Tokenization, Segmentation, and Evaluation

2 February 2022
Assignment 1

Due 10 p.m. on Wednesday.

Worksheet 2

First part due now

Fill in the rest during class today, due tonight
Today

45 min.  Talk about shared reading
30 min.  Split into groups to discuss further reading
Shared reading
Every NLP task involves segmenting language into workable units…

narratives,
paragraphs,
sentences,
words,
morphemes

…which undergo varying degrees of normalization.
Words
What are words?
“cat”
One morning I shot an elephant in my pajamas.
One morning I shot an elephant in my pajamas.

I *didn’t* shoot an elephant.
One morning I shot an elephant in my pajamas.

I didn’t shoot an elephant.

Imma let you finish, but Beyoncé had one of the best videos of all time.
One morning I shot an elephant in my pajamas.

I didn’t shoot an elephant.

Imma let you finish, but Beyoncé had one of the best videos of all time.

I do uh main- mainly business data processing.
One morning I shot an elephant in my pajamas.

I didn’t shoot an elephant.

Imma let you finish, but Beyoncé had one of the best videos of all time.

I do uh main- mainly business data processing.

Have a great day! :-}
Counting words
When we talk about a “word”, we might mean

an abstract *vocabulary item*, or

an individual *occurrence* of a word.
To be or not to be

6 tokens: [to, be, or, not, to, be]

4 types: {be, not, or, to}
How can we use types and tokens to measure vocabulary richness?
# of Unique Words Used Within Artist’s First 35,000 Lyrics

Notes/sources:
All lyrics are via Genius.

pudding.cool/projects/vocabulary
I dumbed down for my audience to double my dollars
They criticized me for it, yet they all yell “holla”
If skills sold, truth be told, I’d probably be
Lyrically Talib Kweli
Truthfully I wanna rhyme like Common Sense
But I did 5 mil – I ain’t been rhyming like Common since

Jay-Z, “Moment of Clarity”
# of Unique Words Used Within Artist's First 35,000 Lyrics

Talib Kweli: 4,078 unique words used

Notes/sources:
All lyrics are via Genius.
# of Unique Words Used Within Artist's First 35,000 Lyrics

3,000 words
4,000
5,000
6,000 words

Common: 5,138 unique words used

Notes/sources:
All lyrics are via Genius.
# of Unique Words Used in 500 Random Samples of 35,000 Lyrics from Country, Rock, Hip Hop

Raw Lyrics Data via John W. Miller
Separating words
Tokenization is splitting a text into the word tokens we want to process.
Why can’t we just use white space to tokenize a text?
That old bell, presage of a train, had just sounded through Oxford station; and the undergraduates who were waiting there, gay figures in tweed or flannel, moved to the margin of the platform and gazed idly up the line.

Max Beerbohm, *Zuleika Dobson*
That old bell, presage of a train, had just sounded through Oxford station; and the undergraduates who were waiting there, gay figures in tweed or flannel, moved to the margin of the platform and gazed idly up the line.

Max Beerbohm, Zuleika Dobson
In a sample of 100 books from Project Gutenberg, we find these tokens when we split on white space:

<table>
<thead>
<tr>
<th>Token</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>earth</td>
<td>889</td>
</tr>
<tr>
<td>earth,</td>
<td>345</td>
</tr>
<tr>
<td>earth.</td>
<td>213</td>
</tr>
<tr>
<td>earth's</td>
<td>76</td>
</tr>
<tr>
<td>earth;</td>
<td>49</td>
</tr>
<tr>
<td>earth.”</td>
<td>19</td>
</tr>
<tr>
<td>earth?</td>
<td>14</td>
</tr>
<tr>
<td>earth!</td>
<td>9</td>
</tr>
<tr>
<td>earth:</td>
<td>8</td>
</tr>
<tr>
<td>earth,”</td>
<td>8</td>
</tr>
<tr>
<td>earth”</td>
<td>6</td>
</tr>
<tr>
<td>earth!”</td>
<td>6</td>
</tr>
<tr>
<td>earth?'</td>
<td>5</td>
</tr>
<tr>
<td>earth.'</td>
<td>4</td>
</tr>
<tr>
<td>earth._</td>
<td>3</td>
</tr>
<tr>
<td>earth—he</td>
<td>3</td>
</tr>
<tr>
<td>earth—</td>
<td>3</td>
</tr>
<tr>
<td>earth)</td>
<td>3</td>
</tr>
<tr>
<td>earth!’</td>
<td>3</td>
</tr>
<tr>
<td>earth—”</td>
<td>2</td>
</tr>
<tr>
<td>earth—“</td>
<td>2</td>
</tr>
<tr>
<td>earth—&quot;</td>
<td>2</td>
</tr>
<tr>
<td>earth—&quot;</td>
<td>2</td>
</tr>
<tr>
<td>earth,”</td>
<td>2</td>
</tr>
<tr>
<td>earth-goddess</td>
<td>2</td>
</tr>
<tr>
<td>earth—as</td>
<td>2</td>
</tr>
<tr>
<td>earth?”</td>
<td>1</td>
</tr>
<tr>
<td>earth;_</td>
<td>1</td>
</tr>
<tr>
<td>earth;&quot;</td>
<td>1</td>
</tr>
<tr>
<td>earth;—</td>
<td>1</td>
</tr>
<tr>
<td>earth._</td>
<td>1</td>
</tr>
<tr>
<td>earth.[5]</td>
<td>1</td>
</tr>
<tr>
<td>earth....</td>
<td>1</td>
</tr>
<tr>
<td>earth.–Thou</td>
<td>1</td>
</tr>
<tr>
<td>earth.”–The</td>
<td>1</td>
</tr>
</tbody>
</table>

...
Can we just strip out all punctuation?
Punctuation can be helpful:

It can mark boundaries for sentences, clauses, parentheticals, asides.

Some punctuation has illocutionary force, like exclamation points and question marks.

Emoticons are strong signals of sentiment.
Most tokenization algorithms for English use regular expressions to segment a string into discrete tokens, e.g.,

```python
>>> text = "That U.S.A. poster-print costs $12.40..."
>>> pattern = r"''(?x)  # set flag to allow verbose regexps
... ([A-Z]\.)+      # abbreviations, e.g., U.S.A.
... | \w+(-\w+)*    # words with optional internal hyphens
... | \$?\d+(\.\d+)?%? # currency and percentages, e.g. $12.40, 82%
... | \.\.\.         # ellipsis
... | \[\[,.;"'?():_'-\] # these are separate tokens; includes ], [ ...

>>> nltk.regexp_tokenize(text, pattern)
['That', 'U.S.A.', 'poster-print', 'costs', '$12.40', '...']
```
But, English is an easy one!

What if you don’t have spaces and punctuation to help you?
Vergilius Augusteus
Susan will be answering your questions at her exclusive album listening party on Saturday. Send in your questions #susanalbumparty Susan HQ
Susan will be answering your questions at her exclusive album listening party on Saturday. Send in your questions #susanalbumparty Susan HQ
Yao Ming reaches finals.
Tokenization’s hard, and the correct behavior depends on what we want to treat as a token.
Consider how to handle clitics.

We can leave them alone:

\[\text{Can’t} \rightarrow \text{Can’t}\]

We can separate them:

\[\text{Can’t} \rightarrow \text{Can’t}\]
\[\text{Can’t} \rightarrow \text{Can’t}\]
\[\text{Can’t} \rightarrow \text{Can’t}\]

We can expand them:

\[\text{Can’t} \rightarrow \text{Can not}\]
What are the dangers of simple tokenization?

Why not leave words like *can’t* alone?
This doesn’t mean we always split distinct units.

German compound nouns like

*Lebensversicherungsgesellschaftsangestellter*

“life insurance company employee”

are treated as individual tokens.
Separating sentences
Why do we want to divide text into sentences?

Why do sentences matter?
“Do you want to go?” asked Jane.

Mr. Collins said he was going.

He lives in the U.S. John, however, lives in Canada.
Normalizing words
What does case folding get us?

What do we lose?
organizes
organized
organizing
organizes → organized → organizing → stemmin → organ
organizes → organized → organizing → organize
the boy’s cars are different colors
→ the boy car be different color

He is reading detective stories
→ he be read detective story
Why might we prefer lemmatization to stemming?
What does lemmatizing get us?
What do we lose?
Whenever we’re making decisions about text, like identifying words or identifying sentence boundaries, we need to think about evaluation.
Reducing the error rate for an application often involves two antagonistic efforts:

- Increasing accuracy or precision
  - That is, minimizing false positives
- Increasing coverage or recall
  - That is, minimizing false negatives
Do you have any questions or concerns from the shared reading?
Further reading
Groups
Acknowledgments

This class incorporates material from:

David Bamman, UC Berkeley
Matt Daniels, “The Largest Vocabulary in Hip Hop”
Jurafsky & Martin, *Speech and Language Processing*, 3rd ed. draft