How does text become data?

Rob Speer
Luminoso
rob@luminoso.com

In this talk

- A whirlwind tour of data-driven NLP techniques
- Copious Python examples
- Don’t worry, code is online:

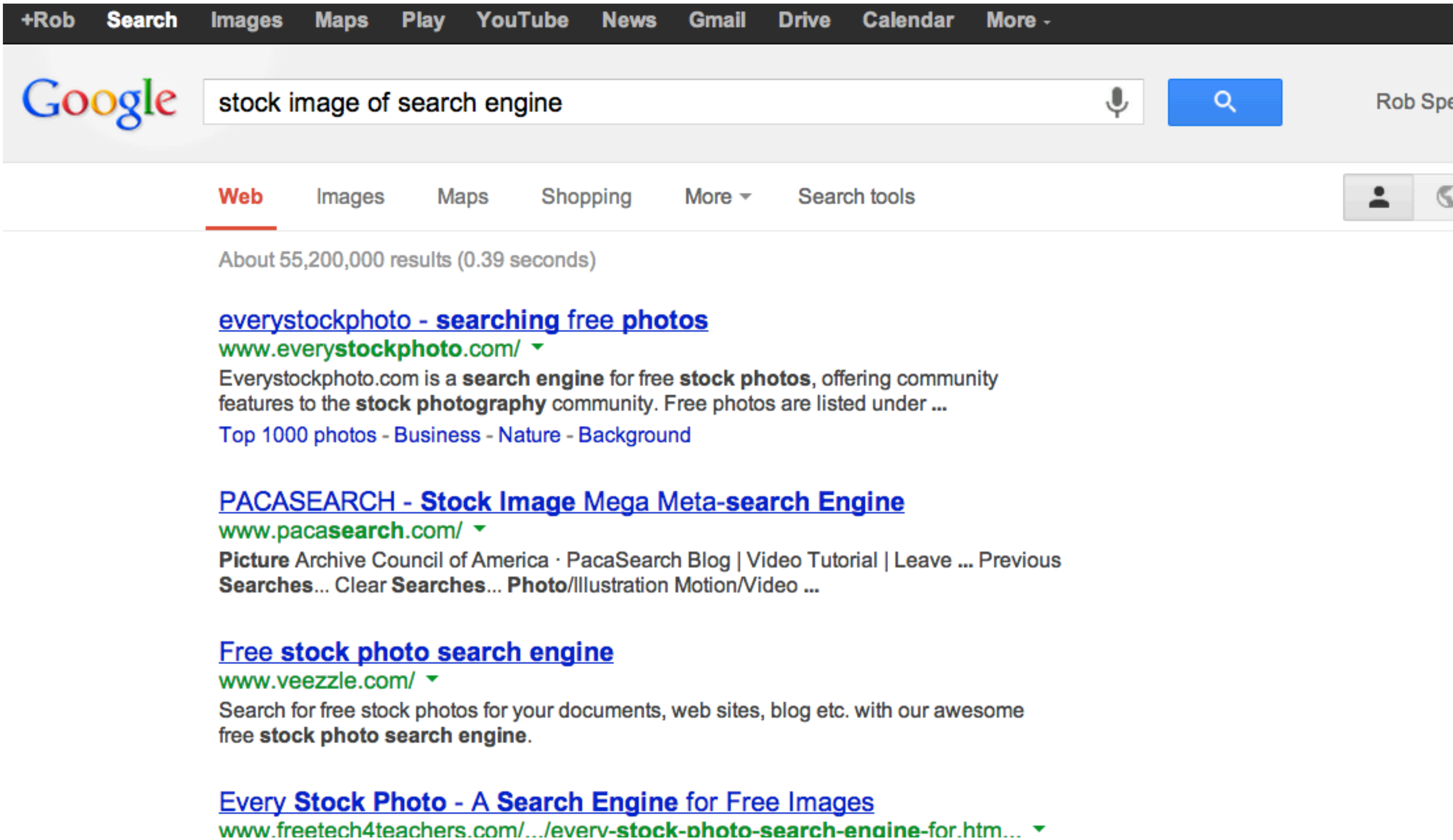
  http://github.com/rspeer/text-as-data

  – This is not what I do at my startup. This is about problems that can be solved in about 20 lines of Python code.

r bæði í lan
nilega ekki upp við velgeng
ekki. Margrét stundar nám við V
u af frítíma sínum í fótboltaæfin
ð sé svona heillandi við fótbolta
er finnst svo gaman að spila fótb
ð fer auðvitað mikill tími í æfin
vini mína. . . ,' Margrét sér fr
um. Hana langar að fara til
ðurlöndin eru

What can you do with text?

Search it

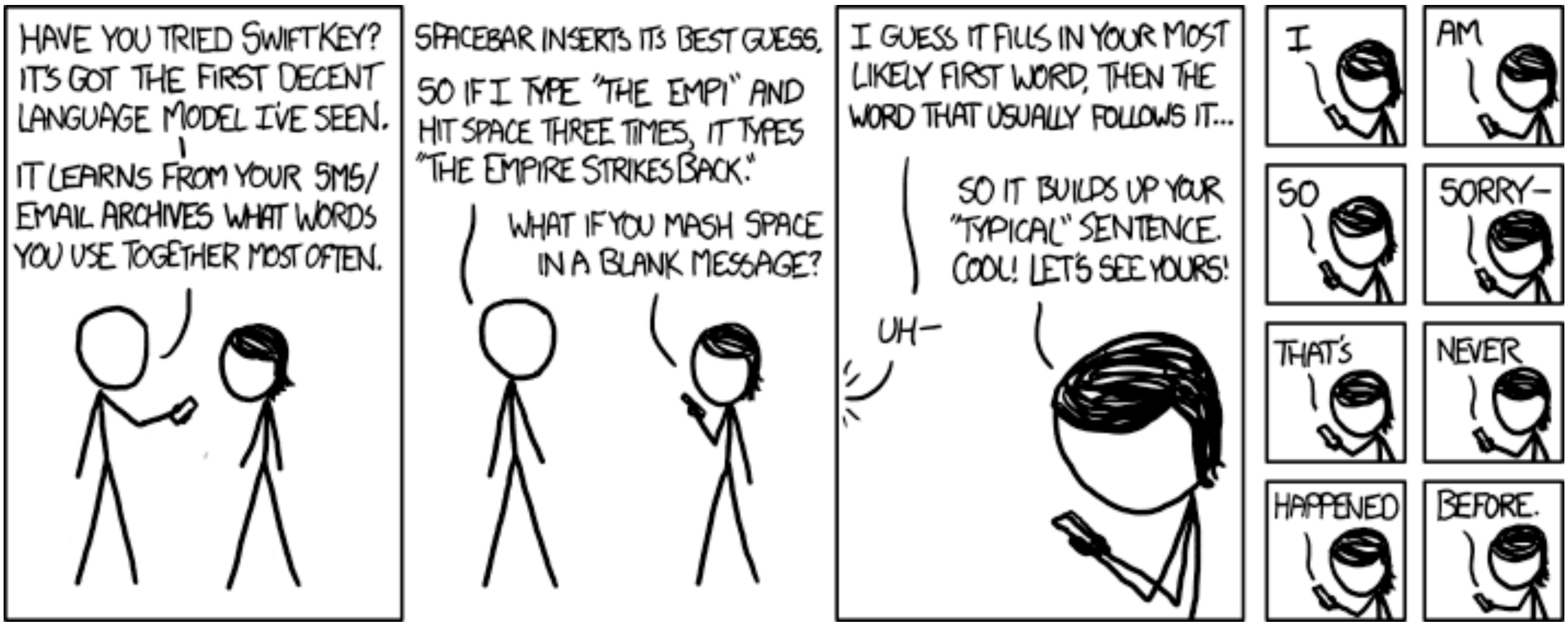

+Rob Search Images Maps Play YouTube News Gmail Drive Calendar More
Google
stock image of search engine
Rob Spe
Web Images Maps Shopping More Search tools
About 55,200,000 results (0.39 seconds)
everystockphoto - searching free photos
www.everystockphoto.com/
Everystockphoto.com is a search engine for free stock photos, offering community features to the stock photography community. Free photos are listed under ...
Top 1000 photos - Business - Nature - Background
PACASEARCH - Stock Image Mega Meta-search Engine
www.pacasearch.com/
Picture Archive Council of America · PacaSearch Blog | Video Tutorial | Leave ... Previous Searches... Clear Searches... Photo/Illustration Motion/Video ...
Free stock photo search engine
www.veezzle.com/
Search for free stock photos for your documents, web sites, blog etc. with our awesome free stock photo search engine.
Every Stock Photo - A Search Engine for Free Images
www.freetech4teachers.com/.../every-stock-photo-search-engine-for.htm...

Classify it

Delete all spam messages now (messages that have been in Spam more than 30 days will be automatically deleted)

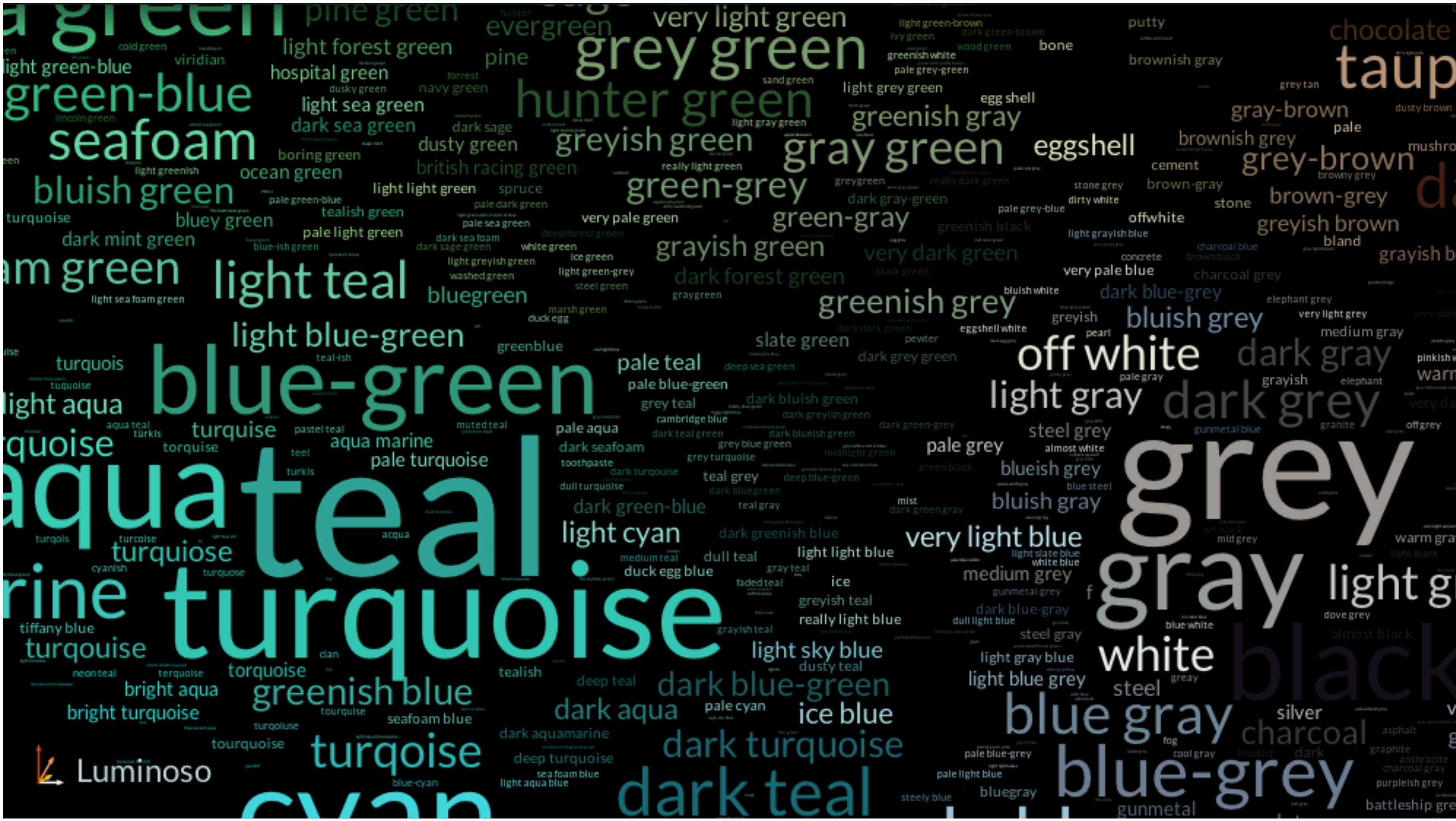
Regal Ecigs	**Trial - Smoke Almost Anywhere!** - Trial - Smoke Almost Anywhere!
Loan Department	**All Credit OK - Up to 1000 dollars** - All Credit OK - Up to 1000 dollars
Next Payday Advance	**Need emergency cash? Let us help you.** - Get the funds you need in 1 hour.
Kohls Summer Savings Gif.	**Are you tired of all the cold weather and ready for Summer? Complete ou**
da_30	**FEEM2013——Submissions due: July 30th,2013** - 2013 International Confe
Harp Mortgage	**President Announced the HARP Program. Save Thousands a Year on Yo**
Jacuzzi Walk In Hot Tubs	**Soak away life's aches and pains with a walk in hot tub** - Soak away life's a
Fingerhut Friends	**Fingerhut: Best Gifts, Low Payments! Open Your Account Today!*** - Finge
The LASIK Vision Institu.	**Looking for more freedom from your glasses? Get Lasik info** - Looking for
KaplanUniversity	**KaplanUniversity online & campus degree programs available** - KaplanUr
Vistaprint Offers	**Buy 250 Premium Business Cards get 250 More Free from Vistaprint!** - Bu
Lifestyle Lift	**Look more beautiful with alot less risk** - Look more beautiful with alot less ri
MetLife Partner	**Get $250k in Term Life Insurance - as low as $16/mo.** - Get $250k in Term L
NextPaydayAdvance	**You Could Have 1,500 Cash Wired Into Your Account - Apply Now** - You C

Predict it



"SwiftKey" from xkcd
http://xkcd.com/1068/

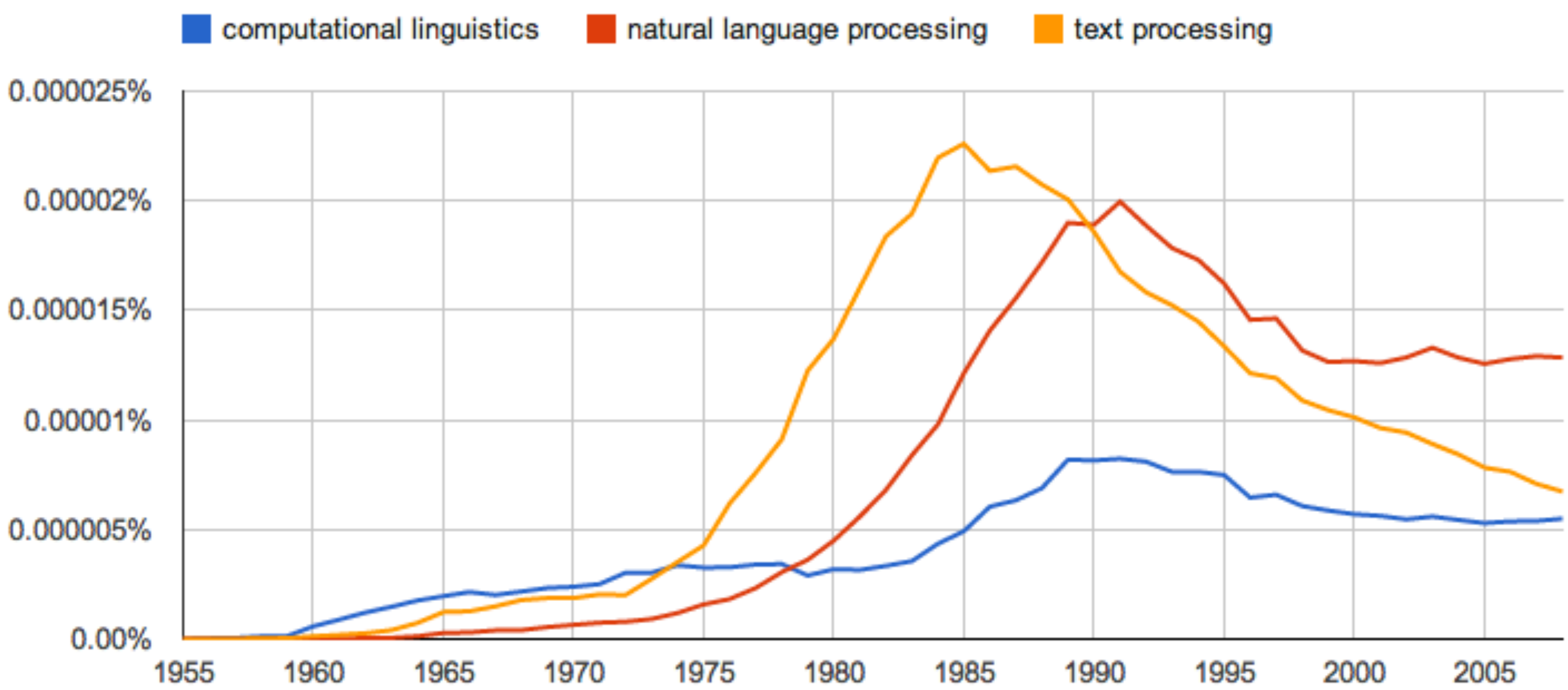
Visualize and explore it



How is the text represented?

Simple word counts

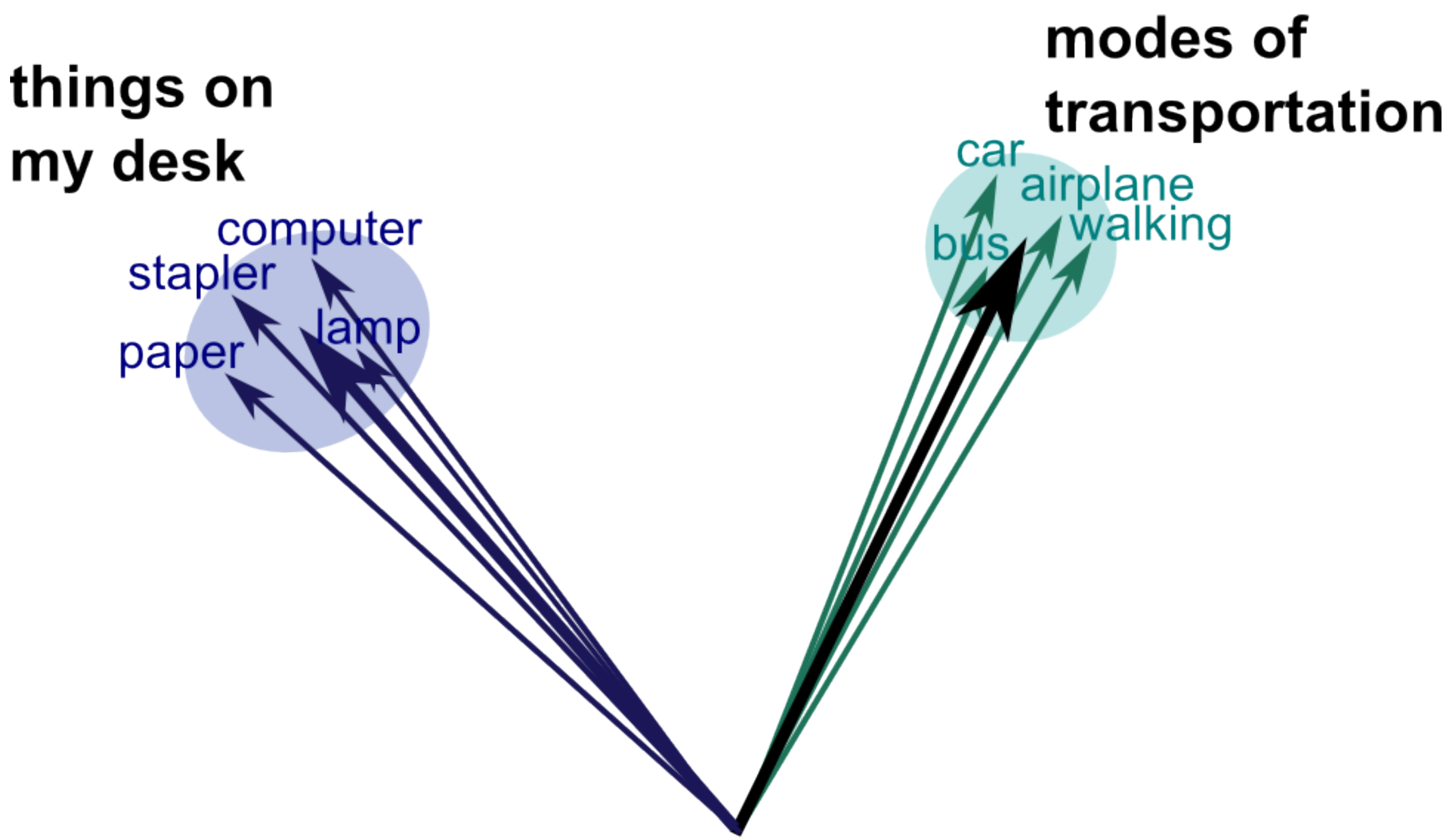
N-gram models



Term-document matrices

	woe	betray	vengeance	death	alas
Julius Caesar	2	1	0	29	8
Hamlet	8	0	2	37	9
Macbeth	2	2	0	20	4

Vector space models



Python example: word splitting and normalizing

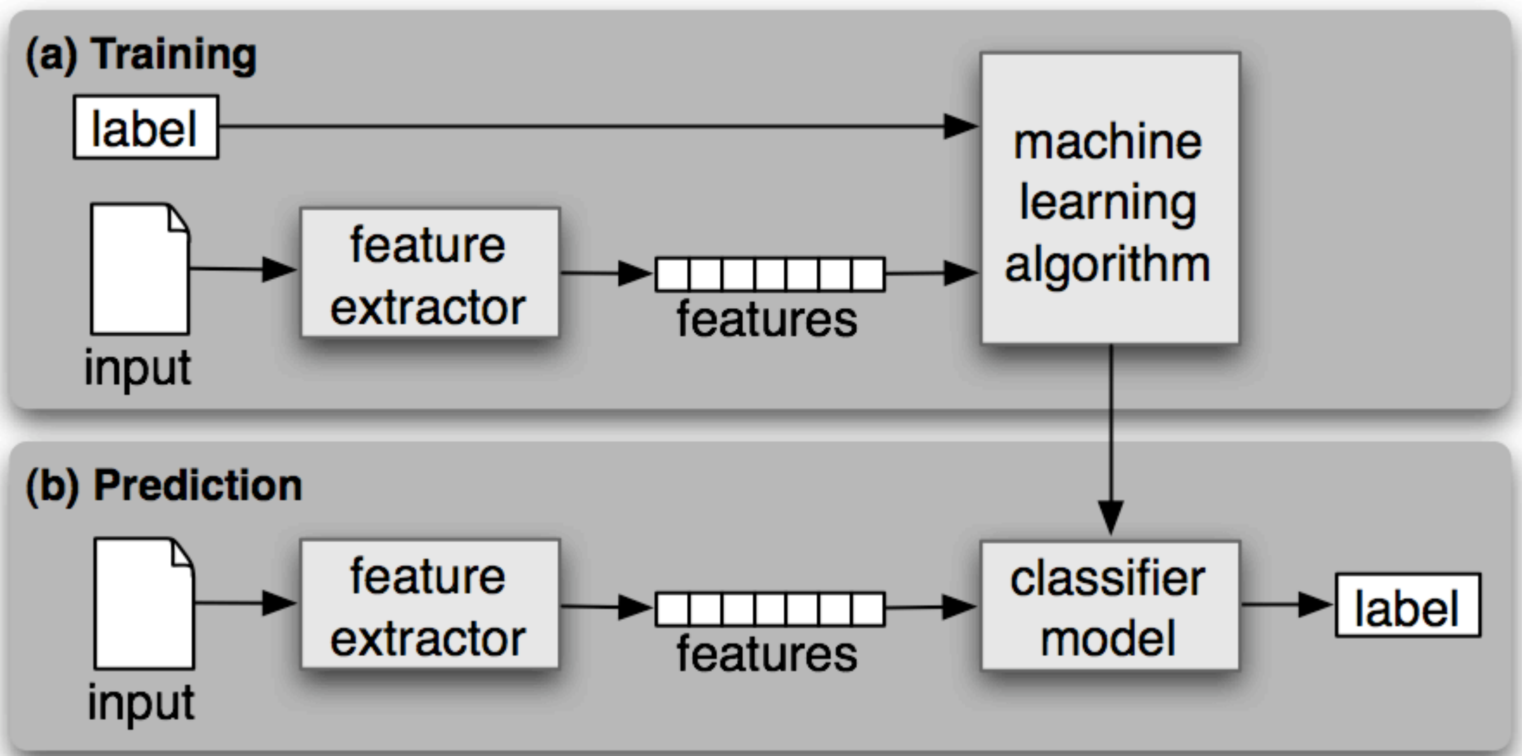
Which N-grams are interesting?

Consider this contingency table:

p(**vice**, **president**)	p(**vice**, ~president)
p(~vice, **president**)	p(~vice, ~president)

Python example: interesting N-grams

Text classification

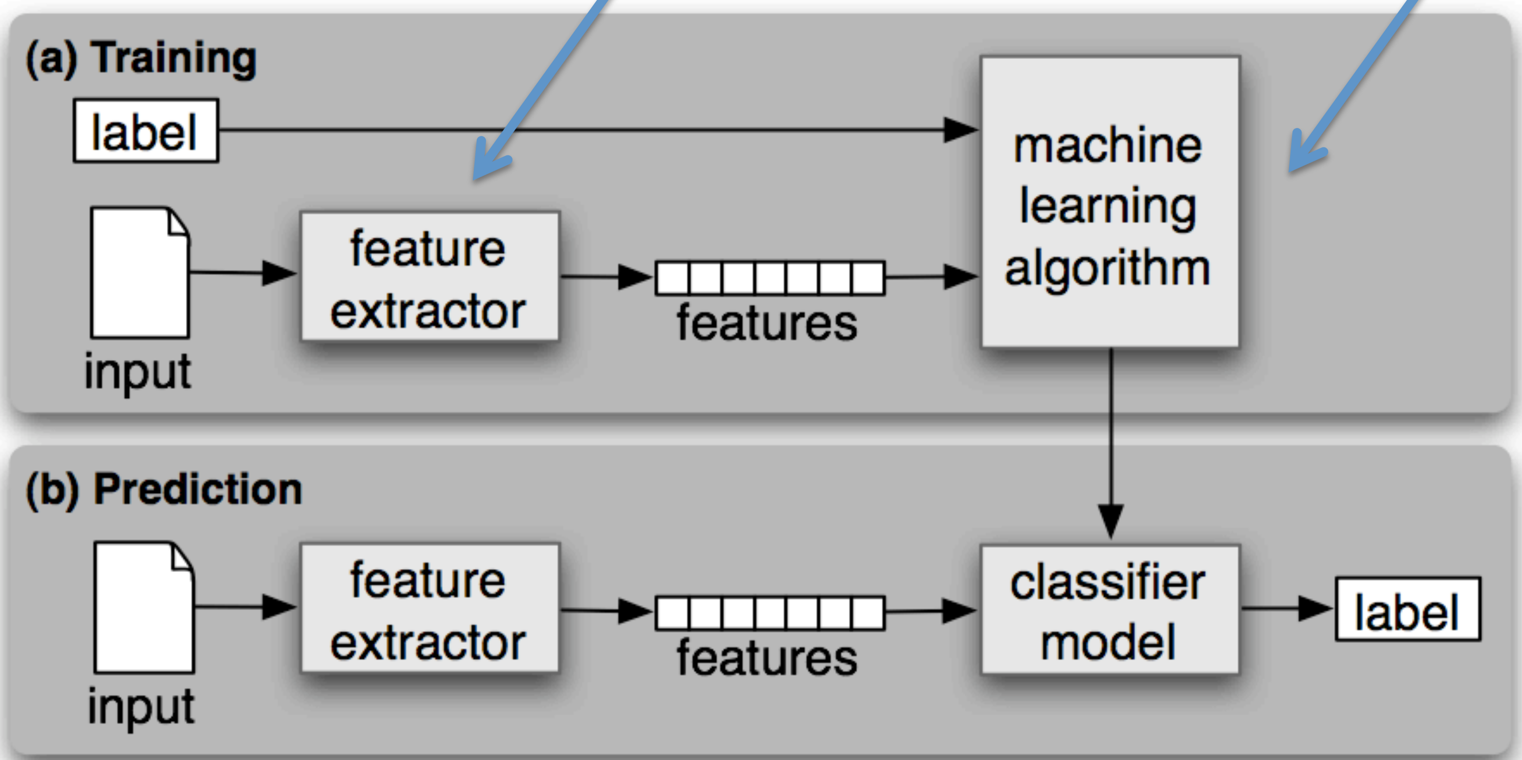


from “Natural Language Processing with Python”,
by Steven Bird, Ewan Klein, and Edward Loper (O’Reilly, 2009)

Text classification

almost certainly
bag-of-words

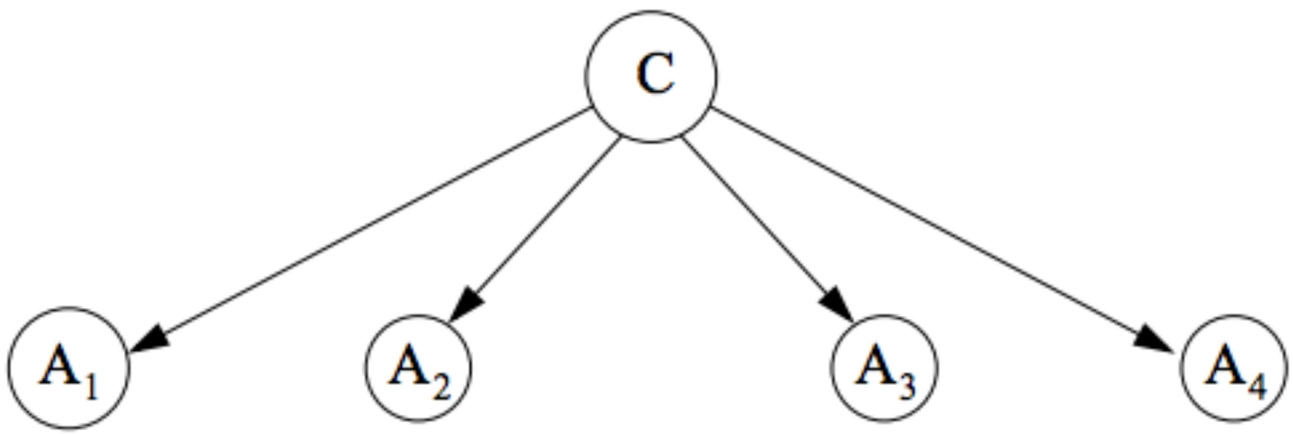
possibly
Naive Bayes



from "Natural Language Processing with Python",
by Steven Bird, Ewan Klein, and Edward Loper (O'Reilly, 2009)

Overview of Naïve Bayes classification

- The probability that a document is in class C depends on its features, A_n
- Assume all features are statistically independent



Python example: Classification with NLTK and scikit-learn

What about stopwords?

- Shouldn't we remove common words such as "the" and "of"?
- It could help
- It could be premature optimization

Text similarity

- Bags of words can tell us how similar documents are

	woe	betray	vengeance	death	alas
Julius Caesar	2	1	0	29	8
Hamlet	8	0	2	37	9
Macbeth	2	2	0	20	4

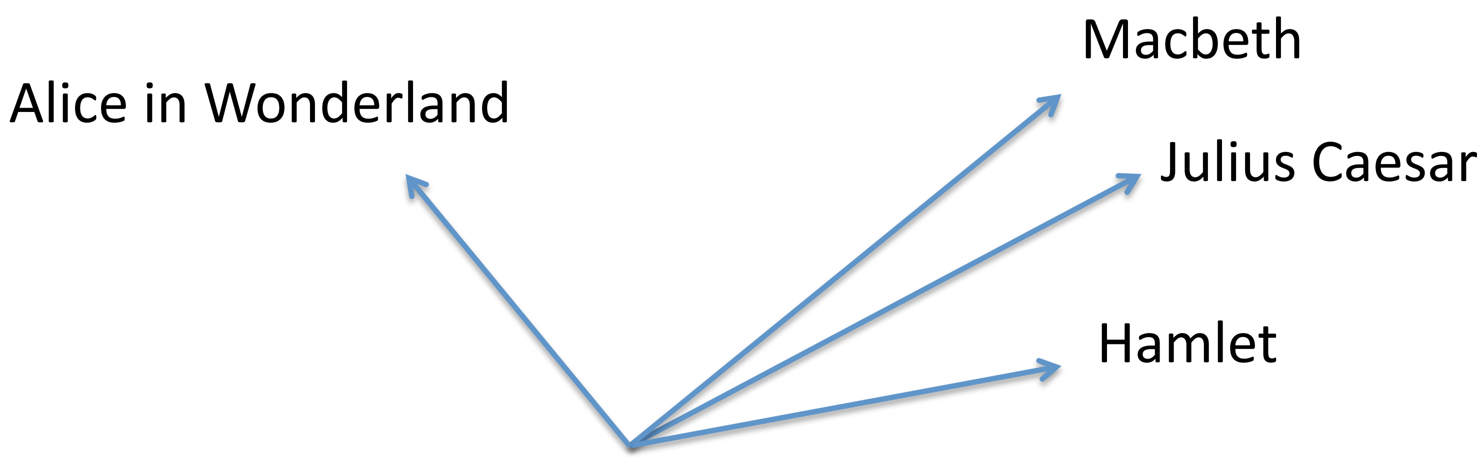
Text similarity

- Bags of words can tell us how similar documents are

	woe	betray	vengeance	death	alas
Julius Caesar	2	1	0	29	8
Hamlet	8	0	2	37	9
Macbeth	2	2	0	20	4
Alice in Wonderland	0	0	0	1	4

Vector-space similarity

- Similar texts have a small angle between them



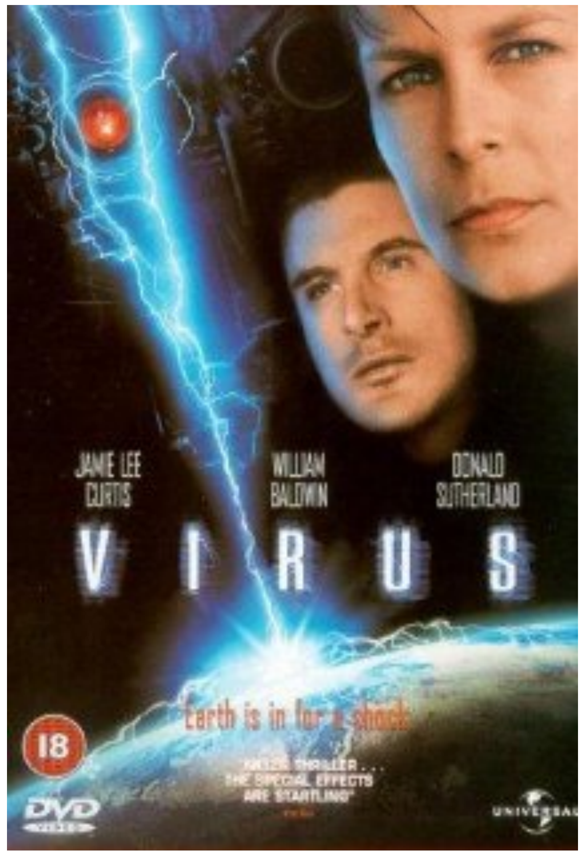
Dimensionality reduction

- Put terms and documents in a lower-dimensional space where we can easily compare them

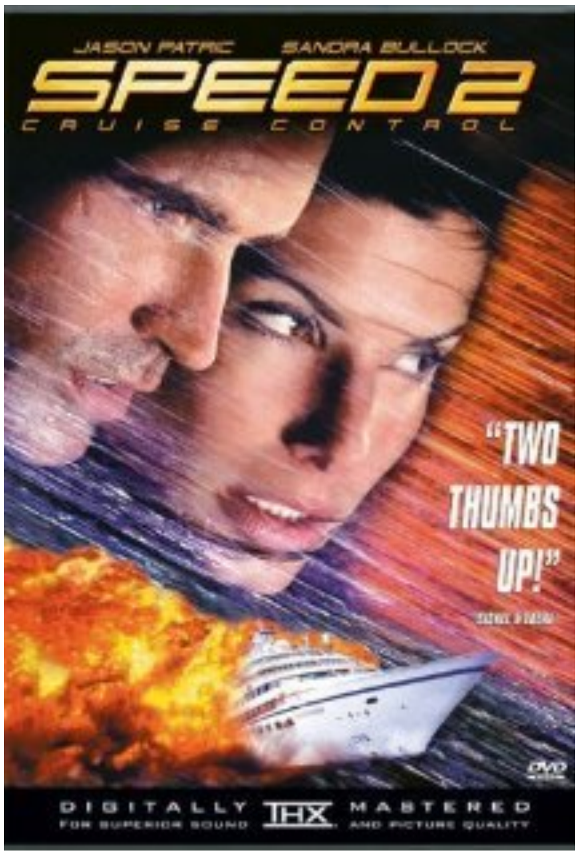
- In NLP, this is called Latent Semantic Analysis or Latent Semantic Inference

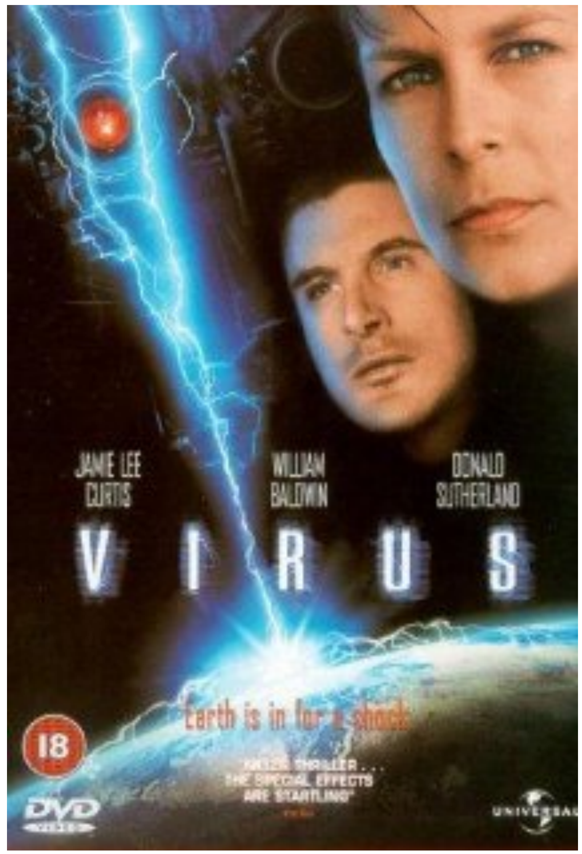
Python example: Unsupervised text similarity using gensim

Similarity of movie reviews

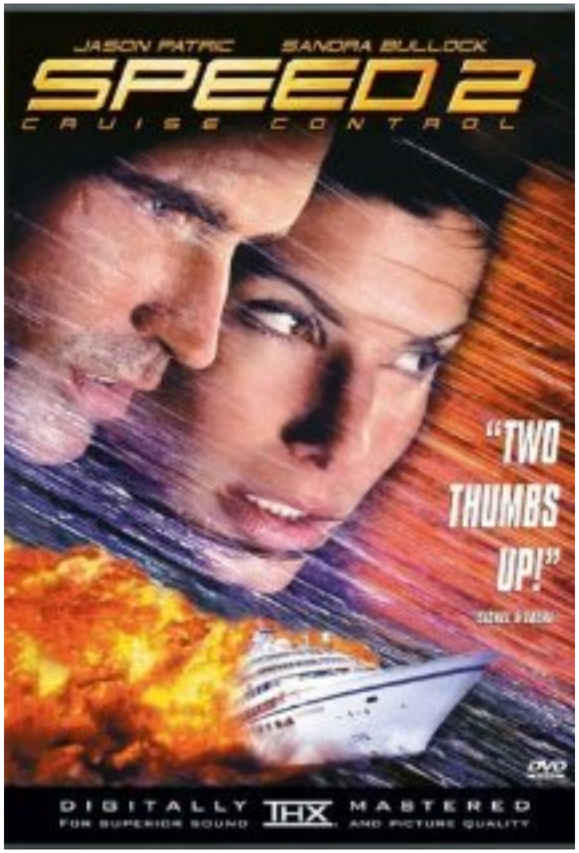


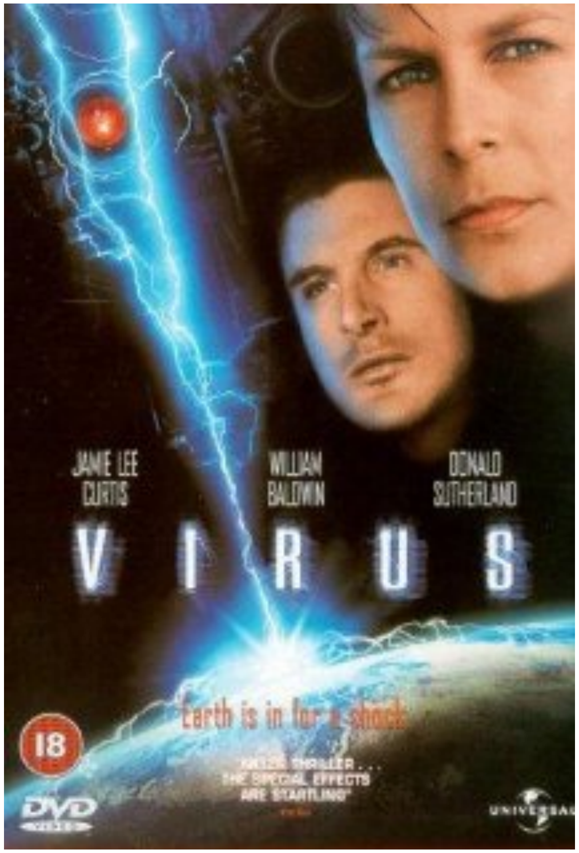
Similarity of movie reviews

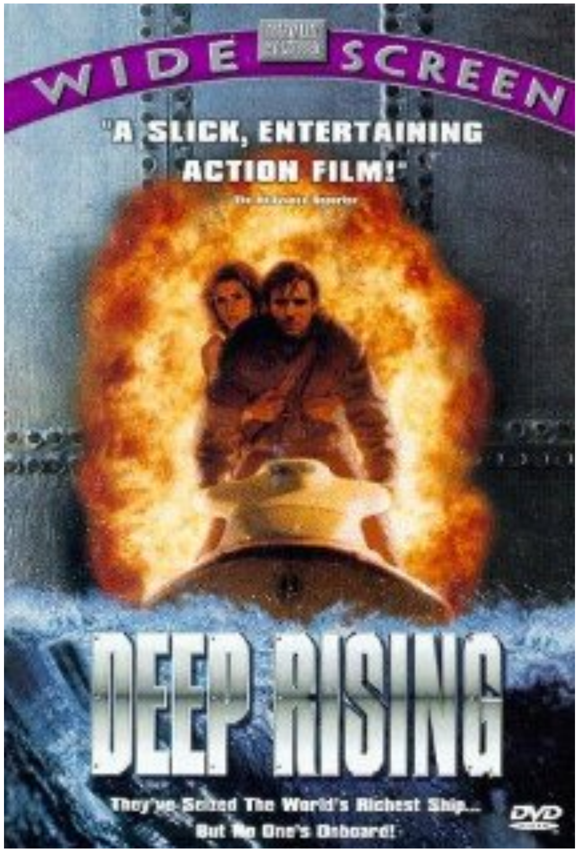




Similarity of movie reviews







Word associations

Word associations

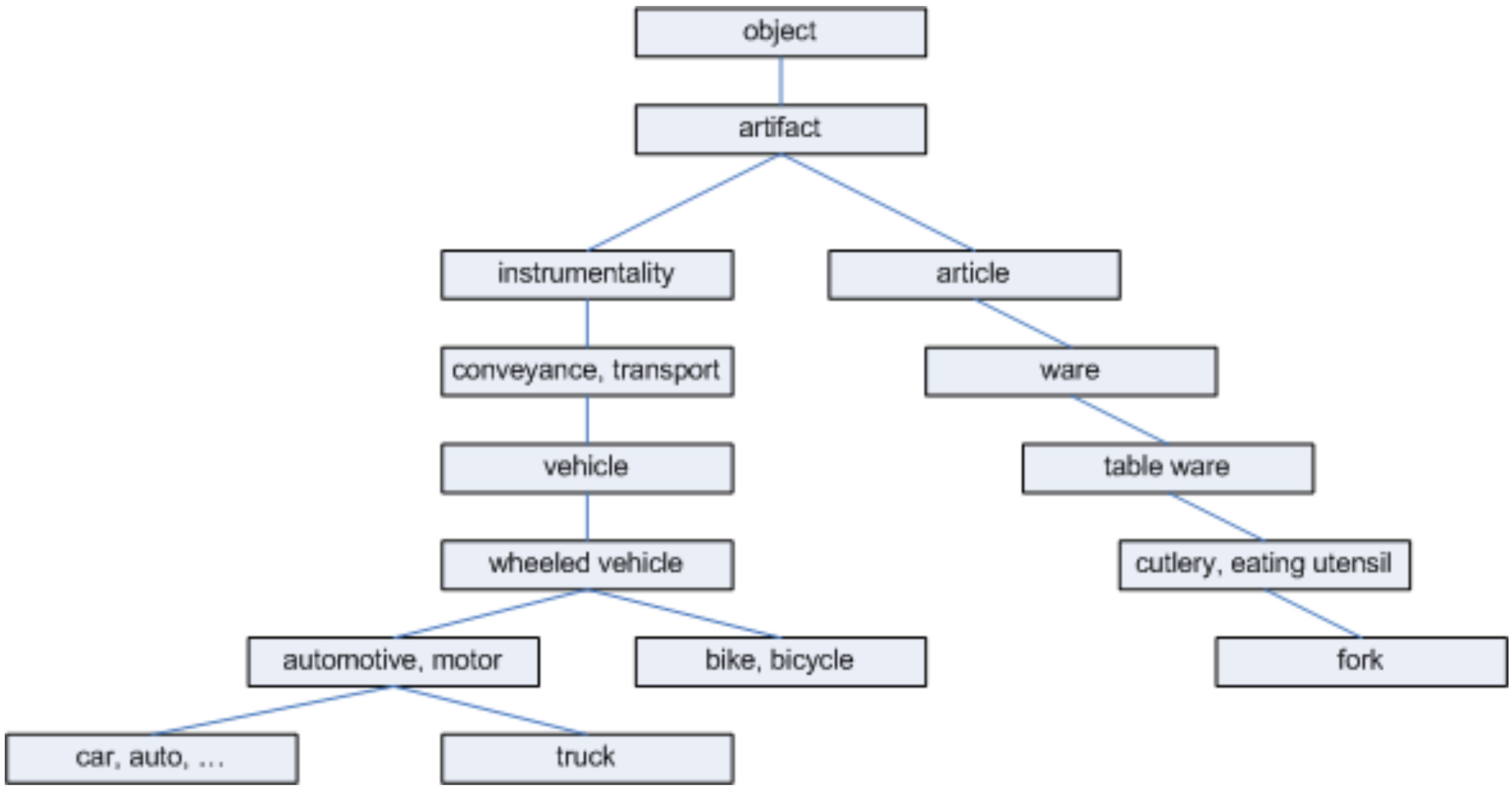
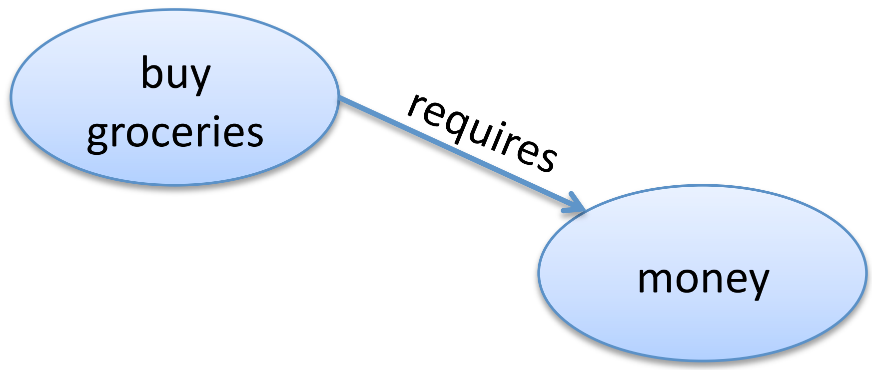
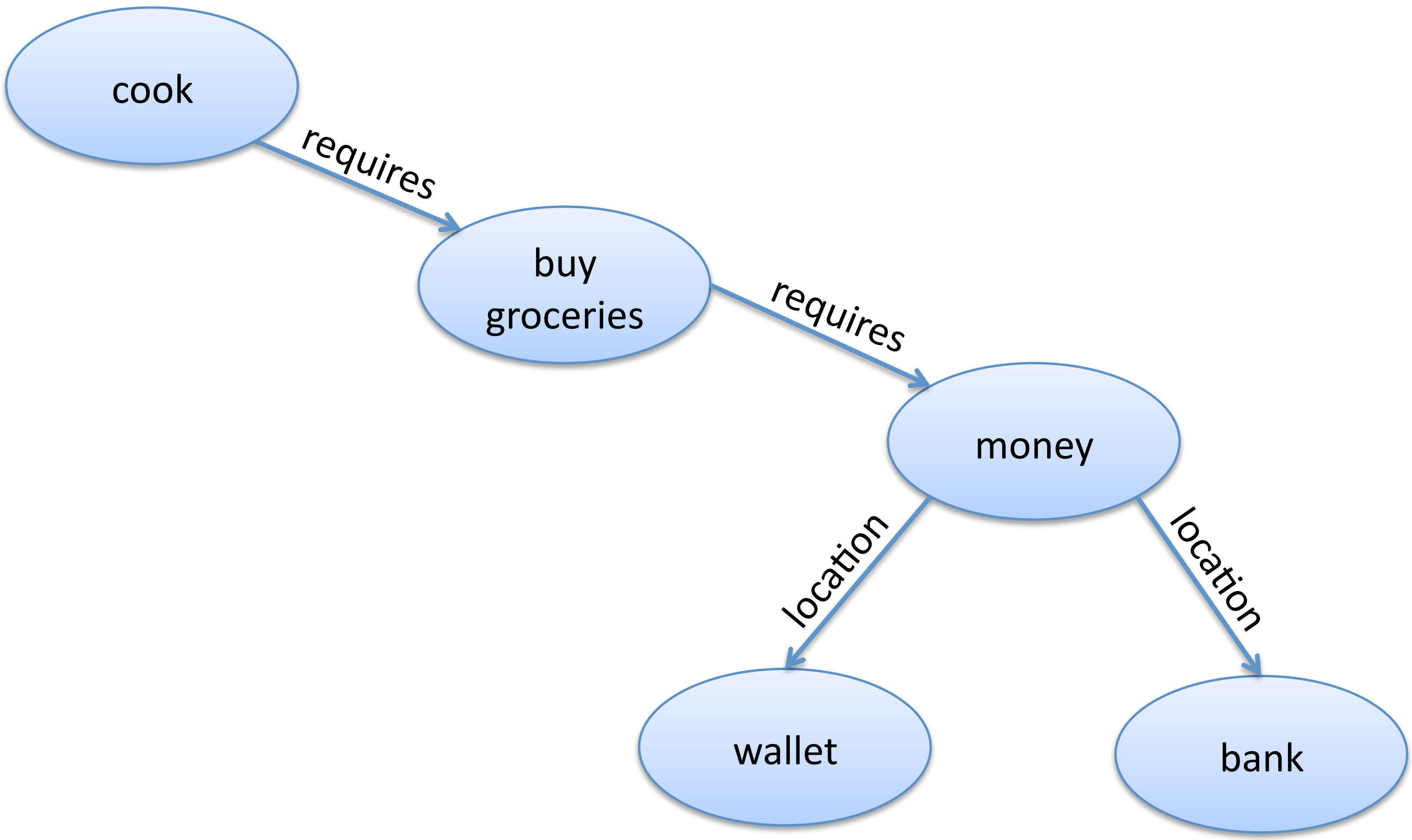
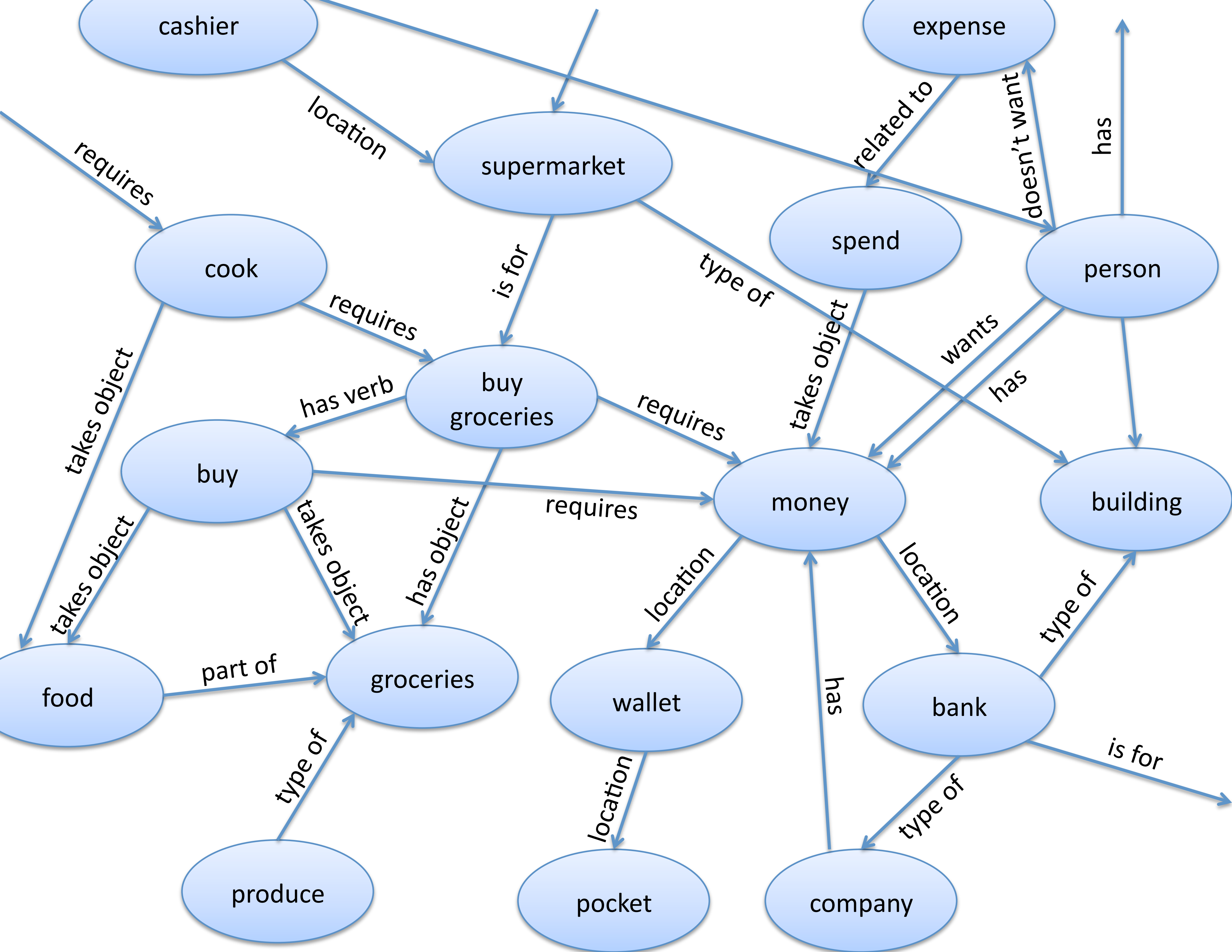


Image source: “WordNet-based semantic similarity measurement”
by Troy Simpson and Thanh Dao, on codeproject.com

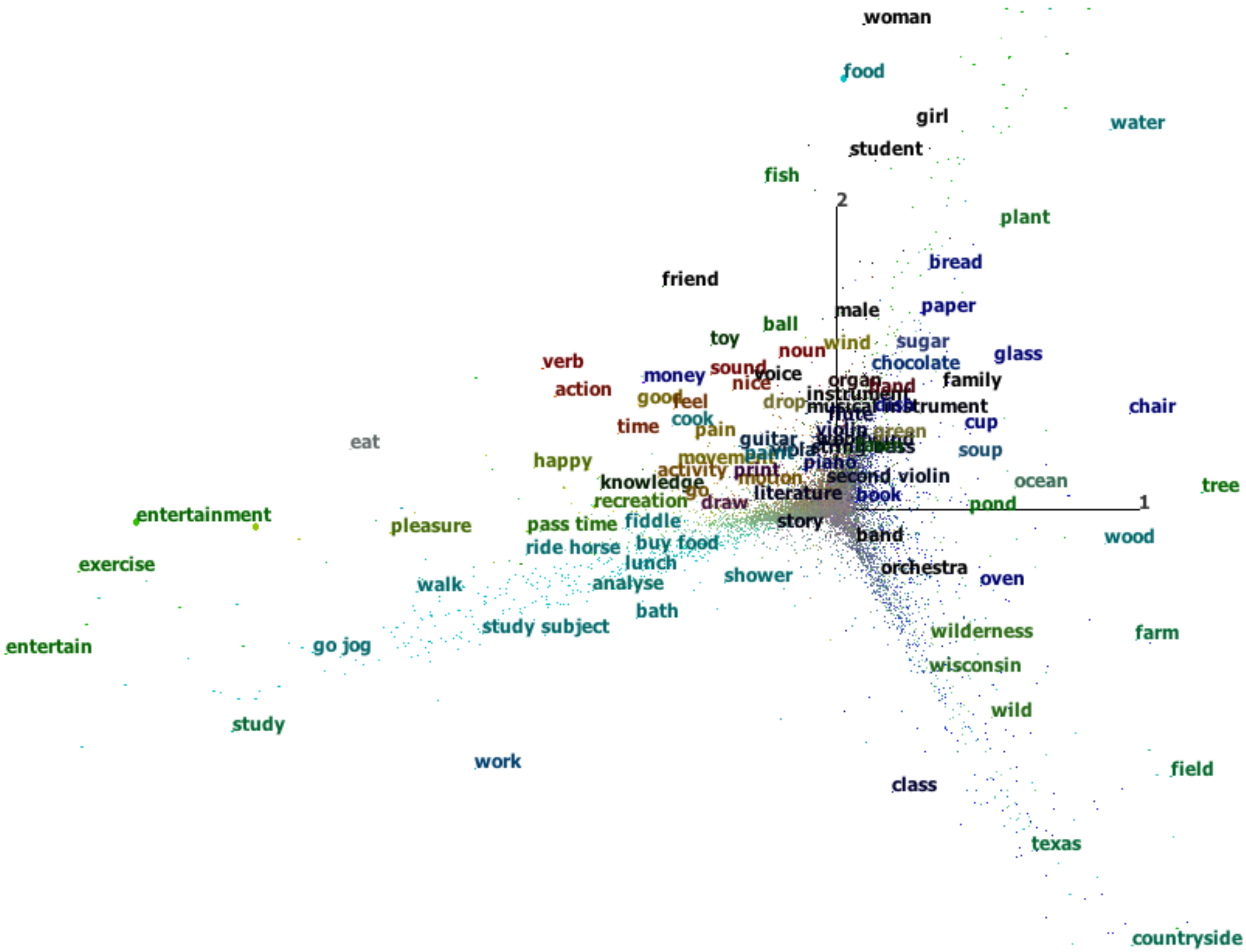
Python example: Querying WordNet


buy
groceries
requires
money


cook
requires
buy
groceries
requires
money
location
location
wallet
bank


cashier
expense
location
related to
doesn't want
has
supermarket
requires
spend
person
cook
is for
type of
requires
takes object
wants
has
buy groceries
has verb
requires
takes object
buy
money
building
requires
takes object
takes object
has object
location
location
type of
food
part of
groceries
wallet
has
bank
is for
type of
location
type of
produce
pocket
company

ConceptNet as a vector space



Python example: Querying ConceptNet

- See API documentation at:
  http://conceptnet5.media.mit.edu

Many incompatible systems

- Supervised text classification
- Unsupervised document similarity
- Domain-general word associations

Many incompatible systems

- Supervised text classification
- Unsupervised document similarity
- Domain-general word associations

- It would be nice if one model could do all of these.

LUMINOSO

NLP with “batteries included”

- **nltk** (the basics)
- **scikit-learn** (classification)
- **gensim** (text similarity)
- Interfaces to **WordNet** and **ConceptNet** (word associations)

What is Python missing?

- A good search index.

What is Python missing?

- A good search index**.**
- Recommendation: use Lucene, or something that uses Lucene.

That's all

Code and slides:

http://github.com/rspeer/text-as-data

Cool things I work on:

http://conceptnet5.media.mit.edu

http://luminoso.com

Extra slides

TF-IDF normalization

- Some documents are longer than others
- Some words appear more than others

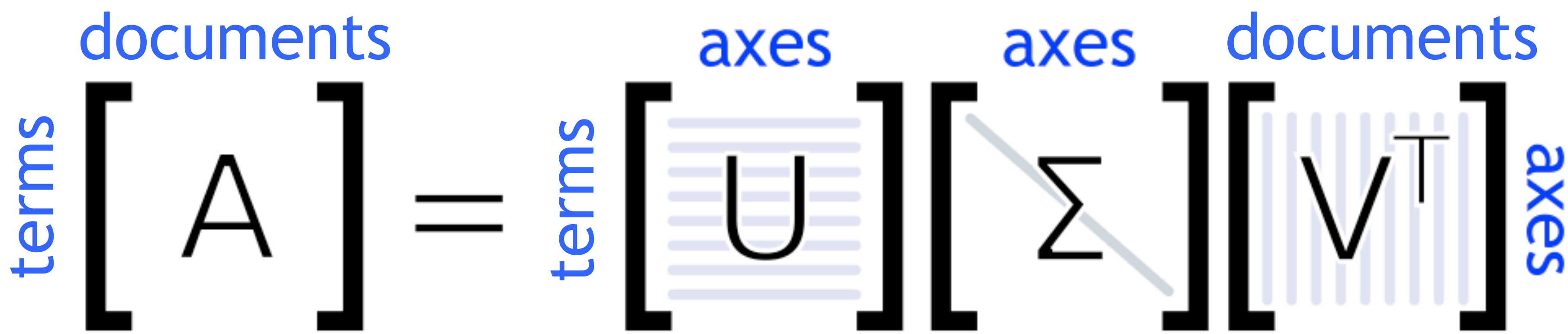
	woe	betray	vengeance	death	alas
Julius Caesar	55.0	32.9	0	0	219.9
Hamlet	38.0	0	73.1	0	171.0
Macbeth	61.4	73.5	0	0	122.7
Alice in Wonderland	0	0	0	0	83.2

(TF-IDF values from NLTK's Project Gutenberg corpus, in micro-bits per word)

TF-IDF normalization

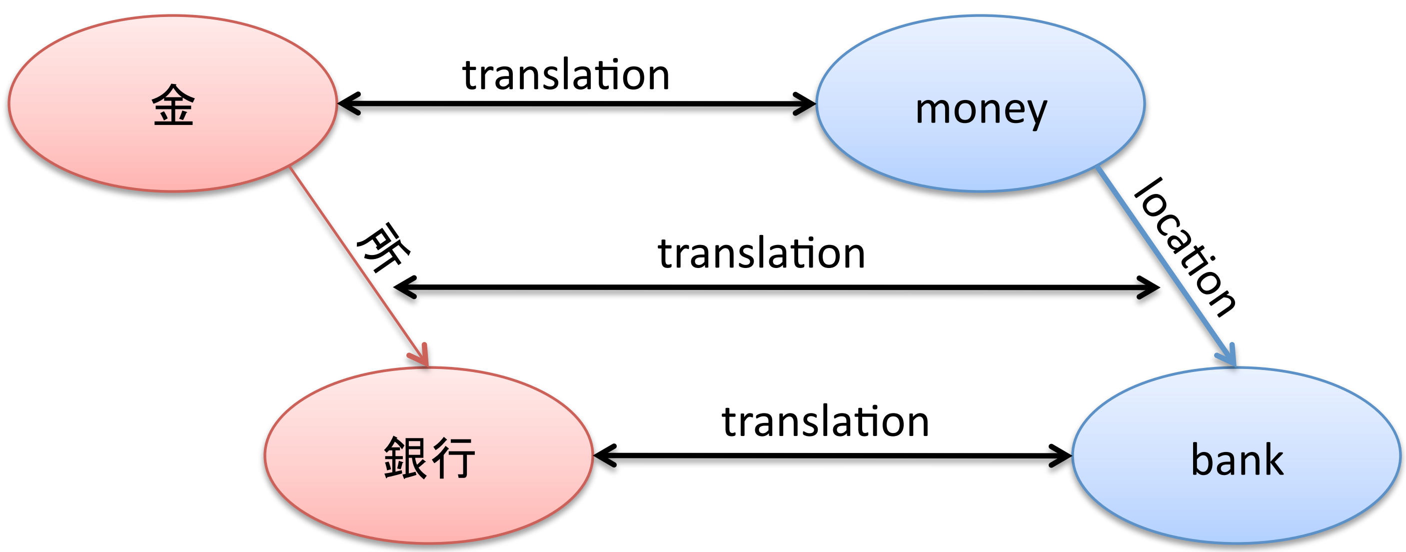
- TF replaces term counts with term frequencies
- IDF tells us how much information we get when a word appears
- In Project Gutenberg:
  - IDF(the) = 0 bits
  - IDF(vengeance) = 1.36 bits
  - IDF(whale) = 2.17 bits
  - IDF(Ishmael) = 3.17 bits

Dimensionality reduction



Dimensionality reduction

$$\text{terms}\overset{\text{documents}}{\begin{bmatrix} \; A \; \end{bmatrix}} \approx \text{terms}\overset{k\text{ axes}}{\begin{bmatrix} U_k \end{bmatrix}}\overset{k\text{ axes}}{\begin{bmatrix} \Sigma_k \end{bmatrix}}\overset{\text{documents}}{\begin{bmatrix} V_k^T \end{bmatrix}}\, k\text{ axes}$$


金
translation
money
所
translation
location
銀行
translation
bank

But Naïve Bayes is so naïve!

- Sure, its fundamental assumption is wrong
- Often, it works anyway
- On NLP tasks, NB is blazingly fast and surprisingly effective

(See “The Optimality of Naive Bayes”, Harry Zhang, AAAI 2004)