TTIC 31190: Natural Language Processing

Kevin Gimpel
Spring 2018

Lecture 1: Introduction; Words

Course Overview

- Second time being offered (first was Winter 2016)
- Designed for first-year TTIC PhD students
- My office hours: 3-4pm Mondays (TTIC 531), or by appointment
- TA: Lifu Tu, TTIC PhD student
- TA office hours: 3-4pm Wednesdays (TTIC 501)

- course had much more interest this year than expected
- if you are not yet registered, it is unlikely you will be able to get a spot
- I have been in touch with you if you’re within the first few spots on the waitlist

Prerequisites

- No course prerequisites, but I will assume:
  - some programming experience (no specific language required)
  - familiarity with basics of calculus, linear algebra, and probability
  - will be helpful to have taken a machine learning course, but not strictly required

Grading

- 3 assignments (15% each)
- midterm exam (15%) (Wed., May 16)
- course project (30%):
  - project proposal (5%)
  - final report (25%)
- class participation, including quizzes (10%)
- no final

Assignments

- mixture of formal exercises, implementation, experimentation, analysis
- first assignment has been posted so that you can have a look at it, due 2 weeks from Wednesday

Project

- Replicate [part of] a published NLP paper, or define your own project
- The project must be done in a group of two
- Each group member will receive same grade
- More details to come

Collaboration Policy

- You are welcome to discuss assignments with others in the course, but solutions and code must be written individually

Lateness Policy

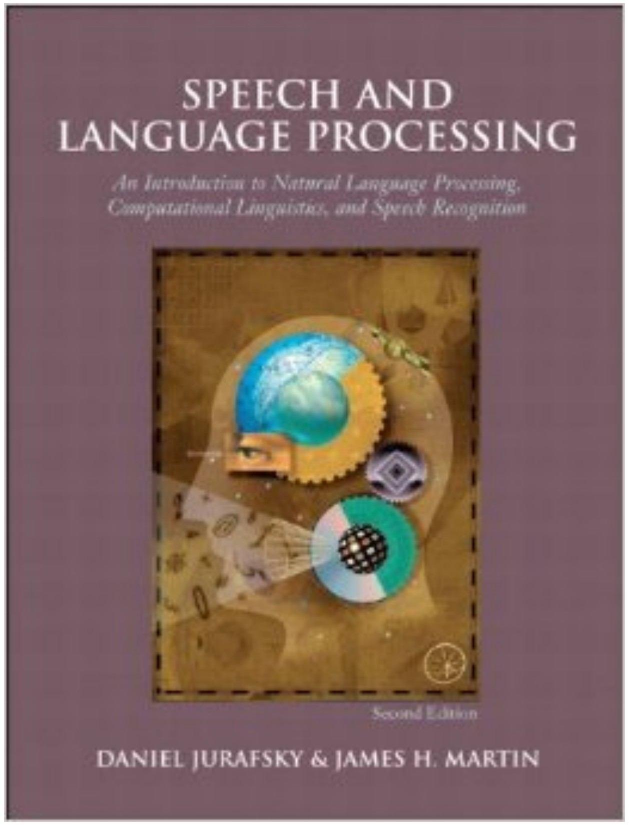
- If you turn in an assignment late, a penalty will be assessed (2% per hour late)

- You will have 4 late days to use as you wish during the quarter

- Late days must be used in whole increments
  - e.g., if you turn in an assignment 6 hours late and want to use a late day to avoid penalty, it will cost an entire late day to do so

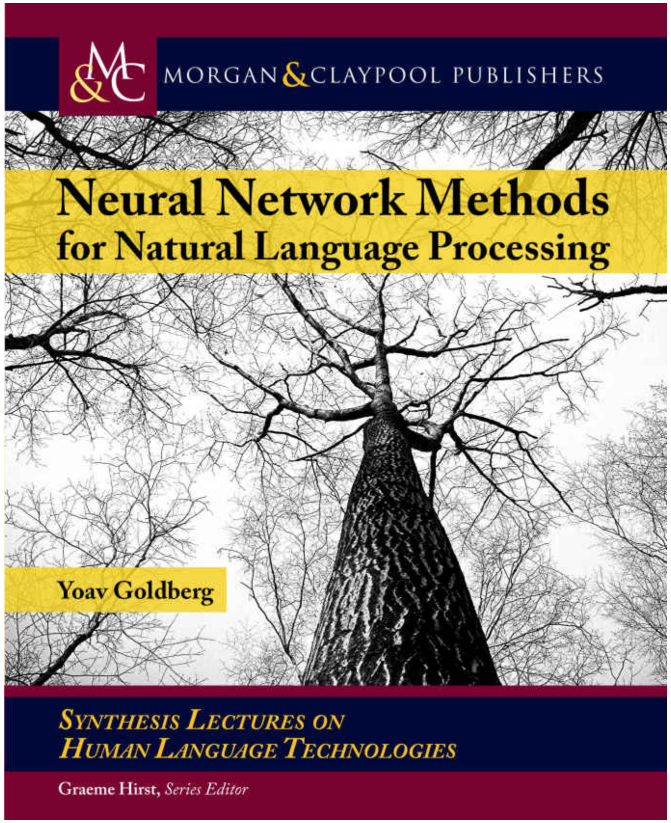
Optional Textbooks (1/2)

- Jurafsky & Martin. *Speech and Language Processing*, 2nd Ed. & 3rd Ed.
- Many chapters of 3rd edition are online
- Copies of 2nd edition available in TTIC library


SPEECH AND
LANGUAGE PROCESSING
An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

Optional Textbooks (2/2)

- Goldberg. *Neural Network Methods for Natural Language Processing*.
- Earlier draft (from 2015) available online
- Two copies on reserve in TTIC library

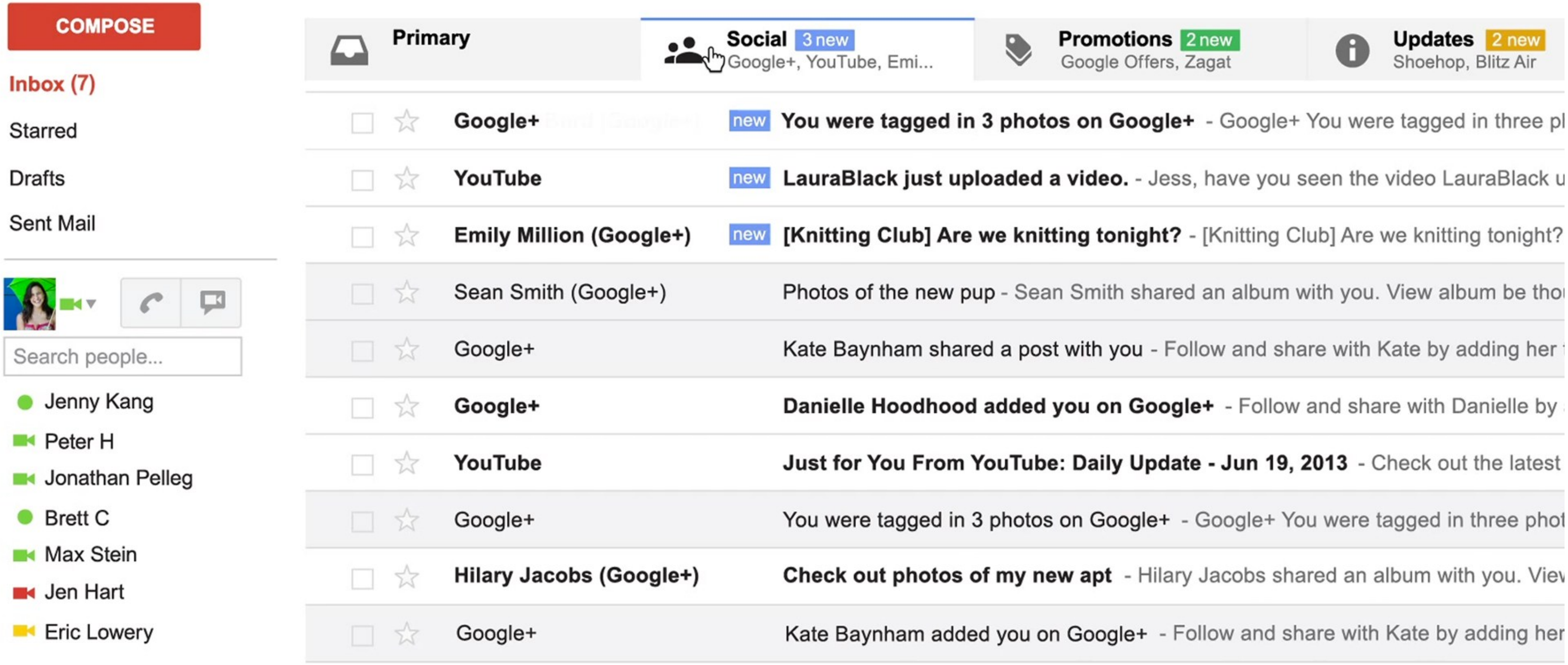

MORGAN & CLAYPOOL PUBLISHERS
Neural Network Methods
for Natural Language Processing
Yoav Goldberg
SYNTHESIS LECTURES ON
HUMAN LANGUAGE TECHNOLOGIES
Graeme Hirst, Series Editor

What is natural language processing?

What is natural language processing?

an experimental computer science research area that includes problems and solutions pertaining to the understanding of human language

Text Classification



Google+ new **You were tagged in 3 photos on Google+** - Google+ You were tagged in three pl

YouTube new **LauraBlack just uploaded a video.** - Jess, have you seen the video LauraBlack u

Emily Million (Google+) new **[Knitting Club] Are we knitting tonight?** - [Knitting Club] Are we knitting tonight?

Sean Smith (Google+) Photos of the new pup - Sean Smith shared an album with you. View album be tho

Google+ Kate Baynham shared a post with you - Follow and share with Kate by adding her

Google+ **Danielle Hoodhood added you on Google+** - Follow and share with Danielle by

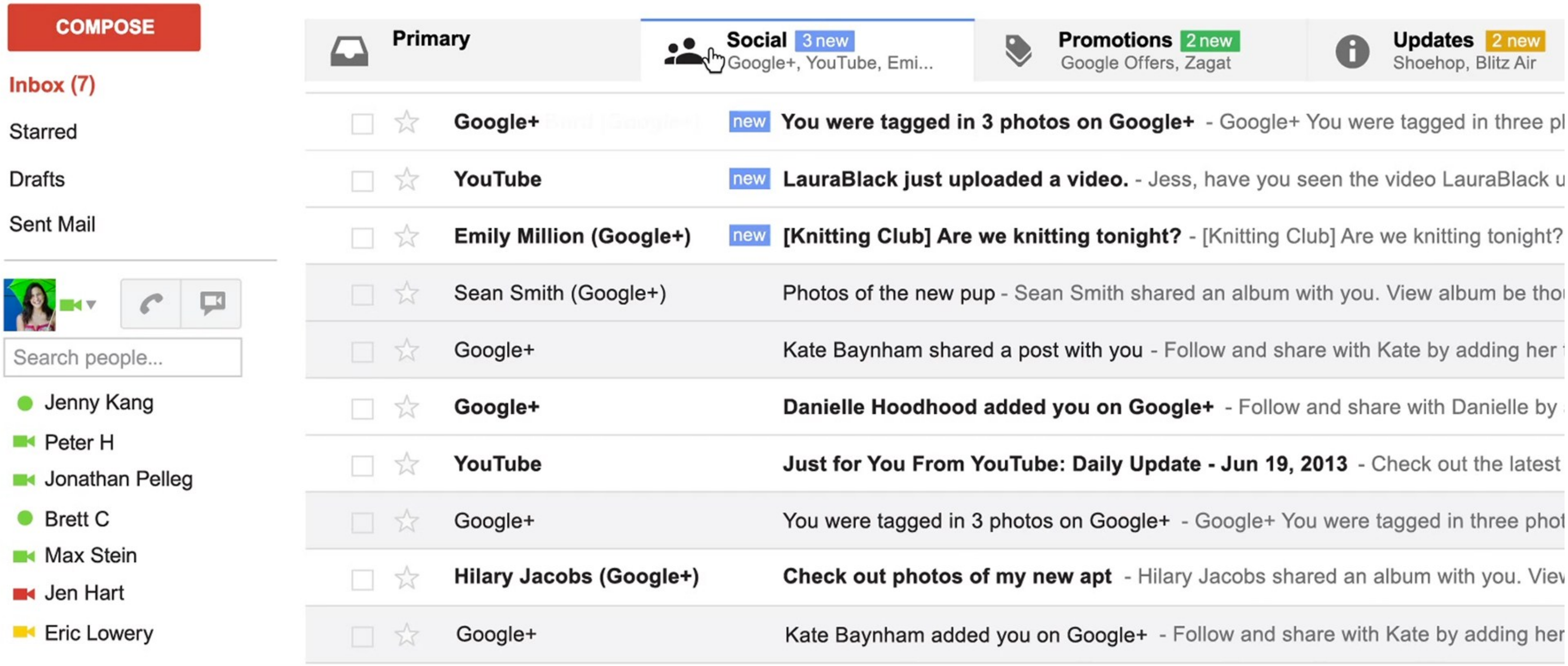
YouTube **Just for You From YouTube: Daily Update - Jun 19, 2013** - Check out the latest

Google+ You were tagged in 3 photos on Google+ - Google+ You were tagged in three phot

Hilary Jacobs (Google+) **Check out photos of my new apt** - Hilary Jacobs shared an album with you. Viev

Google+ Kate Baynham added you on Google+ - Follow and share with Kate by adding her

Text Classification



COMPOSE

Inbox (7)

Starred

Drafts

Sent Mail

Search people...

Jenny Kang

Peter H

Jonathan Pelleg

Brett C

Max Stein

Jen Hart

Eric Lowery

Primary | Social 3 new Google+, YouTube, Emi... | Promotions 2 new Google Offers, Zagat | Updates 2 new Shoehop, Blitz Air

□ ☆ Google+ new You were tagged in 3 photos on Google+ - Google+ You were tagged in three pl

□ ☆ YouTube new LauraBlack just uploaded a video. - Jess, have you seen the video LauraBlack u

□ ☆ Emily Million (Google+) new [Knitting Club] Are we knitting tonight? - [Knitting Club] Are we knitting tonight?

□ ☆ Sean Smith (Google+) Photos of the new pup - Sean Smith shared an album with you. View album be tho

□ ☆ Google+ Kate Baynham shared a post with you - Follow and share with Kate by adding her

□ ☆ Google+ Danielle Hoodhood added you on Google+ - Follow and share with Danielle by

□ ☆ YouTube Just for You From YouTube: Daily Update - Jun 19, 2013 - Check out the latest

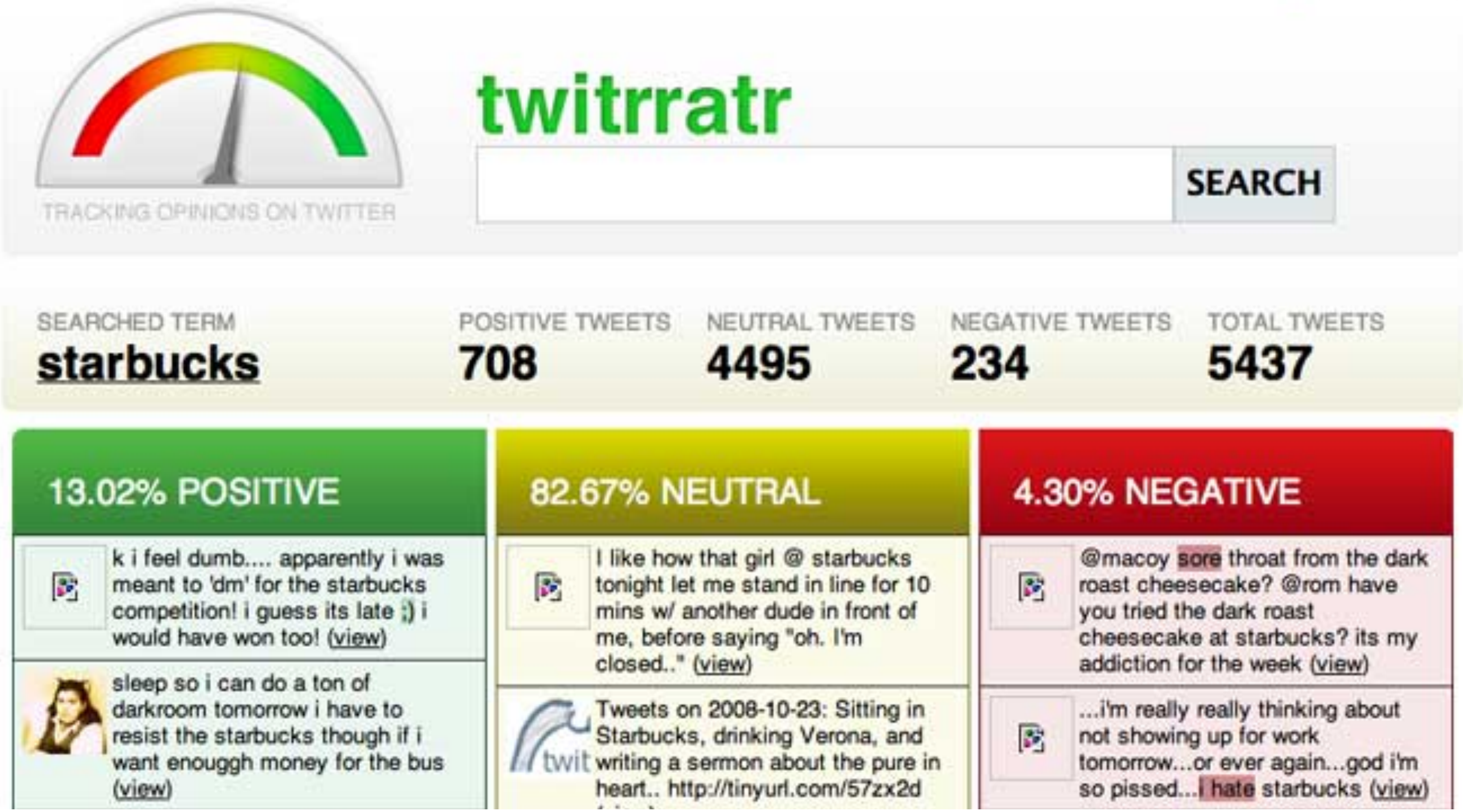
□ ☆ Google+ You were tagged in 3 photos on Google+ - Google+ You were tagged in three phot

□ ☆ Hilary Jacobs (Google+) Check out photos of my new apt - Hilary Jacobs shared an album with you. View

□ ☆ Google+ Kate Baynham added you on Google+ - Follow and share with Kate by adding her

- spam / not spam
- priority level
- category (primary / social / promotions / updates)

Sentiment Analysis



SEARCHED TERM	POSITIVE TWEETS	NEUTRAL TWEETS	NEGATIVE TWEETS	TOTAL TWEETS
starbucks	708	4495	234	5437

13.02% POSITIVE

k i feel dumb.... apparently i was meant to 'dm' for the starbucks competition! i guess its late ;) i would have won too! (view)

sleep so i can do a ton of darkroom tomorrow i have to resist the starbucks though if i want enouggh money for the bus (view)

82.67% NEUTRAL

I like how that girl @ starbucks tonight let me stand in line for 10 mins w/ another dude in front of me, before saying "oh. I'm closed.." (view)

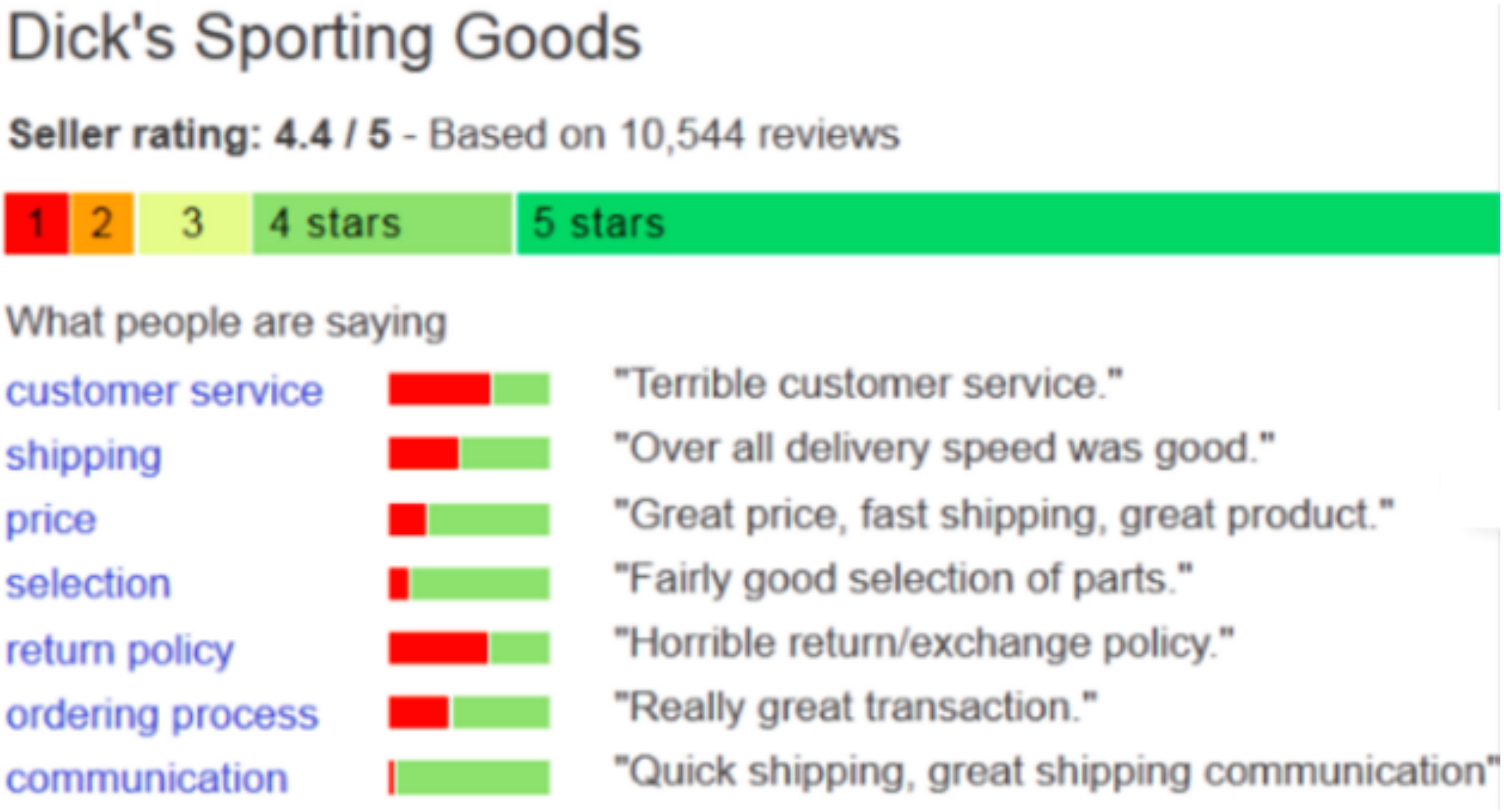
Tweets on 2008-10-23: Sitting in Starbucks, drinking Verona, and writing a sermon about the pure in heart.. http://tinyurl.com/57zx2d

4.30% NEGATIVE

@macoy sore throat from the dark roast cheesecake? @rom have you tried the dark roast cheesecake at starbucks? its my addiction for the week (view)

...i'm really really thinking about not showing up for work tomorrow...or ever again...god i'm so pissed...i hate starbucks (view)

Sentiment Analysis



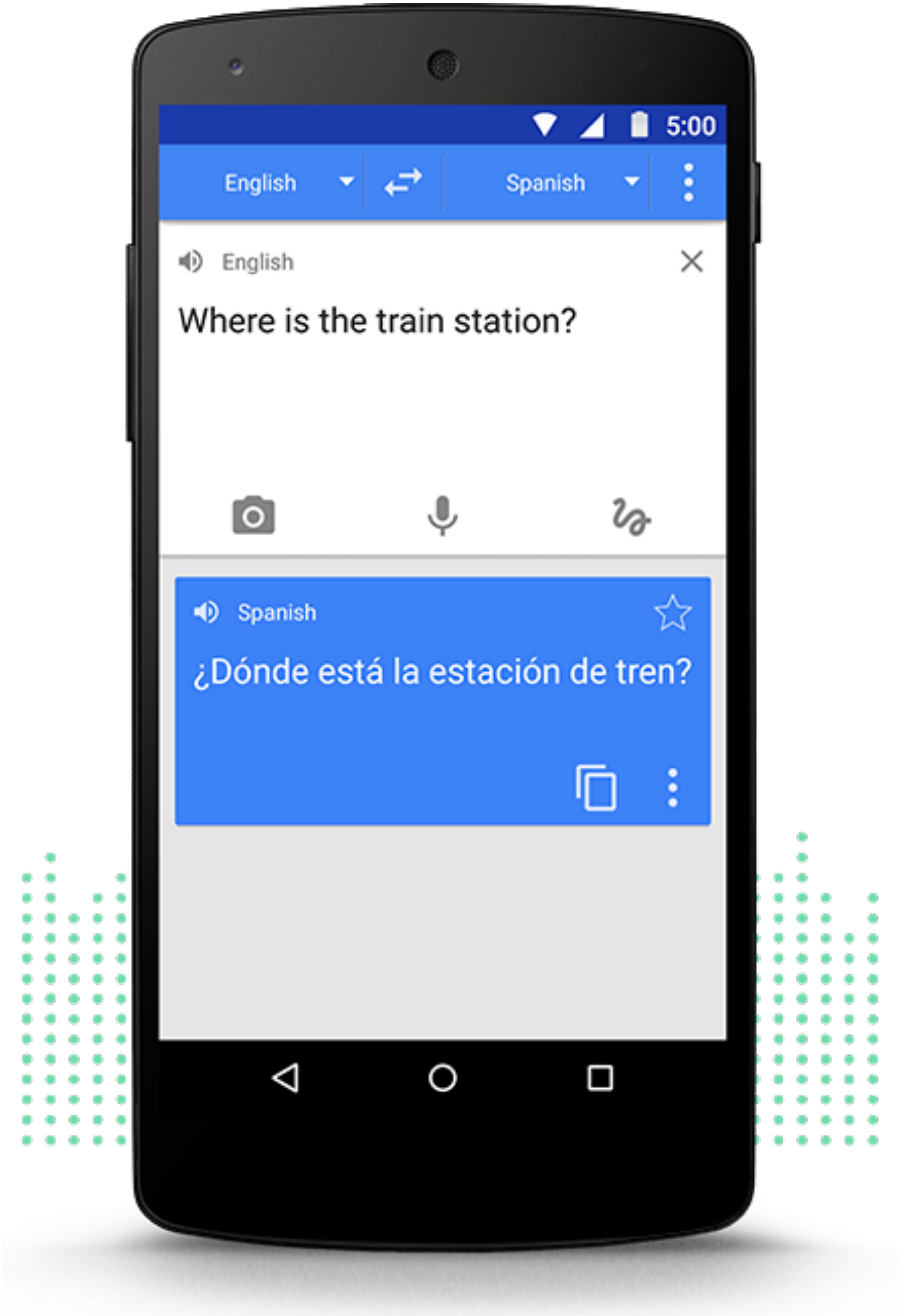
Dick's Sporting Goods

Seller rating: 4.4 / 5 - Based on 10,544 reviews

What people are saying

customer service	"Terrible customer service."
shipping	"Over all delivery speed was good."
price	"Great price, fast shipping, great product."
selection	"Fairly good selection of parts."
return policy	"Horrible return/exchange policy."
ordering process	"Really great transaction."
communication	"Quick shipping, great shipping communication"

Machine Translation



Where is the train station?

¿Dónde está la estación de tren?

Question Answering

Question Answering

amazon alexa

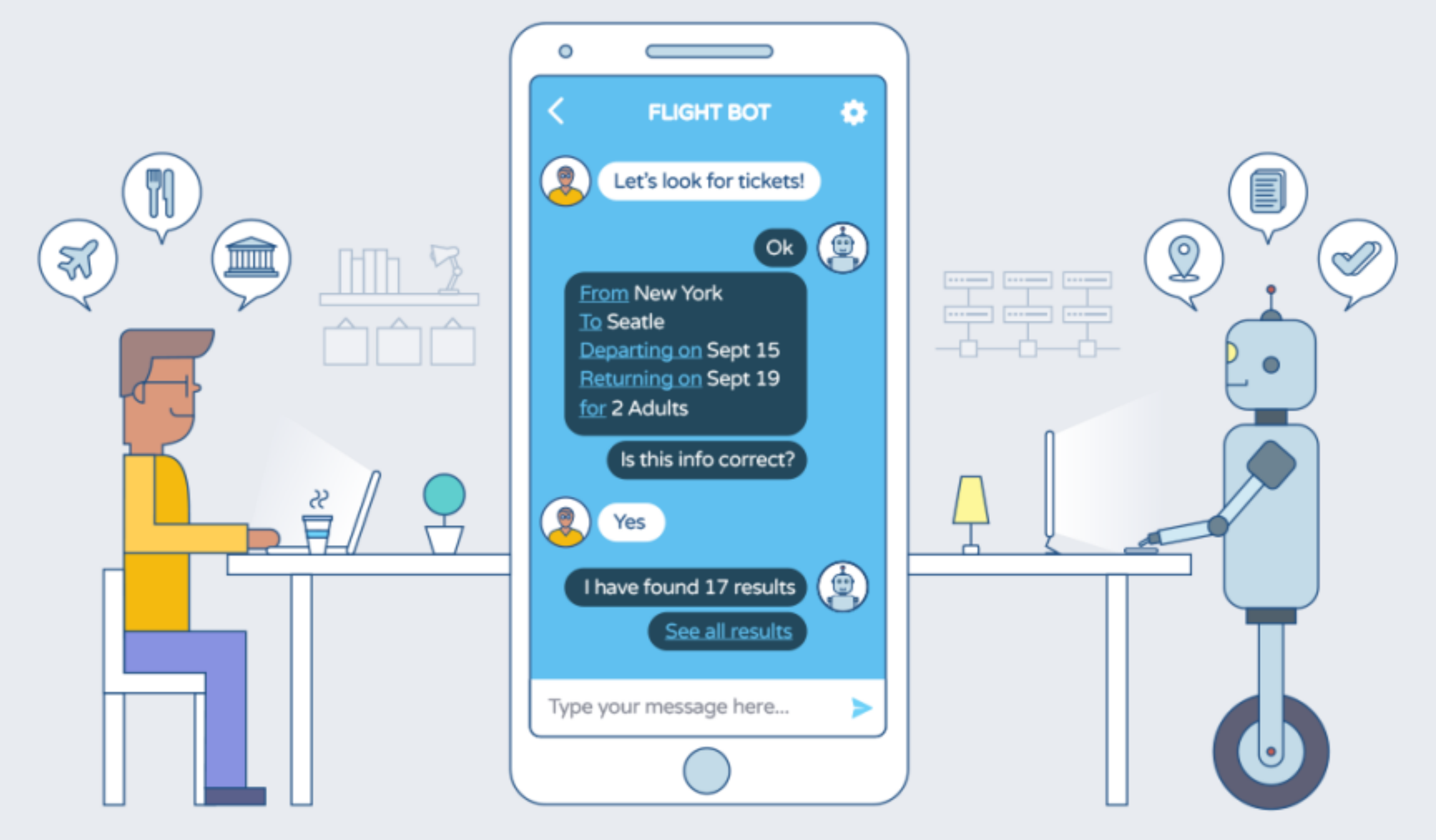
“Alexa, who was President when Barack Obama was nine?”

“Alexa, how’s my commute?

“Alexa, what’s the weather?”

“Alexa, did the 49ers win?”

Dialog Systems



Let's look for tickets!

Ok

From New York
To Seatle
Departing on Sept 15
Returning on Sept 19
for 2 Adults

Is this info correct?

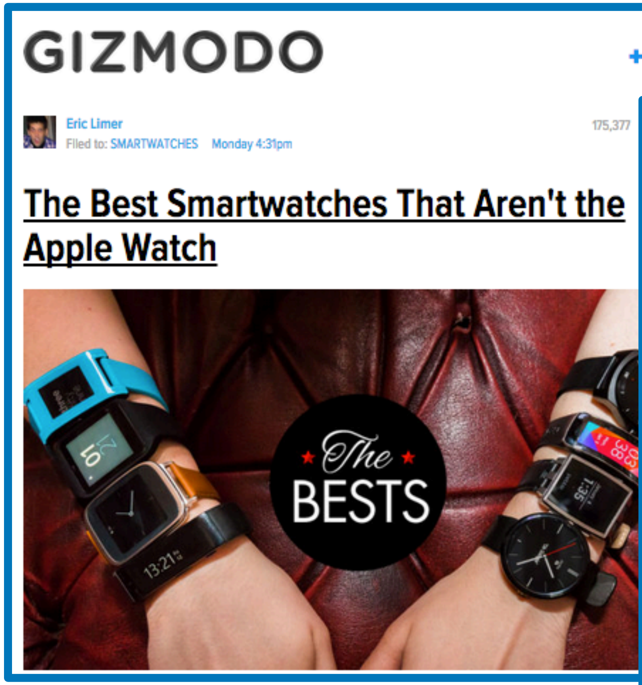
Yes

I have found 17 results

See all results

figure credit: Phani Marupaka

Summarization



GIZMODO

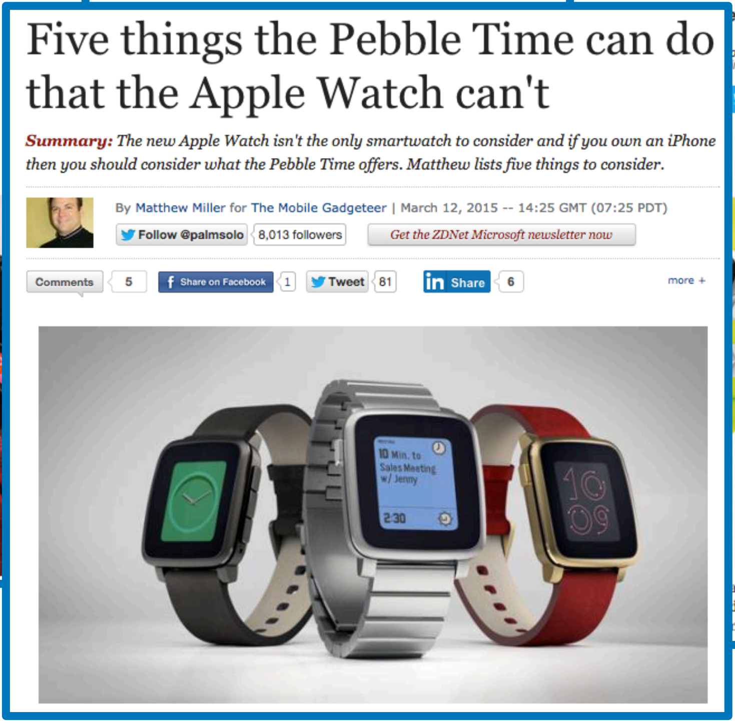
+ FOLLOW

Eric Limer

Filed to: SMARTWATCHES Monday 4:31pm

175,377

The Best Smartwatches That Aren't the Apple Watch



Five things the Pebble Time can do that the Apple Watch can't

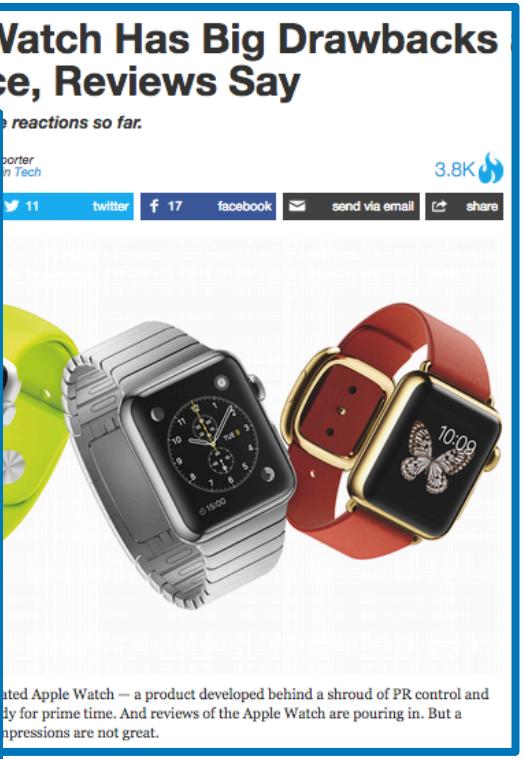
Summary: The new Apple Watch isn't the only smartwatch to consider and if you own an iPhone then you should consider what the Pebble Time offers. Matthew lists five things to consider.

By Matthew Miller for The Mobile Gadgeteer | March 12, 2015 -- 14:25 GMT (07:25 PDT)

Follow @palmsolo 8,013 followers

Get the ZDNet Microsoft newsletter now

Comments 5 | Share on Facebook 1 | Tweet 81 | in Share 6 | more +



Apple Watch Has Big Drawbacks
Interface, Reviews Say

e reactions so far.

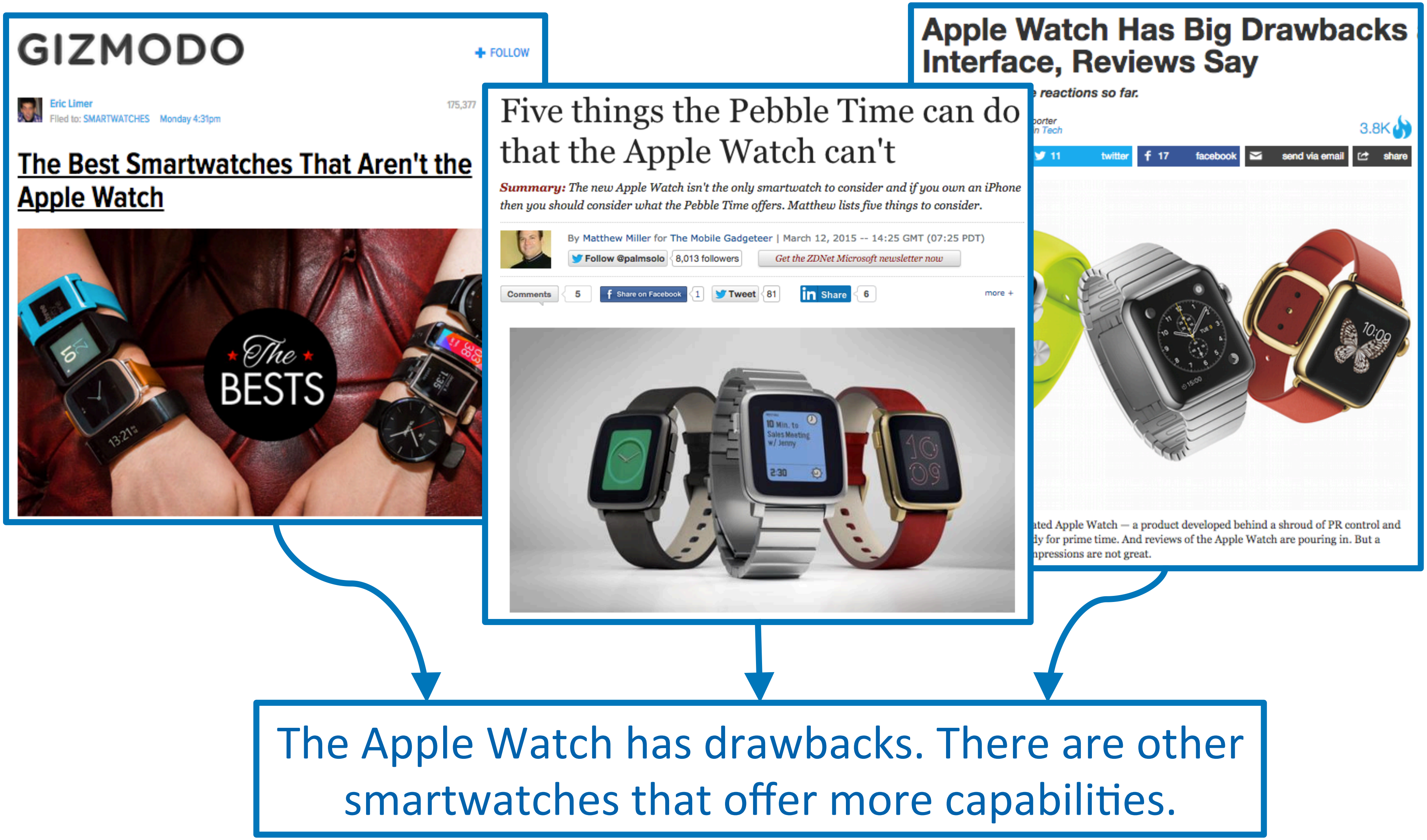
porter
n Tech

3.8K

11 twitter | 17 facebook | send via email | share

ated Apple Watch — a product developed behind a shroud of PR control and dy for prime time. And reviews of the Apple Watch are pouring in. But a npressions are not great.

Summarization



GIZMODO

+ FOLLOW

Eric Limer
Filed to: SMARTWATCHES Monday 4:31pm

175,377

The Best Smartwatches That Aren't the Apple Watch

Five things the Pebble Time can do that the Apple Watch can't

Summary: The new Apple Watch isn't the only smartwatch to consider and if you own an iPhone then you should consider what the Pebble Time offers. Matthew lists five things to consider.

By Matthew Miller for The Mobile Gadgeteer | March 12, 2015 -- 14:25 GMT (07:25 PDT)

Follow @palmsolo 8,013 followers Get the ZDNet Microsoft newsletter now

Comments 5 Share on Facebook 1 Tweet 81 Share 6 more +

Apple Watch Has Big Drawbacks
Interface, Reviews Say

reactions so far.

porter
n Tech

3.8K

11 twitter 17 facebook send via email share

ated Apple Watch — a product developed behind a shroud of PR control and ly for prime time. And reviews of the Apple Watch are pouring in. But a npressions are not great.

The Apple Watch has drawbacks. There are other smartwatches that offer more capabilities.

Part-of-Speech Tagging

Some questioned if Tim Cook ’s first product

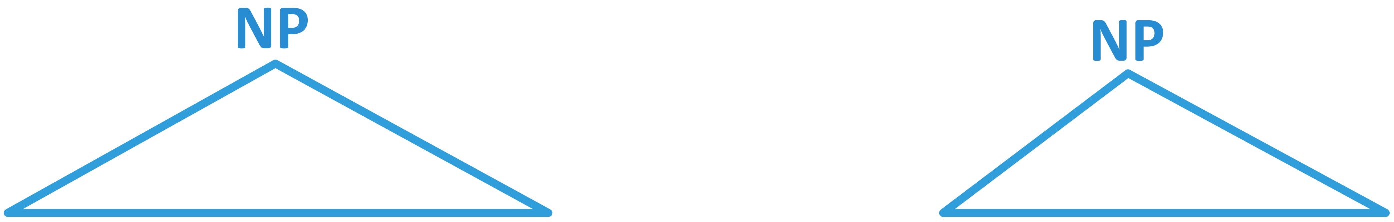
would be a breakaway hit for Apple .

Part-of-Speech Tagging

determiner	verb (past)	prep.	proper noun	proper noun	poss.	adj.	noun
Some	questioned	if	Tim	Cook	’s	first	product

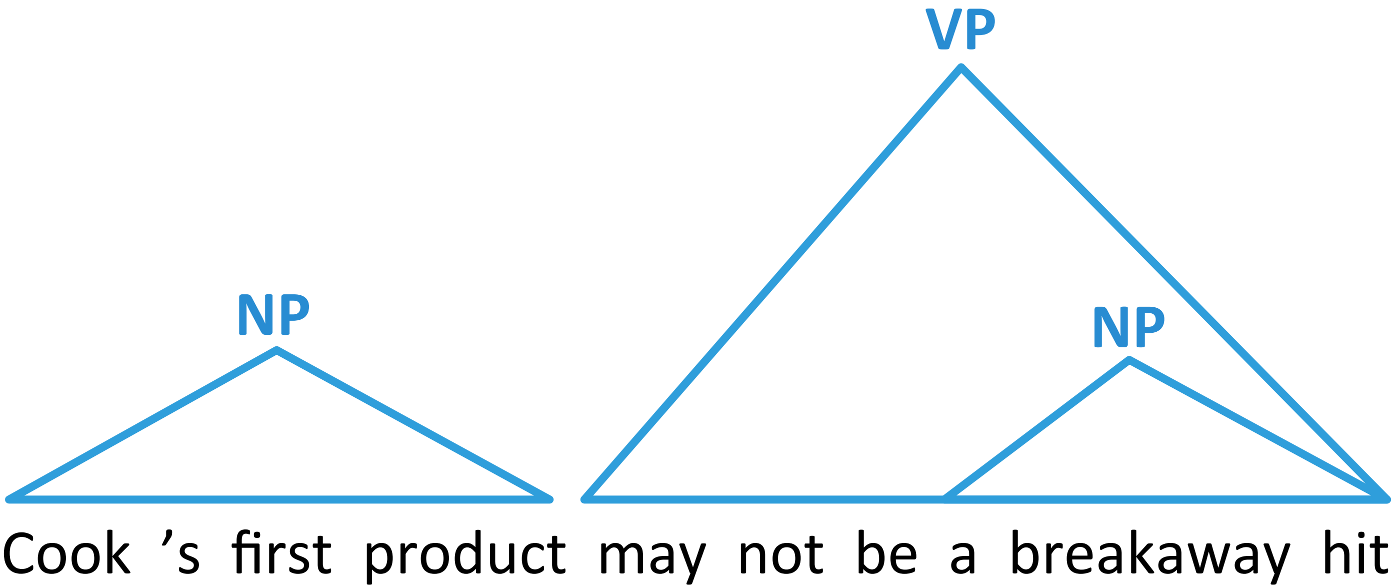
modal	verb	det.	adjective	noun	prep.	proper noun	punc.
would	be	a	breakaway	hit	for	Apple	.

Syntactic Parsing

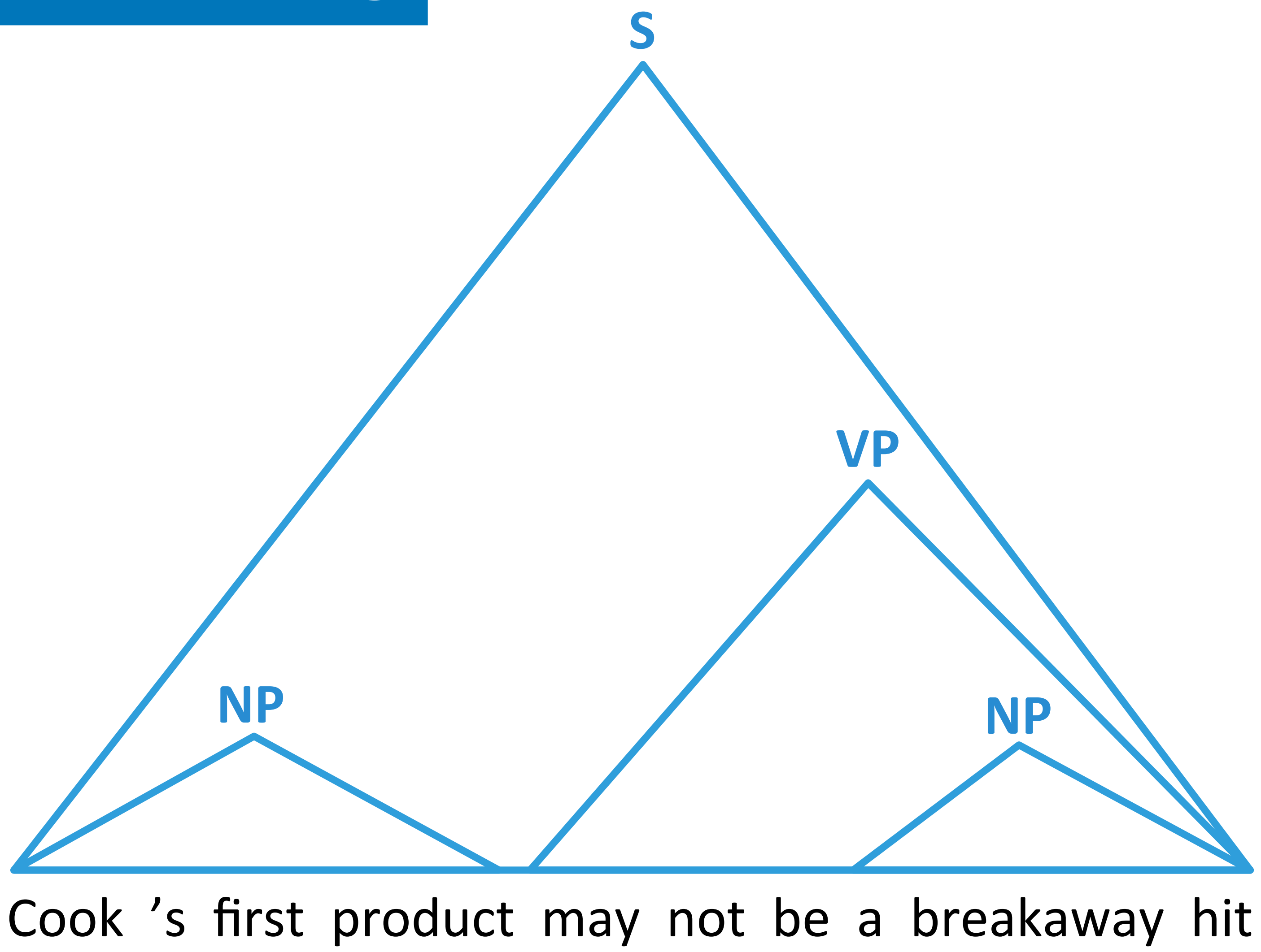


Cook ’s first product may not be a breakaway hit

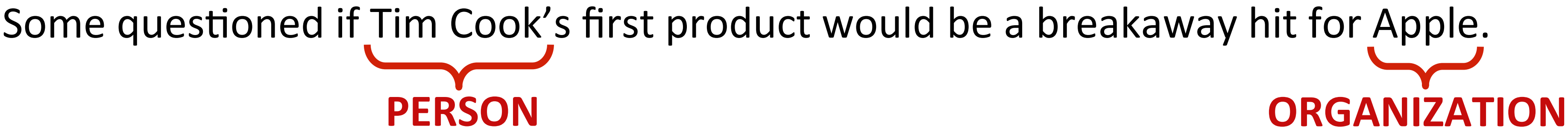
Syntactic Parsing



Syntactic Parsing



Named Entity Recognition



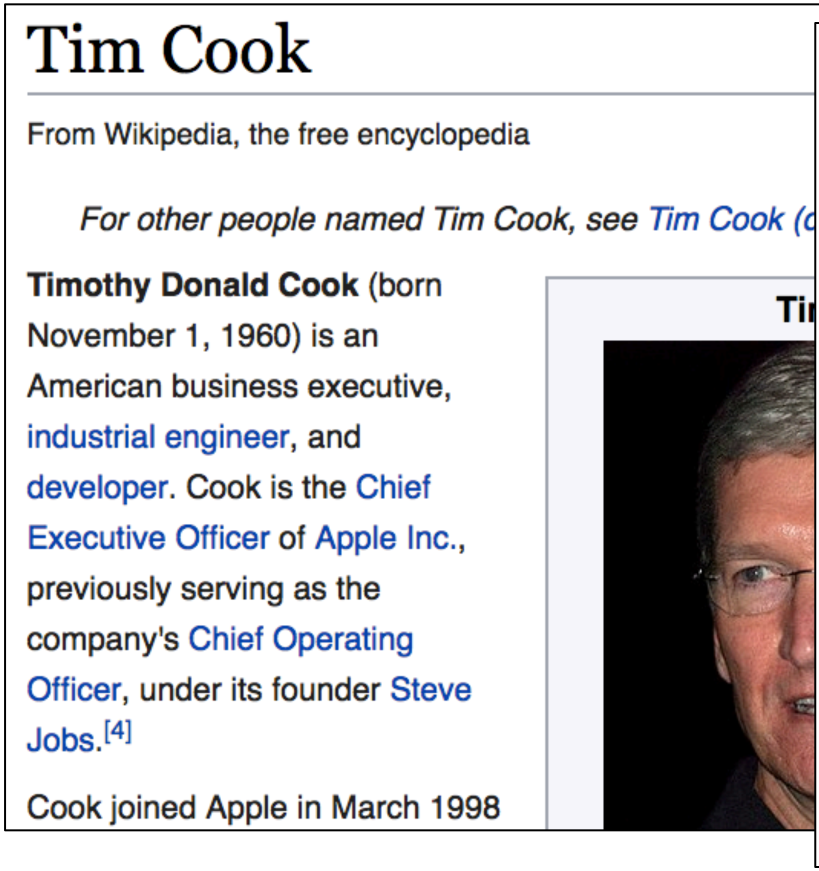
Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

PERSON

ORGANIZATION

Entity Linking

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.



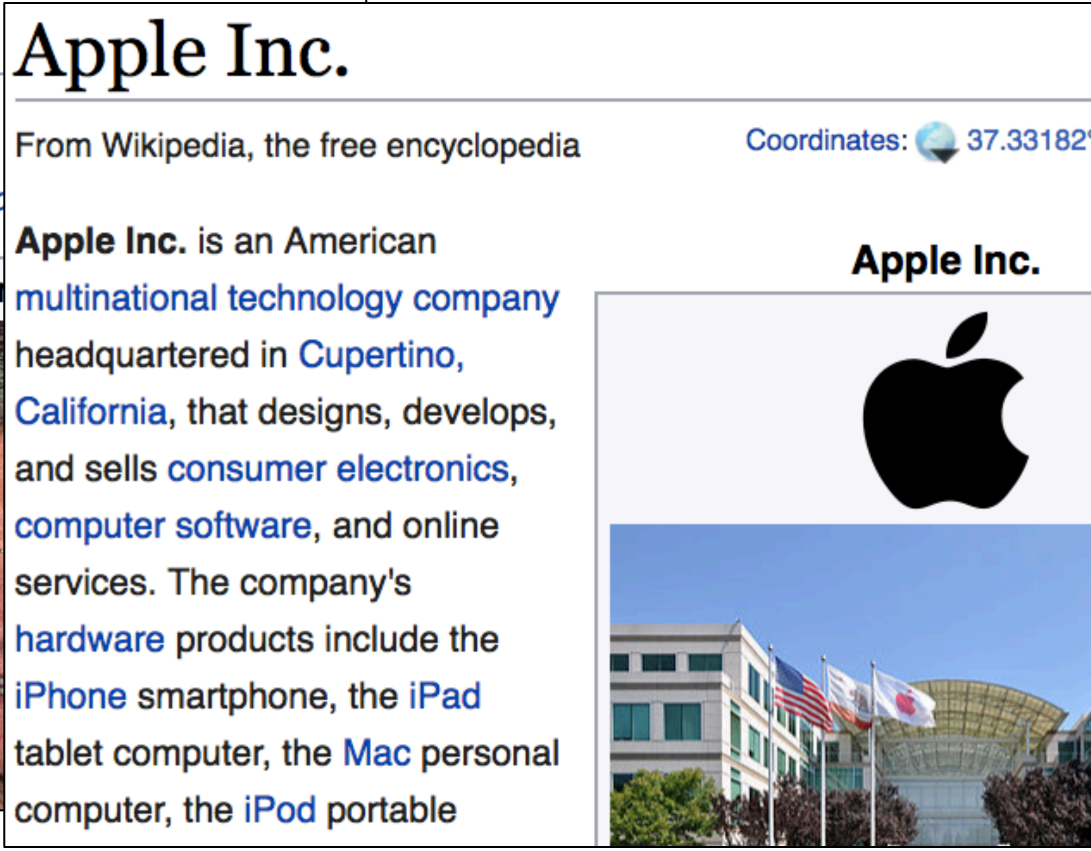
Tim Cook

From Wikipedia, the free encyclopedia

For other people named Tim Cook, see Tim Cook (d

Timothy Donald Cook (born November 1, 1960) is an American business executive, industrial engineer, and developer. Cook is the Chief Executive Officer of Apple Inc., previously serving as the company's Chief Operating Officer, under its founder Steve Jobs.[4]

Cook joined Apple in March 1998



Apple Inc.

From Wikipedia, the free encyclopedia

Coordinates: 37.33182°

Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops, and sells consumer electronics, computer software, and online services. The company's hardware products include the iPhone smartphone, the iPad tablet computer, the Mac personal computer, the iPod portable

Coreference Resolution

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

It's the company's first new device since he became CEO.

Coreference Resolution

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

It's the company's first new device since he became CEO.

Coreference Resolution

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

It's the company's first new device since he became CEO.

Coreference Resolution

Some questioned if Tim Cook's first product would be a breakaway hit for Apple.

??

It's the company's first new device since he became CEO.

“Winograd Schema” Coreference Resolution

The man couldn't lift his son because **he** was so weak.

The man couldn't lift his son because **he** was so heavy.

“Winograd Schema” Coreference Resolution

The man couldn't lift his son because **he** was so weak.

man

The man couldn't lift his son because **he** was so heavy.

son

Reading Comprehension

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

After school, Fritz drew a picture of his bicycle. His uncle said, “Don't draw your bicycle. Ride it!”

…

What did Fritz draw first?

A) the toothpaste

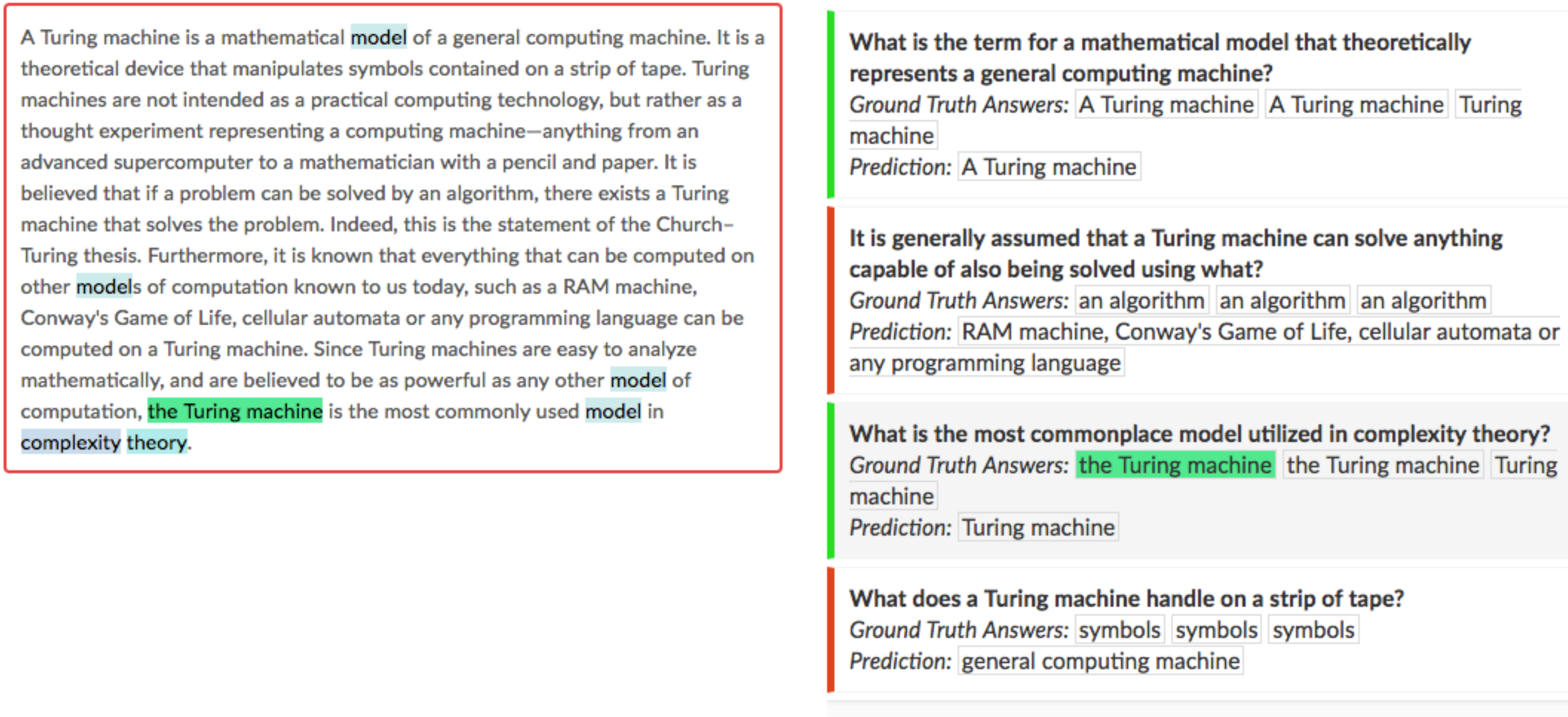
B) his mama

C) cereal and milk

D) his bicycle

MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Reading Comprehension



A Turing machine is a mathematical model of a general computing machine. It is a theoretical device that manipulates symbols contained on a strip of tape. Turing machines are not intended as a practical computing technology, but rather as a thought experiment representing a computing machine—anything from an advanced supercomputer to a mathematician with a pencil and paper. It is believed that if a problem can be solved by an algorithm, there exists a Turing machine that solves the problem. Indeed, this is the statement of the Church–Turing thesis. Furthermore, it is known that everything that can be computed on other models of computation known to us today, such as a RAM machine, Conway's Game of Life, cellular automata or any programming language can be computed on a Turing machine. Since Turing machines are easy to analyze mathematically, and are believed to be as powerful as any other model of computation, the Turing machine is the most commonly used model in complexity theory.

What is the term for a mathematical model that theoretically represents a general computing machine?
Ground Truth Answers: A Turing machine | A Turing machine | Turing machine
Prediction: A Turing machine

It is generally assumed that a Turing machine can solve anything capable of also being solved using what?
Ground Truth Answers: an algorithm | an algorithm | an algorithm
Prediction: RAM machine, Conway's Game of Life, cellular automata or any programming language

What is the most commonplace model utilized in complexity theory?
Ground Truth Answers: the Turing machine | the Turing machine | Turing machine
Prediction: Turing machine

What does a Turing machine handle on a strip of tape?
Ground Truth Answers: symbols | symbols | symbols
Prediction: general computing machine

SQuAD

The Stanford Question Answering Dataset

Sentence Similarity

Input	Output
Other ways are needed. We must find other ways.	4.4
Pakistan bomb victims’ families end protest Pakistan bomb victims to be buried after protest ends	2.6
I absolutely do believe there was an iceberg in those waters. I don't believe there was any iceberg at all anywhere near the Titanic.	1.2

Word Prediction

he bent down and searched the large container, trying to find anything else hidden in it other than the ______

Word Prediction

he turned to one of the cops beside him. “search the entire coffin.” the man nodded and bustled forward towards the coffin.

he bent down and searched the large container, trying to find anything else hidden in it other than the ______

Other language technologies (not typically considered core NLP):

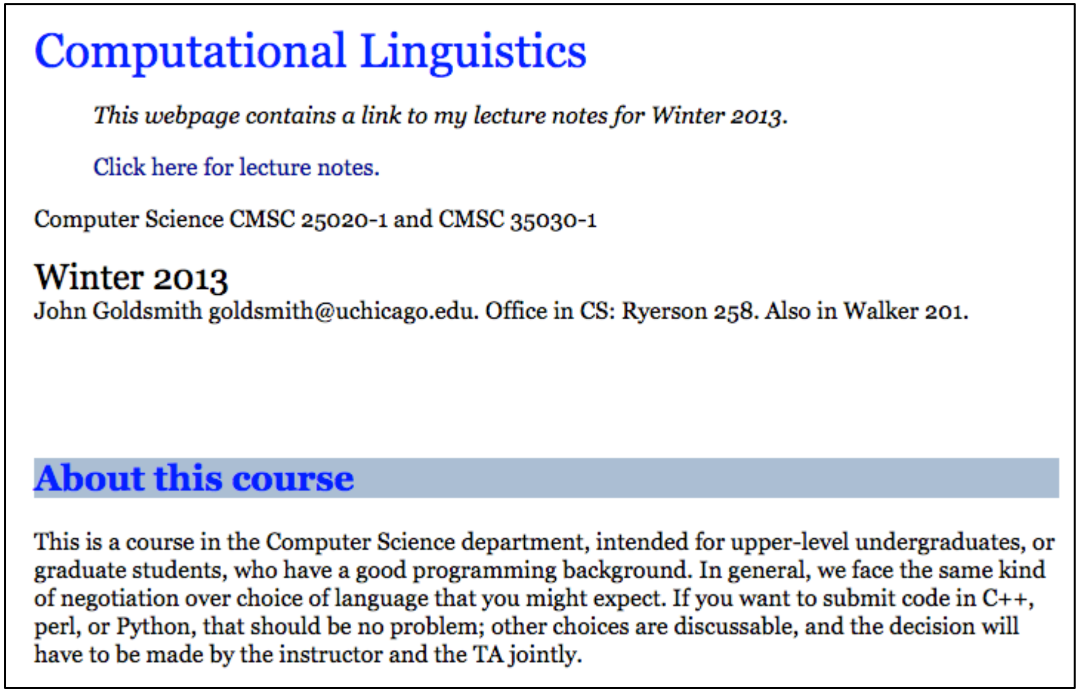
- speech processing (see TTIC 31110)
- information retrieval / web search
- knowledge representation / reasoning

Roadmap

- words, morphology, lexical semantics
- text classification
- simple neural methods for NLP
- language modeling and word embeddings
- recurrent/recursive/convolutional networks in NLP
- sequence labeling, HMMs, dynamic programming
- syntax and syntactic parsing
- semantics, compositionality, semantic parsing
- machine translation and other NLP tasks

Computational Linguistics vs. Natural Language Processing

- how do they differ?



Computational Linguistics

This webpage contains a link to my lecture notes for Winter 2013.

Click here for lecture notes.

Computer Science CMSC 25020-1 and CMSC 35030-1

Winter 2013

John Goldsmith goldsmith@uchicago.edu. Office in CS: Ryerson 258. Also in Walker 201.

About this course

This is a course in the Computer Science department, intended for upper-level undergraduates, or graduate students, who have a good programming background. In general, we face the same kind of negotiation over choice of language that you might expect. If you want to submit code in C++, perl, or Python, that should be no problem; other choices are discussable, and the decision will have to be made by the instructor and the TA jointly.

Computational Linguistics vs. Natural Language Processing

- English is a “head-final” language: the head of a noun phrase comes at the end

- computational linguistics is about **linguistics**
  - **computational** is a modifier
- natural language processing is about **processing**
  - **natural language** is a modifier

Computational Linguistics vs. Natural Language Processing

- many people think of the two terms as synonyms

- computational linguistics is more inclusive; more likely to include sociolinguistics, cognitive linguistics, and computational social science

- NLP is more likely to use machine learning and involve engineering / system-building

Is NLP Science or Engineering?

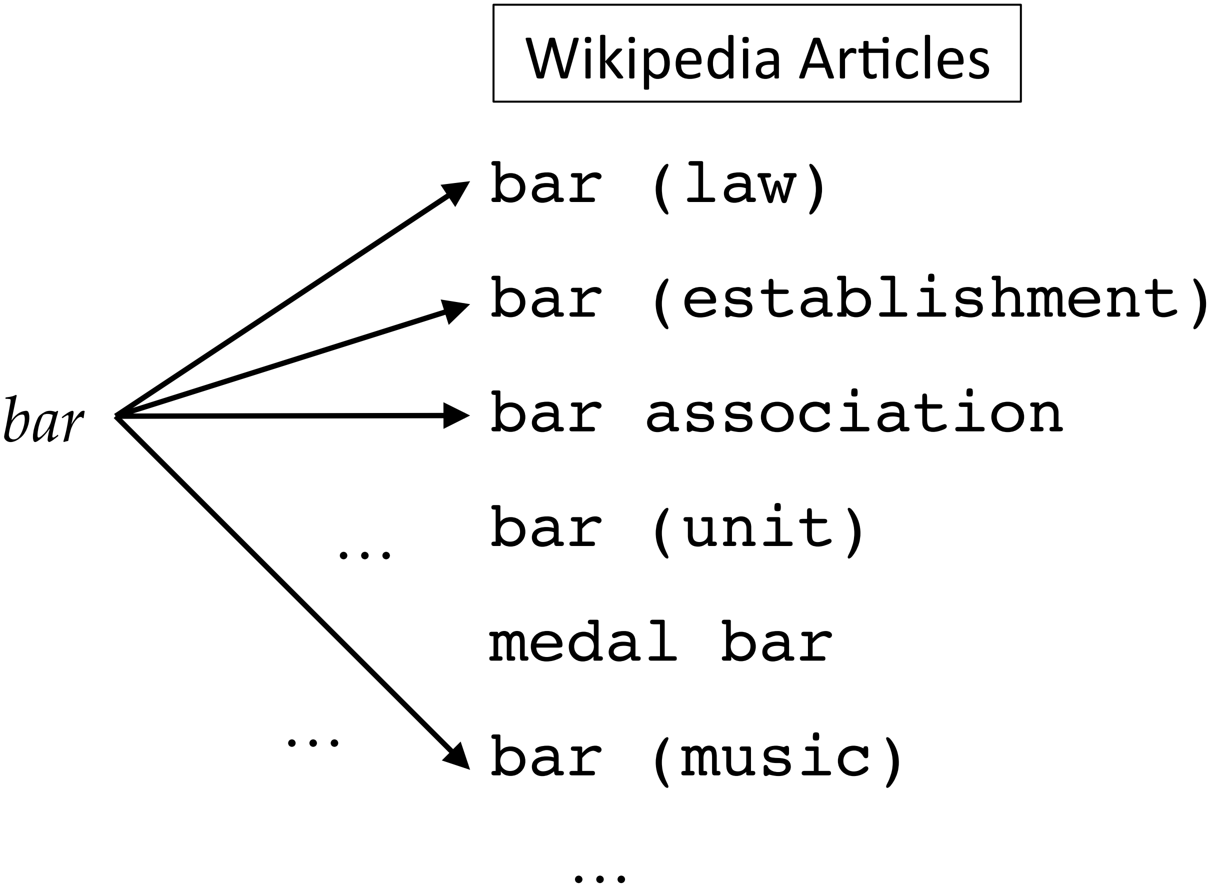
- goal of NLP is to develop technology, which takes the form of engineering
- though we try to solve today's problems, we seek principles that will be useful for the future
- if science, it's not linguistics or cognitive science; it's the science of computational processing of language
- I like to think of NLP as the science of engineering solutions to problems involving natural language

Why is NLP hard?

- ambiguity and variability of linguistic expression:
  - **variability**: many forms can mean the same thing
  - **ambiguity**: one form can mean many things

- many different kinds of variability and ambiguity
- each NLP task must address distinct kinds

Example: Hyperlinks in Wikipedia



Wikipedia Articles

bar

`bar (law)`

`bar (establishment)`

`bar association`

`bar (unit)`

`medal bar`

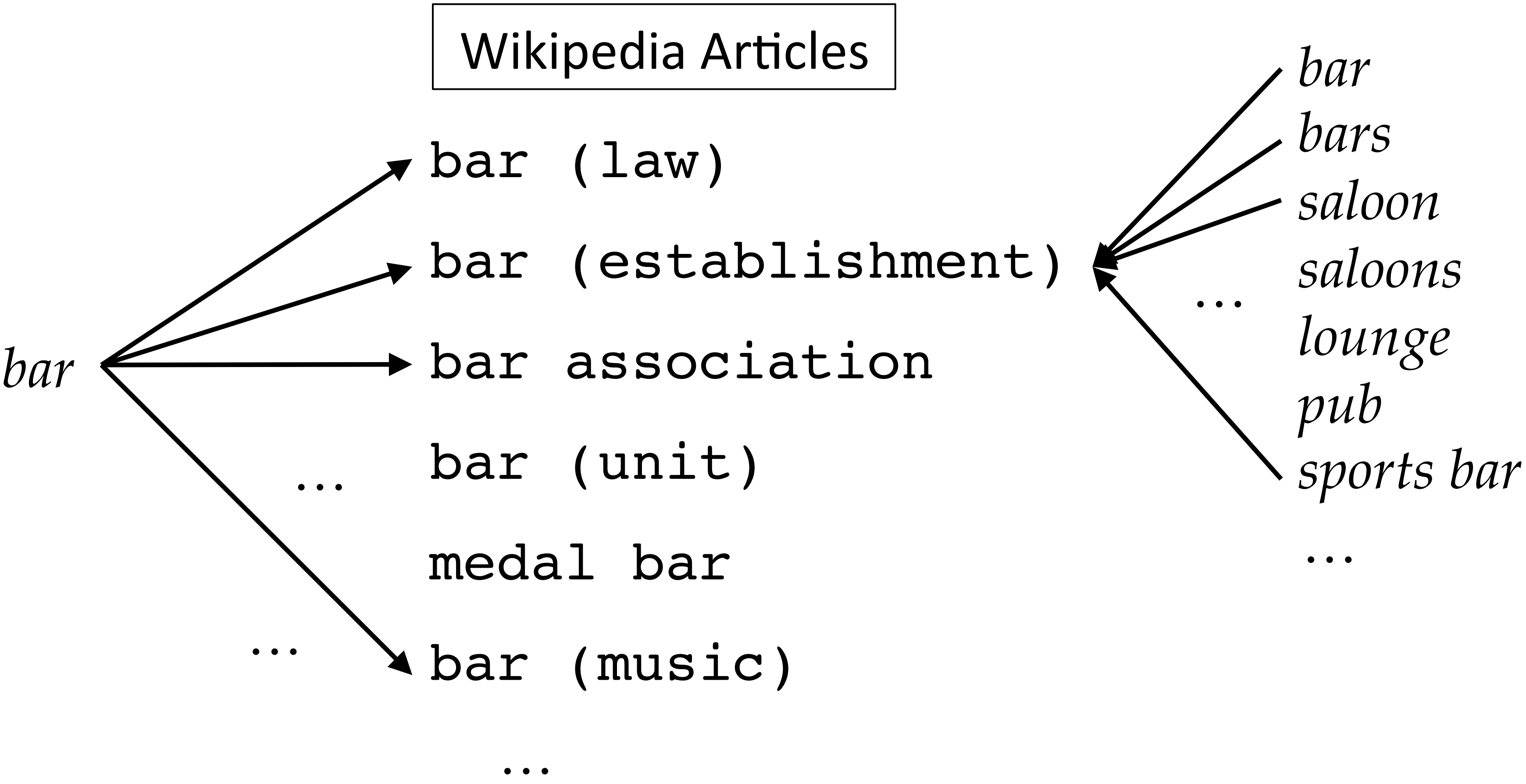
`bar (music)`

…

…

…

Example: Hyperlinks in Wikipedia



Wikipedia Articles

bar

bar (law)

bar (establishment)

bar association

bar (unit)

medal bar

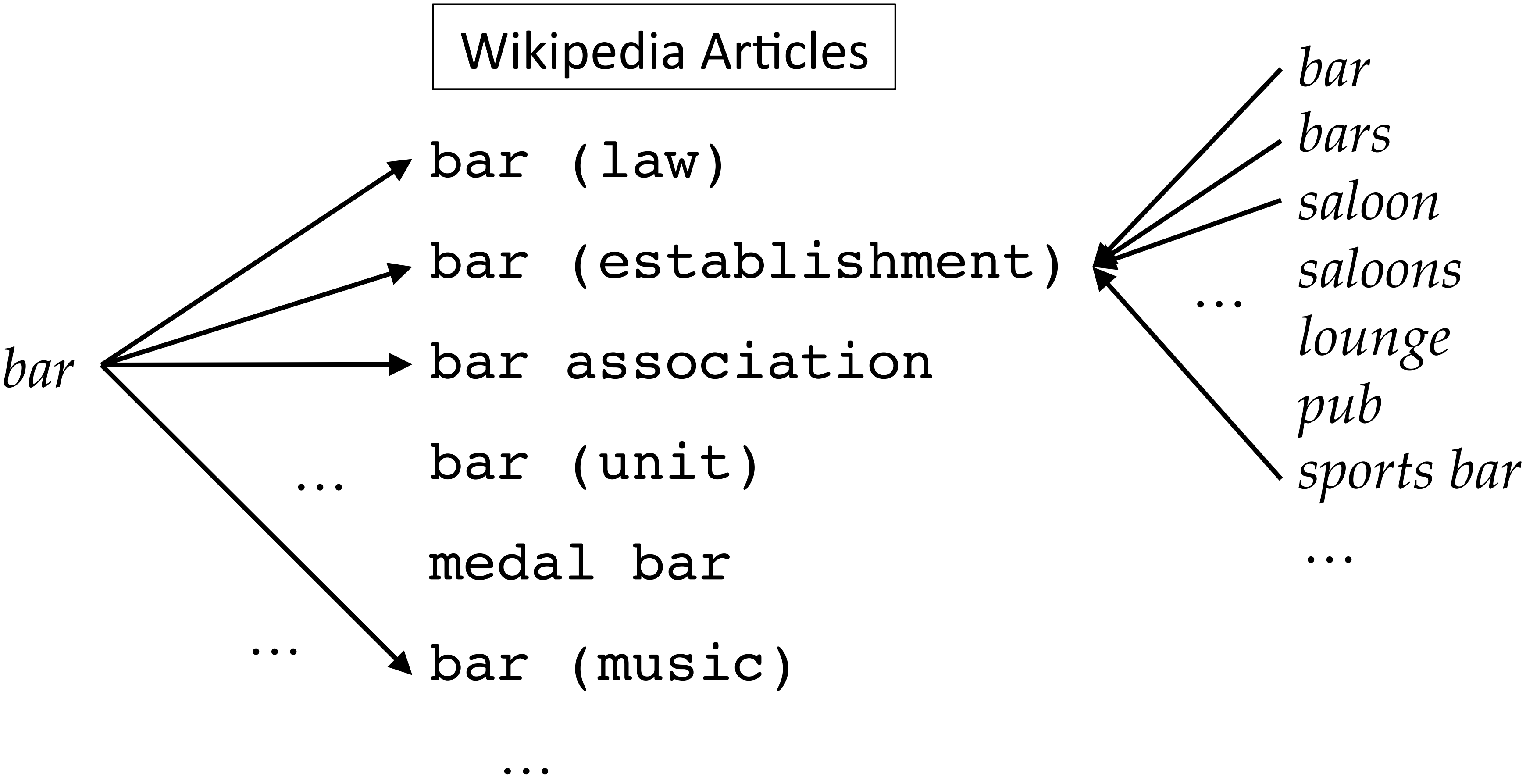
bar (music)

…

bar
bars
saloon
saloons
lounge
pub
sports bar
…

Ambiguity

Variability



Wikipedia Articles

bar

`bar (law)`

`bar (establishment)`

`bar association`

`bar (unit)`

`medal bar`

`bar (music)`

…

…

…

bar

bars

saloon

saloons

lounge

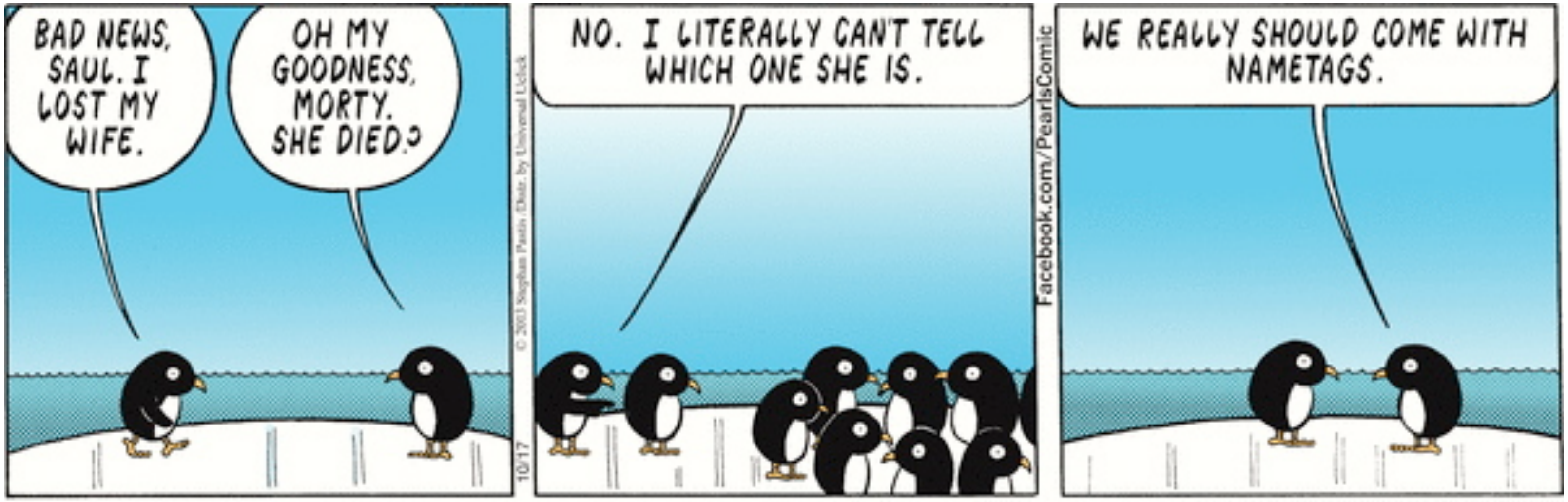
pub

sports bar

…

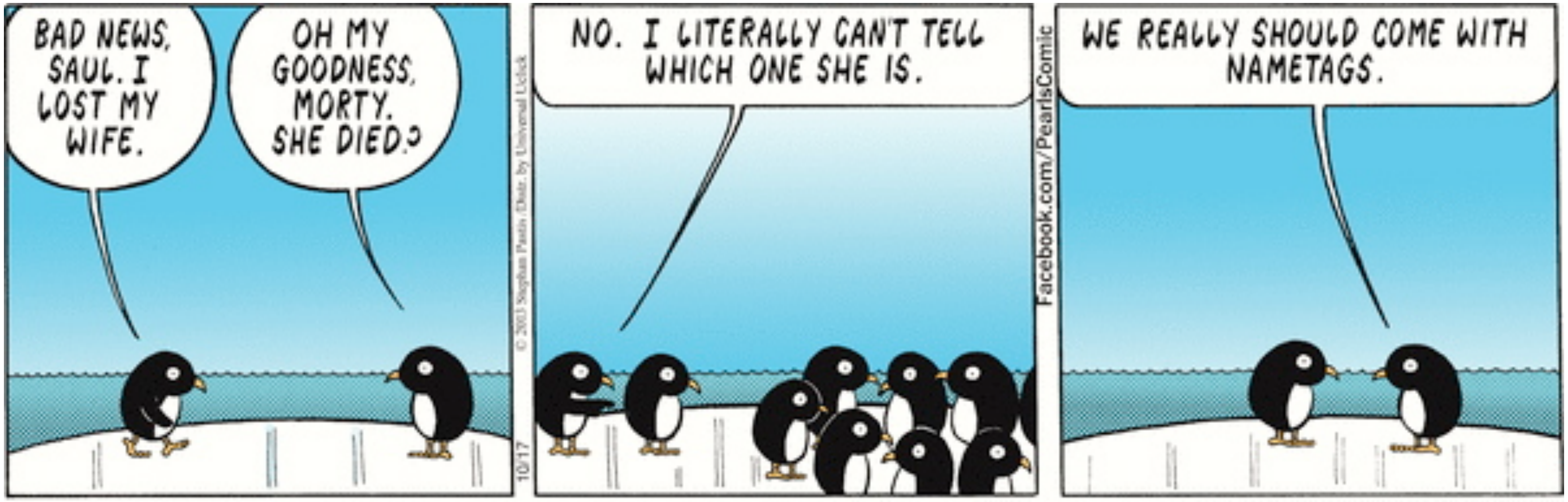
…

Word Sense Ambiguity

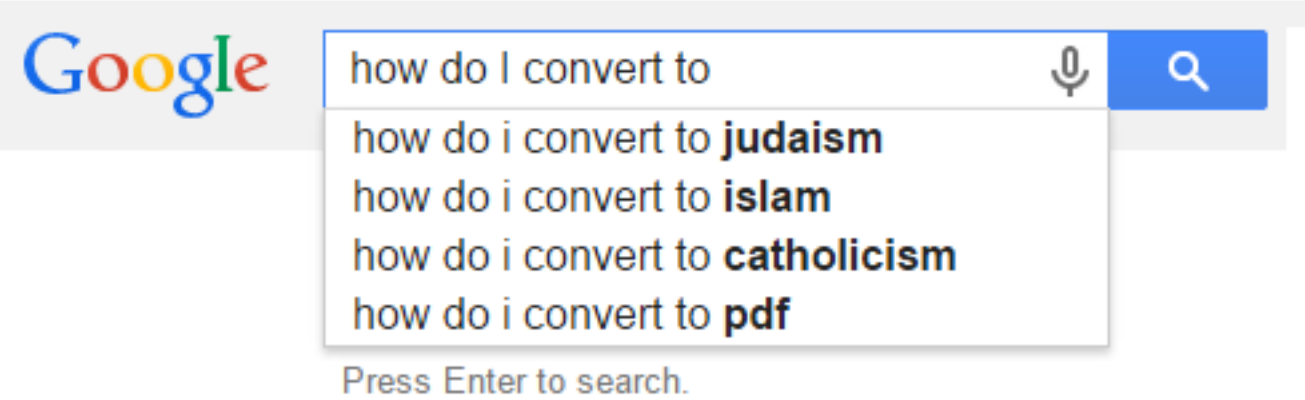

BAD NEWS, SAUL. I LOST MY WIFE.
OH MY GOODNESS, MORTY. SHE DIED?
© 2013 Stephan Pastis /Distr. by Universal Uclick
10/17
NO. I LITERALLY CAN'T TELL WHICH ONE SHE IS.
Facebook.com/PearlsComic
WE REALLY SHOULD COME WITH NAMETAGS.

credit: A. Zwicky

Word Sense Ambiguity


BAD NEWS, SAUL. I LOST MY WIFE.
OH MY GOODNESS, MORTY. SHE DIED?
© 2013 Stephan Pastis /Distr. by Universal Uclick
10/17
NO. I LITERALLY CAN'T TELL WHICH ONE SHE IS.
Facebook.com/PearlsComic
WE REALLY SHOULD COME WITH NAMETAGS.

credit: A. Zwicky



how do I convert to

how do i convert to **judaism**

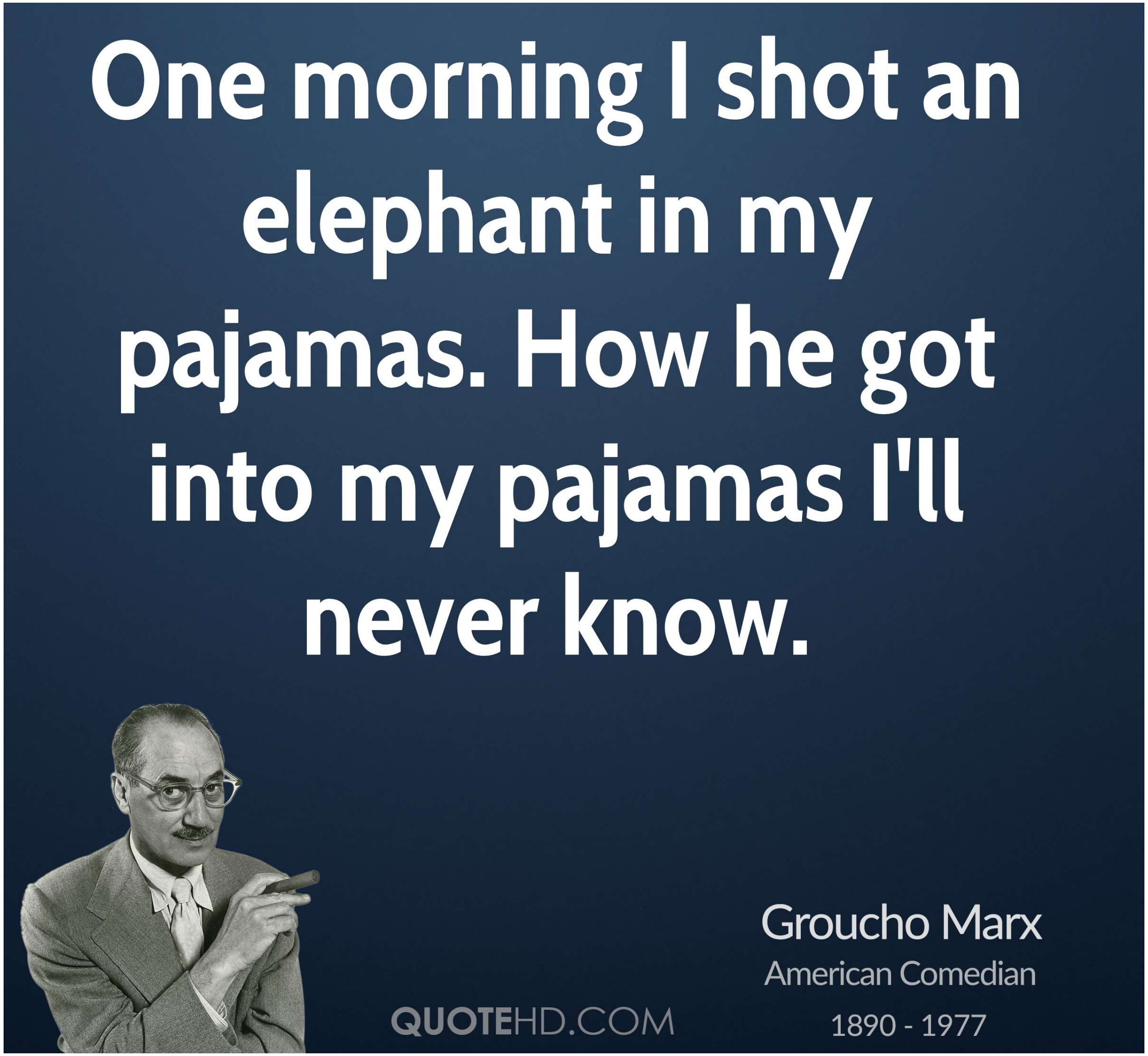
how do i convert to **islam**

how do i convert to **catholicism**

how do i convert to **pdf**

Press Enter to search.

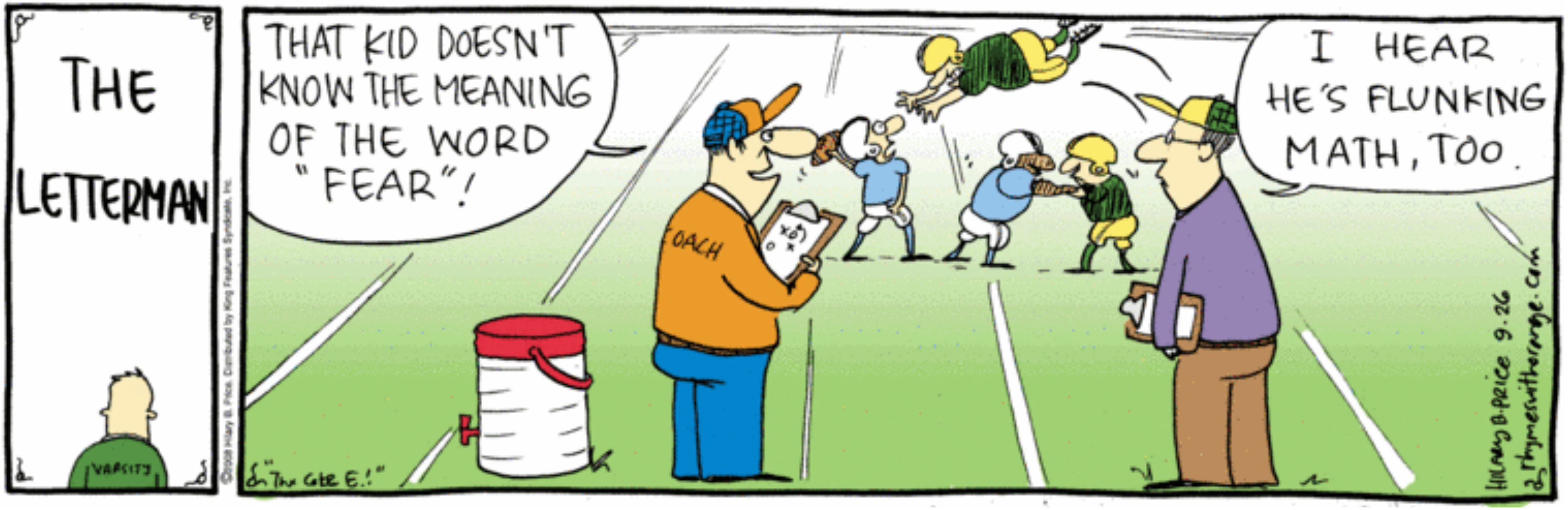
Attachment Ambiguity



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

Groucho Marx
American Comedian
1890 - 1977

Meaning Ambiguity


THE LETTERMAN
VARSITY
THAT KID DOESN'T KNOW THE MEANING OF THE WORD "FEAR"!
COACH
I HEAR HE'S FLUNKING MATH, TOO.
"The Gate E.!"
©2008 Hilary B. Price. Distributed by King Features Syndicate, Inc.
Hilary B. Price 9.26
rhymeswithorange.com

Roadmap

- words, morphology, lexical semantics
- text classification
- simple neural methods for NLP
- language modeling and word embeddings
- recurrent/recursive/convolutional networks in NLP
- sequence labeling, HMMs, dynamic programming
- syntax and syntactic parsing
- semantics, compositionality, semantic parsing
- machine translation and other NLP tasks

Words

- what is a word?
- tokenization
- morphology
- lexical semantics

What is a word?

Tokenization

- tokenization: convert a character stream into words by adding spaces
- for certain languages, highly nontrivial
- e.g., Chinese word segmentation is a widely-studied NLP task

Tokenization

- for other languages (English), tokenization is easier but is still not always obvious
- the data for your homework has been tokenized:
  - punctuation has been split off from words
  - contractions have been split

Intricacies of Tokenization

- separating punctuation characters from words?
  - *, ” ? !* → always separate
  - *.* → when shouldn’t we separate it?

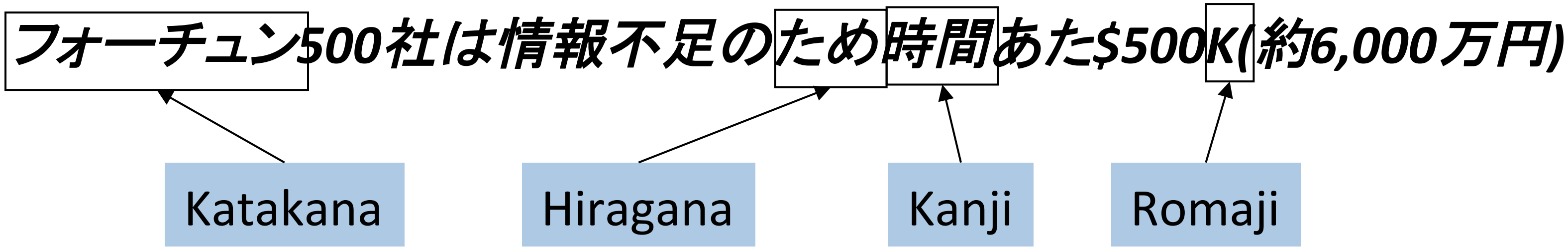
Intricacies of Tokenization

- separating punctuation characters from words?
  - *, ” ? !* → always separate
  - . → when shouldn’t we separate it?
    - *Dr., Mr., Prof., U.S., etc.*

Intricacies of Tokenization

- separating punctuation characters from words?
  - *, ” ? !* → always separate
  - *.* → when shouldn’t we separate it?
    - *Dr., Mr., Prof., U.S., etc.*
- English contractions:
  - *isn’t, aren’t, wasn’t,…* → *is n’t, are n’t, was n’t,…*
  - but how about these: *can’t, won’t* → *ca n’t, wo n’t*
  - *ca* and *wo* are then different forms from *can* and *will*

- Chinese and Japanese: no spaces between words:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova now lives in US southeastern Florida
- Further complicated in Japanese, with multiple alphabets intermingled
  - Dates/amounts in multiple formats



フォーチュン500社は情報不足のため時間あた$500K(約6,000万円)

J&M/SLP3

Removing Spaces?

- tokenization is usually about adding spaces
- but might we also want to remove spaces?
- what are some English examples?

Removing Spaces?

- tokenization is usually about adding spaces
- but might we also want to remove spaces?
- what are some English examples?
  - names?
    - New York → NewYork
  - non-compositional compounds?
    - hot dog → hotdog
  - other artifacts of our spacing conventions?
    - New York-Long Island Railway → ?

Types and Tokens

- once text has been tokenized, let's count the words
- types: entries in the vocabulary
- tokens: instances of types in a corpus
- example sentence: *If they want to go , they should go .*
  - how many types?
  - how many tokens?

Types and Tokens

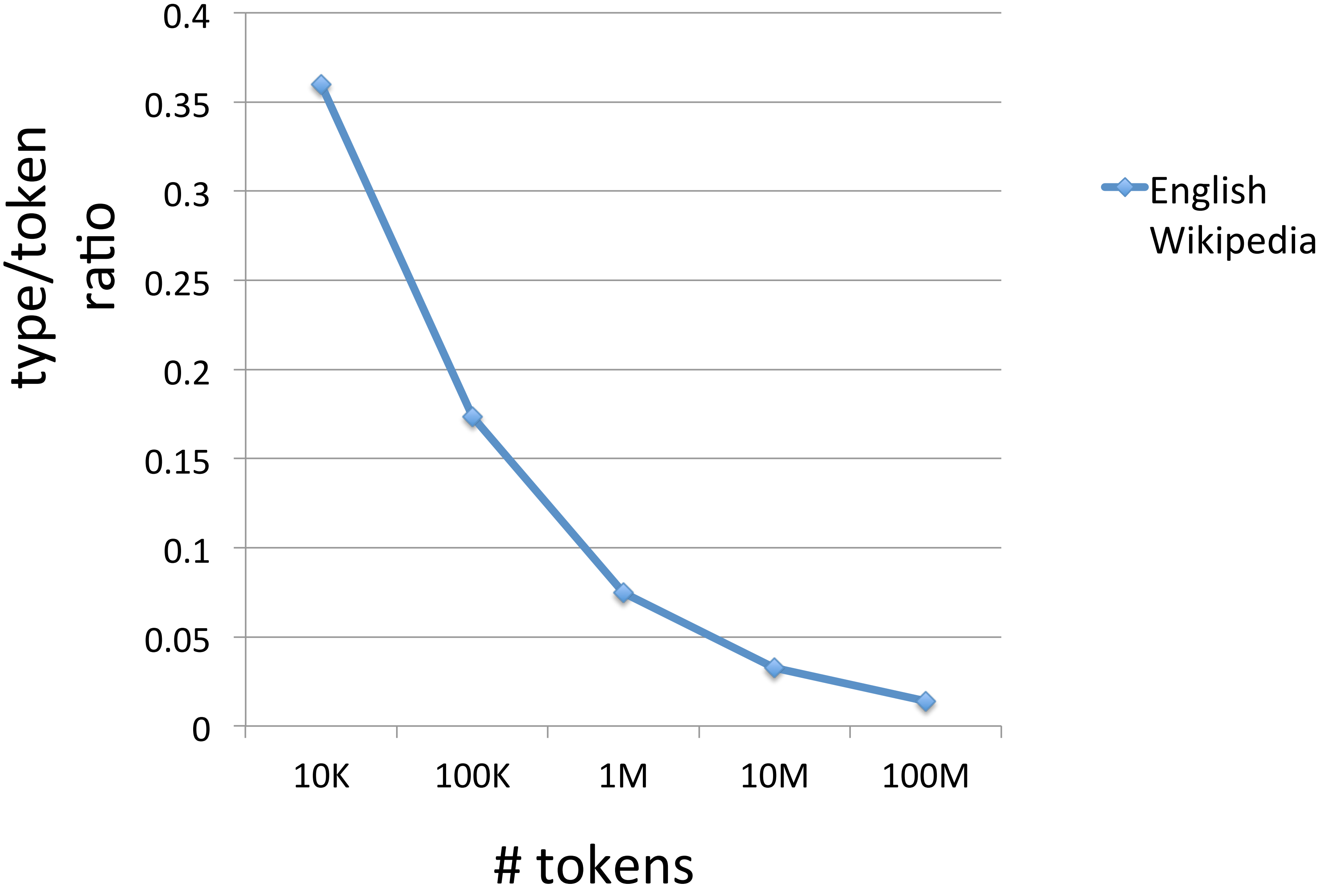
- once text has been tokenized, let's count the words
- types: entries in the vocabulary
- tokens: instances of types in a corpus
- example sentence: *If they want to go , they should go .*
  - how many types? 8
  - how many tokens? 10
- type/token ratio: useful statistic of a corpus (here, 0.8)

Types and Tokens

- once text has been tokenized, let’s count the words
- types: entries in the vocabulary
- tokens: instances of types in a corpus
- example sentence: *If they want to go , they should go .*
  - how many types? 8
  - how many tokens? 10
- type/token ratio: useful statistic of a corpus (here, 0.8)
- as we add data, what happens to the type-token ratio?

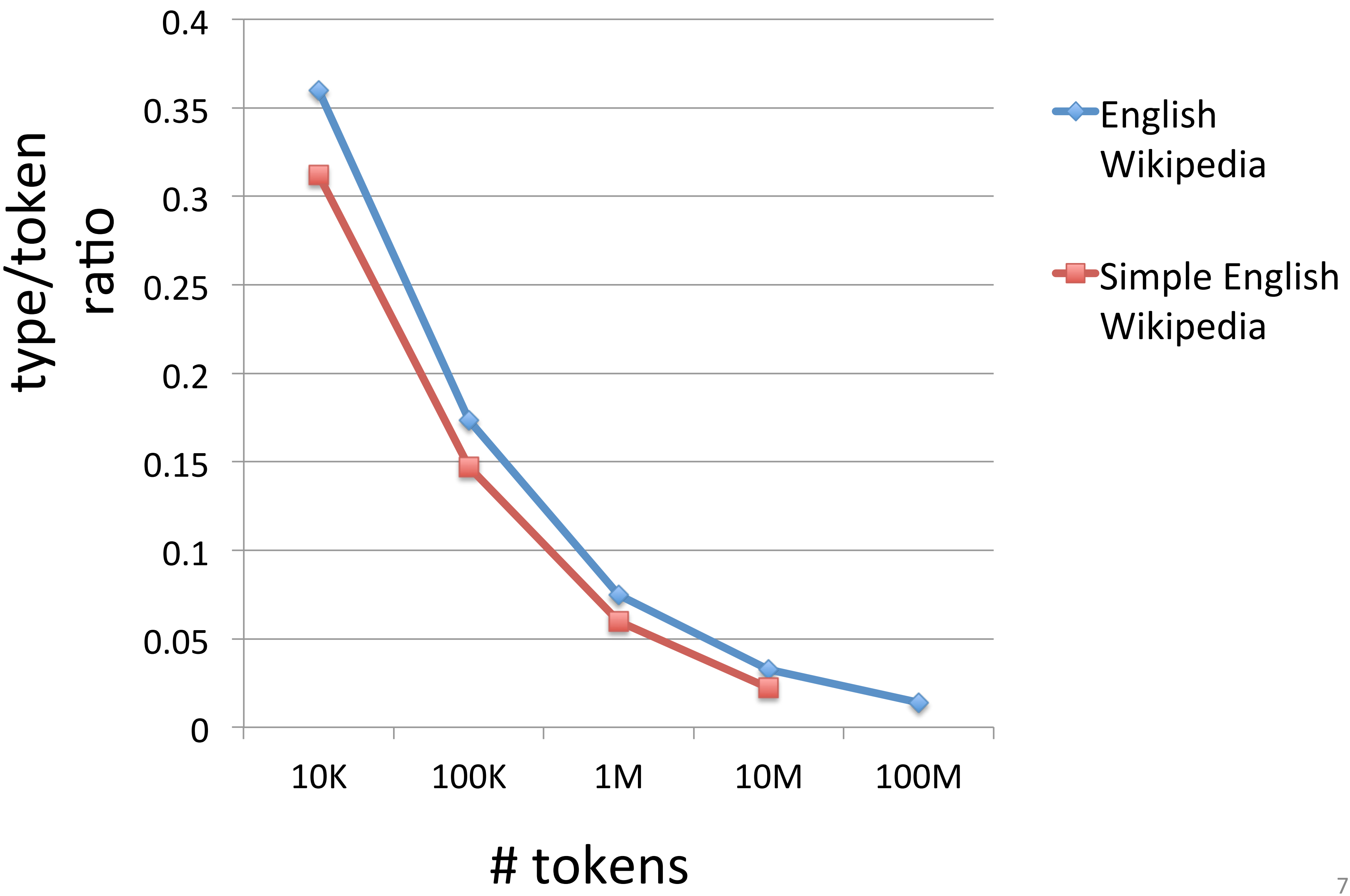
- How will the type/token ratio change when adding more data?

More data → Lower type/token ratio



- What has a higher type/token ratio,

Simple English Wikipedia or English Wikipedia?

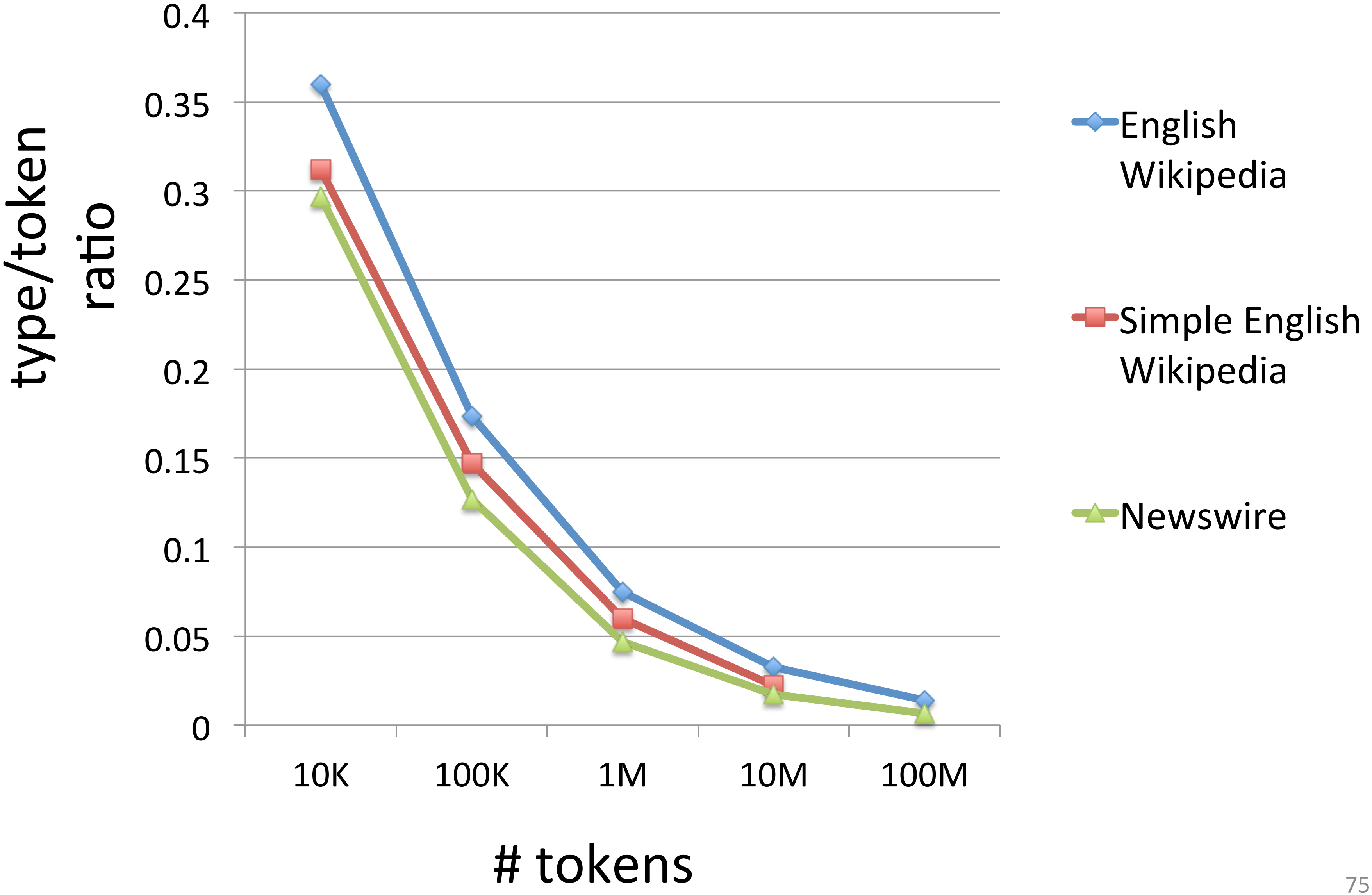

type/token ratio
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10K
100K
1M
10M
100M
# tokens
English Wikipedia
Simple English Wikipedia

- What has a higher type/token ratio,

Simple English Wikipedia or English Wikipedia?

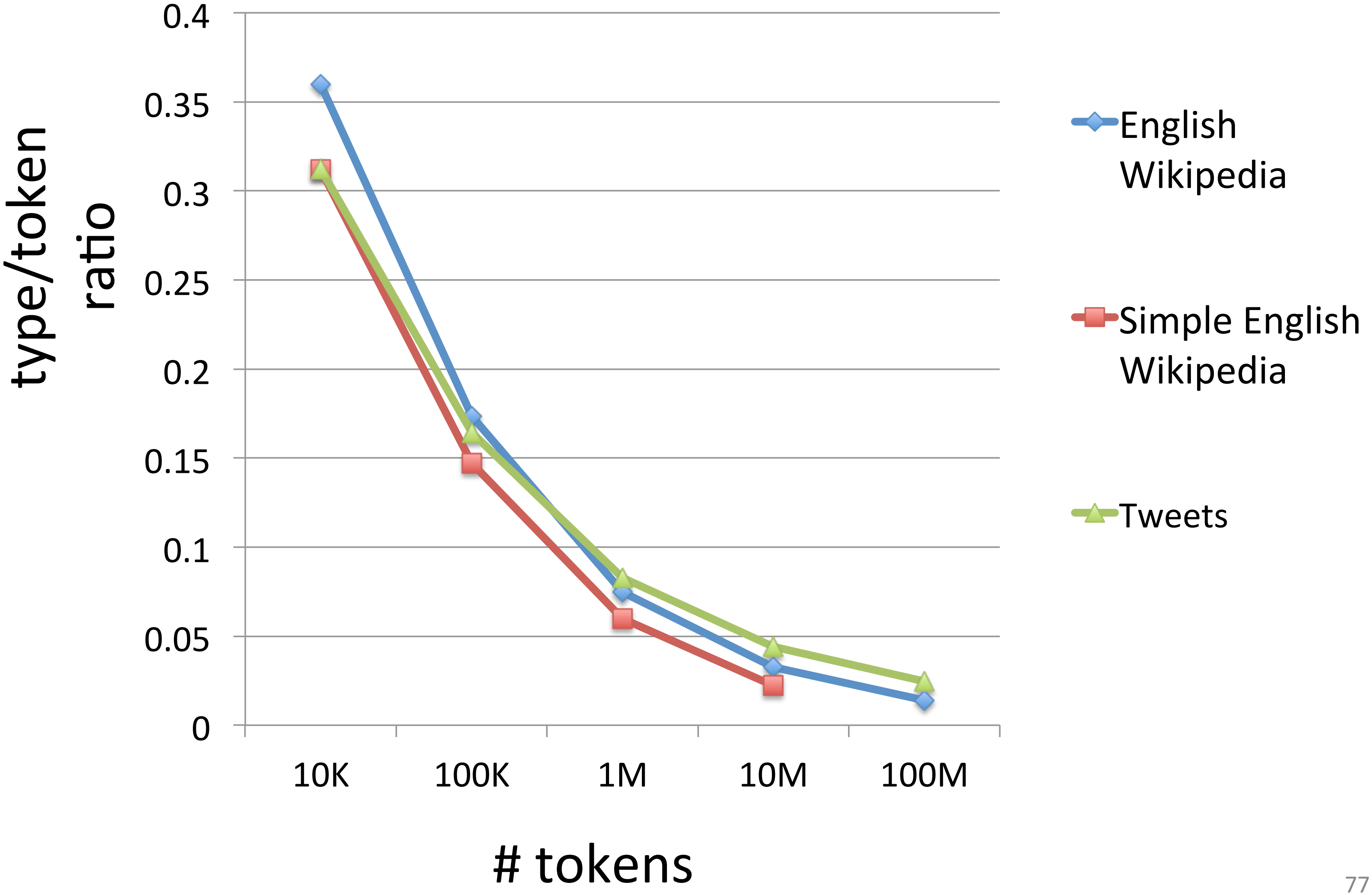
  - English Wikipedia
  - type/token ratio is one measure of complexity

- How about Wikipedia vs Newswire?



type/token ratio
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10K
100K
1M
10M
100M
# tokens
English Wikipedia
Simple English Wikipedia
Newswire

- Wikipedia vs Simple English Wikipedia?
  - Wikipedia
- Wikipedia vs Newswire?
  - Wikipedia
- Wikipedia vs Tweets?


type/token ratio
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10K
100K
1M
10M
100M
# tokens
English Wikipedia
Simple English Wikipedia
Tweets

- Wikipedia vs Simple English Wikipedia?
  - Wikipedia
- Wikipedia vs Newswire?
  - Wikipedia
- Wikipedia vs Tweets?
  - Tweets (once you have 1 million or more tokens)

“really” on Twitter

```
224571 really             50 reallllllly      15 reallllyy
  1189 rly                48 reeeeeally       15 reallllllllly
  1119 realy              41 reeally          15 reaallly
   731 rlly               38 really2          14 reeeeeeally
   590 reallly            37 reaaaaally       14 reallllyyyy
   234 realllly           35 reallyyyyy       13 reeeaaally
   216 reallyy            31 reely            12 rreally
   156 relly              30 realllyyy        12 reaaaaaally
   146 reallllly          27 realllyy         11 reeeeallly
   132 rily               27 reaaly           11 reeeallly
   104 reallyyy           26 realllyyyy       11 realllllyyy
    89 reeeally           25 realllllllly     11 reaallyy
    89 realllllly         22 reaaallly        10 reallyreallyreally
    84 reaaally           21 really-          10 reaaaly
    82 reaally            19 reeaally          9 reeeeeeeally
    72 reeeeally          18 reallllyyy        9 reallys
    65 reaaaally          16 reaaaallly        9 really-really
    57 reallyyyy          15 realyy            9 r)eally
    53 rilly              15 reallyreally      8 reeeaally
```

“really” on Twitter

```
8 reallyyyyyyy
8 reallyyyyyy
8 realky
7 relaly
7 reeeeeeeeeally
7 reeeealy
7 reeeeaaally
7 reallllllyyy
7 realllllllllllly
7 reaaaaaaally
7 raelly
7 r3ally
6 r-really
6 reeeaaalllyyy
6 reeeaaallly
6 reeeaaaally
6 realyl
6 r-e-a-l-l-y
6 realllyyyyy
6 reallllllllllly
6 reaaaaaallly
5 rrrreally
5 rrly
5 rellly
5 reeeeeeeeally
5 reeeeaally
5 reeeeaaallly
5 reeallyyy
5 realllllllllllly
5 reallllllllllllly
5 reaalllyy
5 reaaaalllly
5 reaaaaallly
4 rllly
4 reeeeeeeeeeeally
4 reeealy
4 reeaaaally
4 realllllyyyy
4 realllllllyyyy
4 reaalllyyy
4 reaalllly
4 reaaalllyy
4 reaaalllly
4 reaaaaly
3 reeeeealllly
3 reeeealllly
3 reeeeaaaaally
3 reeeaallly
3 reeeaaallllyyy
3 reealy
3 reeallly
3 reeaaly
3 reeaalllyyy
3 reeaalllly
3 reeaaallly
3 reallyyyyyyyyy
3 reallyl
```

“really” on Twitter

```
3 really)
3 r]eally
3 realluy
3 reallllyyyyy
3 reallllllyyyyyyy
3 reallllllyyyy
3 reallllllyy
3 reallllllllllllllllly
3 realiy
3 reaallyyyy
3 reaallllly
3 reaaallyy
3 reaaaallyy
3 reaaaallllly
3 reaaaaaly
3 reaaaaaaaally
3 r34lly
2 rrreally
2 rreeaallyy
2 rlyyyy
2 rlyyy
2 reqally
2 rellyy
2 rellys
2 reeely
2 reeeeeealy
2 reeeeeallly
2 reeeeeaally
2 reeeeeaaally
2 reeeeeaaallllly
2 reeeeallyyy
2 reeeeallllyyy
2 reeeeaaallllyyyy
2 reeeeaaalllly
2 reeeeaaaally
2 reeeeaaaalllyyy
2 reeeallyy
2 reeallyy
2 reeaallyy
2 reeaalllyy
2 reeaallly
2 reeaaally
2 reaqlly
2 realyyy
2 reallyyyyyyyyyyyy
2 reallyyyyyyyy
2 realllyyyyyy
2 reallllyyyyyy
2 realllllyyyyyy
2 realllllyy
2 reallllllyyyyy
2 realllllllyyyyy
2 realllllllyy
2 realllllllllllllllly
2 reallllllllllllllllllly
```

```
1 rrrrrrrrrrrrrrreeeeeeeeeeeeaaaaaallllllllyyyyyy
1 rrrrrrrrrreally
1 rrrrrreeeeeeeaaaalllllyyyyyyy
1 rrrrrrealy
1 rrrrrreally
…
1 re-he-he-heeeeally
1 re-he-he-he-ealy
1 reheheally
1 reelllyy
1 reellly
1 ree-hee-heally
…
1 reeeeeeeeeaally
1 reeeeeeeeeaaally
1 reeeeeeeeeaaaaaalllyyy
1 reeeeeeeeeaaaaaaallllllllyyyyyyyy
1 reeeeeeeeeaaaaaaallllllllyyyyyyyy
1 reeeeeeeeeaaaaaaaaalllllllllyyyyyyyy
1 reeeeeeeeaaaaaaaalllllyyyyyy
```

```
1 reallyreallyreallyreallyreallyreallyreallyreallyreallyreally
  reallyreallyreallyreallyreallyreallyreally
1 reallyreallyreallyreallyreallyr33lly
1 really/really/really
1 really(really
…
1 reallllllllyyyy
1 realllllllllyyyyyy
1 realllllllllyyyyy
1 realllllllllyyyy
1 realllllllllyyy
1 reallllllllllyyyyy
1 reallllllllllllyyyyyy
1 realllllllllllllllllllly
1 realllllllllllllllllllllly
1 realllllllllllllllllllllllyyyyy
1 realllllllllllllllllllllllllllly
1 reallllllllllllllllllllllllllllly
1 realllllllllllllllllllllllllllllllllly
1 realllllllllllllllllllllllllllllllllllllllllllllly
1 realllllllllllllllllllllllllllllllllllllllllllllllllllllllly
1 reallllllllllllllllllllllllllllllllllllllllllllllllllllllllllll
  llllllly
```

How many words are there?

- how many English words exist?
- when we increase the size of our corpus, what happens to the number of types?

How many words are there?

- how many English words exist?
- when we increase the size of our corpus, what happens to the number of types?
  - a bit surprising: vocabulary continues to grow in any actual dataset
  - you'll just never see all the words
  - in 1 million tweets, 15M tokens, 600k types
  - in 56 million tweets, 847M tokens, ? types

How many words are there?

- how many English words exist?
- when we increase the size of our corpus, what happens to the number of types?
  - a bit surprising: vocabulary continues to grow in any actual dataset
  - you'll just never see all the words
  - in 1 million tweets, 15M tokens, 600k types
  - in 56 million tweets, 847M tokens, 11M types