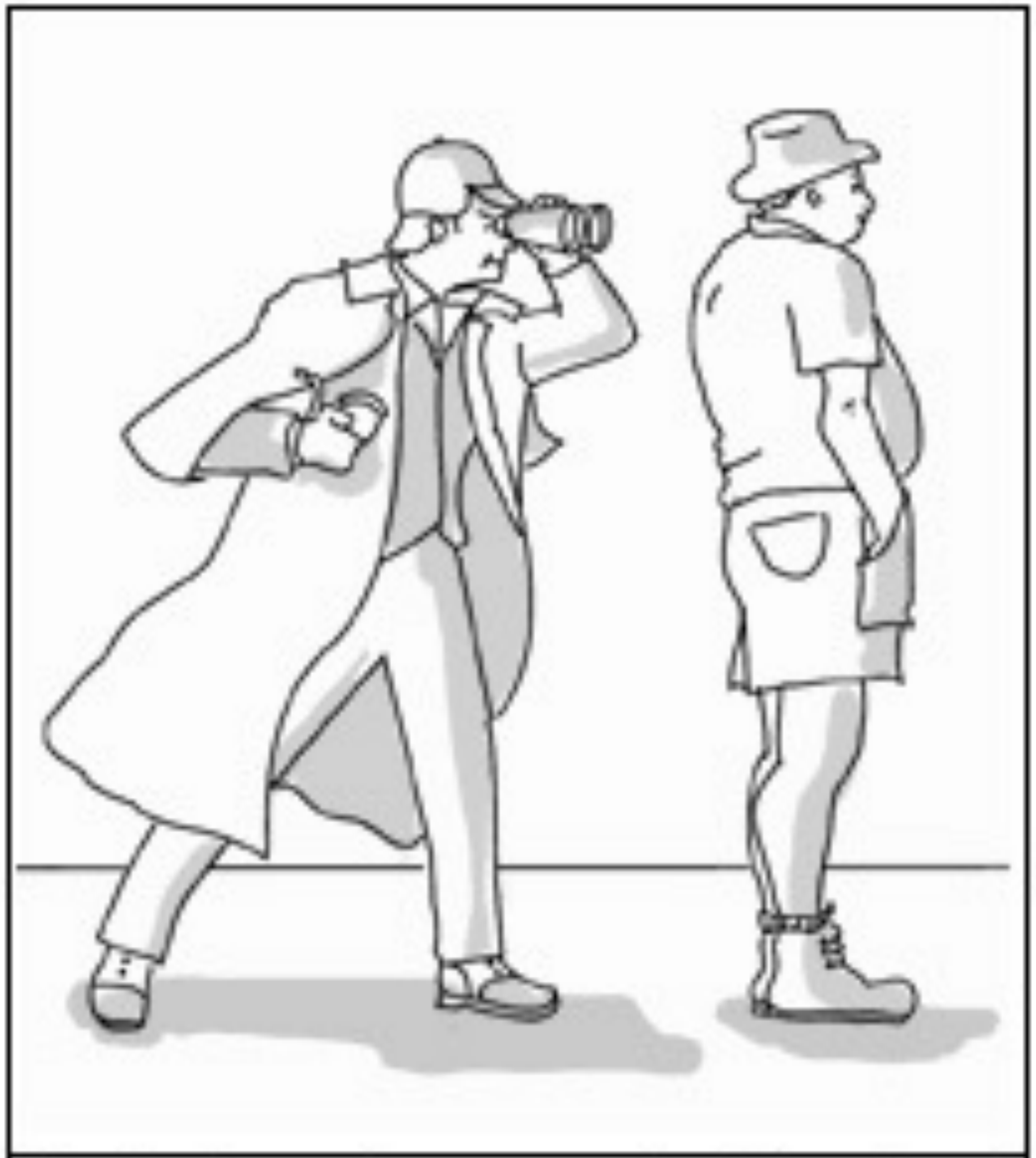
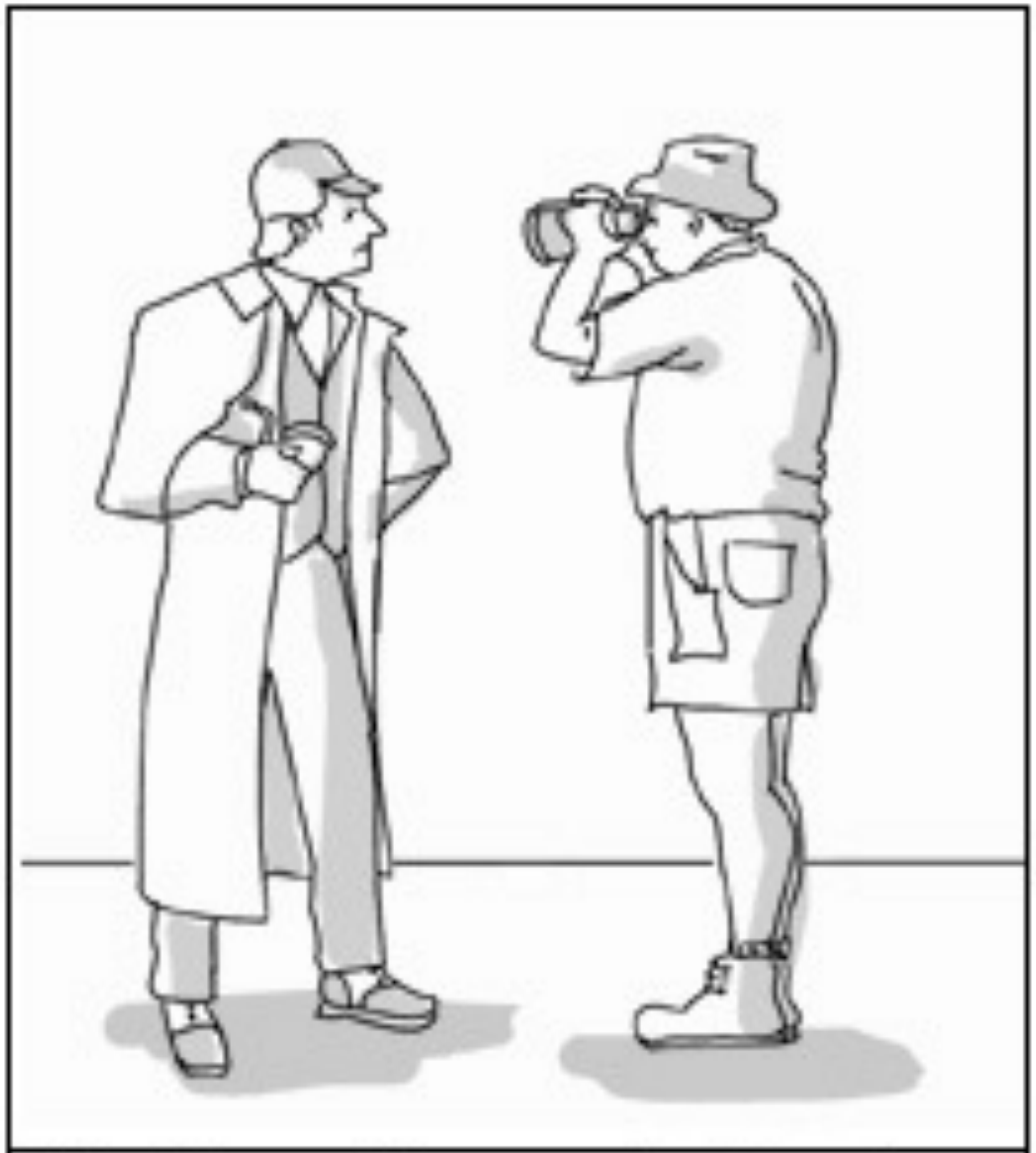
Wouldn’t it be cool if computers UNDERSTAND our language

Language is Ambiguous



Sherlock saw the man using binoculars.



Sherlock saw the man using binoculars.

But we can still get some Data Science done

Classification

Similarity

Search

Sentiment Analysis

This talk

- A quick trip to frequently use data-drive NLP techniques with hands on exercises.
    - Text Munging
    - Bags of Word Naive Bayes
    - TF-IDF
    - Vector Space Model

Text Munging

- Tokenization
- Stemming:
  - Porter Stemmer - set of rules for normalizing tokens e.g.
    - SSES —> SS (caresses —> caress)
    - ATIONAL —> ATE (relatiional —> relate)
- Lemmatization:
  - first determining the part of speech of a word, and applying different normalization rules for each part of speech
- Removing Stop Words

NLTK Tokenizer

- Punkt Sentence Tokenizer
  - divides a text into a list of sentences using an unsupervised algorithm for abbreviation words, collocations, and words that start sentences.

Stop Words

- Most common, short function words that do not contain important information, not useful as text features.

- Not a fixed list, can be discovered from corpora.

- Example: http://www.textfixer.com/resources/common-english-words.txt

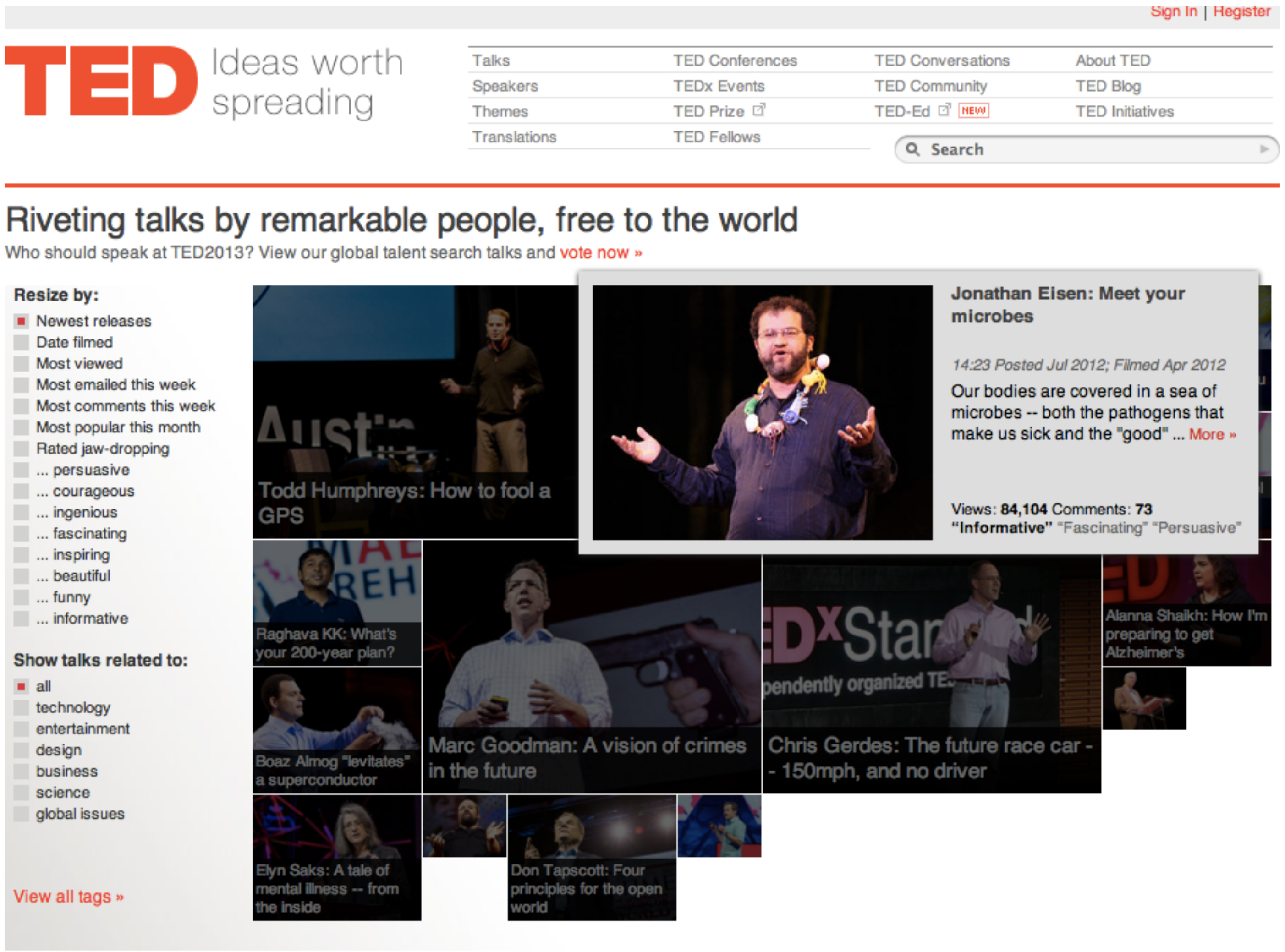
- NLTK: nltk.corpus.stopwords.words('english')

Python Tools

- Regular Expression Module (**re**)
- Natural Language Toolkit (**nltk**, nlpnet)
- Vector Space Model (word2vec, glove)
- Deep Neural Nets: (**Gensim**)

Data Sets

- Data Set 1: TED CLDC Corpus
  - MultiLingual Document Classification
  - Computational Linguistic Group, Oxford UK , http://www.clg.ox.ac.uk/tedcorpus



Riveting talks by remarkable people, free to the world

Who should speak at TED2013? View our global talent search talks and vote now »

Resize by:
Newest releases
Date filmed
Most viewed
Most emailed this week
Most comments this week
Most popular this month
Rated jaw-dropping
... persuasive
... courageous
... ingenious
... fascinating
... inspiring
... beautiful
... funny
... informative
Show talks related to:
all
technology
entertainment
design
business
science
global issues
View all tags »
Todd Humphreys: How to fool a GPS
Raghava KK: What's your 200-year plan?
Boaz Almog "levitates" a superconductor
Marc Goodman: A vision of crimes in the future
Chris Gerdes: The future race car - 150mph, and no driver
Alanna Shaikh: How I'm preparing to get Alzheimer's
Elyn Saks: A tale of mental illness -- from the inside
Don Tapscott: Four principles for the open world

Jonathan Eisen: Meet your microbes

14:23 Posted Jul 2012; Filmed Apr 2012

Our bodies are covered in a sea of microbes -- both the pathogens that make us sick and the "good" ... More »

Views: 84,104 Comments: 73

"Informative" "Fascinating" "Persuasive"

- Data set 2: Ted Talks Transcript Corpus
  - TED Lecture Recommendation Project
  - Idiap Research Institute (EPFL) https://www.idiap.ch/dataset/ted

Bag of Words and Naive Bayes

- Term frequency in “bag of words” can be used to compare how similar documents are

	woe	betray	revenge	death	alas
Julius Caesar	***.046***	***.018***	***.139***	***0***	***.159***
Hamlet	***.142***	***0***	***.287***	***0***	***.110***
Macbeth	***.053***	***.041***	***.120***	***0***	***.082***
Alice in Wonderland	***0***	***0***	***0***	***0***	***.054***

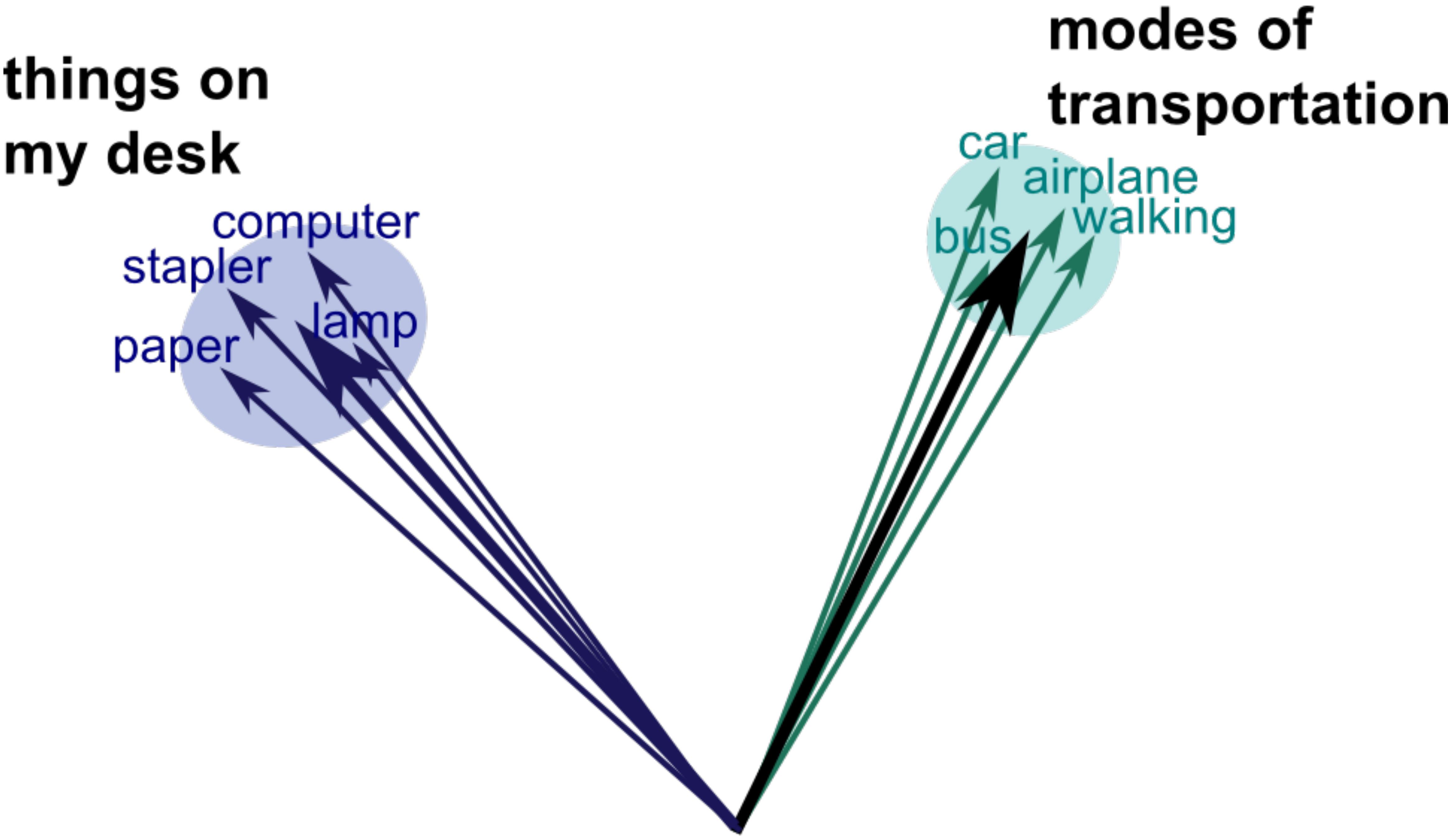
- Naive Bayes predicts probability that a document is in class C based on its features (individual word).
- assumes that all features are statistically independent

Rob Speer & Catherine Havasi
Luminoso Technologies

TF-IDF : Term Frequency Inverse Document Frequency

- TF normalizes term (token) counts with frequencies
- IDF is the number of documents in corpus containing that term.
- IDF tells us how much information (bits) we get when that term (token) appear.

Vector space model



Rob Speer & Catherine Havasi
Luminoso Technologies

Vector space representation

- individual word as vector
- document as collection of these vectors is space
- enables linear algebra on text
    - e.g. King - Man + Woman = Queen

Important Topics We don’t have time to cover

- Topic Modeling (unsupervised document clustering)
- http://radimrehurek.com/gensim/tutorial.html
- Semantic Analysis
- Summarization
- Machine translation
- Knowledge representation and Reasoning

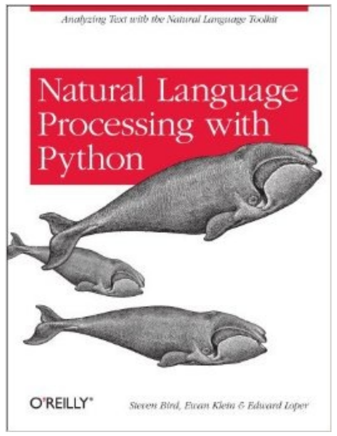
gensim

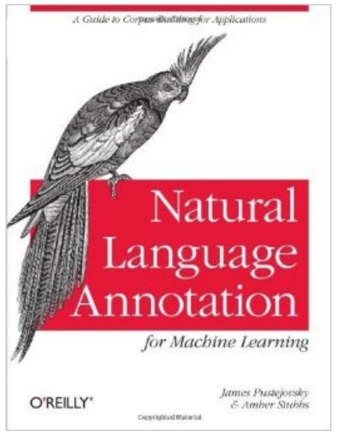
topic modelling for humans

Recommended Books and Tutorials

Natural Language Processing with Python

http://www.nltk.org/book/


Analyzing Text with the Natural Language Toolkit
Natural Language Processing with Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper

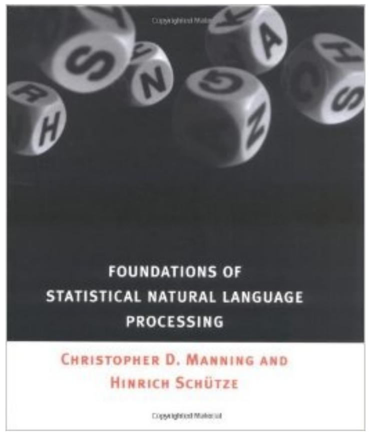

A Guide to Corpus-Building for Applications
Natural Language Annotation
for Machine Learning
O'REILLY
James Pustejovsky & Amber Stubbs

Natural Language Annotation for Machine Learning

https://www.safaribooksonline.com/library/view/natural-language-annotation/9781449332693/

Foundations of Statistical Natural Language Processing

http://nlp.stanford.edu/fsnlp/

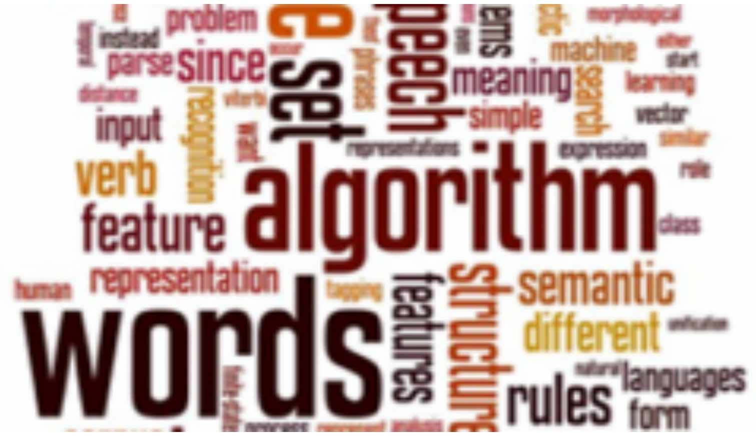

FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING
CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE

Recommended Online Courses/Lectures

Natural Language Processing

Michael Collins,
Columbia University

https://www.coursera.org/course/nlangp



Natural Language Processing

Christopher Manning
Stanford University

https://www.coursera.org/course/nlp

Deep Learning for NLP without magic

Richard Socher, Yoshua Bengio, Christopher Manning

http://techtalks.tv/talks/deep-learning-for-nlp-without-magic-part-1/58414/

http://techtalks.tv/talks/deep-learning-for-nlp-without-magic-part-2/58415/