Introduction to Computation for the Humanities and Social Sciences

CS 3

Chris Tanner

Lecture 15

New Tweet, Who Dis?

Lecture 15

- Regular Expressions — Continued
- Sentiment Analysis
- Unigrams / Bigrams / Language Models

Regex — Recap

- You define a regular expression (a String of text representing a pattern), and **findall()** or **sub()** tries to repeatedly fit it to a specified text, in a left-to-right fashion.

Regex — Recap

- **[ ]** denotes that you want to match any of the characters within it

- **+** denotes that the previous character or [ ] block should appear 1 or more times

- ***** same as above but 0 or more times (meaning it's optional)

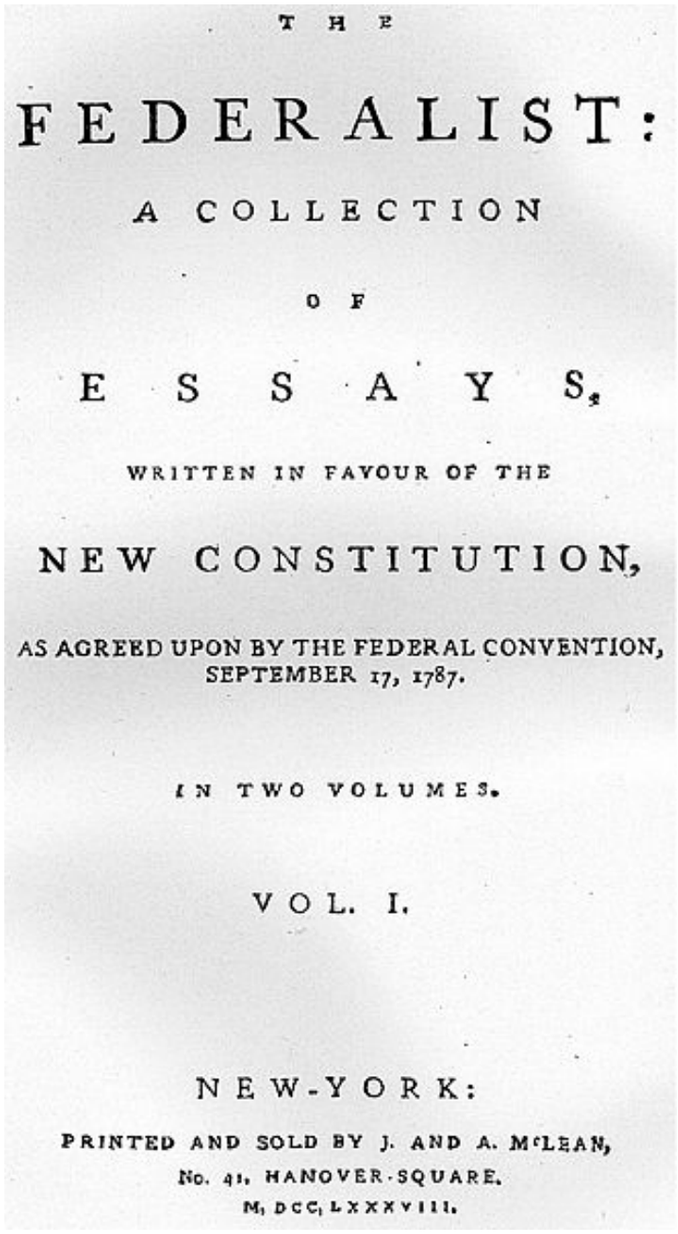
- **?** denotes the prev. character or [ ] block should appear 0 or 1 times, and it also designates a non-greedy approach

REAL-TIME CODING

let's work through the homework together

Text Comparison

Can you identify the authors by comparing the style of writing?


THE
FEDERALIST:
A COLLECTION
OF
ESSAYS,
WRITTEN IN FAVOUR OF THE
NEW CONSTITUTION,
AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.
IN TWO VOLUMES.
VOL. I.
NEW-YORK:
PRINTED AND SOLD BY J. AND A. M'LEAN,
No. 41, HANOVER-SQUARE.
M, DCC, LXXXVIII.

Text Comparison

- How do you discuss analytical differences between different texts?
- You need computable metrics that have a grounded meaning

Text Analysis

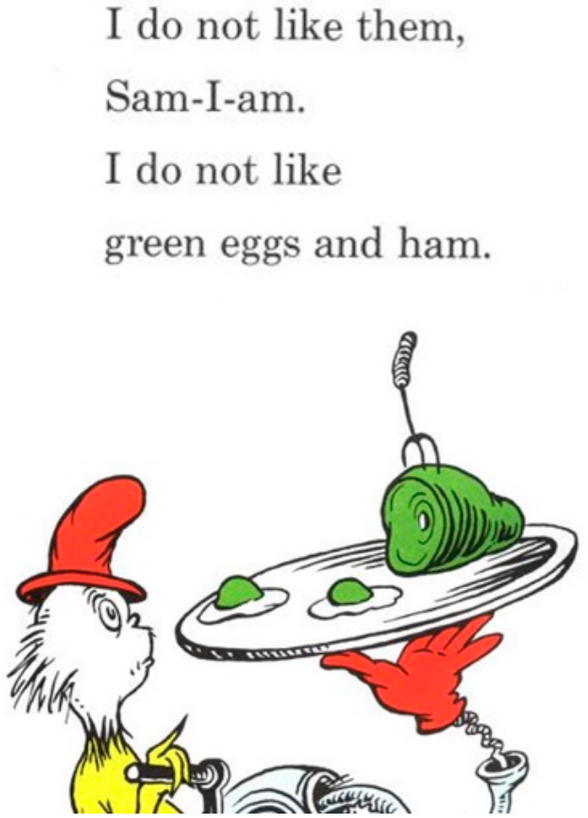
Unique Words (Ratio)

- Fraction of words that occur only a single time
- Measures richness of vocabulary
- Shakespeare's plays include 28,829 different words, 12,493 which occur only a single time
- 1 in 70 words in plays written by Shakespeare are single occurrence words!

Text Analysis

Word Length (Ratio)

- Ratios between short words to medium-length words, and long words to medium-length words
- Measures difficulty of vocabulary



Emotional Sentiment

- Words are assigned an emotional sentiment score

- Whether a text has a positive or negative sentiment is simply a summation over the words' individual sentiment scores

Text Analysis

Emotional Sentiment

- **positive_words** = {“happy”, “good”, “great”, “amazing” …}
- **negative_words** = {“bad”, “horrible”, “crappy”, “poor”, “janky”}

Text Analysis

Emotional Sentiment

- **positive_words** = {“happy”, “good”, “great”, “amazing” …}

- **negative_words** = {“bad”, “horrible”, “crappy”, “poor”, “janky”}

$$sentence\ \ score = \frac{(\#poswords - \#negwords)}{\#totalwords}$$

Language Models

- a **Language Model** is a representation of the language used by a given entity (e.g., a person or a genre or any other well-defined author of text)

- Typically, these models are represented by statistics that are largely focused on just word counts

Unigram Language Models

- The most simple approach involves counting all of the individual words that an author uses (i.e., unigram counts).

- You've already done this before! **word_counts = {}**

- Then, you can construct probabilities by just dividing by the total # of words the author uses. e.g., maybe she/he uses the words:
  'reductionist' = 0.01% of the time
  'imperialism' = 0.04% of the time
  'capital' = 0.06% of the time
  etc

Unigram Language Models

- With these probabilities, you can do several fun things:
    - compare the model of one author to a different author
    - measure the likelihood that a given body of text was generated from a given user

Unigram Language Models

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.02%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.06%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.1%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.09%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.08%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

0.31%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

1.02%

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Unigram Language Models

"C

mi

bo

ec

political discourse had grown too vitriolic"

If we assume each word is independent from another, we could just multiply every word's probability to come up w/ how likely the entire sequence came from the particular author

Unigram Language Models

Weaknesses with this unigram model?

Bi-gram Language Models

- The word order doesn't matter
- Context is way too narrow
- Instead, we could look at every pair of two consecutive words! That'll help some!

Bi-gram Language Models

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models — Let's first count all bigrams!

1

"Coming less than two weeks before the midterm elections, the discovery of the pipe bombs reverberated across a country already on edge, stirring anew questions about whether political discourse had grown too vitriolic"

Bi-gram Language Models

For your Project #2, we only require you to build a dictionary of the bi-gram counts. You do not need to further turn those into probabilities.

LAB TIME