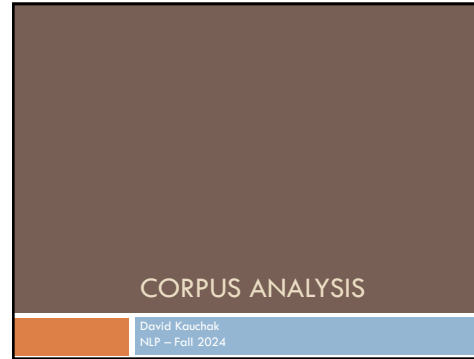
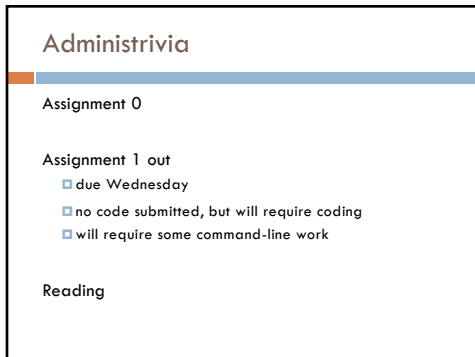




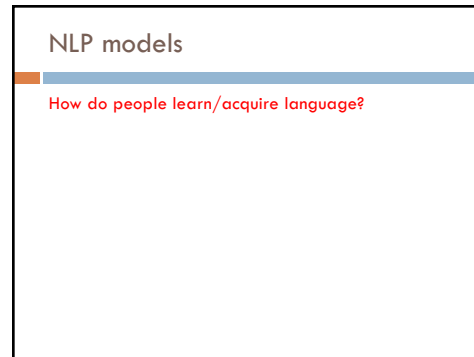
1



2



3



4

## NLP models

A lot of debate about how human's learn language

- ▣ Rationalist (e.g. Chomsky)
- ▣ Empiricist

From my perspective (and many people who study NLP)...

- ▣ I don't care :)

Strong AI vs. weak AI: don't need to accomplish the task the same way people do, just the same task

- ▣ Machine learning
- ▣ Statistical NLP

5

## Vocabulary

Word

- ▣ a unit of language that native speakers can identify
- ▣ words are the blocks from which sentences are made

Sentence

- ▣ a string of words satisfying the grammatical rules of a language

Document

- ▣ A collection of sentences

Corpus

- ▣ A collection of related texts

6

## Corpus examples

Any you've seen or played with before?

7

## Corpus characteristics

What are some defining characteristics of corpora?

8

## Corpus characteristics

- monolingual vs. parallel
- language
- annotated (e.g. parts of speech, classifications, etc.)
- source (where it came from)
- size

9

## Corpus examples

- Linguistic Data Consortium
  - ▣ <http://www.ldc.upenn.edu/Catalog/byType.jsp>
- Dictionaries
  - ▣ WordNet – 206K English words
  - ▣ CELEX2 – 365K German words
- Monolingual text
  - ▣ Gigaword corpus
    - 4M documents (mostly news articles)
    - 1.7 trillion words
    - 11GB of data (4GB compressed)
  - ▣ Enron e-mails
    - 517K e-mails

10

## Corpus examples

- Monolingual text continued
  - ▣ Twitter
  - ▣ Chatroom
  - ▣ Many non-English resources
- Parallel data
  - ▣ ~10M sentences of Chinese-English and Arabic-English
  - ▣ Europarl
    - ~25M sentence pairs with English with 21 different languages
  - ▣ 260K sentences of English Wikipedia—Simple English Wikipedia

11

## Corpus examples

- Annotated
  - ▣ Brown Corpus
    - 1M words with part of speech tag
  - ▣ Penn Treebank
    - 1M words with full parse trees annotated
  - ▣ Other treebanks
    - Treebank refers to a corpus annotated with trees (usually syntactic)
    - Chinese: 51K sentences
    - Arabic: 145K words
    - many other languages...
    - BLIPP: 300M words (automatically annotated)

12

### Corpora examples

Many others...

- ▣ Spam and other text classification
- ▣ Google n-grams
  - 2006 (24GB compressed!)
  - 13M unigrams
  - 300M bigrams
  - ~1B 3,4 and 5-grams
- ▣ Speech
- ▣ Video (with transcripts)

13

### Corpus analysis

Corpora are important resources

Often give examples of an NLP task we'd like to accomplish

Much of NLP is data-driven!

A common and important first step to tackling many problems is analyzing the data you'll be processing

14

### Corpus analysis

What types of questions might we want to ask?

How many...

- ▣ documents, sentences, words

On average, how long are the:

- ▣ documents, sentences, words

What are the most frequent words? pairs of words?

How many different words are used?

Data set specifics, e.g. proportion of different classes?

...

15

### Corpora issues

Somebody gives you a file and says there's text in it

Issues with obtaining the text?

- ▣ text encoding
- ▣ language recognition
- ▣ formatting (e.g. web, xml, ...)
- ▣ misc. information to be removed
  - header information
  - tables, figures
  - footnotes

16

## A rose by any other name...

### Word

- a unit of language that native speakers can identify
- words are the blocks from which sentences are made

### Concretely:

- We have a stream of characters
- We need to break into words
- **What is a word?**
- **Issues/problem cases?**
- **Word segmentation/tokenization?**

17

## Tokenization issues: ‘

Finland's capital...

?

18

## Tokenization issues: ‘

Finland's capital...

Finland                      Finland ' s

Finland ' s                      Finland s

Finland s                      Finland's

**What are the benefits/drawbacks?**

19

## Tokenization issues: ‘

Aren't we ...

?

20

### Tokenization issues: '

**Aren't we ...**

Aren't      Aren t

Are n't      Aren t

Are not

21

### Tokenization issues: hyphens

**Hewlett-Packard**      **state-of-the-art**

**co-education**      **lower-case**

**take-it-or-leave-it**      **26-year-old**

?

22

### Tokenization issues: hyphens

**Hewlett-Packard**      **state-of-the-art**

**co-education**      **lower-case**

Keep as is

merge together

- HewlettPackard
- stateoftheart

What are the benefits/drawbacks?

Split on hyphen

- lower case
- co education

23

### More tokenization issues

Compound nouns: San Francisco, Los Angeles, ...

- One token or two?

Numbers

- Examples
  - Dates: 3/12/91
  - Model numbers: B-52
  - Domain specific numbers: PGP key - 324a3df234cb23e
  - Phone numbers: (800) 234-2333
  - Scientific notation: 1.456 e-10

24

## Tokenization: language issues

*Lebensversicherungsgesellschaftsangestellter*

'life insurance company employee'

Opposite problem we saw with English (San Francisco)

German compound nouns are not segmented

German retrieval systems frequently use a **compound splitter** module

25

## Tokenization: language issues

莎拉波娃现在居住在美国东南部的佛罗里达。

Where are the words?

thisissue

Many character based languages (e.g., Chinese characters) have no spaces between words

- A word can be made up of one or more characters
- There is ambiguity about the tokenization, i.e. more than one way to break the characters into words
- Word segmentation problem
- can also come up in speech recognition

26

## Word counts: Tom Sawyer

How many words?

- 71,370 total
- 8,018 unique

Is this a lot or a little? How might we find this out?

- Random sample of news articles: 11K unique words

What does this say about Tom Sawyer?

- Simpler vocabulary (colloquial, audience target, etc.)

27

## Word counts

What are the most frequent words?

What types of words are most frequent?

Word	Frequency
the	3332
and	2972
a	1775
to	1725
of	1440
was	1161
it	1027
in	906
that	877
he	877
I	783
his	772
you	686
Tom	679
with	642


28

### Word counts

	Word Frequency	Frequency of frequency
	1	3993
	2	1292
8K words in vocab	3	664
71K total occurrences	4	410
	5	243
how many occur	6	199
once? twice?	7	172
	8	131
	9	82
	10	91
	11-50	540
	51-100	99
	> 100	102

29

### Zipf's "Law"



George Kingsley Zipf  
1902-1950

The frequency of the occurrence of a word is inversely proportional to its frequency of occurrence ranking

Their relationship is log-linear, i.e. when both are plotted on a log scale, the graph is a straight line

30

### Zipf's law

At a high level:

- a few words occur very frequently
- a medium number of elements have medium frequency
- many words occur very infrequently

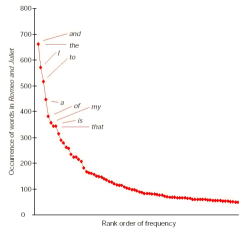
31

### Zipf's law

$$f = C \frac{1}{r}$$

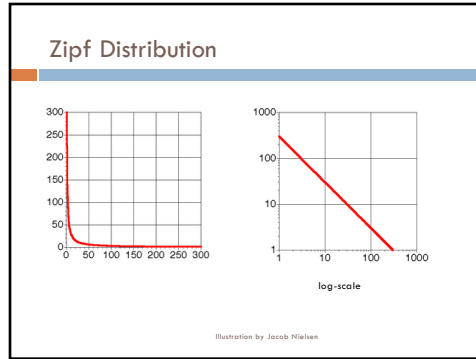
The product of the frequency of words (f) and their rank (r) is approximately constant

Constant is corpus dependent, but generally grows roughly linearly with the amount of data

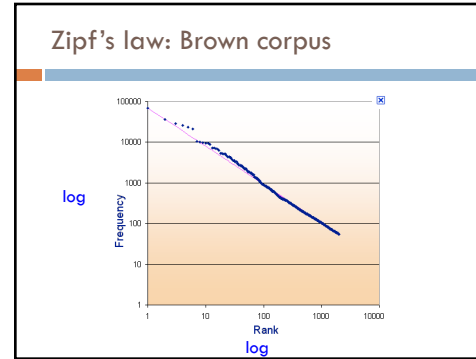


32





33



34

### Zipf's law: Tom Sawyer

Word	Frequency	Rank
the	3332	1
and	?	2

$$f = C \frac{1}{r}$$


---


$$C = f * r = 3332$$

$$f = 3332 * \frac{1}{2} = 1666$$

35

### Zipf's law: Tom Sawyer

Word	Frequency	Rank
the	3332	1
and	2972	2

$$f = C \frac{1}{r}$$


---


$$C = f * r = 3332$$

$$f = 3332 * \frac{1}{2} = 1666$$

36

Zipf's law: Tom Sawyer

Word	Frequency	Rank
the	*****	1
and	2972	2
a	?	3

$$f = C \frac{1}{r}$$


---


$$C = f * r$$

$$= 2972 * 2$$

$$= 5944$$

$$f = 5944 * \frac{1}{3}$$

$$= 1981$$

37

Zipf's law: Tom Sawyer

Word	Frequency	Rank
the	*****	1
and	2972	2
a	1775	3

$$f = C \frac{1}{r}$$


---


$$C = f * r$$

$$= 2972 * 2$$

$$= 5944$$

$$f = 5944 * \frac{1}{3}$$

$$= 1981$$

38

Zipf's law: Tom Sawyer

Word	Frequency	Rank
he	877	10
friends	?	800

$$f = C \frac{1}{r}$$


---


$$C = f * r$$

$$= 877 * 10$$

$$= 8770$$

$$f = 8770 * \frac{1}{800}$$

$$= 10.96$$

39

Zipf's law: Tom Sawyer

Word	Frequency	Rank
he	877	10
friends	10	800

$$f = C \frac{1}{r}$$


---


$$C = f * r$$

$$= 877 * 10$$

$$= 8770$$

$$f = 8770 * \frac{1}{800}$$

$$= 10.96$$

40

## Zipf's law: Tom Sawyer

Word	Frequency	Rank	$C = f \cdot r$
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
Oh	116	90	10440
two	104	100	10400
name	21	400	8400
group	13	600	7800
friends	10	800	8000
family	8	1000	8000
sins	2	3000	6000
Applausive	1	8000	8000

What does this imply about C/zipf's law? How would you pick C?

41

## Sentences

## Sentence

- a string of words satisfying the grammatical rules of a language

## Sentence segmentation

- How do we identify a sentence?
- Issues/problem cases?
- Approach?

42

## Sentence segmentation: issues

## A first answer:

- something ending in a: . ? !
- gets 90% accuracy

Dr. Dave gives us just the right amount of homework.

Abbreviations can cause problems

43

## Sentence segmentation: issues

## A first answer:

- something ending in a: . ? !
- gets 90% accuracy

The scene is written with a combination of unbridled passion and sure-handed control: In the exchanges of the three characters and the rise and fall of emotions, Mr. Weller has captured the heartbreaking inexorability of separation.

sometimes: ; ; and – might also denote a sentence split

44

### Sentence segmentation: issues

A first answer:

- something ending in a . ? !
- gets 90% accuracy

"You remind me," she remarked, "of your mother."

Quotes often appear outside the ending marks

45

### Sentence segmentation

Place initial boundaries after: . ? !

Move the boundaries after the quotation marks, if they follow a break

Remove a boundary following a period if:

- it is a known abbreviation that doesn't tend to occur at the end of a sentence (Prof., vs.)
- it is preceded by a known abbreviation and not followed by an uppercase word

46

### Sentence length

What is the average sentence length, say for news text? 23

Length	percent	cumul. percent
1-5	3	3
6-10	8	11
11-15	14	25
16-20	17	42
21-25	17	59
26-30	15	74
31-35	11	86
36-40	7	92
41-45	4	96
46-50	2	98
51-100	1	99.99
101+	0.01	100

47

### A real-world example

Patterns of Speech: 75 Years of the State of the Union Addresses

In 2011, President Obama was the first modern president to use the words "jobs", "investment" and "deficit" in a State of the Union speech. The other words were in longer speeches. Below is a histogram of the number of these words used in each State of the Union address from 1937 to 2011.

**'jobs'**

**'invest'**

**'deficit'**

<http://graphics.nytimes.com/packages/html/interactive/2011/01/25/ny/politics/state-of-the-union-words-usage.html>

48