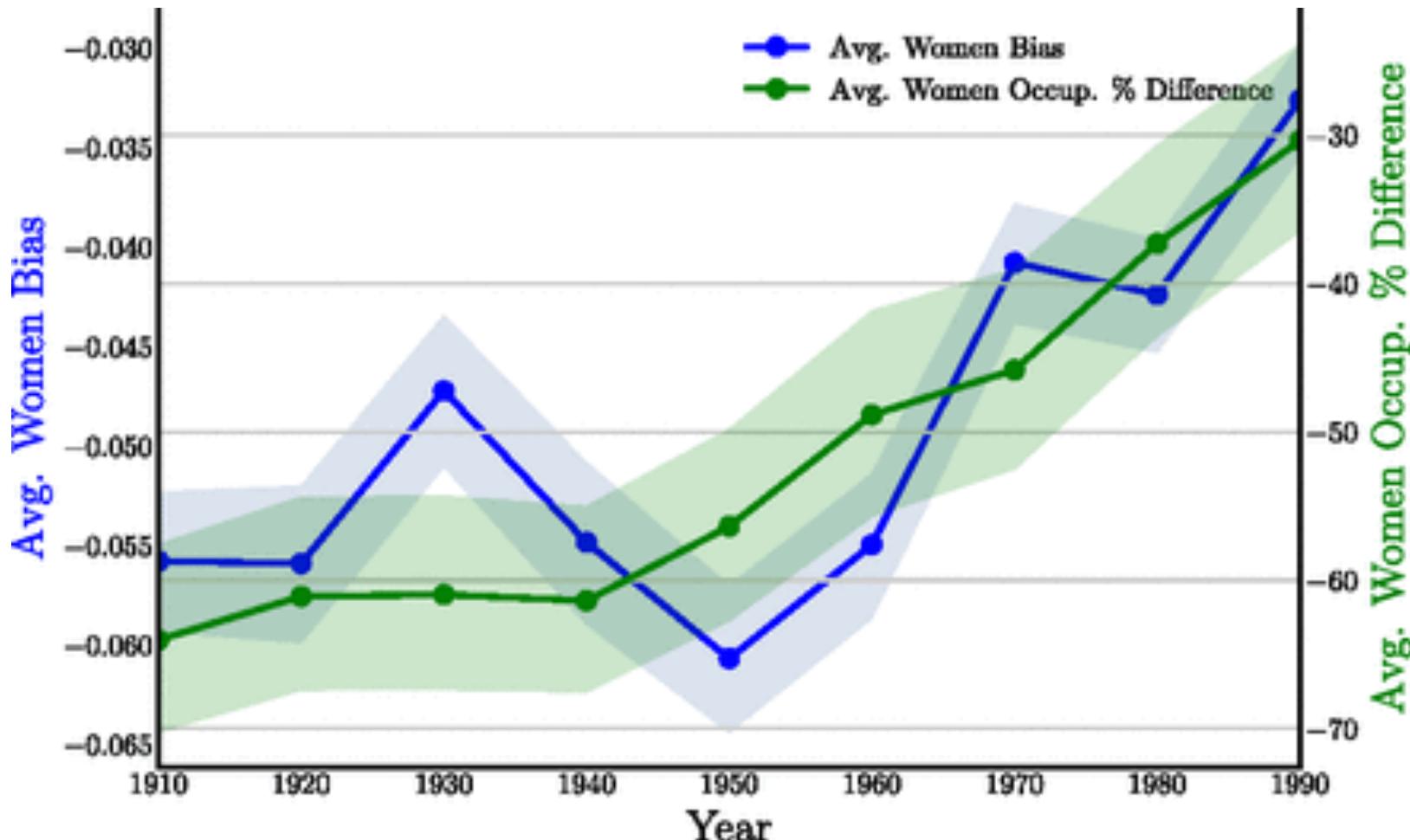
Monitoring market change from text data



Market/Volatility prediction



Monitoring societal changes



Information Extraction (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
...

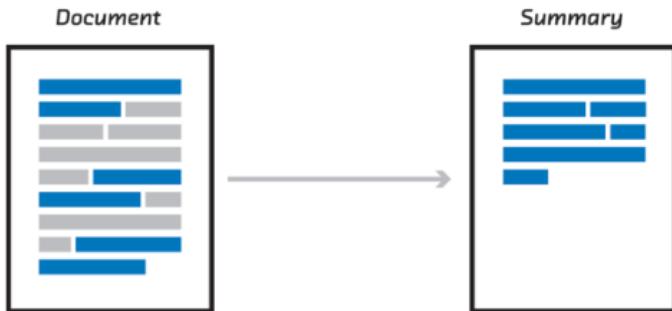
textual abstract:
summary for human

structured knowledge extraction:
summary for machine

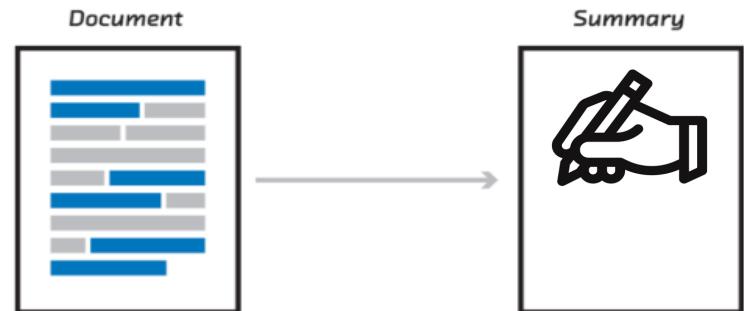
Summarization

- 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 Jgraphics. #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 acceptable. 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 .

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.

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.

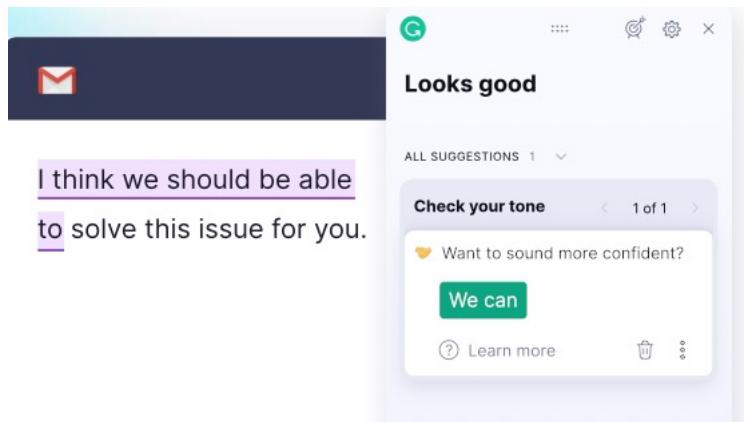
Content generator / Writing assistant / Spell checker

If you'd told me a year ago that today I would finish a marathon, I would have laughed. Your support had a huge affect on me!

Replace the word
effect

Let me tell you a story about a little robot named Jarvis. Jarvis is a robot that can do anything. He's been programmed to write blog posts, articles, scripts and even books. He has an advanced artificial intelligence system that makes him the perfect writer for you.

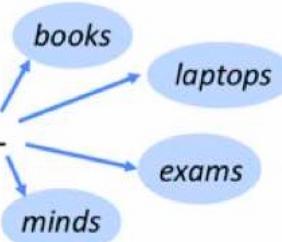
He'll take your ideas and turn them into something beautiful. You don't have to worry about grammar or spelling mistakes because he does all the work for you. Jarvis can even make your content more exciting!



A screenshot of the LanguageTool website. It features a text input area with a purple border containing the 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." Below the text area are dropdown menus for "English" and "American" and a "Check Text" button. A red box at the bottom contains the message: "English has incomplete support in LanguageTool. Would you like to help?"

Language Modeling

the students opened their _____



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

Look more:

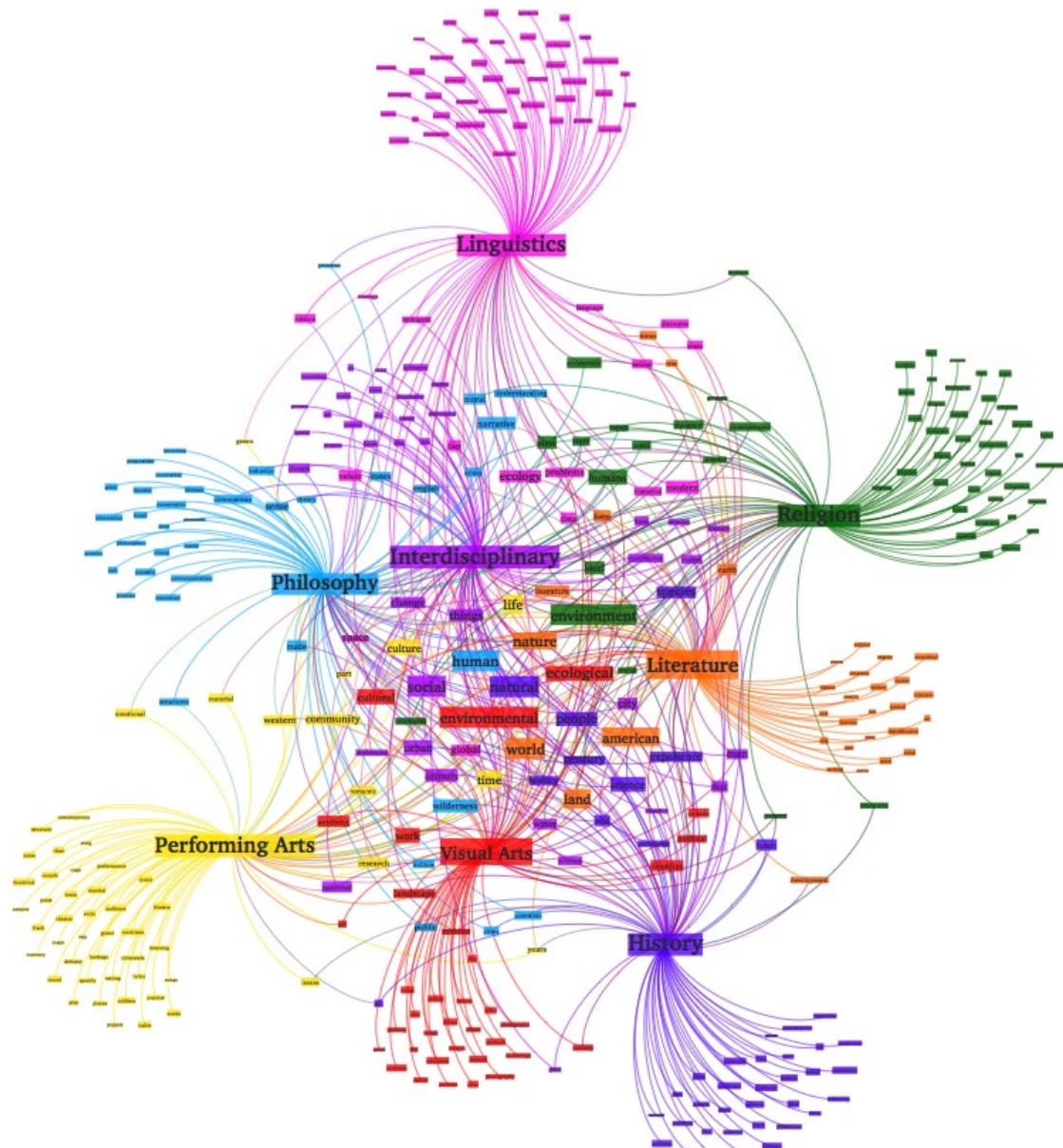
<https://taktotransformer.com>

<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.

Topic modeling / Clustering



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 above the text:

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

Named Entity Recognition

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

Relation Extraction

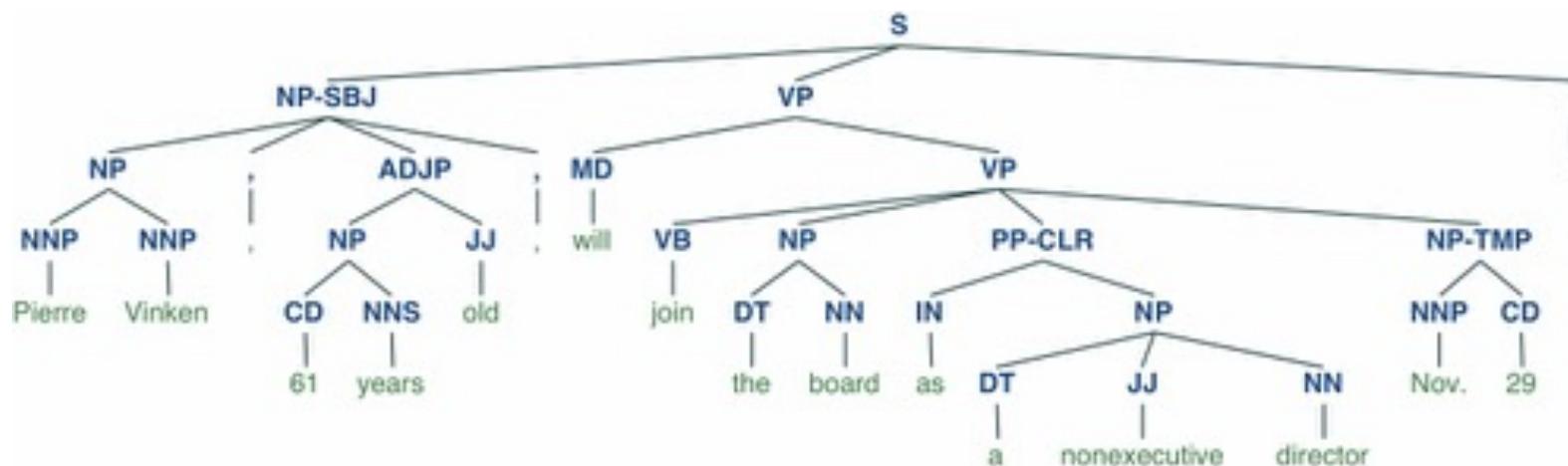
CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by \$6 per round trip on flights to some cities also served by lower-cost carriers. **American Airlines**, a **unit of AMR**, immediately matched the move, **spokesman Tim Wagner** said. **United**, a **unit of 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

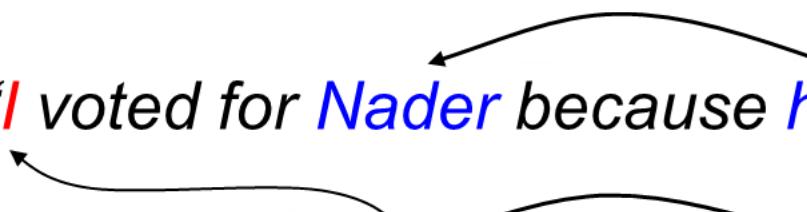
- “A Part-Of-Speech Tagger is a piece of software that reads text in some language and assigns parts of speech to each word, such as noun, verb, adjective”



Coreference Resolution

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

“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](#) ▾

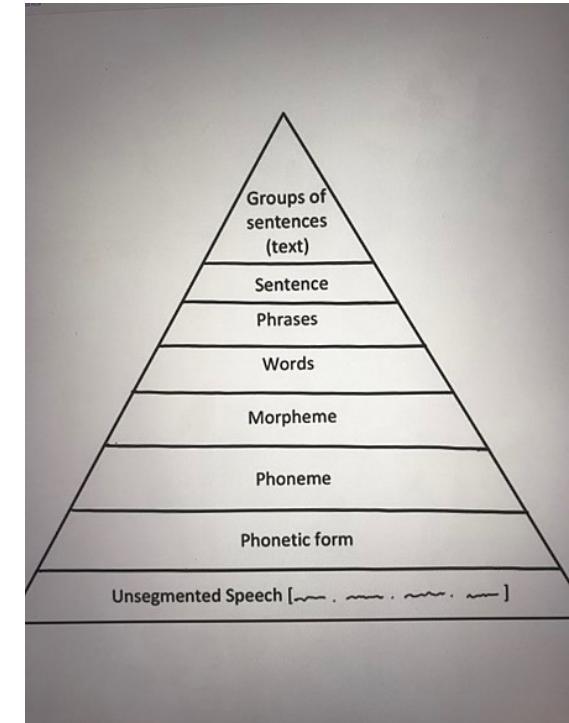
Agenda

- 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



"I MISS THE GOOD OLD DAYS WHEN ALL WE HAD TO WORRY ABOUT WAS NOUNS AND VERBS."

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”



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

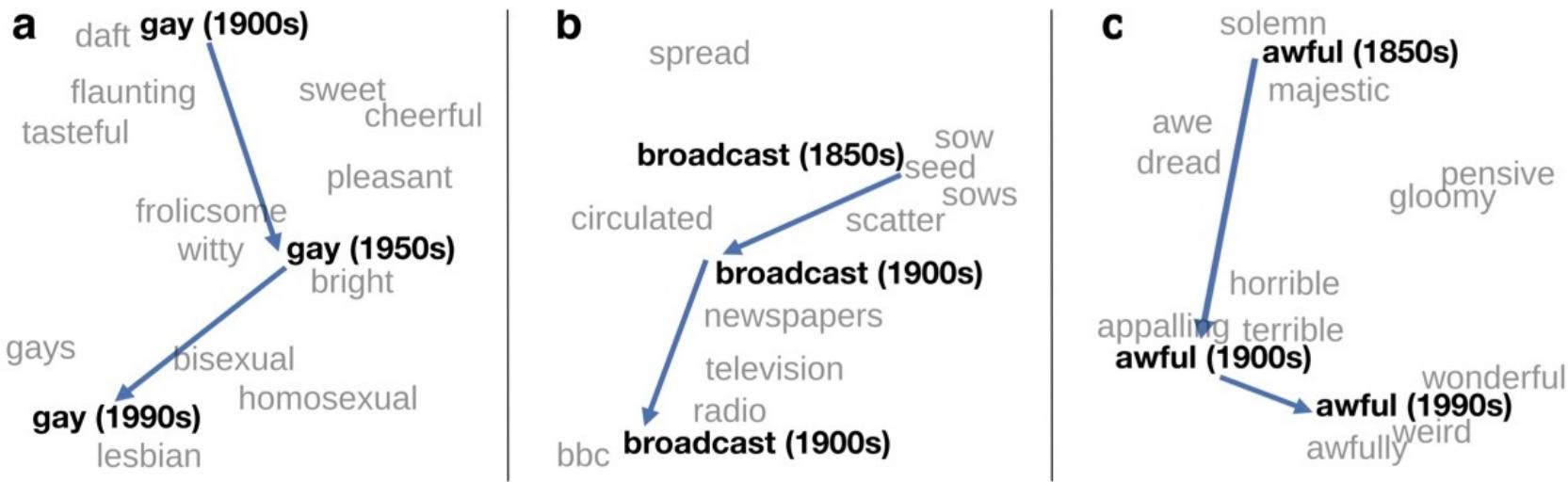


Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.² **a**, The word *gay* shifted from meaning “cheerful” or “frolicsome” to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to “casting out seeds”; with the rise of television and 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/7ROeIYvo8f0?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

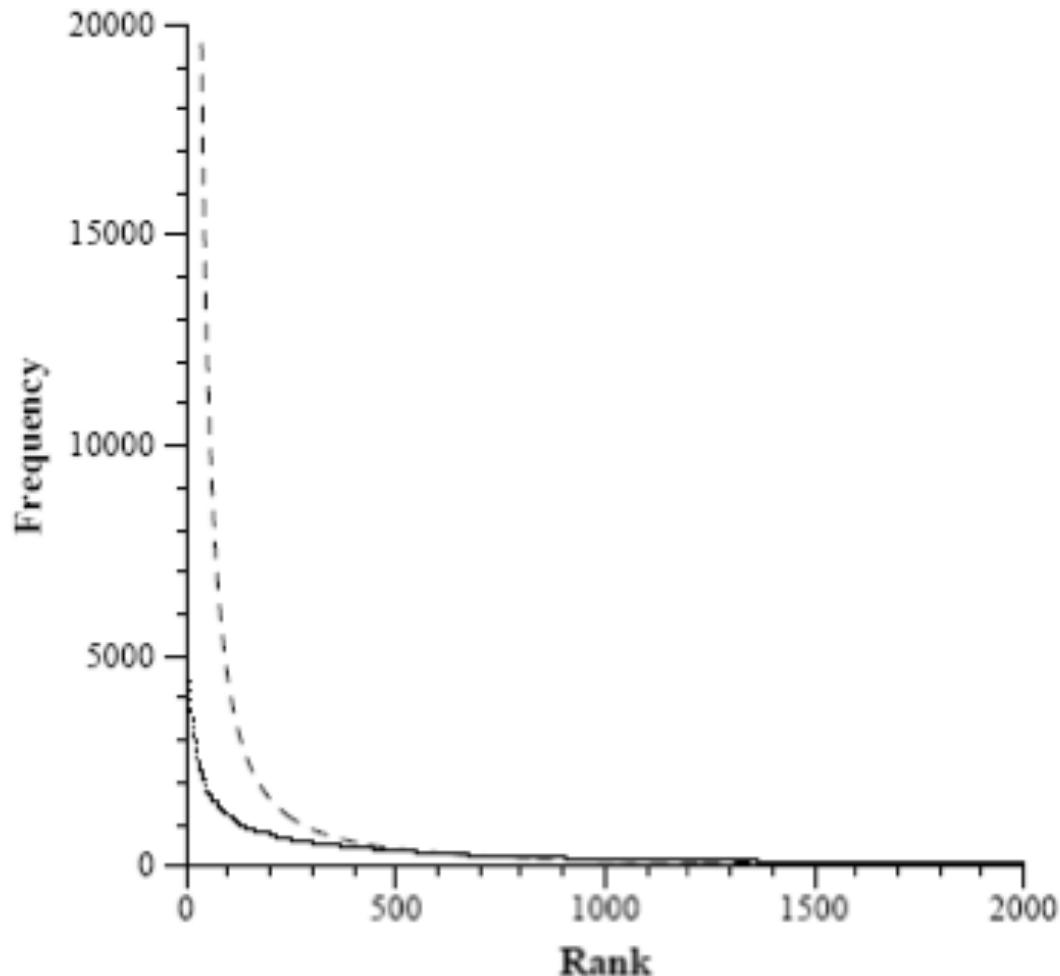
- Sort the vocabularies of a dictionary based on their frequencies in a text corpus
 - E.g. the results from the [Brown Corpus](#):

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*

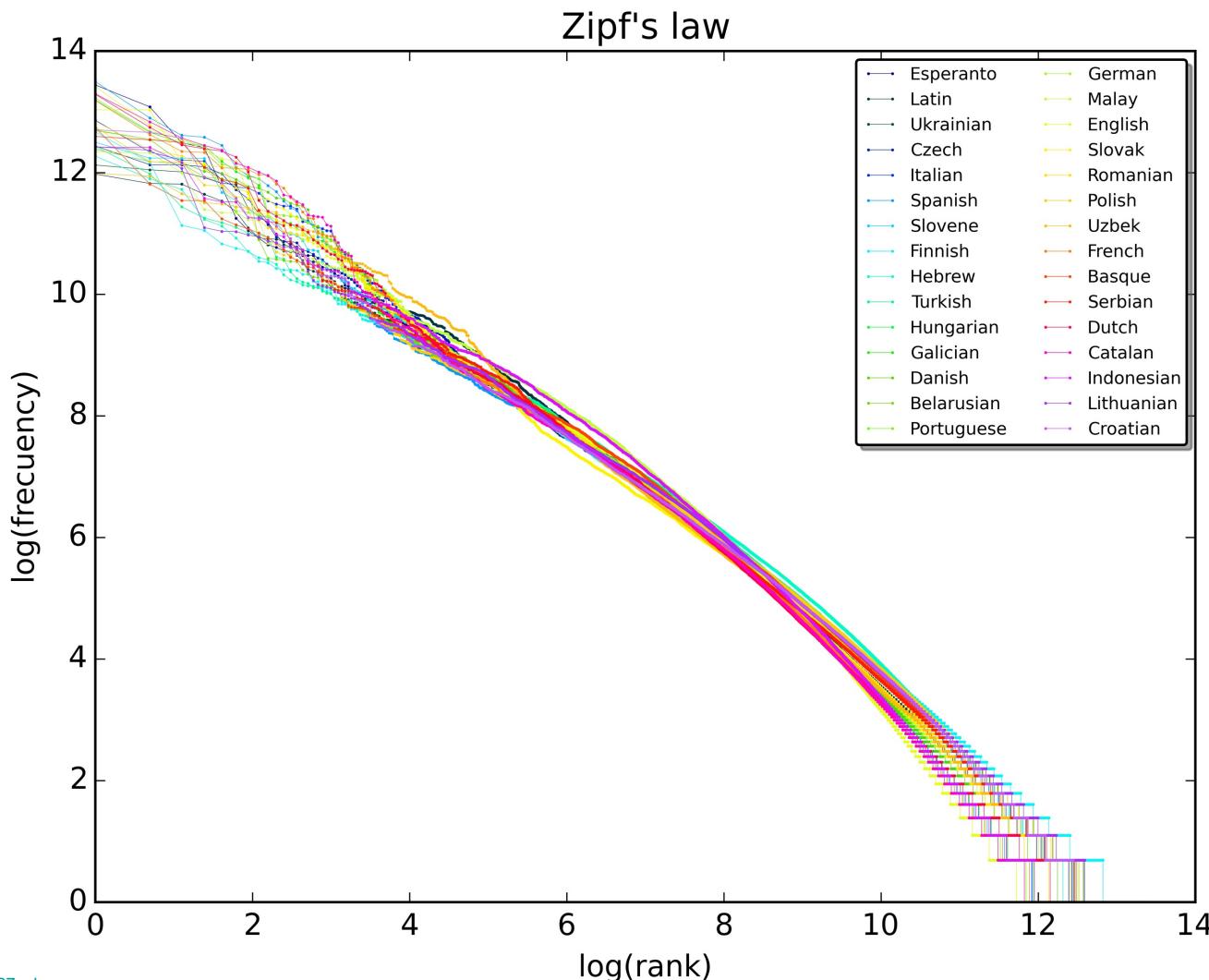
Zipf's Law

- The plot of rank versus frequency looks like this:



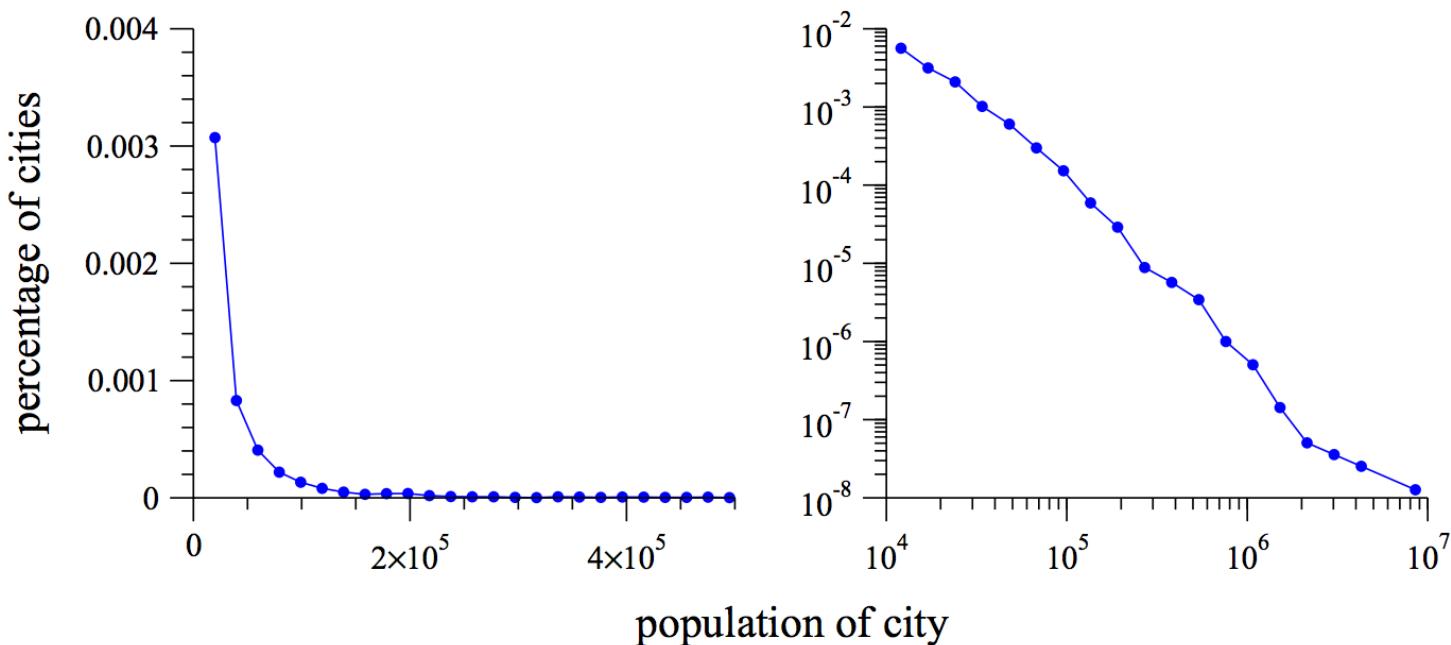
Zipf's Law

- Applying logarithm to both axes:



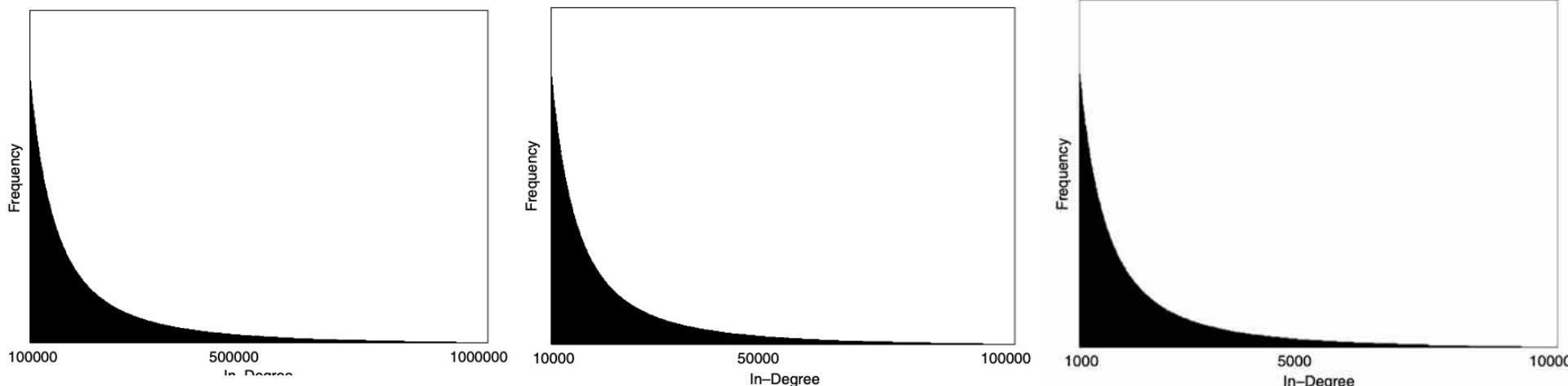
Power Law in social science

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



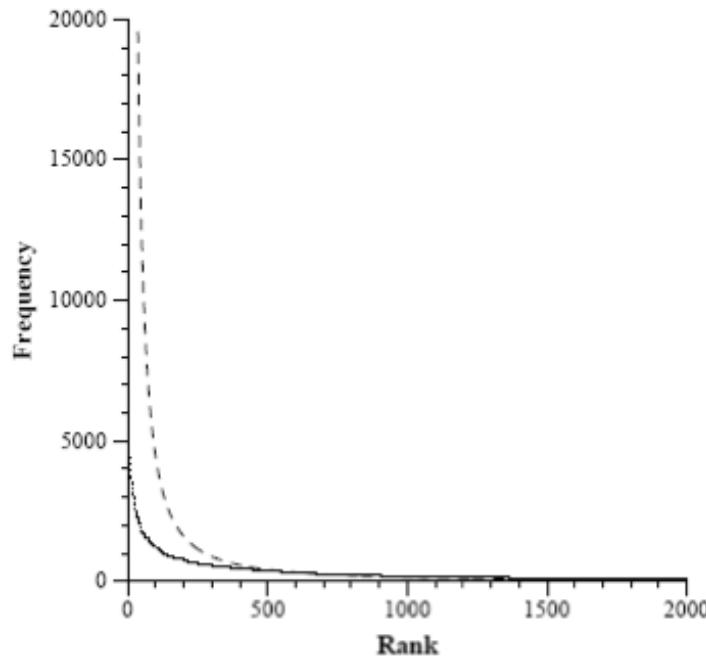
Power Law in web

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



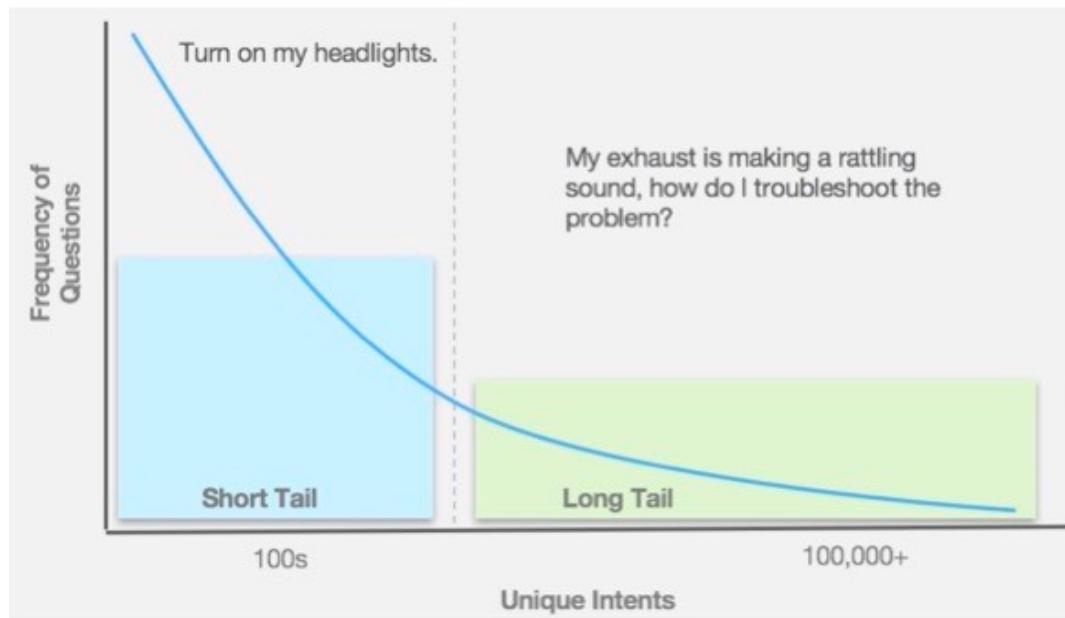
What does Zipf' Law tell us?

- 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?

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

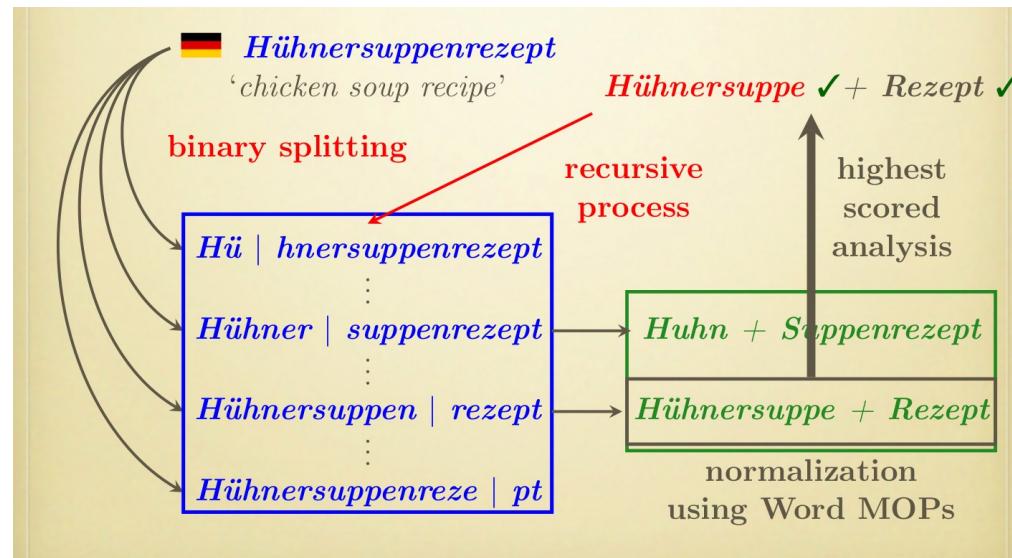
- Normalization harmonizes the written forms of the words with same meanings
- Some examples:
 - deleting periods
 - ***U.S.A.*** → ***USA***
 - deleting hyphens
 - ***anti-discriminatory*** → ***antidiscriminatory***
 - Accents
 - French ***résumé*** → ***resume***
 - Umlauts
 - German: ***Tuebingen*** → ***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):



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* → *ss*
 - *ies* → *i*
 - *ational* → *ate*
 - *tional* → *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
- Lemmatizer reduces inflectional/variant forms to base forms.
Examples:
 - **am, are, is → be**
 - **car, cars, car's, cars' → car**
 - ***the boy's cars are different colors* → *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