Programming, Problem Solving, and Algorithms

CPSC203, 2019 W1

Announcements

~~"Problem of the Day" continues!~~

I think canvas is live. POTD 1-27 recorded + MT.

Final exam: 12/11 noon, here.

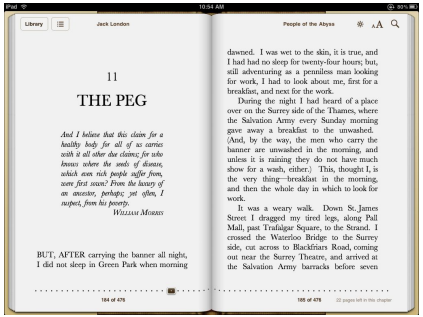
Today:

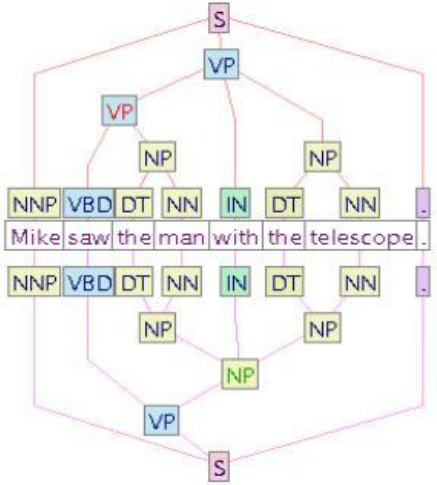
Visualizing Literature

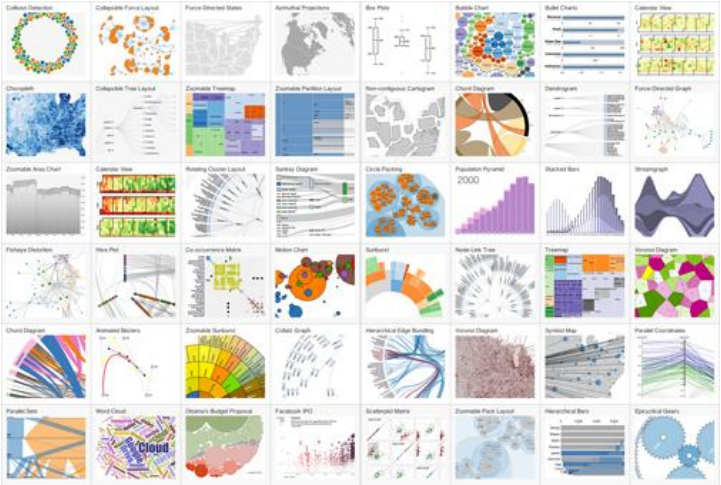
Natural Language Processing

Named Entity Recognition

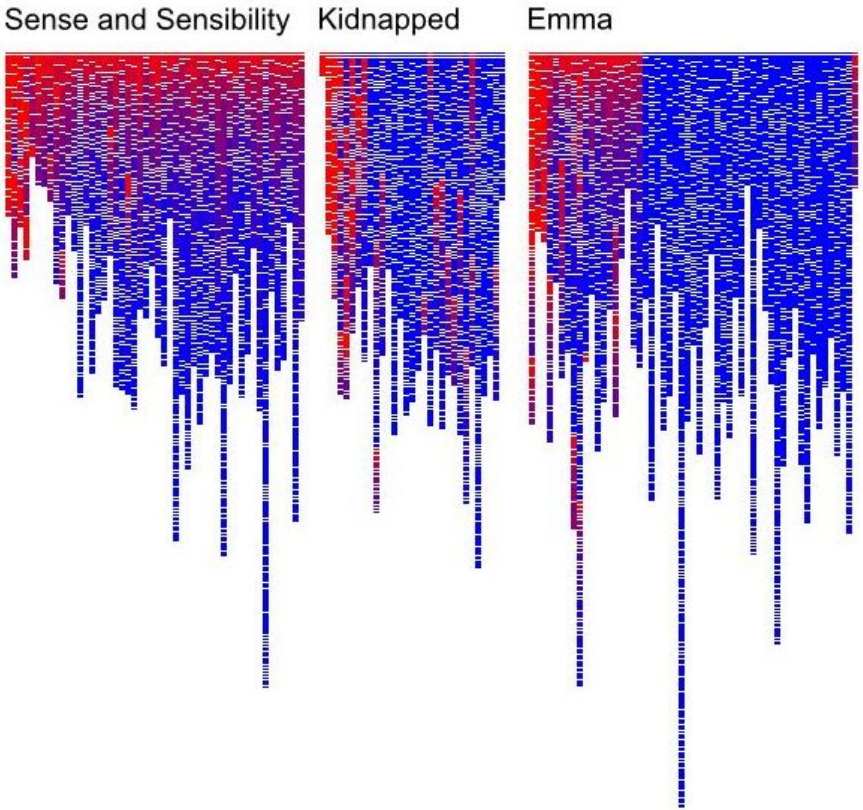
Visualizing Literature





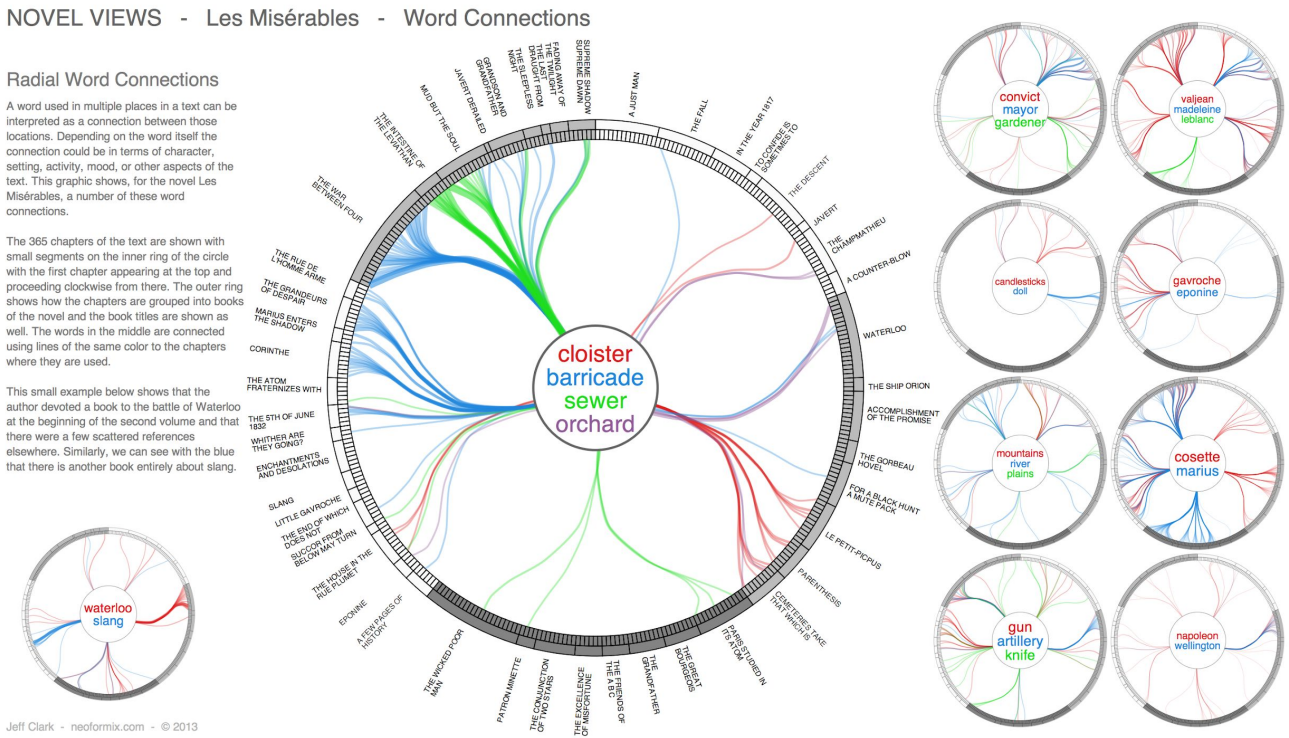


Example



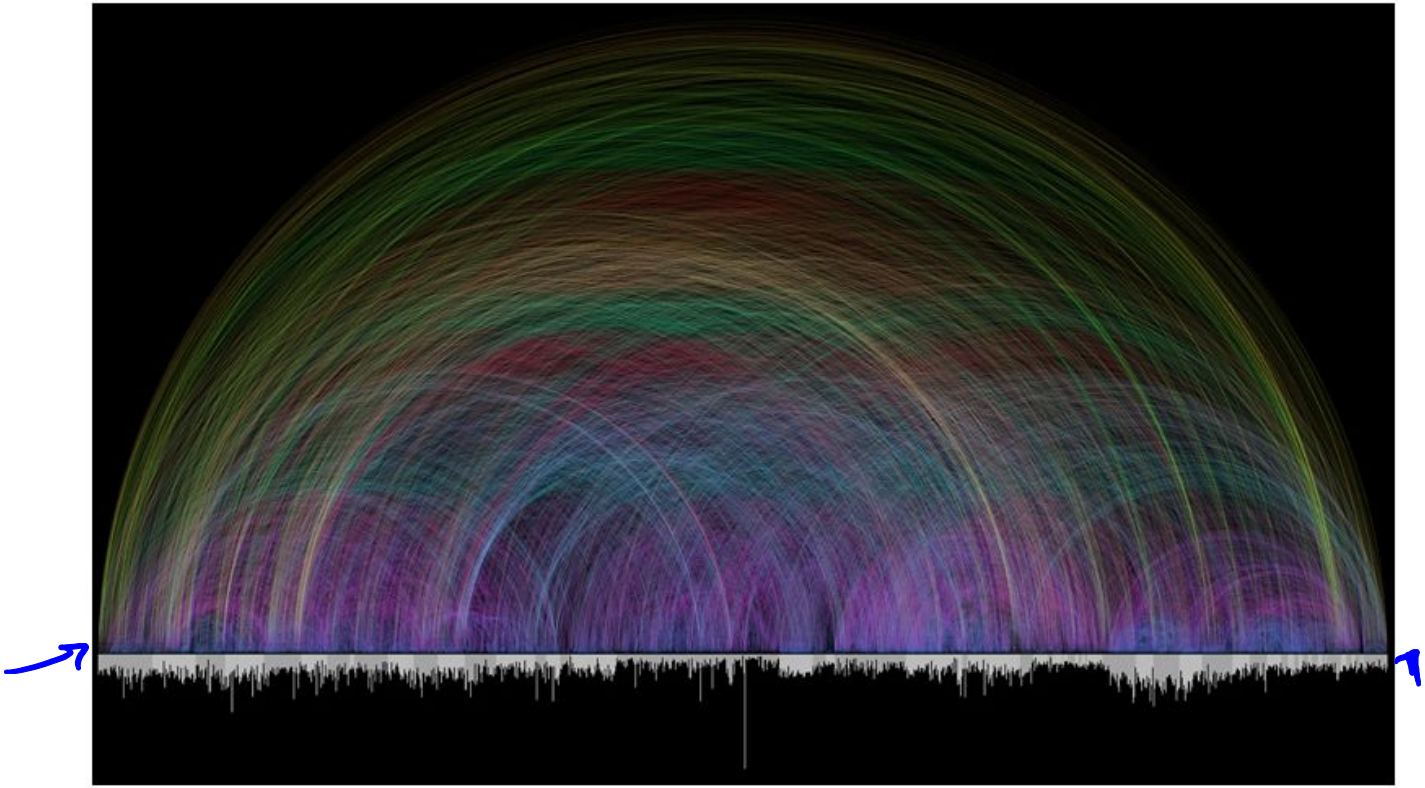
http://datamining.typepad.com/data_mining/2011/09/visualizing-lexical-novelty-in-literature.html

Example



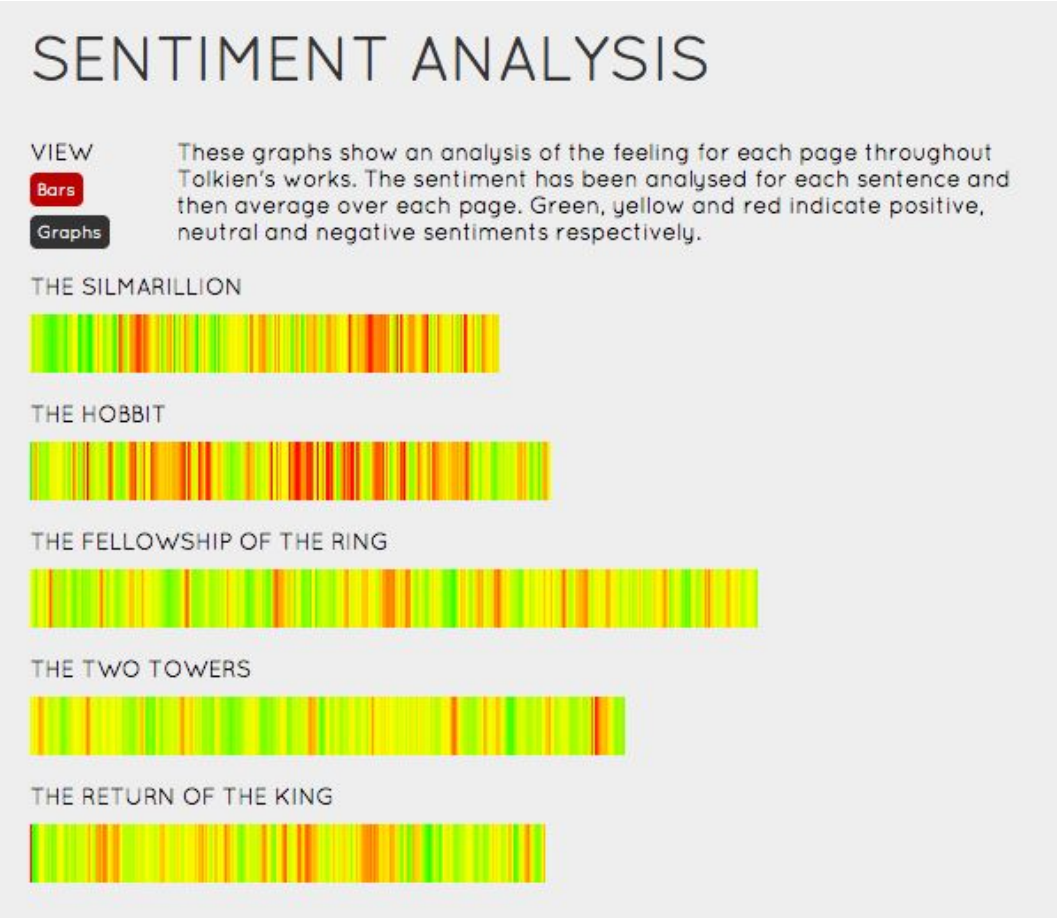
http://neoformix.com/2013/NovelViews.html

Example



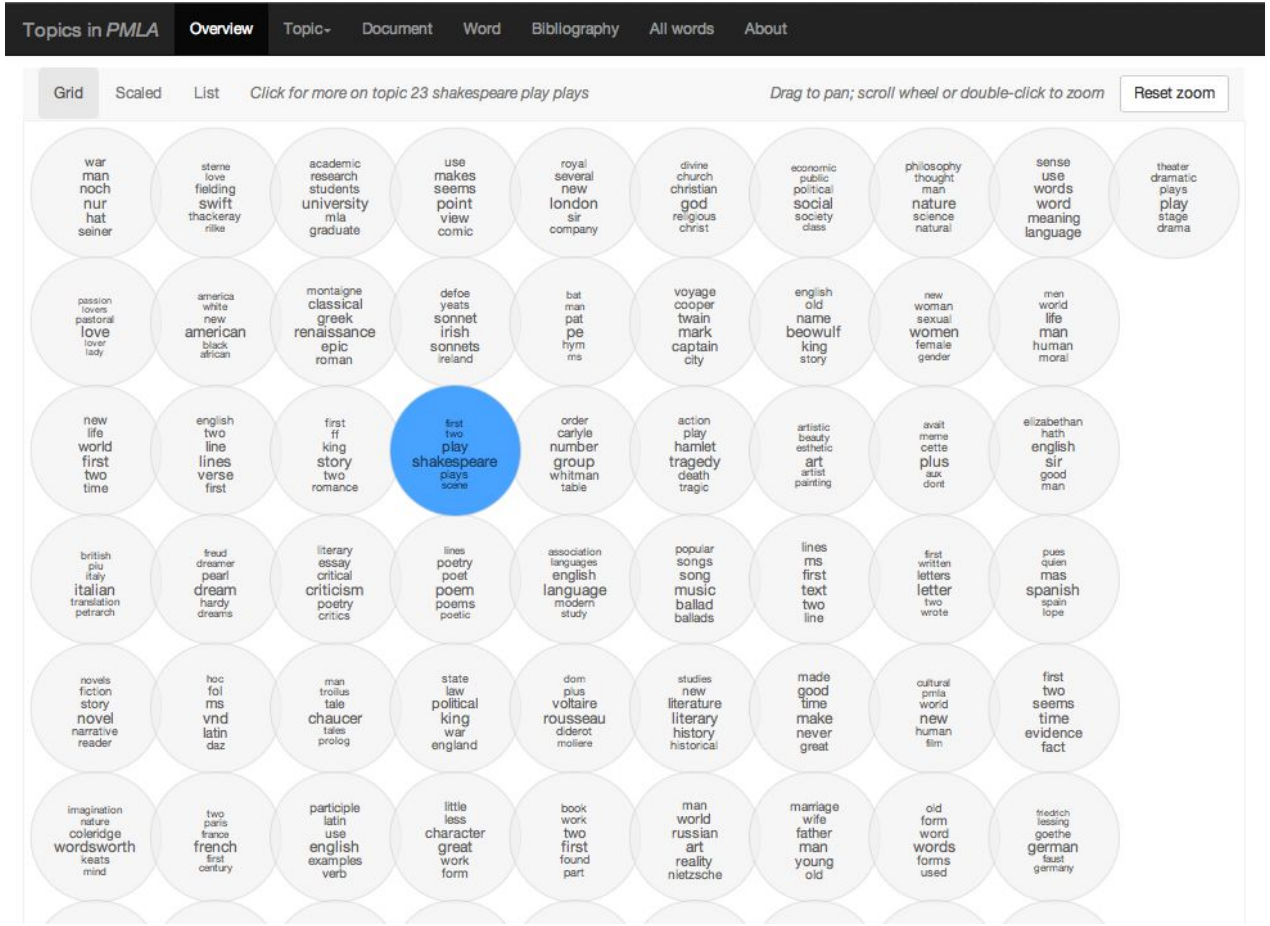
http://www.chrisharrison.net/index.php/Visualizations/BibleViz

Example



http://lotrproject.com/statistics/books/

Example



http://agoldst.github.io/dfr-browser/

Example

http://www.emelynbaker.com/westeros.html

Example

The New York Times

http://www.nytimes.com/newsgraphics/2013/12/30/year-in-interactive-storytelling/

How do we begin?

```
textRaw = open('[redacted]hp.txt').read()
```

returns a string.

We want to analyze the data by word or by sentences or by paragraphs or by chapter books

can do this using nltk's "tokenizer"

natural language toolkit

Tokenization

Once + Future King

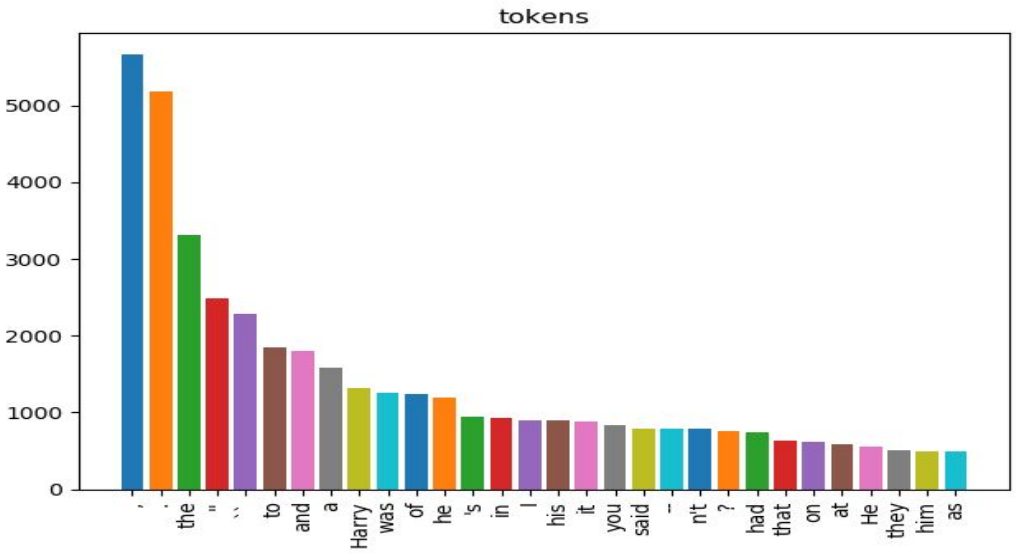
Translate: “Astrology. The governess was always\ngetting muddled with her astrolabe, and when she got specially muddled she would take it out\nof the Wart by rapping his knuckles. She did not rap Kay's knuckles, because when Kay grew\nolder”

Into: ['Astrology.', 'The', 'governess', 'was', 'always', 'getting', 'muddled', 'with', 'her', 'astrolabe', ',', 'and', 'when', 'she', 'got', 'specially', 'muddled', 'she', 'would', 'take', 'it', 'out', 'of', 'the', 'Wart', 'by', 'rapping', 'his', 'knuckles.', 'She', 'did', 'not', 'rap', 'Kay', "'s", 'knuckles', ',', 'because', 'when', 'Kay', 'grew', 'older']

http://text-processing.com/demo/tokenize/

Python Demo

The python script in “LecHP” was assembled from examples in Ch1-3 of the NLTK book. http://www.nltk.org/book/



Pre-processing

a necessary evil... address by getting to know your data.

```
begged so hard, cried even, I had to let him stay. It
turned out okay. My mother got rid of the vermin and
he's a born mouser. Even catches the occasional rat.
Sometimes, when I clean a kill, I feed Buttercup the
entrails. He has stopped hissing at me.

Entrails. No hissing. This is the closest we will ever
come to love.



3 | P a g e



The Hunger Games - Suzanne Collins



I swing my legs off the bed and slide into my hunting
boots. Supple leather that has molded to my feet. I
```

gutenberg.org

A feasible sequence… for data cleaning

lower case

eliminate punctuation

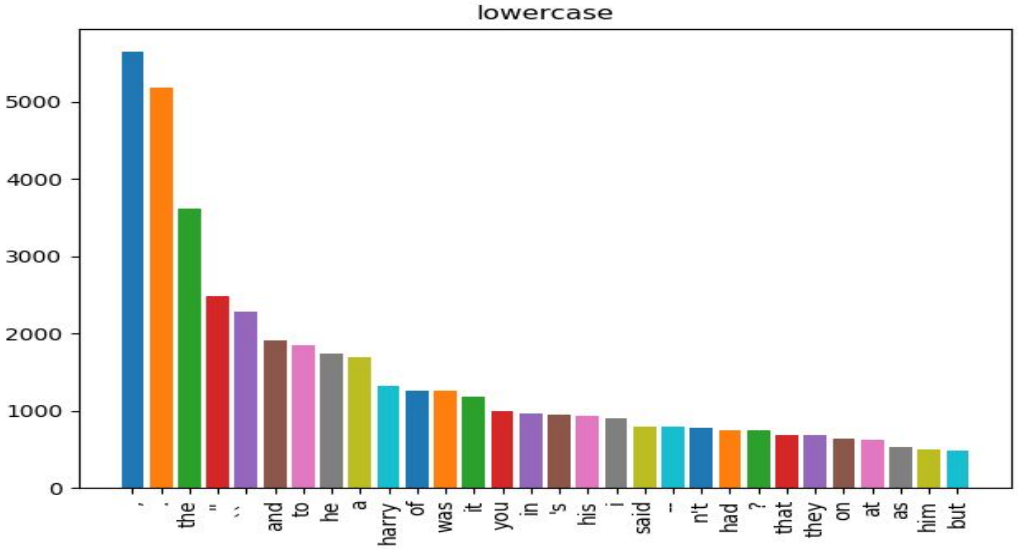
remove stop words

stem

Unify tally for “Valor” and “valor”

Depending on task, may not want to do this… caps are useful for detecting “named entities.”

The != the



A feasible sequence...

lower case

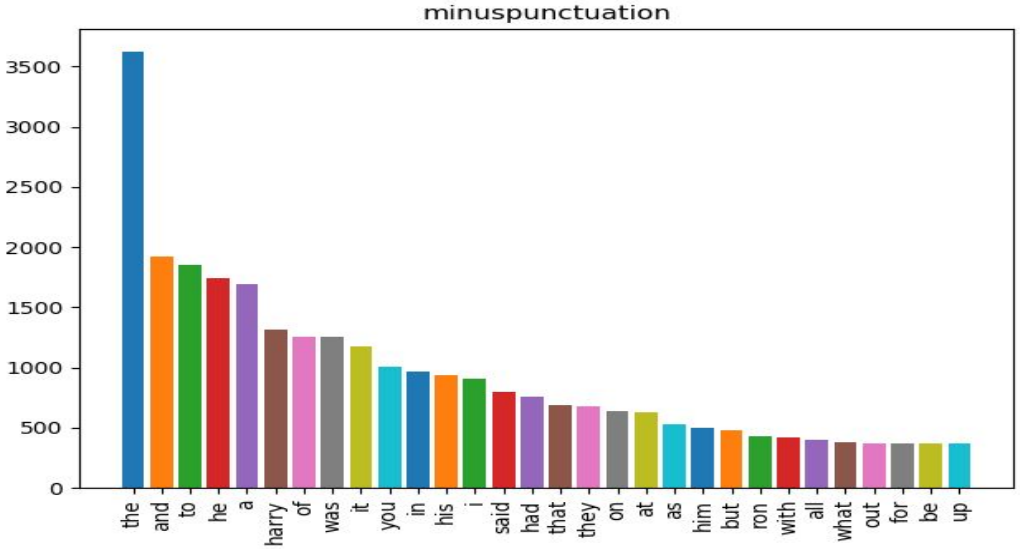
eliminate punctuation

remove stop words

stem

Punct tokenizer leaves periods at end of sentences: “father.”

amazingly, it works fine for “Dr.”, “$3.50”, “!”



A feasible sequence...

lower case

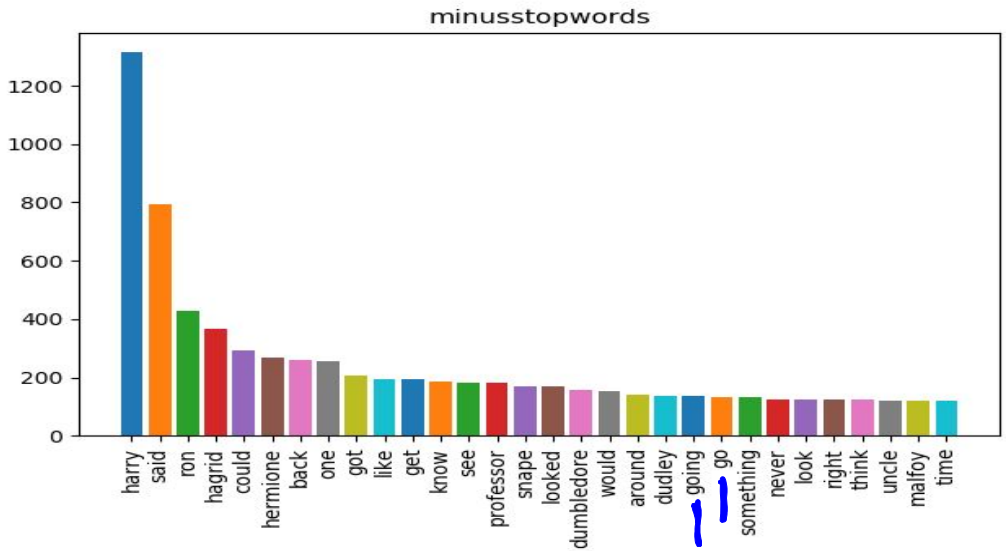
eliminate punctuation

remove stop words

stem

List of common, unhelpful words compiled by nltk from large corpora. We keep words that aren't in that list.

More sophisticated approach is called tf-idf...



A feasible sequence...

lower case

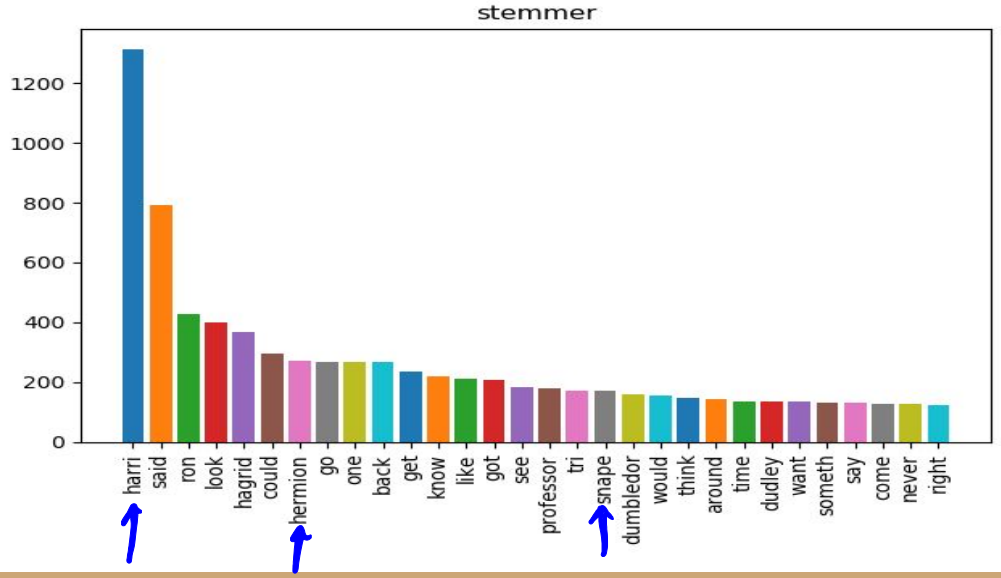
eliminate punctuation

remove stop words

stem

“goes” -> “go”
“running” -> “run”
“eaten” -> “eat”

NLTK provides the stemmer



Parts of Speech...

Middle School Grammar:

Noun, Verb, Adjective, Adverb, Preposition, Conjunction, Pronoun, Interjection

open

closed

Subdivide classes!

ex. Noun → proper

→ common

UPenn POS

POS inference is hard!

wordNet is resource for defn + pos given any word.

Plug in the curling iron. JJ

I want to learn to play curling. NN

The ribbon is curling around the Maypole. VBG

Search space

ex. Plays well with others. $2 \cdot 4 = 8$

NNS	UH	IN	NNS
VBZ	JJ		
	NN		
	RB		

# of different labellings is product of possible word senses.

Word senses come from lexicon (like wordNet)

Search space

Only 11% of English words have more than one POS, but...

they tend to be very common words.

I know that he is honest.

That movie was fantastic!

I wouldn't go that far.

Accuracy and expectations...

- modern POS taggers achieve 97% accuracy. w00t!

- tagging w most frequent POS gets 90%.

- humans achieve 98% agreement.

One last example:

Bill	saw	that	man	yesterday.
NNP	NN	DT	NN	NN
VB	VB	IN	VB	RB
NN				

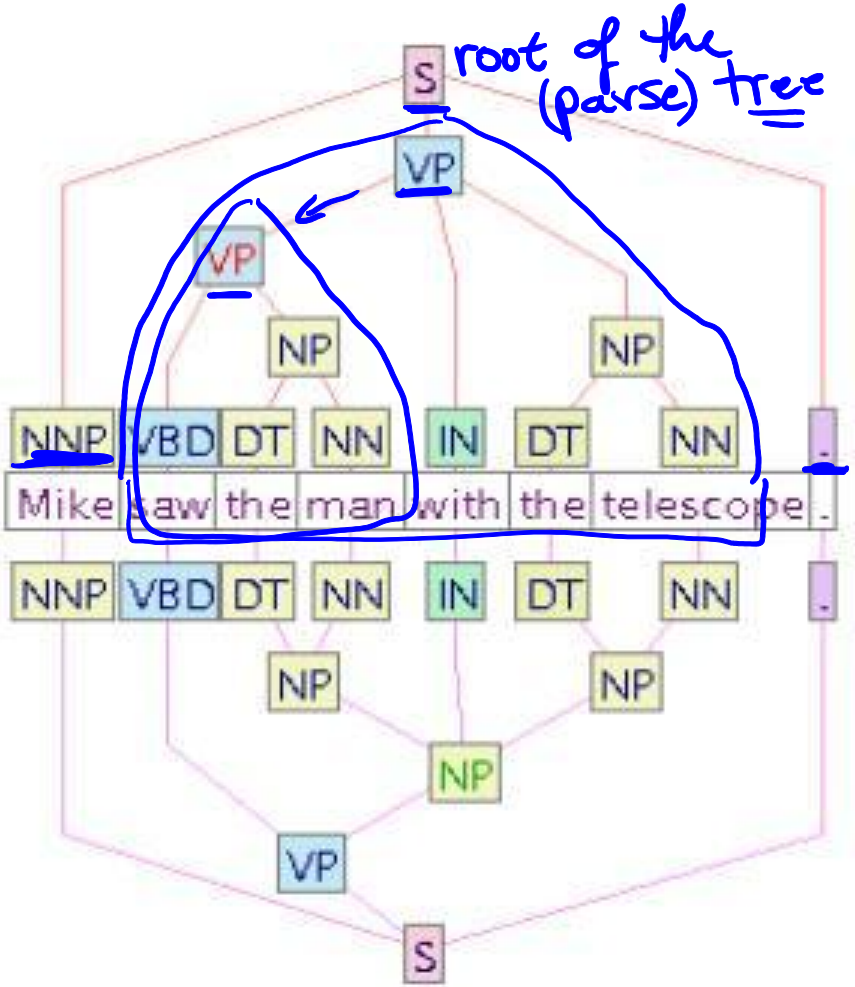
- man is rarely a verb
- VB never follows DT

96.6% on known words
86.8% on unknown words

POS tagging in NLTK

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

Ch5: Categorizing and Tagging Words



Named Entity Recognition (NER)

1. Underline all of the proper nouns (named entities) in this text..

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

Typical categories of entities are PERSON, LOCATION, ORGANIZATION. Think about how you might discover each of the entities using a program.

NLTK NER discovery...

2. Repo LecHP contains a file called test.py. Modify and execute this file to answer the following questions. In each case, sketch an example of the output, and explain it briefly in English.

a. if `textRaw` is the string above, what is the result of

```
sents = sent_tokenize(text)
```

b. if `sents` is the result of part a, what is the result of

```
sentWords = [word_tokenize(s) for s in sents if s]
```

NLTK NER discovery…

c. if `sentWords` is the result of part b, what is the result of

```
sentWordsPOS = [pos_tag(s) for s in sentWords]
```

d. if `sentWordsPOS` is the result of part c, what is the result of

```
sentWordsNER = [ne_chunk(s) for s in sentWordsPOS]
```

e. if `sentWordsNER` is the result of part d, what is the result of

```
subtrees = [chunk.subtrees() for chunk in sentWordsNER]
```

NLTK NER discovery...

f. if `subtrees` is the result of part e, what is the result of

```
entities = [[ s for s in st if s.label() == "PERSON"] for st in subtrees]
```

g. if `entities` is the result of part f, what is the result of

```
entities = [[ ' '.join(c[0] for c in s.leaves()) for s in st] for st in entities]
```

3. Write python code that would extract all the verbs from the text above. The answer to problem 2c will help you!

4. (challenge) Write a function `personVerbs(person, text)`, that returns a list of all the verbs that occur in sentences that also contain `person`.

POTD #38 Tue

https://github.students.cs.ubc.ca/cpsc203-2019w-t1/potd36

Describe any snags you run into:

1. Line ___: _____
2. Line ___: _____
3. Line ___: _____
4. Line ___: _____
5. Line ___: _____

ToDo for next class...

POTD: Continue every weekday! Submit to repo.

Reading: TLACS Ch 10 & 12 (lists and dictionaries)

References:

https://www.youtube.com/watch?v=wsSEKm-rU6U

https://github.com/gboeing/osmnx-examples/tree/master/notebooks

https://gist.github.com/psychemedia/b49c49da365666ba9199d2e27d002d07