https://www.youtube.com/watch?v=bScFi6DaaM

Administrivia

Assignment 0

Assignment 1 out
- due Wednesday
- no code submitted, but will require coding
- will require some command-line work

Reading

NLP models

How do people learn/acquire language?
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

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

Corpus examples

Any you’ve seen or played with before?

Corpus characteristics

What are some defining characteristics of corpora?
Corpus characteristics

- Monolingual vs. parallel
- Language
- Annotated (e.g. parts of speech, classifications, etc.)
- Source (where it came from)
- Size

Corpus examples

- Linguistic Data Consortium
  - [http://www.ldc.upenn.edu/Catalog/byType.jsp](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

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

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)
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)

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

Corpora 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?

...
A rose by any other name…

<table>
<thead>
<tr>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>a unit of language that native speakers can identify</td>
</tr>
<tr>
<td>words are the blocks from which sentences are made</td>
</tr>
</tbody>
</table>

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

Tokenization issues: ‘

Finland's capital…

Tokenization issues: ‘

Finland’s capital…

Tokenization issues: ‘

Aren't we …

What are the benefits/drawbacks?
Tokenization issues: ' Aren’t we …

- 'Aren’t'
- Aren’t
- Aren’t
- Aren't
- Are n’t
- Are not

Tokenization issues: hyphens

- **Hewlett-Packard**
- **state-of-the-art**
- **co-education**
- **lower-case**
- **take-it-or-leave-it**
- **26-year-old**

Tokenization issues: hyphens

- **Hewlett-Packard**
- **state-of-the-art**
- **co-education**
- **lower-case**

More tokenization issues

- **Compound nouns:** San Francisco, Los Angelos, …
  - 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
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

Many character based languages (e.g. Chinese) 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

莎拉波娃现在居住在美国东南部的佛罗里达。Where are the words?

thisissue

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.)

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
</tr>
<tr>
<td>to</td>
<td>1725</td>
</tr>
<tr>
<td>of</td>
<td>1440</td>
</tr>
<tr>
<td>was</td>
<td>1161</td>
</tr>
<tr>
<td>it</td>
<td>1027</td>
</tr>
<tr>
<td>in</td>
<td>906</td>
</tr>
<tr>
<td>that</td>
<td>877</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
</tr>
<tr>
<td>this</td>
<td>783</td>
</tr>
<tr>
<td>his</td>
<td>772</td>
</tr>
<tr>
<td>you</td>
<td>686</td>
</tr>
<tr>
<td>Tom</td>
<td>679</td>
</tr>
<tr>
<td>with</td>
<td>642</td>
</tr>
</tbody>
</table>

What are the most frequent words?

What types of words are most frequent?
Word counts

<table>
<thead>
<tr>
<th>Word Frequency</th>
<th>Frequency of frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3993</td>
</tr>
<tr>
<td>2</td>
<td>1292</td>
</tr>
<tr>
<td>3</td>
<td>664</td>
</tr>
<tr>
<td>4</td>
<td>410</td>
</tr>
<tr>
<td>5</td>
<td>243</td>
</tr>
<tr>
<td>6</td>
<td>199</td>
</tr>
<tr>
<td>7</td>
<td>172</td>
</tr>
<tr>
<td>8</td>
<td>131</td>
</tr>
<tr>
<td>9</td>
<td>82</td>
</tr>
<tr>
<td>10</td>
<td>91</td>
</tr>
<tr>
<td>11-50</td>
<td>540</td>
</tr>
<tr>
<td>51-100</td>
<td>99</td>
</tr>
<tr>
<td>&gt; 100</td>
<td>102</td>
</tr>
</tbody>
</table>

| 8K words in vocab |
| 71K total occurrences |
| how many occur once? twice? |

Zipf’s “Law”

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.

At a high level:
- a few words occur very frequently
- a medium number of elements have medium frequency
- many words occur very infrequently

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.
**Zipf Distribution**

**Zipf’s law: Brown corpus**

**Zipf’s law: Tom Sawyer**

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
</tr>
<tr>
<td>end</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

\[ f = C \frac{1}{r} \]

\[ C = f \cdot r \]
\[ = 3332 \quad f = 3332 \cdot \frac{1}{2} \]

**Zipf’s law: Tom Sawyer**

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
</tr>
<tr>
<td>end</td>
<td>2972</td>
<td>2</td>
</tr>
</tbody>
</table>

\[ f = C \frac{1}{r} \]

\[ C = f \cdot r \]
\[ = 3332 \quad f = 3332 \cdot \frac{1}{2} \]
### Zipf’s law: Tom Sawyer

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>2972</td>
<td>1</td>
</tr>
<tr>
<td>and</td>
<td>1775</td>
<td>2</td>
</tr>
<tr>
<td>a</td>
<td>877</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ f = C \frac{1}{r} \]

\[ C = f \times r \]

\[ f = 2972 \times 2 \]

\[ f = 5944 \]

\[ f = 1981 \]

\[ f = 877 \times 10 \]

\[ f = 8770 \]

\[ f = 10.96 \]
**Zipf’s law: Tom Sawyer**

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Rank</th>
<th>C = r * p</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
<td>3332</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>2</td>
<td>5944</td>
</tr>
<tr>
<td>a</td>
<td>1773</td>
<td>3</td>
<td>5235</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>10</td>
<td>8770</td>
</tr>
<tr>
<td>but</td>
<td>410</td>
<td>20</td>
<td>8200</td>
</tr>
<tr>
<td>be</td>
<td>294</td>
<td>30</td>
<td>8820</td>
</tr>
<tr>
<td>Oh</td>
<td>116</td>
<td>90</td>
<td>10540</td>
</tr>
<tr>
<td>two</td>
<td>104</td>
<td>100</td>
<td>10400</td>
</tr>
<tr>
<td>name</td>
<td>21</td>
<td>400</td>
<td>8400</td>
</tr>
<tr>
<td>group</td>
<td>13</td>
<td>600</td>
<td>7800</td>
</tr>
<tr>
<td>friends</td>
<td>10</td>
<td>800</td>
<td>8000</td>
</tr>
<tr>
<td>family</td>
<td>8</td>
<td>1000</td>
<td>8000</td>
</tr>
<tr>
<td>sir</td>
<td>2</td>
<td>3000</td>
<td>6000</td>
</tr>
<tr>
<td>Applausive</td>
<td>1</td>
<td>8000</td>
<td>8000</td>
</tr>
</tbody>
</table>

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

**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?

**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

**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
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

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

Sentence length

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

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

A real-world example