NLP for Social Good

Splash 2018

What is it?

…and why do we care?

What is “NLP”?

“**Natural-language processing** (**NLP**) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data.”

- Wikipedia

https://en.wikipedia.org/wiki/Natural-language_processing

What is “NLP”?

“Natural-language processing (NLP) is an area of **computer science and artificial intelligence** concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data.”

- Wikipedia

https://en.wikipedia.org/wiki/Natural-language_processing

What is “NLP”?

“Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the **interactions between computers and human (natural) languages**, in particular how to program computers to fruitfully process large amounts of natural language data.”

- Wikipedia

https://en.wikipedia.org/wiki/Natural-language_processing

What is “NLP”?

“Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular **how to program computers** to fruitfully process large amounts of natural language data.”

- Wikipedia

https://en.wikipedia.org/wiki/Natural-language_processing

What is “NLP”?

“Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully **process large amounts of natural language data**.”

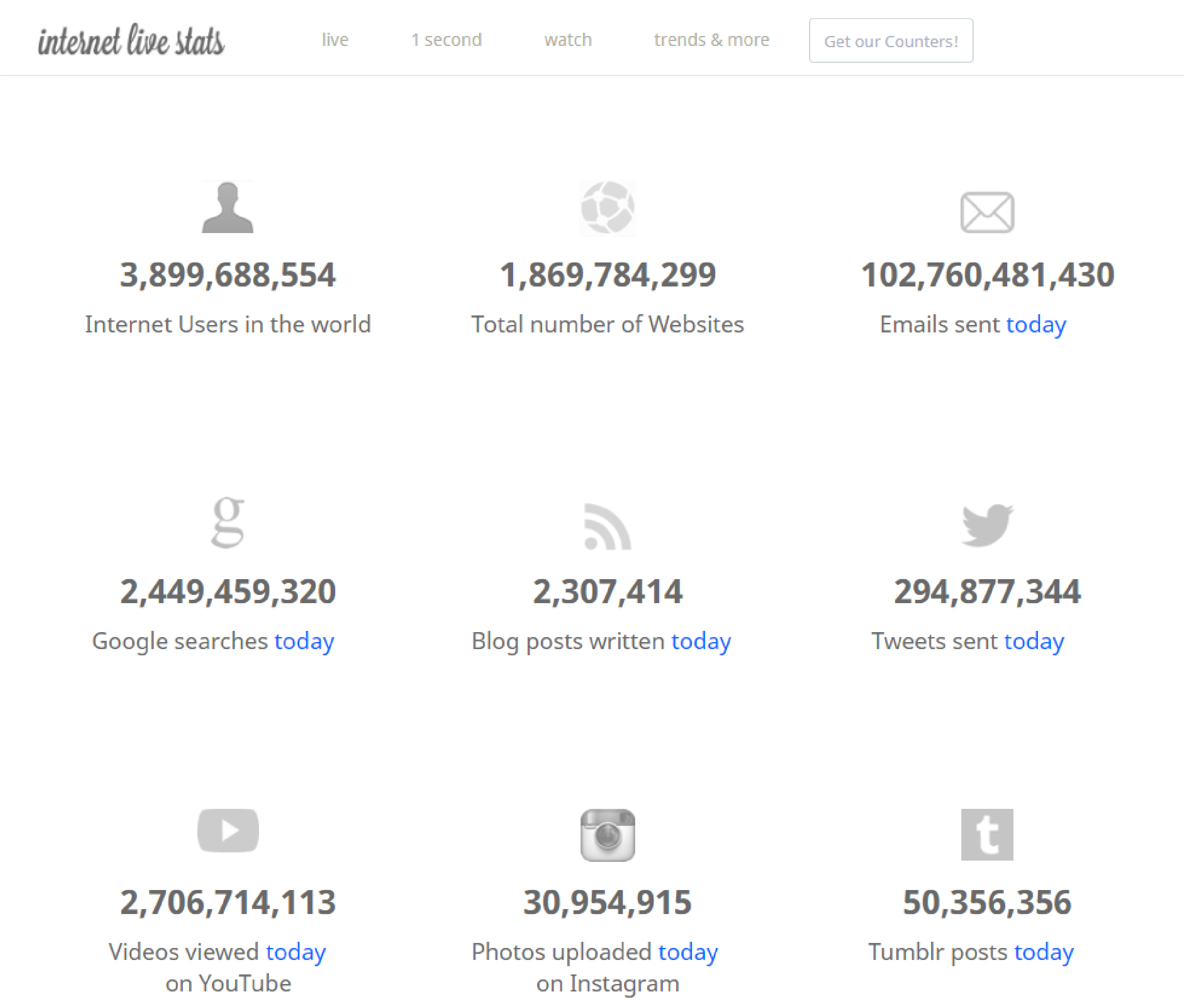
- Wikipedia

https://en.wikipedia.org/wiki/Natural-language_processing

Why do we care?

“Analysts at Gartner (gated) estimate that **upward of 80%** of enterprise data today is **unstructured**.”

https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#13919ec4493a



3,899,688,554
Internet Users in the world

1,869,784,299
Total number of Websites

102,760,481,430
Emails sent today

2,449,459,320
Google searches today

2,307,414
Blog posts written today

294,877,344
Tweets sent today

2,706,714,113
Videos viewed today on YouTube

30,954,915
Photos uploaded today on Instagram

50,356,356
Tumblr posts today

http://www.internetlivestats.com/

What can we work on?

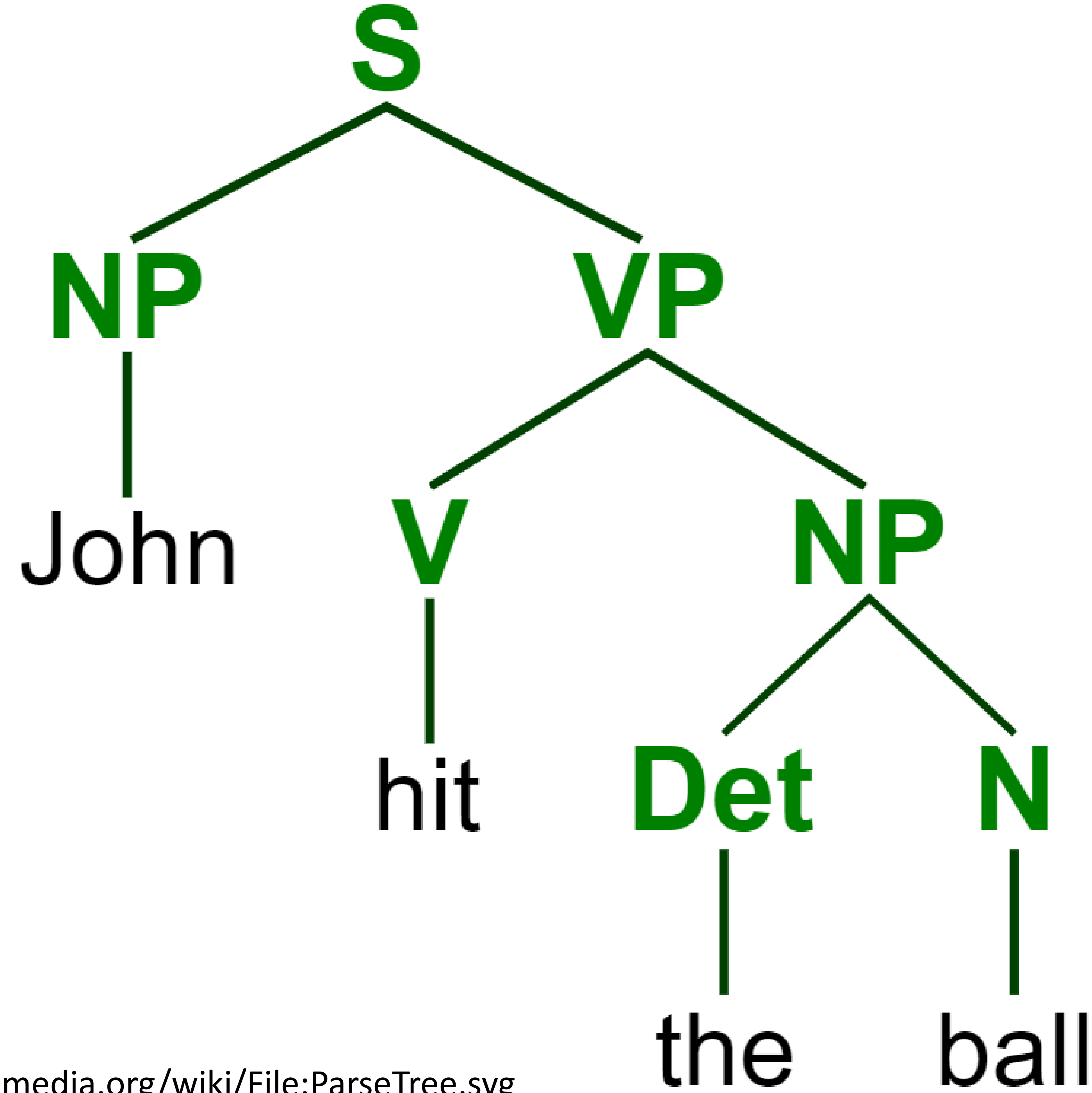
What is “NLP”?

“Natural-language processing (NLP) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data.

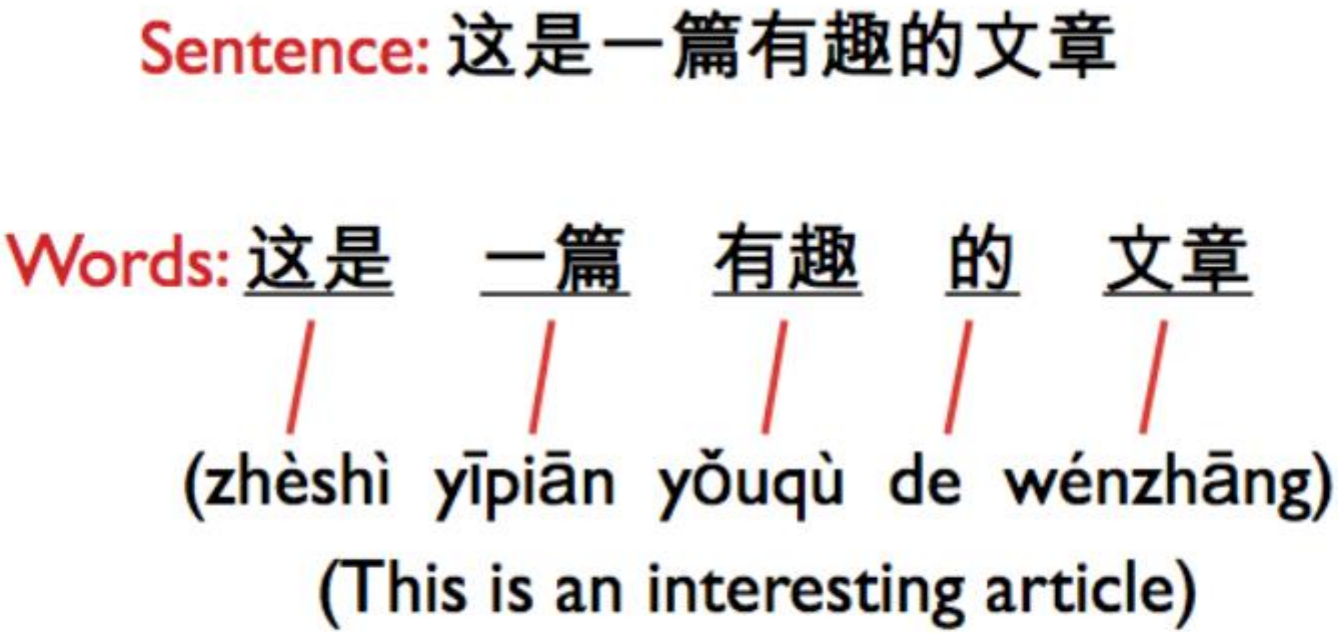
Challenges in natural-language processing frequently involve **speech recognition**, **natural-language understanding**, and **natural-language generation**.”

- Wikipedia

https://en.wikipedia.org/wiki/Natural-language_processing



https://commons.wikimedia.org/wiki/File:ParseTree.svg



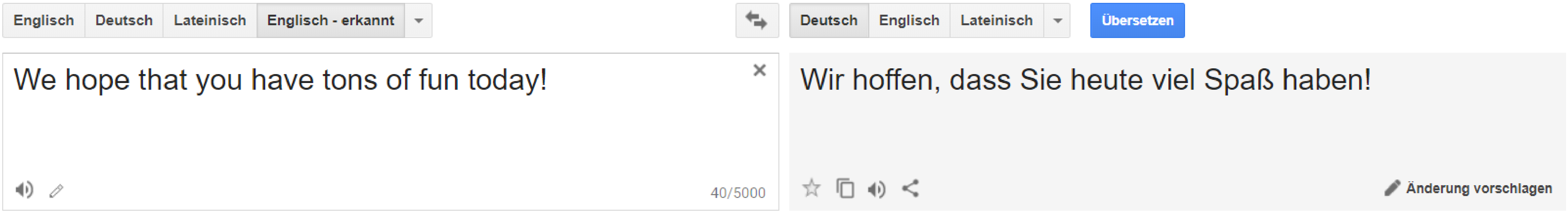
Sentence: 这是一篇有趣的文章

Words: 这是 一篇 有趣 的 文章

(zhèshì yīpiān yǒuqù de wénzhāng)

(This is an interesting article)

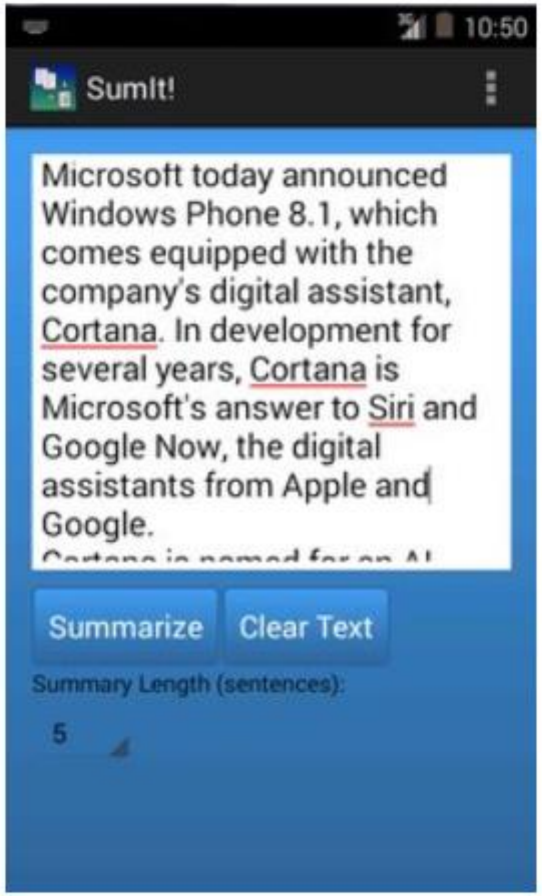
http://iconictranslation.com/2014/10/language-challenge-7-chinese/



We hope that you have tons of fun today!

Wir hoffen, dass Sie heute viel Spaß haben!

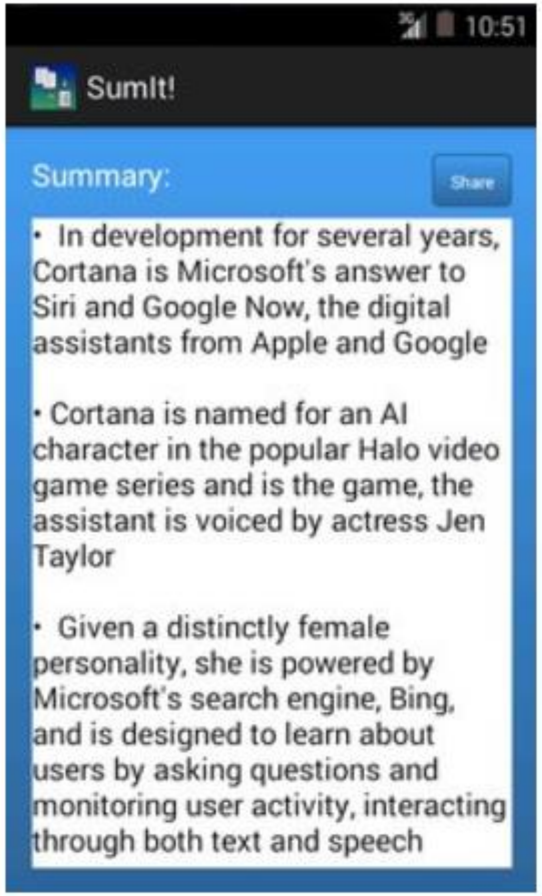
translate.google.com



Microsoft today announced Windows Phone 8.1, which comes equipped with the company's digital assistant, Cortana. In development for several years, Cortana is Microsoft's answer to Siri and Google Now, the digital assistants from Apple and Google.
Cortana is named for an AI

Summary Length (sentences):

5



Summary:

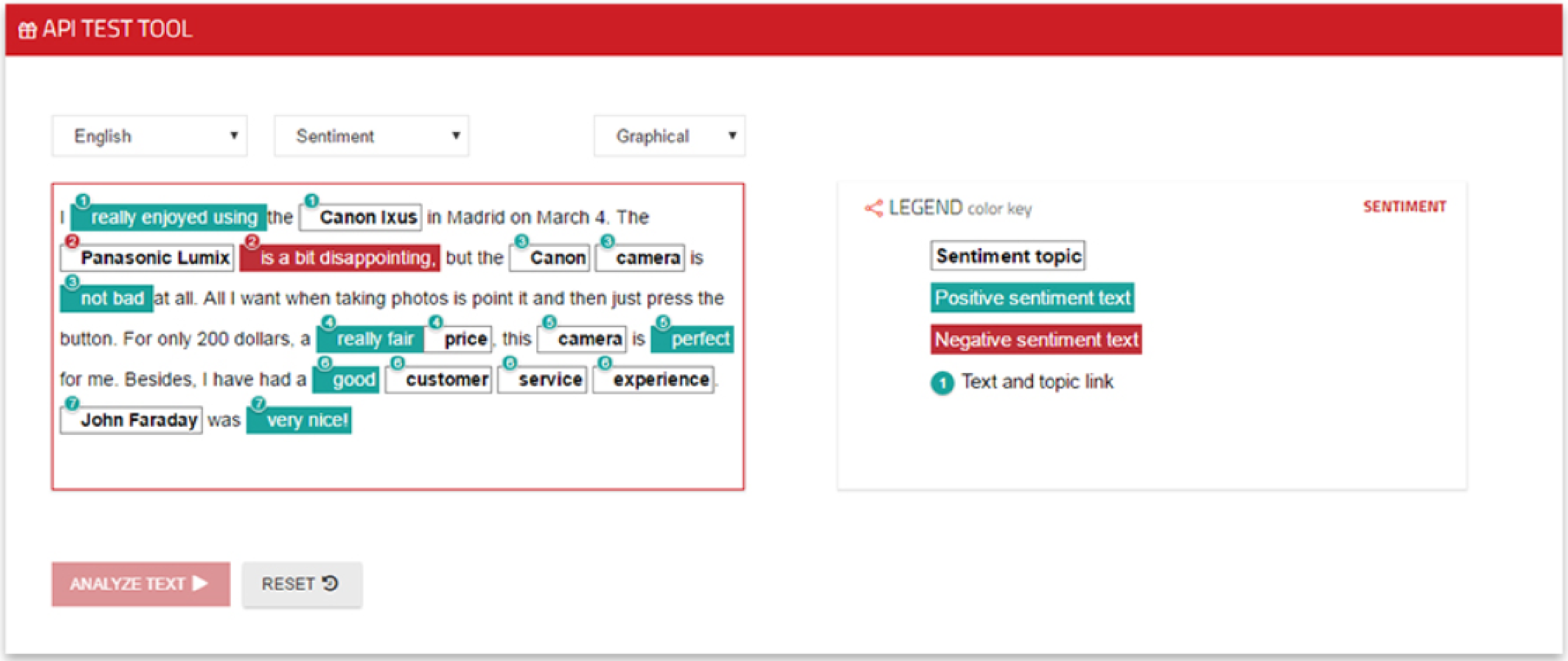
- In development for several years, Cortana is Microsoft's answer to Siri and Google Now, the digital assistants from Apple and Google

- Cortana is named for an AI character in the popular Halo video game series and is the game, the assistant is voiced by actress Jen Taylor

- Given a distinctly female personality, she is powered by Microsoft's search engine, Bing, and is designed to learn about users by asking questions and monitoring user activity, interacting through both text and speech

https://apkpure.com/sumit-text-summarization/com.karimo.sumit_final

Bitext Topic-based Sentiment Analysis service provides polarity, sentiment scoring, sentiment text and, most important, sentiment topic identification out of raw data with over 90% accuracy by relying on Deep Linguistic Analysis and in our in-house parsing technology.



English ▾ Sentiment ▾ Graphical ▾

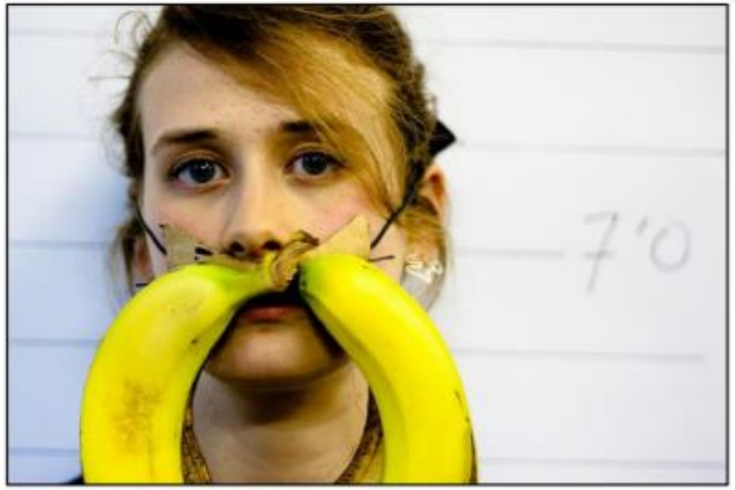
I really enjoyed using the Canon Ixus in Madrid on March 4. The Panasonic Lumix is a bit disappointing, but the Canon camera is not bad at all. All I want when taking photos is point it and then just press the button. For only 200 dollars, a really fair price, this camera is perfect for me. Besides, I have had a good customer service experience. John Faraday was very nice!

LEGEND color key

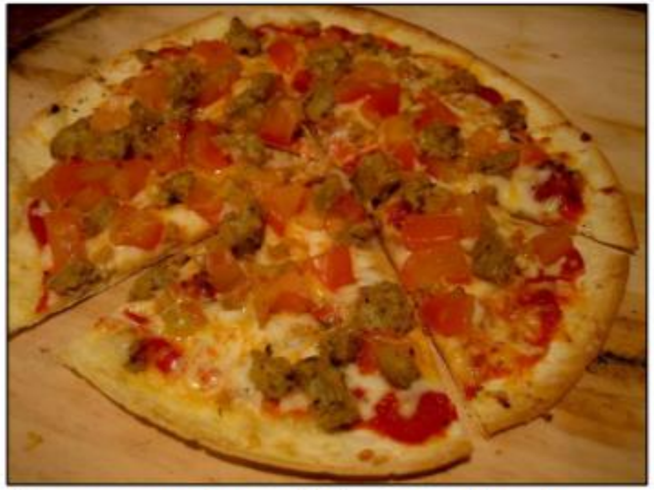
Sentiment topic
Positive sentiment text
Negative sentiment text
1 Text and topic link

I want to try it →

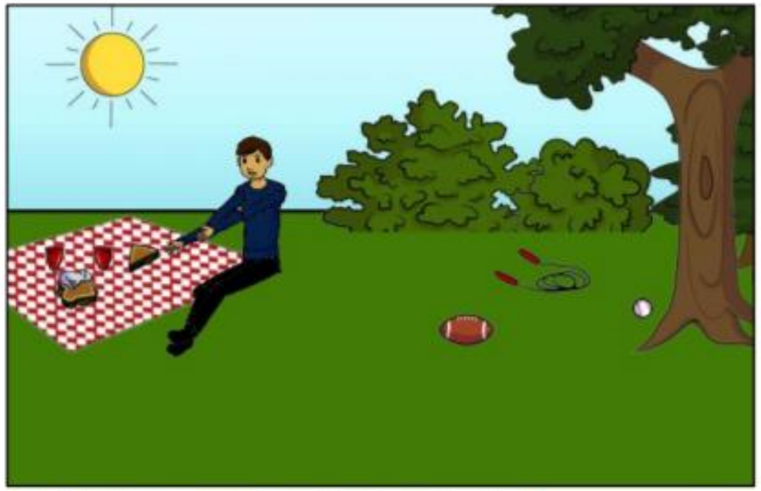
https://www.bitext.com/sentiment-analysis/



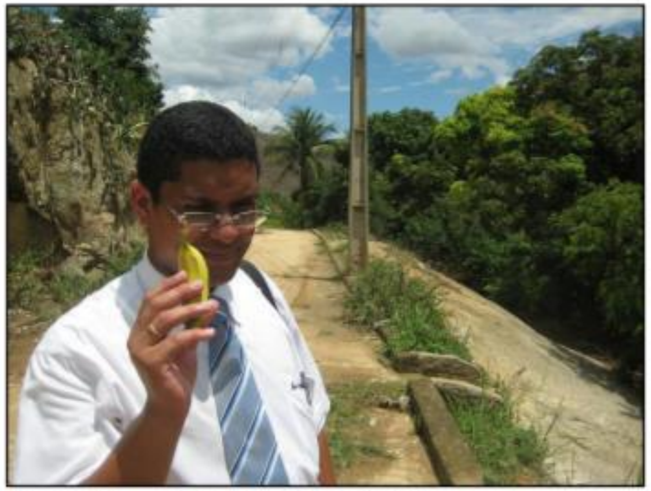
What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



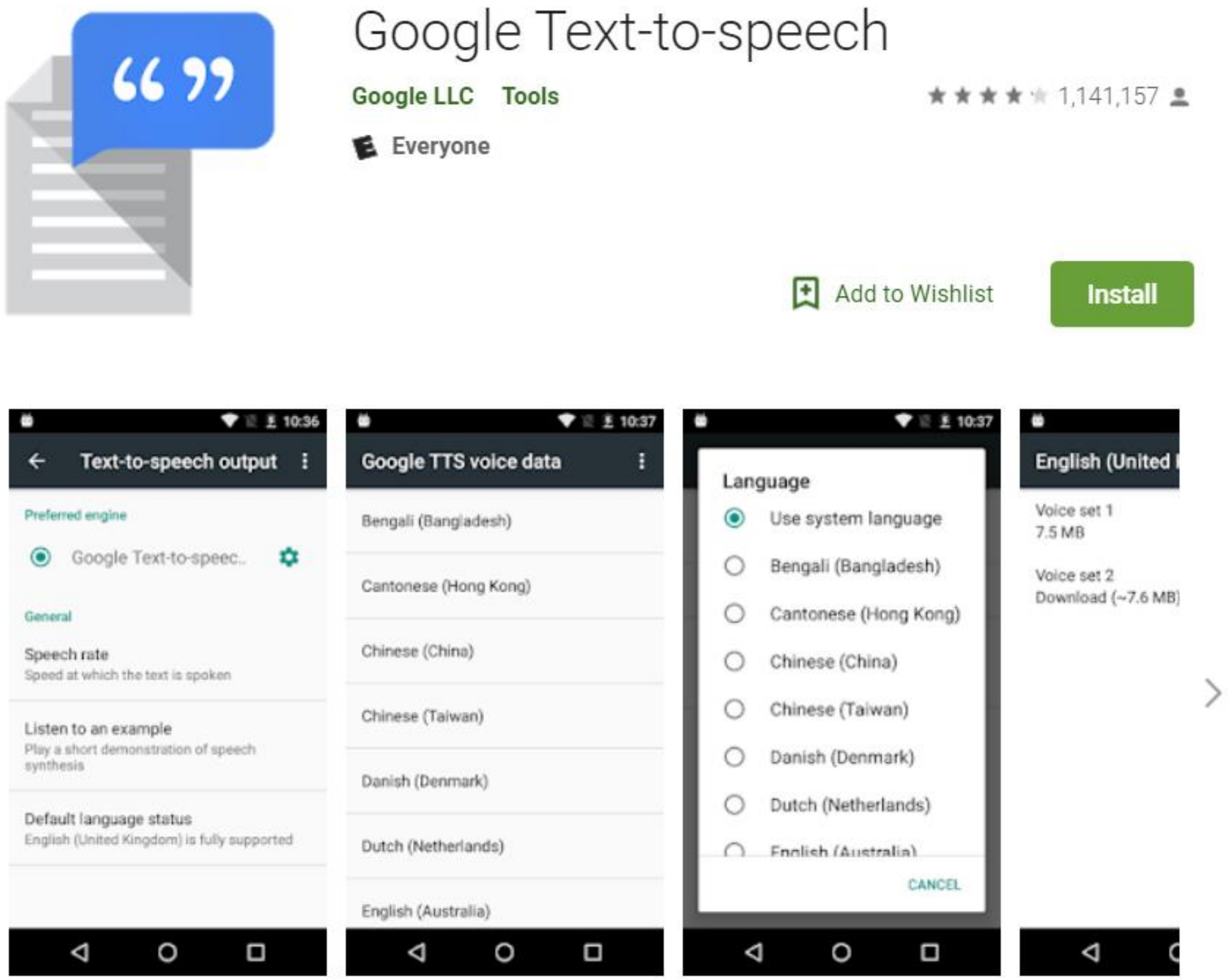
Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

A. Agrawal *et al.*, “VQA: Visual Question Answering,” *arXiv:1505.00468 [cs]*, May 2015.

https://matlab1.com/support-vector-machine-speech-recognition/



Google Text-to-speech

Google LLC Tools

★★★★☆ 1,141,157

Everyone

Add to Wishlist

Install

10:36

Text-to-speech output

Preferred engine

◉ Google Text-to-speec..

General

Speech rate
Speed at which the text is spoken

Listen to an example
Play a short demonstration of speech synthesis

Default language status
English (United Kingdom) is fully supported

10:37

Google TTS voice data

Bengali (Bangladesh)

Cantonese (Hong Kong)

Chinese (China)

Chinese (Taiwan)

Danish (Denmark)

Dutch (Netherlands)

English (Australia)

10:37

Language

◉ Use system language

○ Bengali (Bangladesh)

○ Cantonese (Hong Kong)

○ Chinese (China)

○ Chinese (Taiwan)

○ Danish (Denmark)

○ Dutch (Netherlands)

○ English (Australia)

CANCEL

English (United

Voice set 1
7.5 MB

Voice set 2
Download (~7.6 MB)

https://play.google.com/store/apps/details?id=com.google.android.tts&hl=en_US

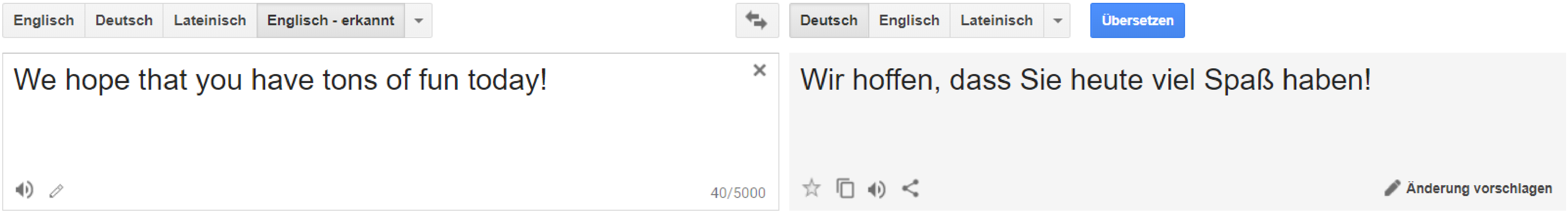
What makes it difficult?

What is “NLP”?

“**Natural-language processing** (**NLP**) is an area of computer science and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to fruitfully process large amounts of natural language data.”

- Wikipedia

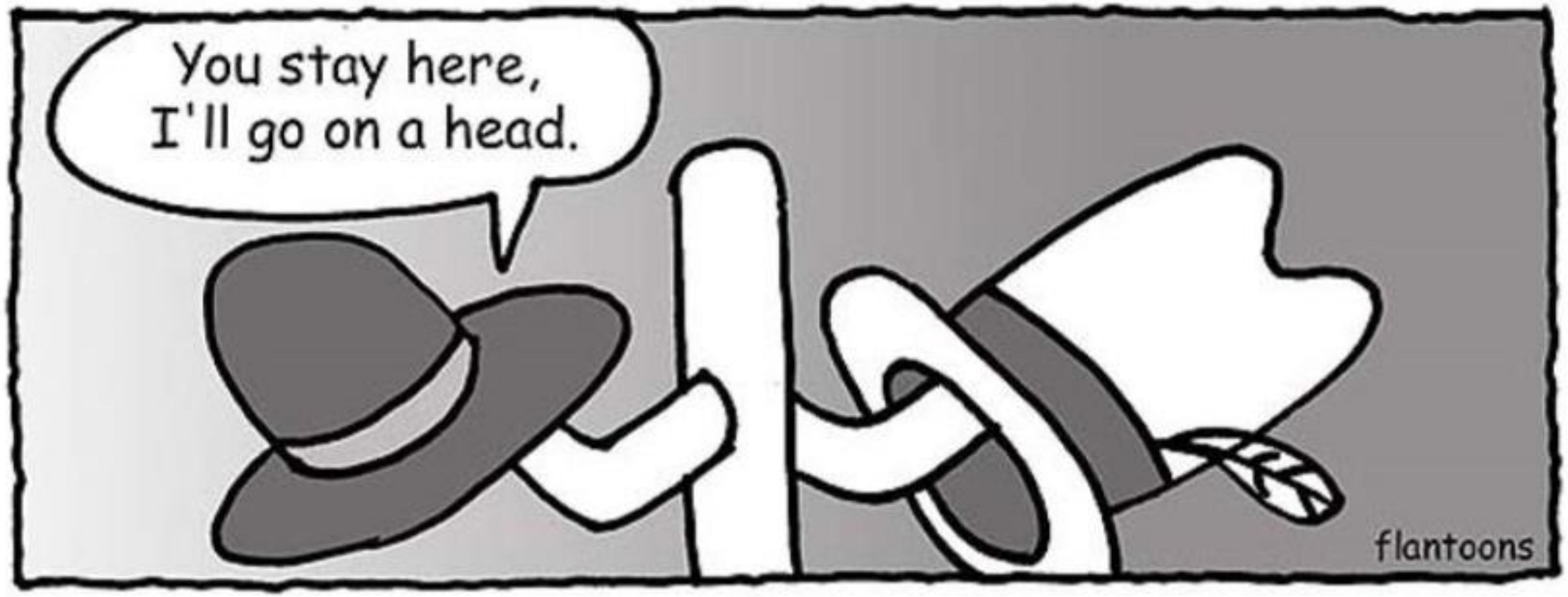
https://en.wikipedia.org/wiki/Natural-language_processing



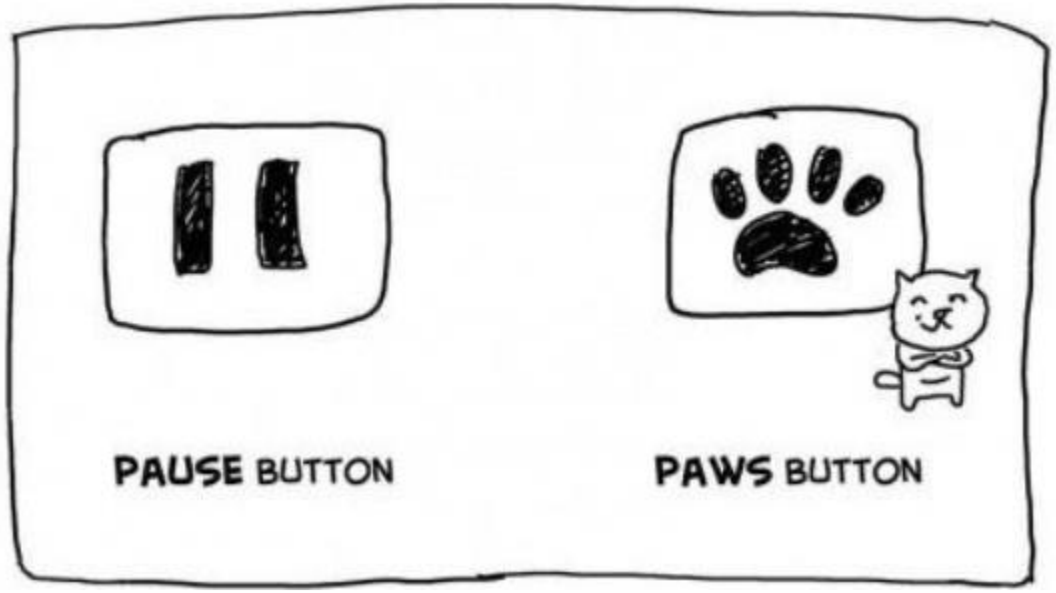
We hope that you have tons of fun today!

Wir hoffen, dass Sie heute viel Spaß haben!

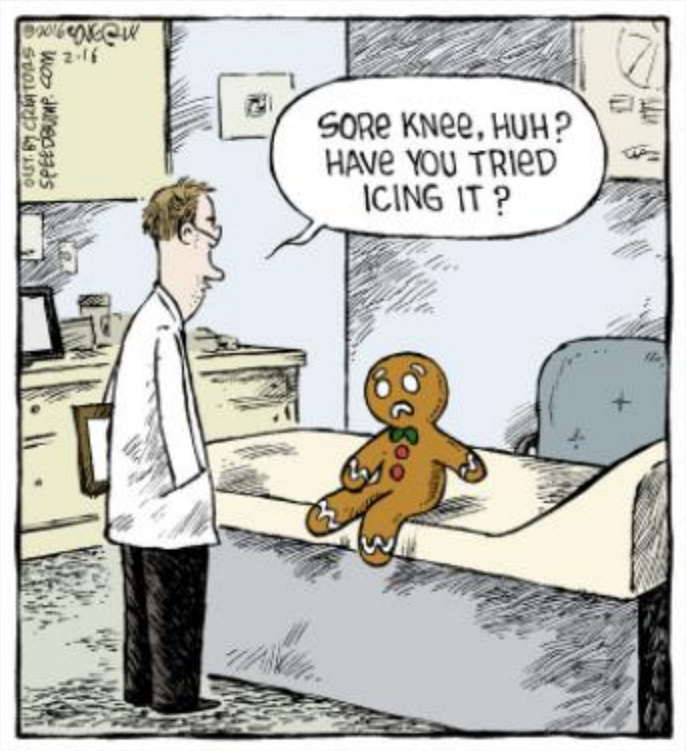
translate.google.com



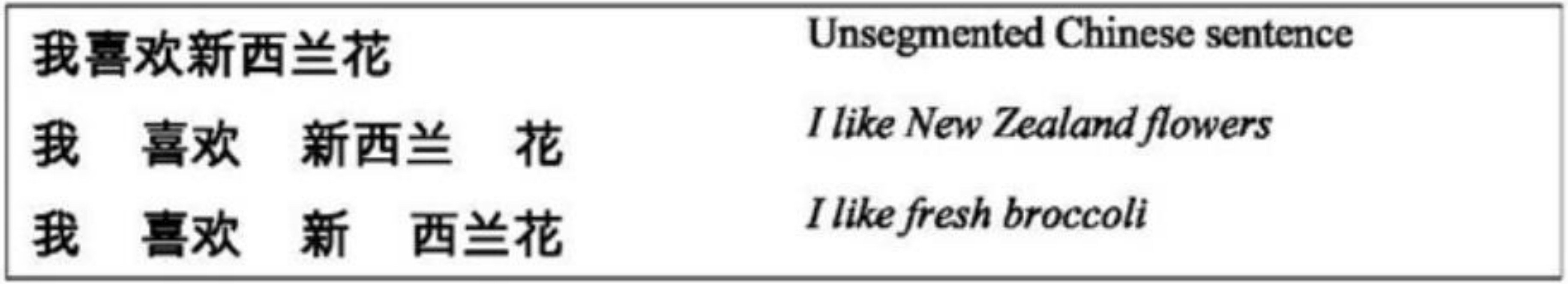
https://www.ecenglish.com/en/social/blog/ec-central/2015/11/23/vocab-review-homophones



https://i.pinimg.com/originals/0a/92/15/0a921501fa2c6ad60f2e7cde0c8e90a4.jpg

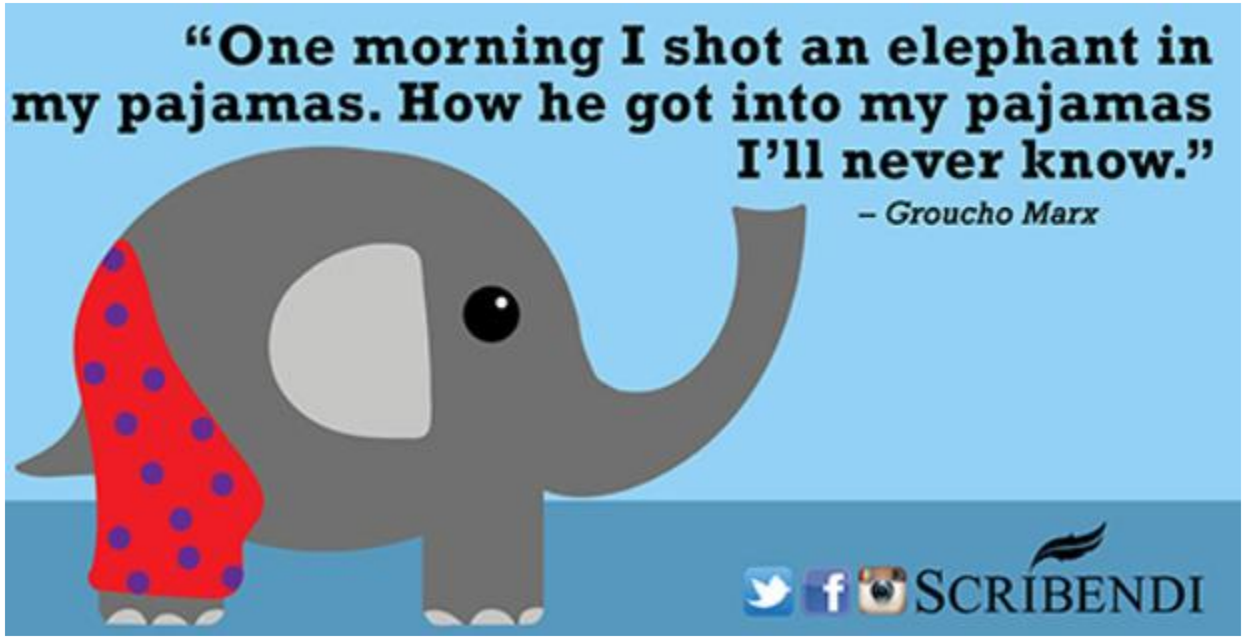


https://www.creators.com/read/speed-bump/02/16/165650



我喜欢新西兰花	Unsegmented Chinese sentence
我 喜欢 新西兰 花	*I like New Zealand flowers*
我 喜欢 新 西兰花	*I like fresh broccoli*

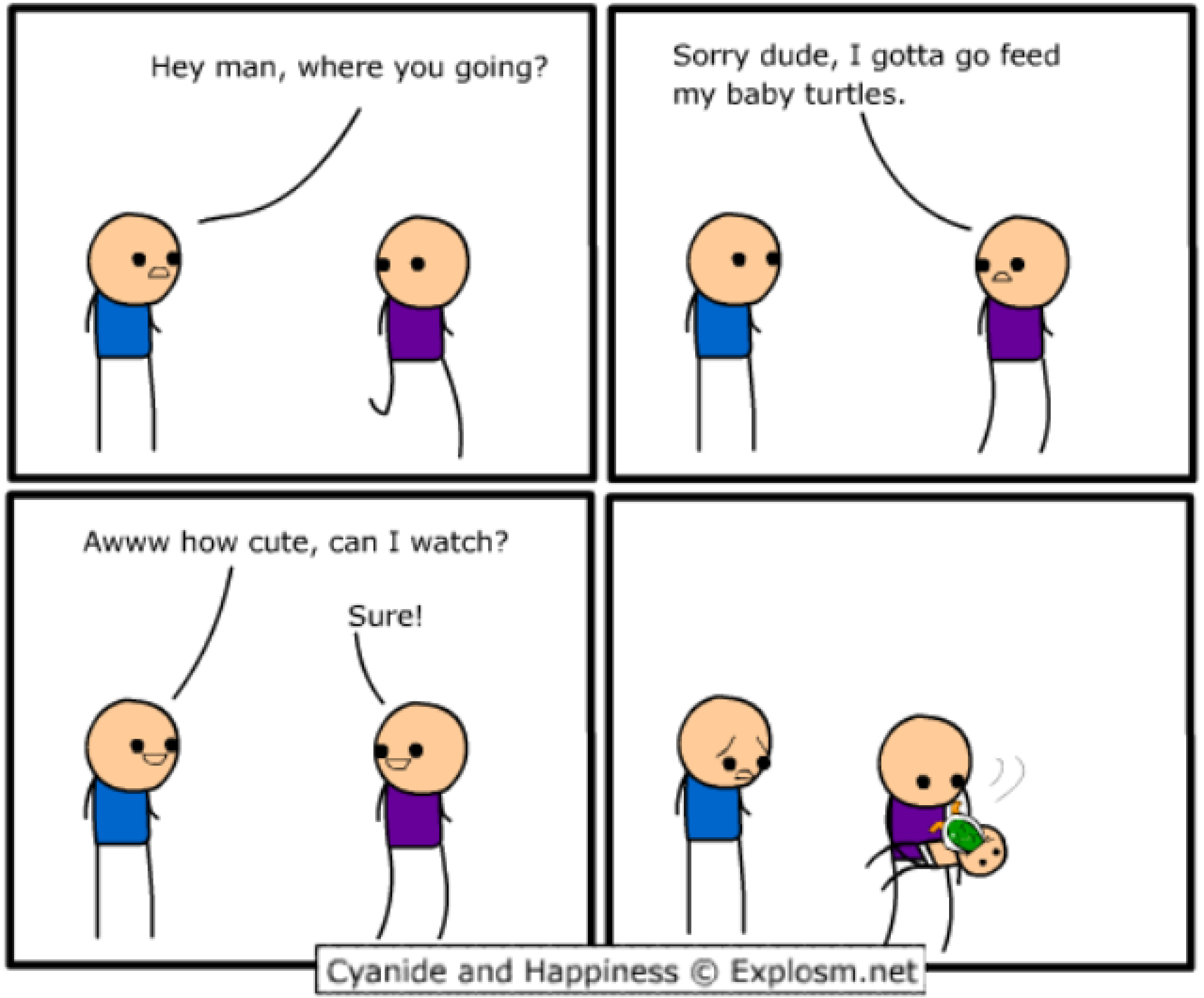
http://what-when-how.com/how-to-build-a-digital-library/word-segmentation-and-sorting-digital-library/



"One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know."

– *Groucho Marx*

http://onthesannyside.blogspot.com/2015/08/fun-with-dangling-modifiers.html


Hey man, where you going?
Sorry dude, I gotta go feed my baby turtles.
Awww how cute, can I watch?
Sure!
Cyanide and Happiness © Explosm.net

http://explosm.net/comics/1206

KIDS CAN BE SO CRUEL.
memaitic.net
WE CAN? THANKS MOM!
mematic.net

https://imgur.com/r/thesimpsons/1ZKxV5O

How does it work?

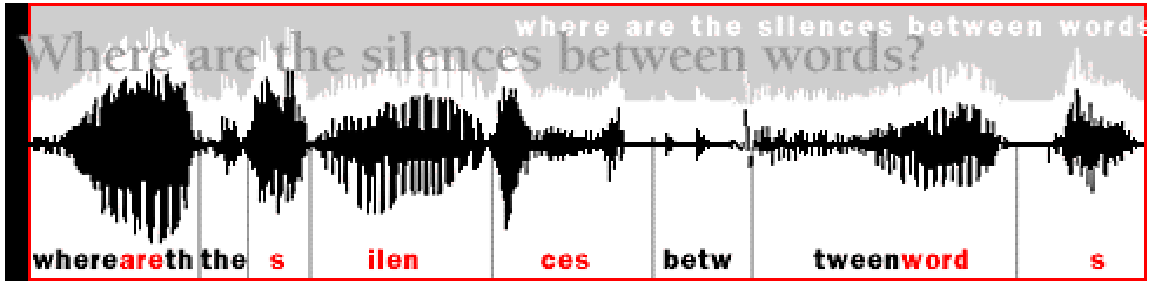
…and how can we make it work?

Hangman

- Why did you guess the letters in the order that you did?

Human Learning

- Statistical learning: learning how to understand and speak language by analyzing how others speak
  - Word frequencies, word order, recurring grammatical patterns
- Segmentation: 8-month-old infants react differently to pseudowords than non-words after 2 minutes of exposure to nonsense speech
- Examples:
  - “Powerful tea”
  - “Strong rain”
  - “



http://akenney.github.io/mind-cascade/language/

Natural Language Understanding

- Break speech down into letter / sound chunks
- Compare to recorded phonemes and pick likeliest combinations
  - Thank back to Hangman
  - Garden path sentences
- Get the important question words

Natural Language Generation

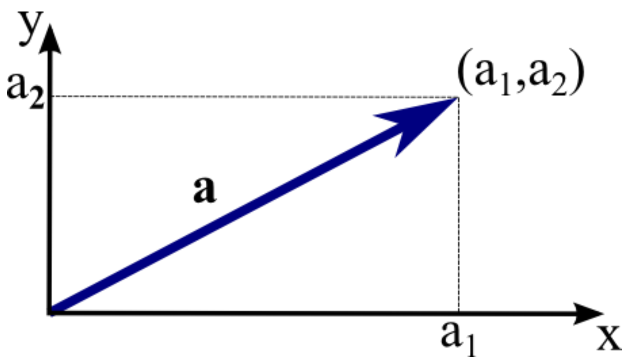
- Check the internet for response to question (or perform an action)
- Form sentence around answer information using grammatical rules
- Use speech to text!

Representing Words

- Computers can model human statistical learning
- “ “You shall know a word by the company it keeps.” -- John Rupert Firth
- Given one word, what others are most likely to occur around it?

Representing Words

- Represent words as vectors
  - Vectors are like coordinates in higher dimensions
  - Like an arrow in multidimensional space
- These vectors are created by attempting to predict the contexts in which a given word is likely to appear
- The spatial relationships between word vectors often contain meaning about words!



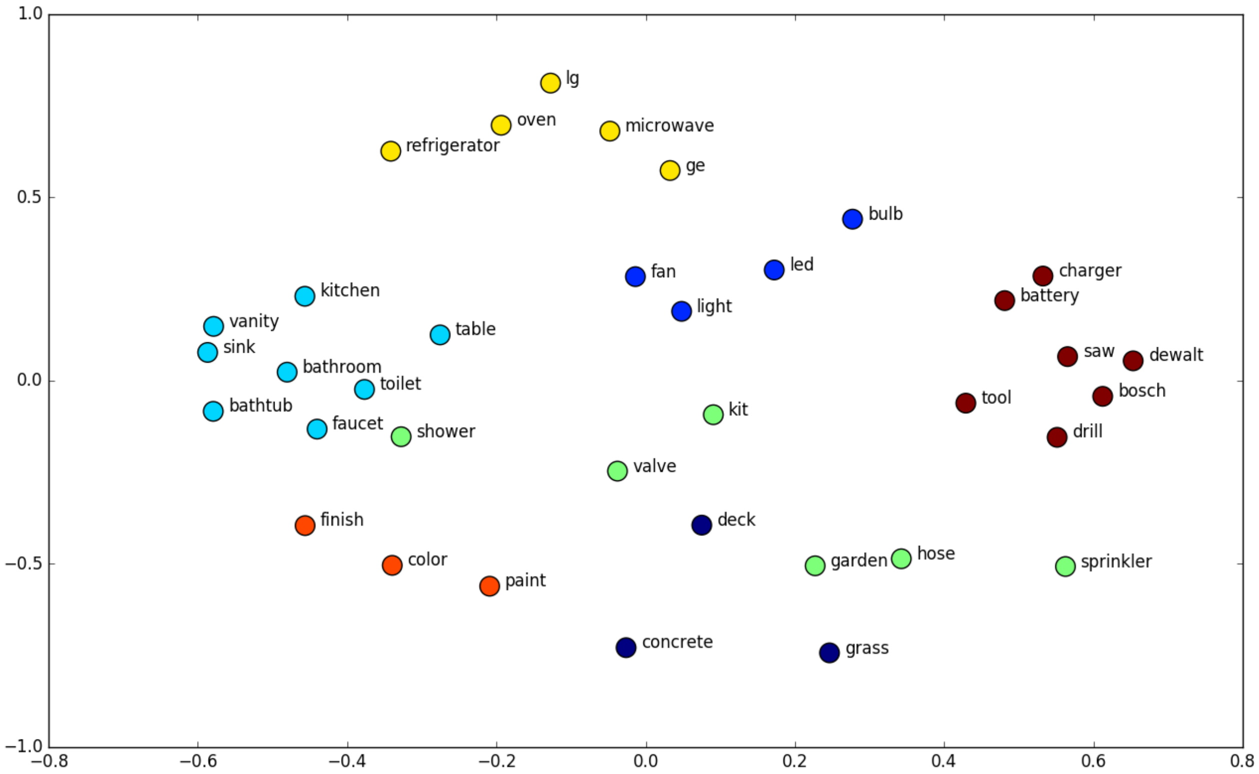
W("pair")=(0.2,−0.4,0.7,...)
W("pear")=(0.0,−0.1,0.1,...)

https://mathinsight.org/vectors_cartesian_coordinates_2d_3d

http://sanjaymeena.io/tech/word-embeddings/

Word Patterns

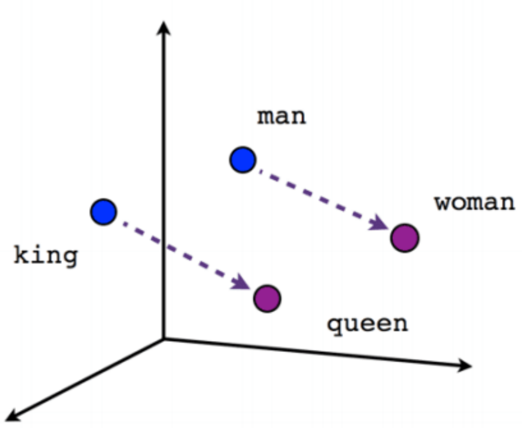
- The computer doesn't *know* anything about the meaning of the words at all, but some interesting patterns emerge…
- Words that have similar meanings are usually very close in space



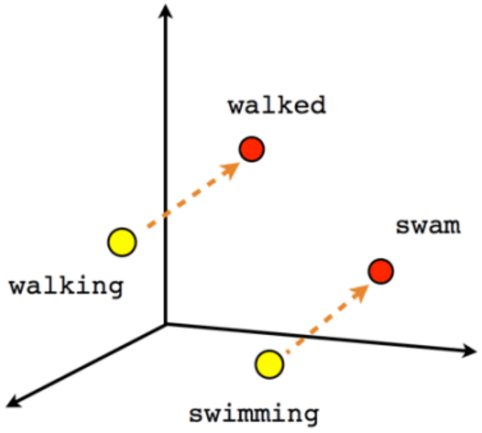
http://suriyadeepan.github.io/2016-06-28-easy-seq2seq/

Word Patterns

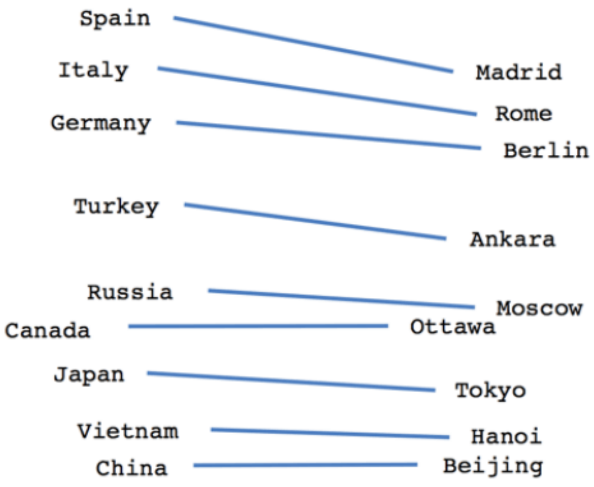
- Words with parallel relationships have matching spatial relationships



Male-Female



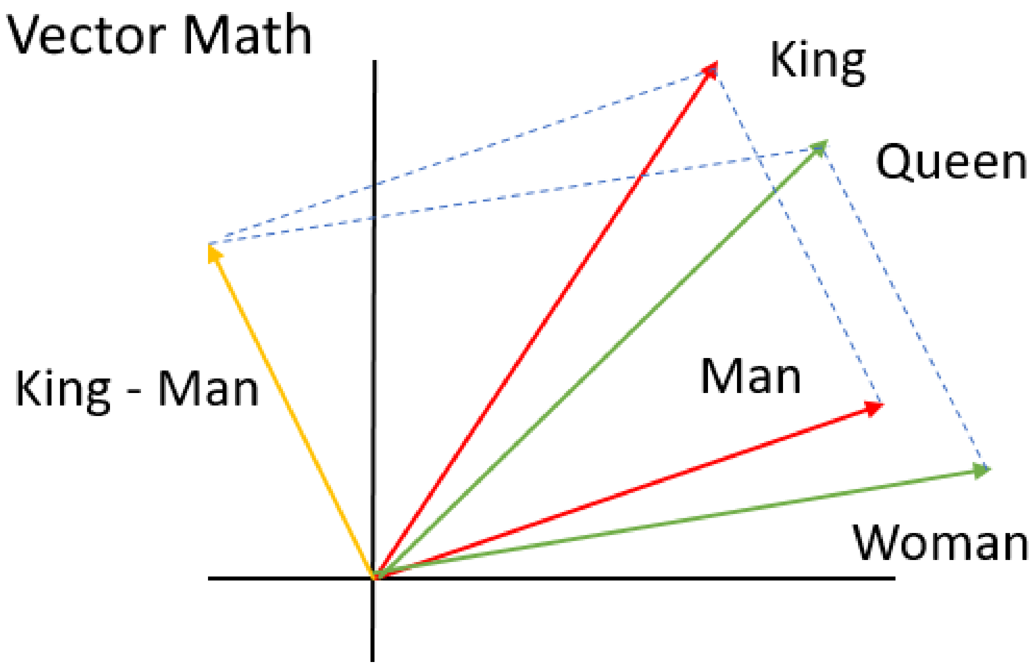
Verb tense



Country-Capital

http://sanjaymeena.io/tech/word-embeddings/

Cool Pattern: Vector Math

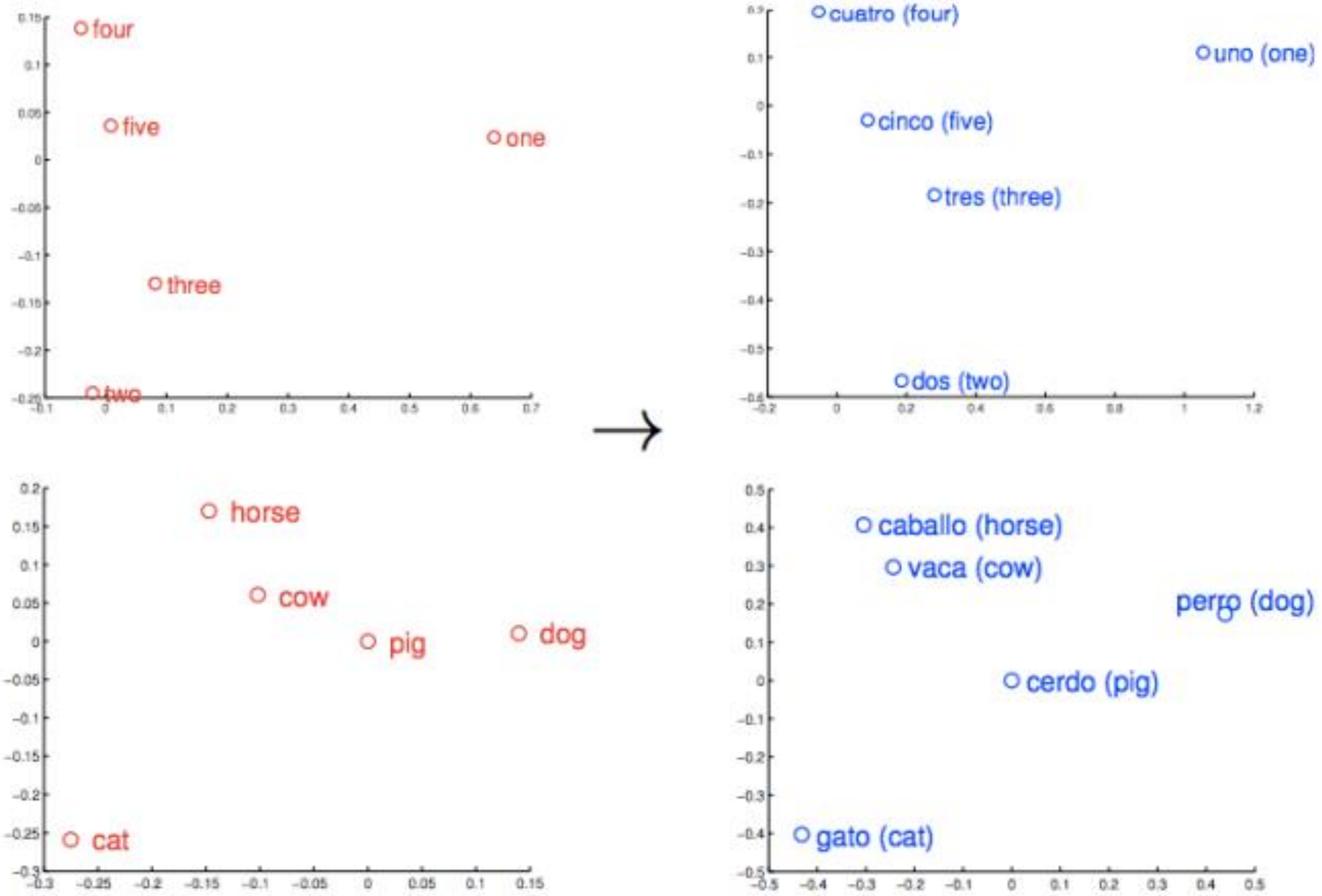


- Analogies
  - King is to Man as Queen is to….
- (King – Man) = (Queen – Woman)
- Just based on solving for closest vector (minimum distance), we can get 60% accuracy at solving analogies
- Machine learning is NOT PERFECT, there is always some degree of error

https://blogs.mathworks.com/loren/2017/09/21/math-with-words-word-embeddings-with-matlab-and-text-analytics-toolbox/

Cool Pattern: Translation Similarity

- Same-meaning words tend to have very similar vectors across languages
- Comparing these “language spaces” helps us perform translation



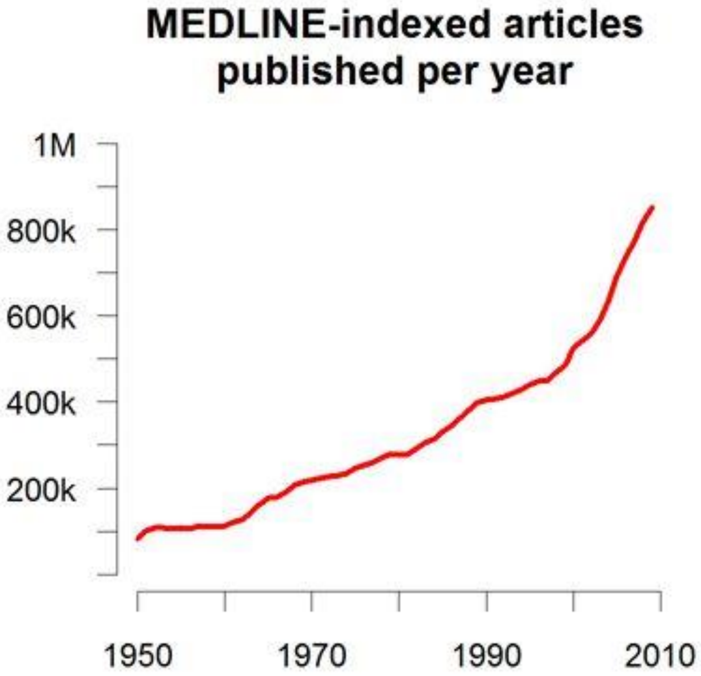
https://deeplearning4j.org/word2vec.html

Where do we use it?

…and how do we use it?

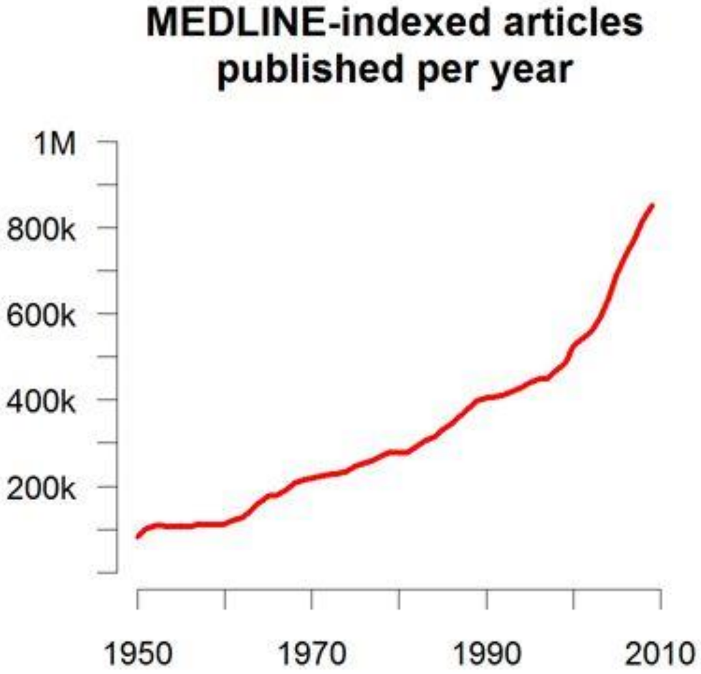
Medicine

- More complete and accurate patient records – computer assisted coding



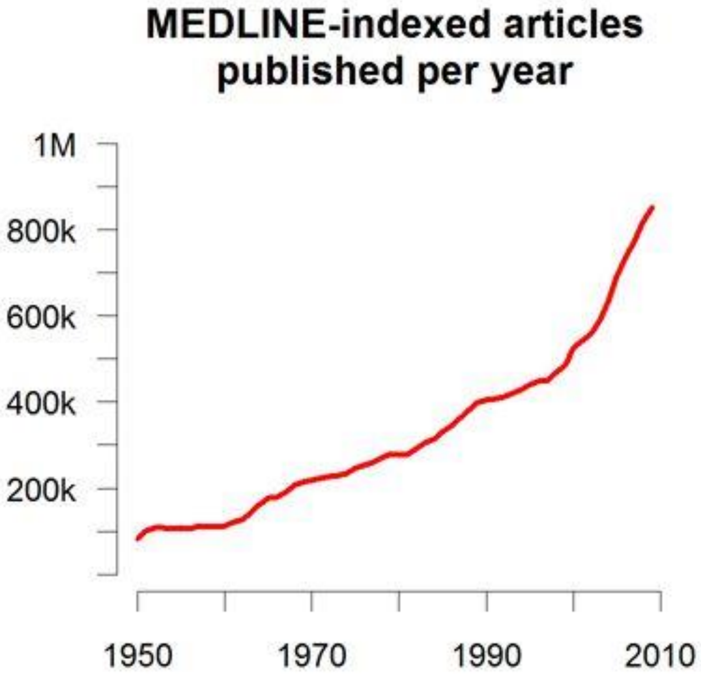
Medicine

- More complete and accurate patient records – computer assisted coding
- Efficiency gains: doctors can dictate their notes



Medicine

- More complete and accurate patient records – computer assisted coding
- Efficiency gains: doctors can dictate their notes
- Clinical decision support: IBM Watson can learn from entirety of Medline literature; impossible for doctors to read even a fraction of this



Censorship in China

Figure 4. Events with Highest and Lowest Censorship Magnitude

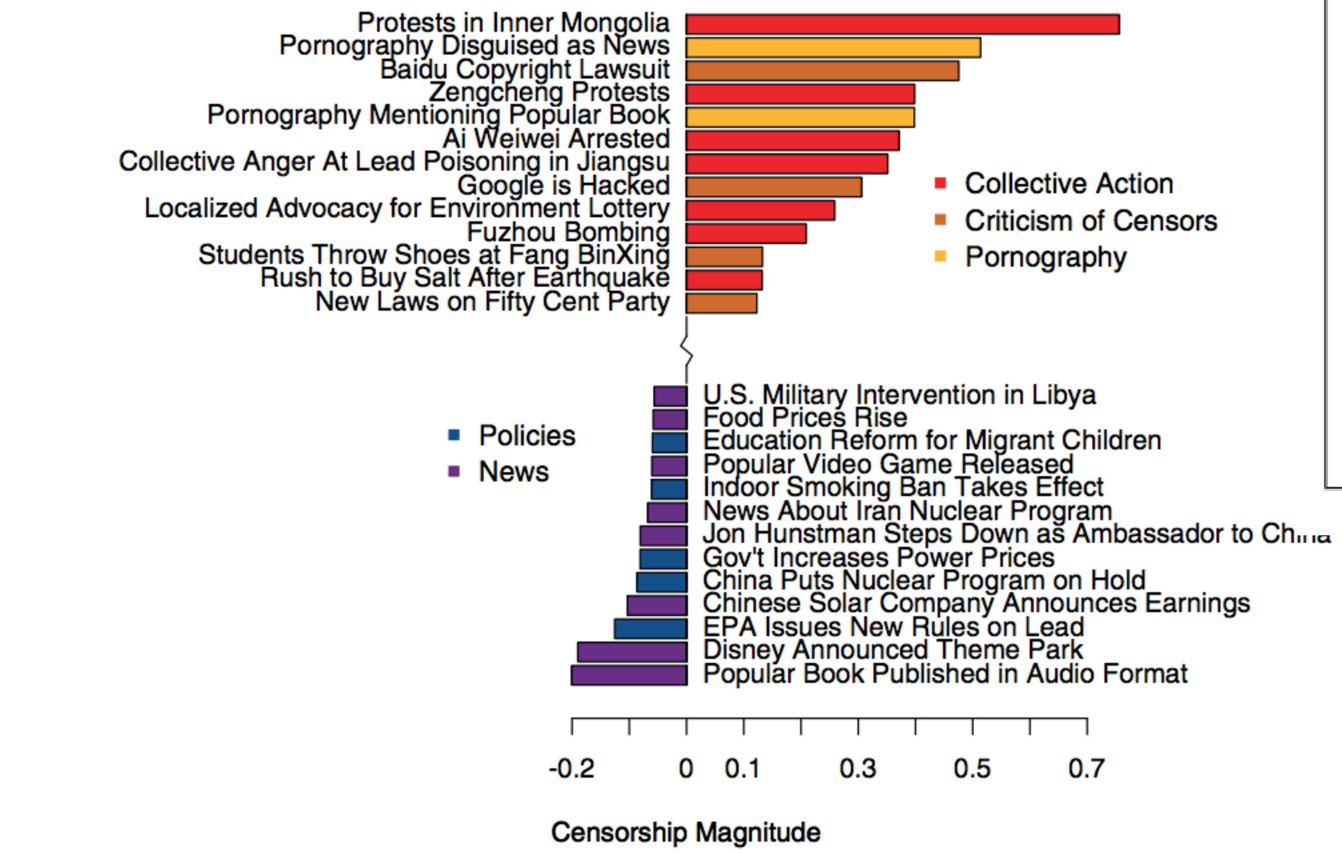
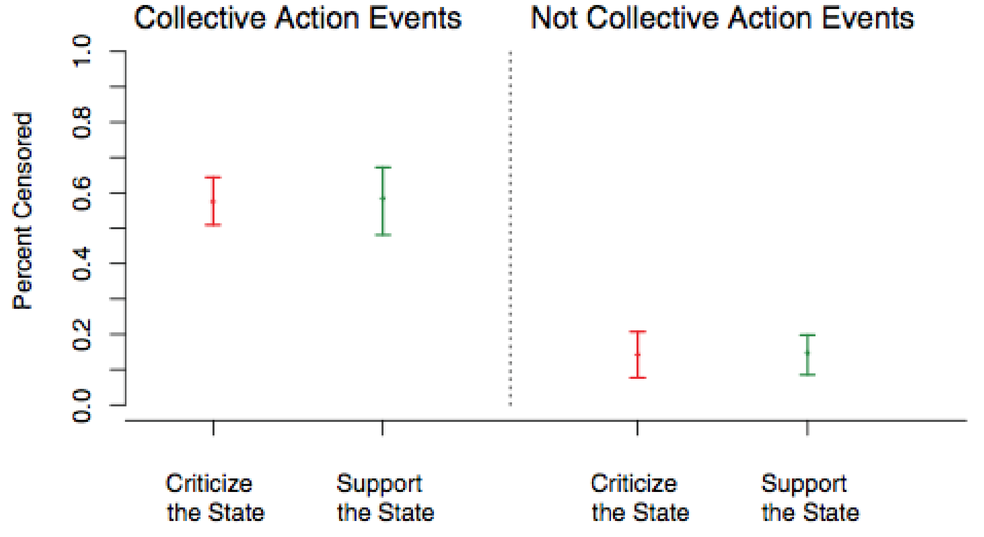
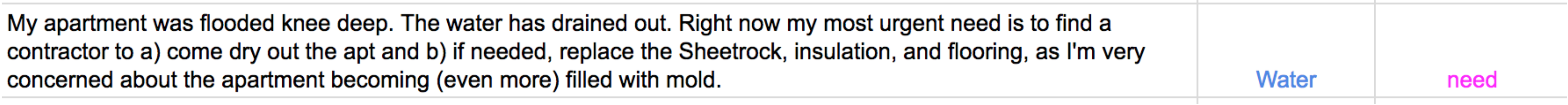


Figure 9. Content of All Censored Posts (Regardless of Topic Area)

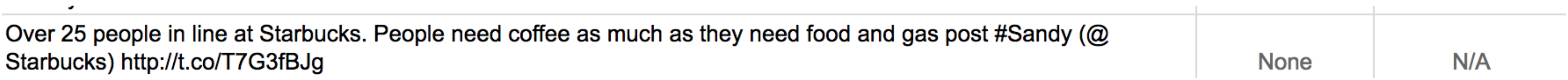


King et al. 2013

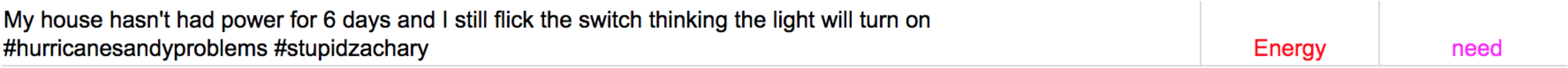
Disaster Relief and Resource Allocation - Twitter



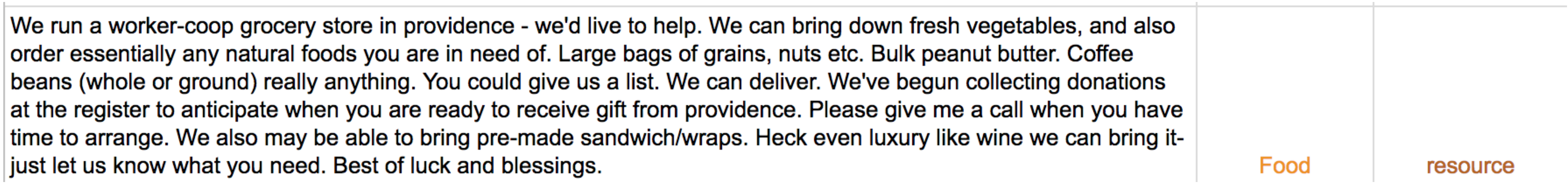
My apartment was flooded knee deep. The water has drained out. Right now my most urgent need is to find a contractor to a) come dry out the apt and b) if needed, replace the Sheetrock, insulation, and flooring, as I'm very concerned about the apartment becoming (even more) filled with mold.	Water	need



Over 25 people in line at Starbucks. People need coffee as much as they need food and gas post #Sandy (@ Starbucks) http://t.co/T7G3fBJg	None	N/A



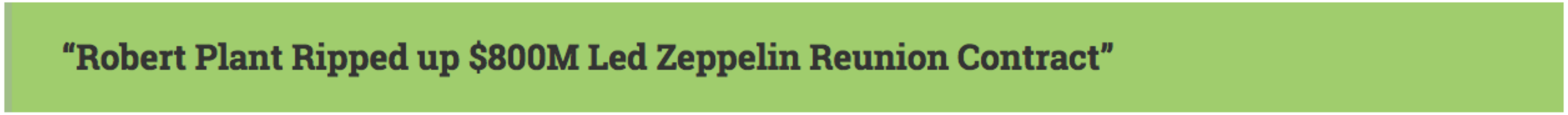
My house hasn't had power for 6 days and I still flick the switch thinking the light will turn on #hurricanesandyproblems #stupidzachary	Energy	need



We run a worker-coop grocery store in providence - we'd live to help. We can bring down fresh vegetables, and also order essentially any natural foods you are in need of. Large bags of grains, nuts etc. Bulk peanut butter. Coffee beans (whole or ground) really anything. You could give us a list. We can deliver. We've begun collecting donations at the register to anticipate when you are ready to receive gift from providence. Please give me a call when you have time to arrange. We also may be able to bring pre-made sandwich/wraps. Heck even luxury like wine we can bring it-just let us know what you need. Best of luck and blessings.	Food	resource

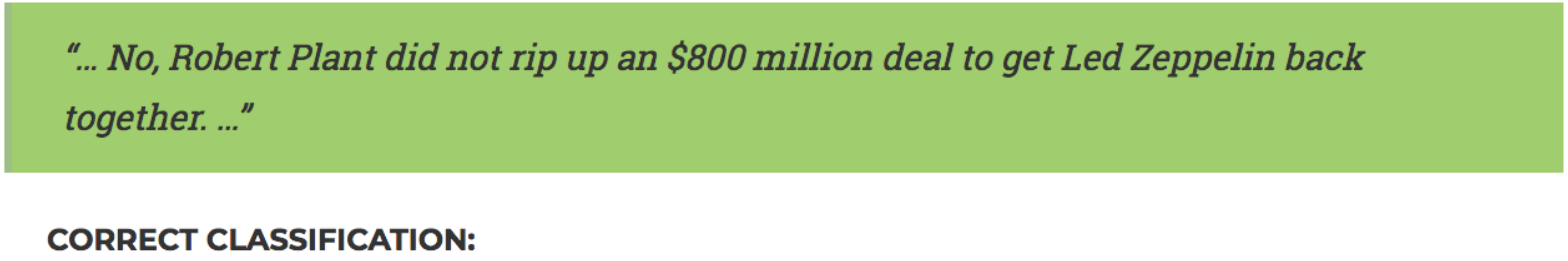
Fake News Challenge

Headline:



“Robert Plant Ripped up $800M Led Zeppelin Reunion Contract”

Claims:



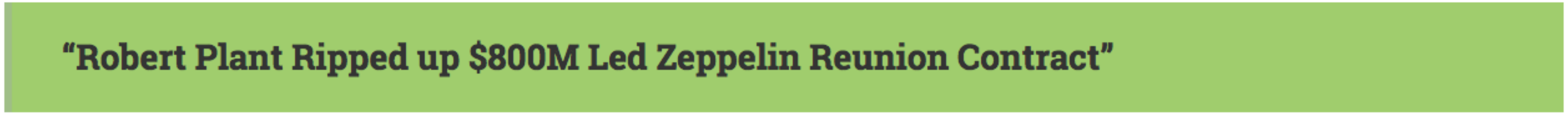
“… No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together. …”

CORRECT CLASSIFICATION:

Goal: train computer to recognize claims associated with headline as IN AGREEMENT, IN DISAGREEMENT, RELATED or UNRELATED

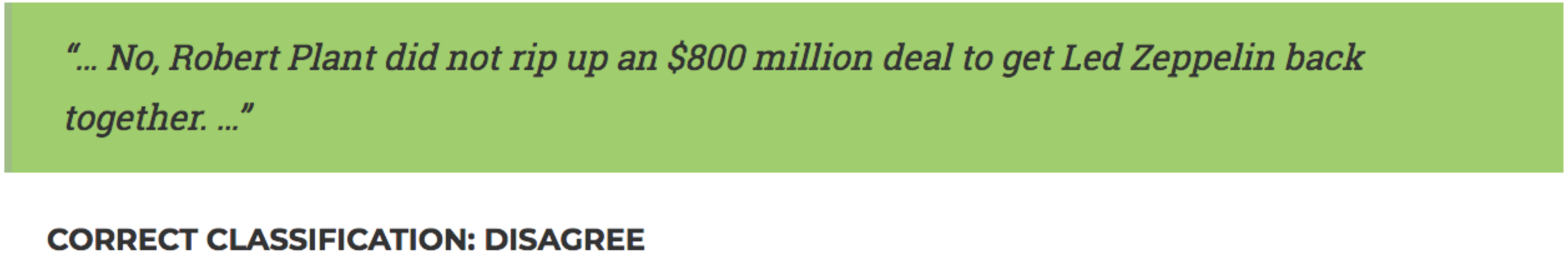
Fake News Challenge

Headline:



“Robert Plant Ripped up $800M Led Zeppelin Reunion Contract”

Claims:



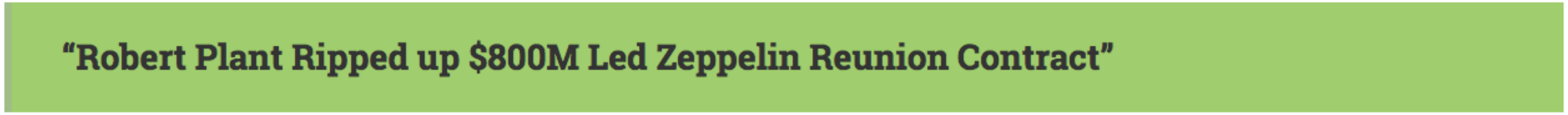
“... No, Robert Plant did not rip up an $800 million deal to get Led Zeppelin back together. ...”

CORRECT CLASSIFICATION: DISAGREE

Goal: train computer to recognize claims associated with headline as IN AGREEMENT, IN DISAGREEMENT, RELATED or UNRELATED

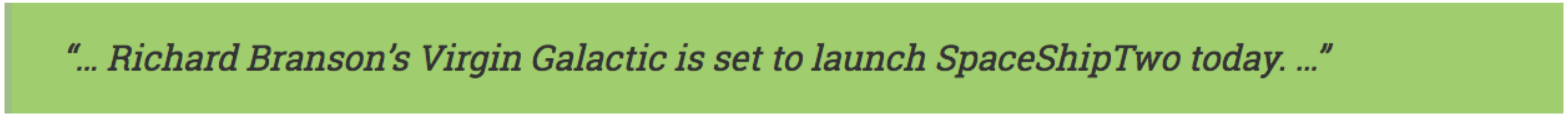
Fake News Challenge

Headline:



“Robert Plant Ripped up $800M Led Zeppelin Reunion Contract”

Claims:



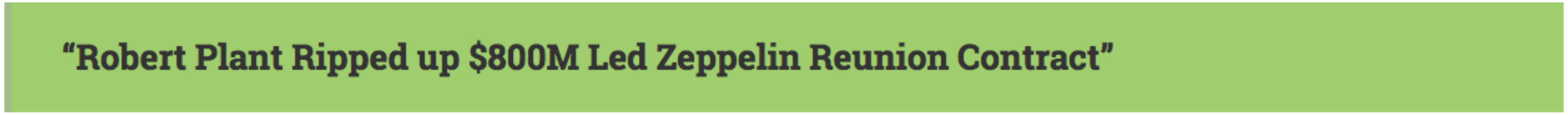
“... Richard Branson’s Virgin Galactic is set to launch SpaceShipTwo today. ...”

CORRECT CLASSIFICATION:

Goal: train computer to recognize claims associated with headline as IN AGREEMENT, IN DISAGREEMENT, RELATED or UNRELATED

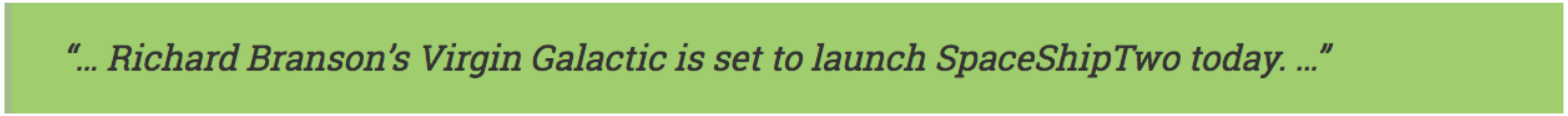
Fake News Challenge

Headline:

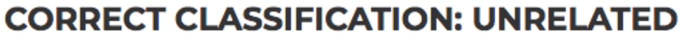


“Robert Plant Ripped up $800M Led Zeppelin Reunion Contract”

Claims:



“… Richard Branson’s Virgin Galactic is set to launch SpaceShipTwo today. …”

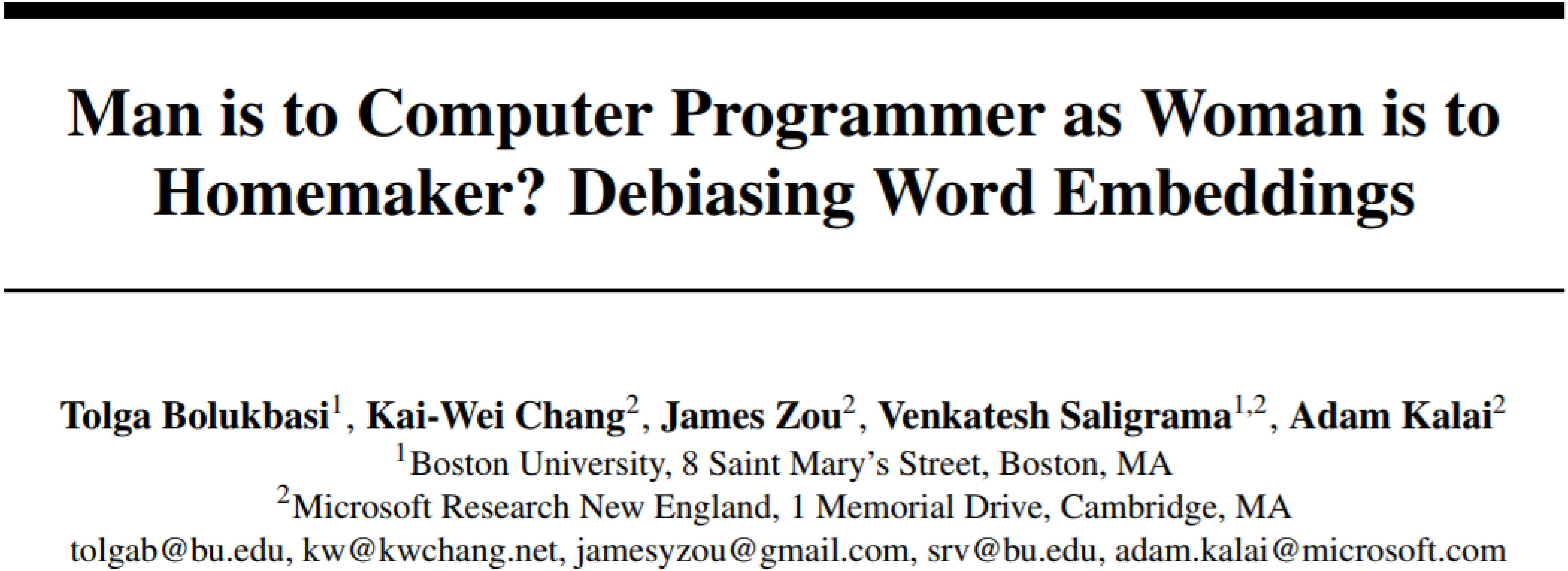


CORRECT CLASSIFICATION: UNRELATED

Goal: train computer to recognize claims associated with headline as IN AGREEMENT, IN DISAGREEMENT, RELATED or UNRELATED

What else could we do?

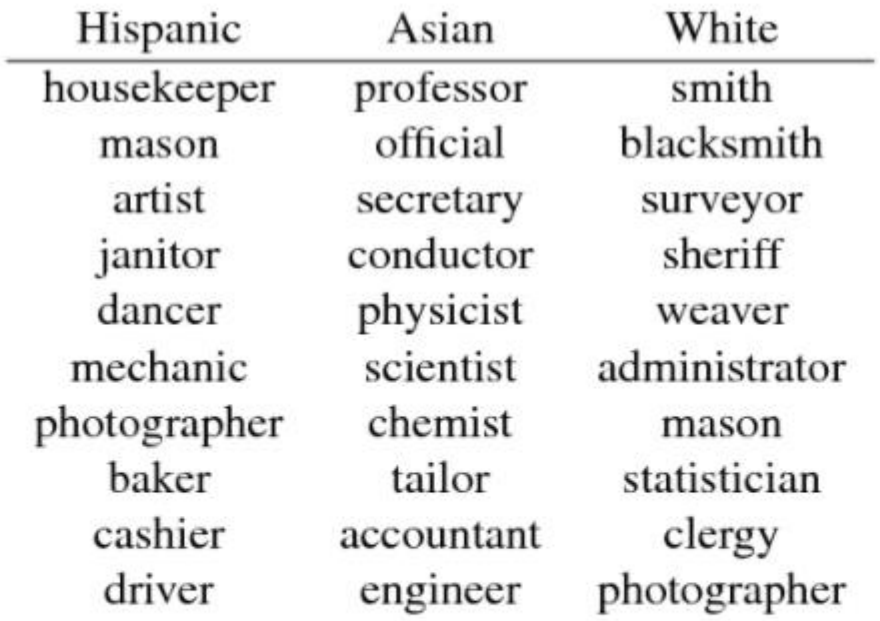
…and what should we do?



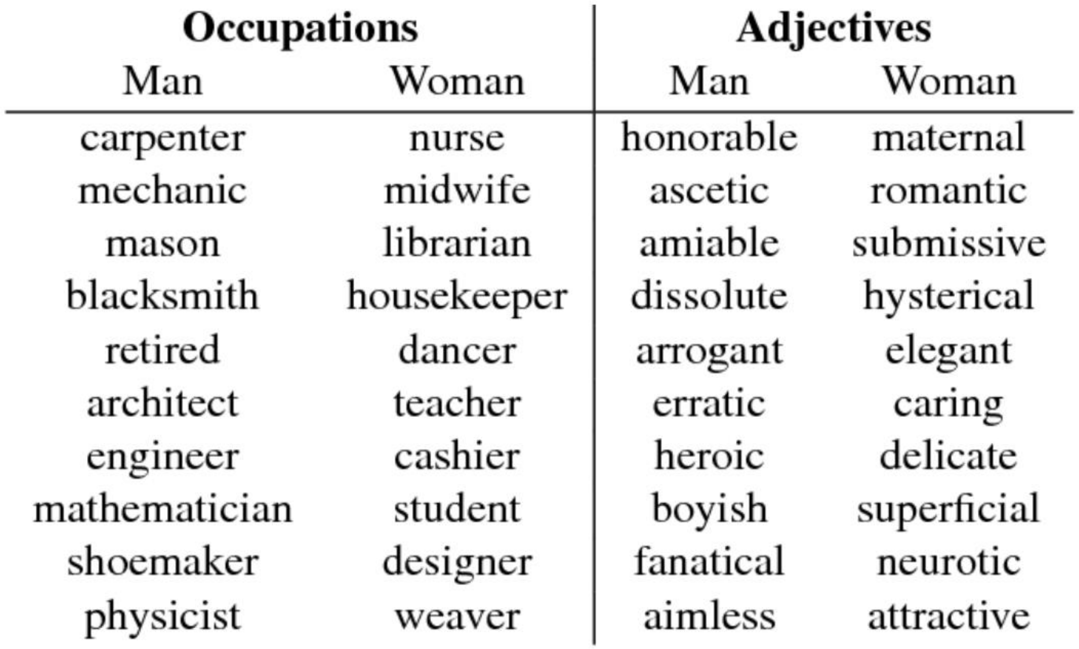
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf



Hispanic	Asian	White
housekeeper	professor	smith
mason	official	blacksmith
artist	secretary	surveyor
janitor	conductor	sheriff
dancer	physicist	weaver
mechanic	scientist	administrator
photographer	chemist	mason
baker	tailor	statistician
cashier	accountant	clergy
driver	engineer	photographer



Occupations		**Adjectives**	
Man	Woman	Man	Woman
carpenter	nurse	honorable	maternal
mechanic	midwife	ascetic	romantic
mason	librarian	amiable	submissive
blacksmith	housekeeper	dissolute	hysterical
retired	dancer	arrogant	elegant
architect	teacher	erratic	caring
engineer	cashier	heroic	delicate
mathematician	student	boyish	superficial
shoemaker	designer	fanatical	neurotic
physicist	weaver	aimless	attractive

https://www.semanticscholar.org/paper/Word-Embeddings-Quantify-100-Years-of-Gender-and-Garg-Schiebinger/56e8985fa61037662650803d48177d752e9ce557

What can we do in the future?

Questions?

…and answers?