Can "something" Point to Dementia?

Metis Project - Natural Language Processing

Mailei Vargas

February 2019

Project Motivation

Our speech and writing can offer a lot of information and clues into how our brain is functioning. Dementia is often linked to language according to researchers at the University of Toronto. There is currenly big efforts to find computational solutions that encoporates natural language processing and machine learning to detect early signs of dementia, particularly Alzheimer's. If dementia/Alzheimer's is detected before a person begins to exhibit human noticable symptoms, then it can be treated sooner to slow the progression with the proper medications.

A famous study called the "Nun Study" was the catalyst for this field of study. A group of nearly 700 nuns were used to find patterns in their habits (no pun intended) and environments to determine if there was a correlation between the nuns who did and didn't get Alzheimer's. There was a suprising discovery when researchers analyzed the nuns' autobigraphical essays written in their 20s in order to be accepted into the convent. With an astonishing 85% accuracy, nuns that had far less grammatical complexity and idea density in their essays were most likely to have gotten Alzheimer's.

Overview

For this project, I focused on three authors, Iris Murdoch (who did have Alzheimer's), P.D. James (who did not have Alzheimer's) and Agatha Christie (who was suspected of having Alzheimer's) and analyzed three of their books. One book each from the beginning, middle, and end of their careers.

I focused on the some of the following measurable characteristics of their writings:

- parts of speech (nouns, pronouns, verbs, adjectives, etc.) and the quantity
- Vocabulary size
    - Length of words
    - size of unique vocabulary used for each book
- use of indefinite pronouns
- Idea density

    - Length of the sentences
- Sentiment
- Clustering

Tools

Below is a list of the libraries, sites, and tools that aided in the completion of this project.

Websites/References	Tools	Libraries for Analysis/Modeling	Custom functions scripts
NLP and Dementia	Python	numpy/scipy	Preprocessing.py
NLP to detect Alzheimers	Jupiter Notebook	pandas	
Nun Study	PyCharm	matplotlib/seaborn	
Detecting Linguistic Characteristics...	GitHub	NLTK	
Project Gutenburg	Tableau	Gensim	
		sklearn	
		pyLDAvis	

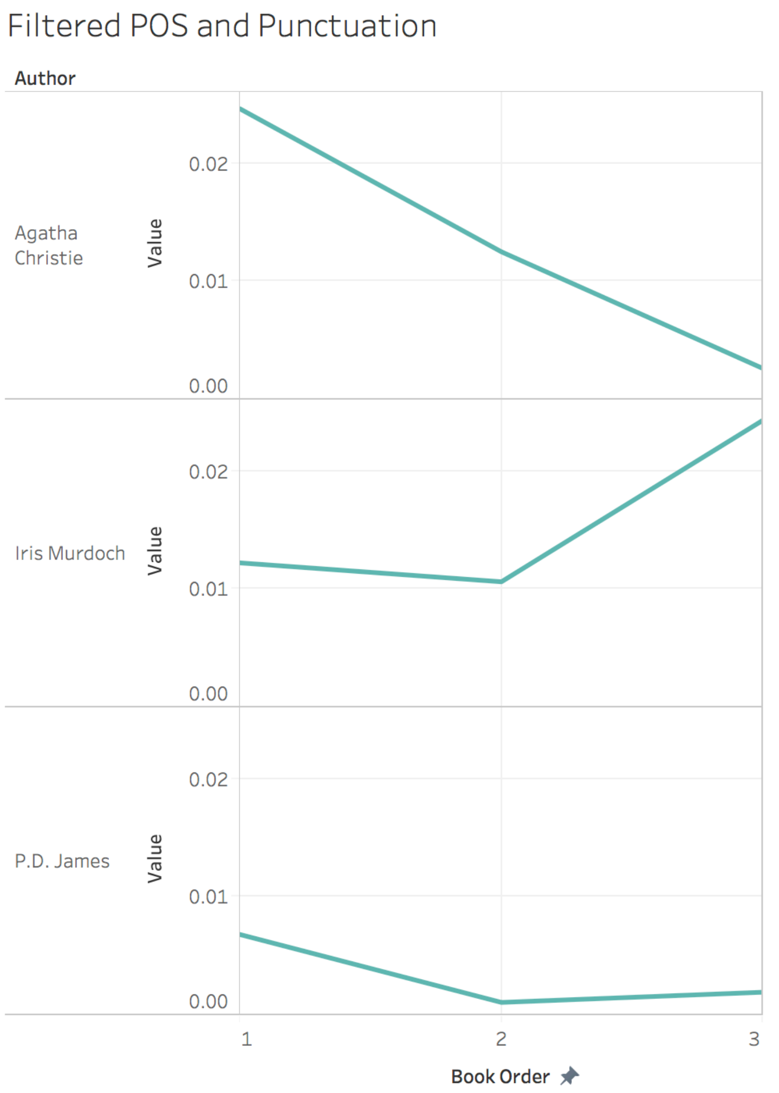
Data Preprocessing

The following "features" were extracted from 9 novels (3 each per author).

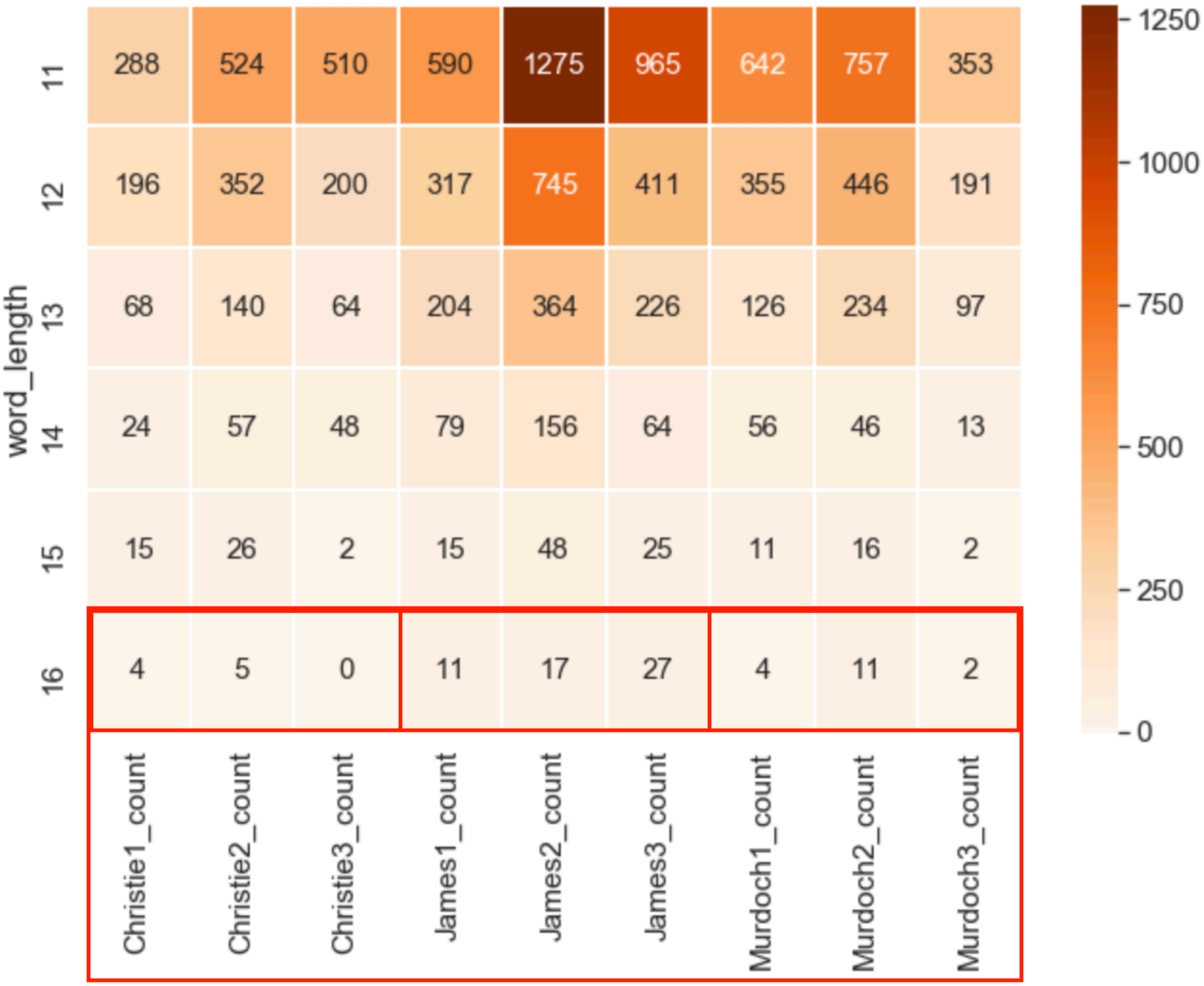
- counts of all various punctuations
- Tokenizing, tagging and counting vocabulary usage, parts of speech, and sentence lenth
- Get sentiment for each sentence and average it over all sentences in the book
- Do the above with and without stop words for analysis and topic modeling respectively

Data Exploration, Visulizations, and Results

Much time was spent on going through all of the parts of speech and punctuation in hopes of discovering either an increase or decrease in usage. For example, the below graph demonstrates usage of exclamation points and its trend over the 3 books per author. I honed in on the parts of speech that had the most significant increase/decrease and made an attempt to determine if these changes were statistically significant using a **chi squared** test. All seemed to result in low p-values that were suspiciously statistically significant.

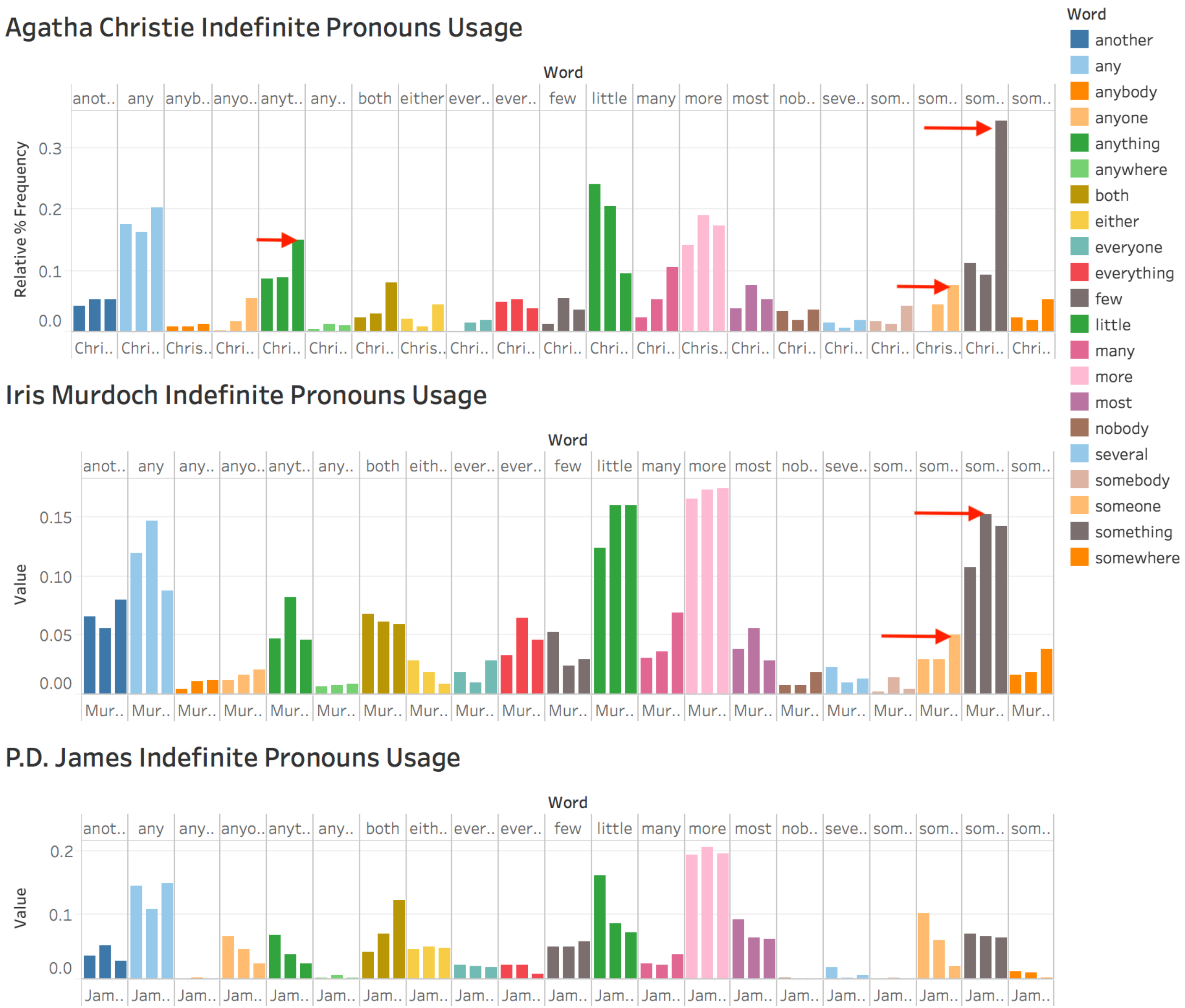


I compared the length of the words used and the frequency of use over time and created a heatmap in order to see any differences. It appears that the last row of words of length 16 characters shows a decline for the two authors who did have Alzheimers and an increase for the author who did not. This is not necessarily significant, but still interesting to note.



word_length	Christie1_count	Christie2_count	Christie3_count	James1_count	James2_count	James3_count	Murdoch1_count	Murdoch2_count	Murdoch3_count
11	288	524	510	590	1275	965	642	757	353
12	196	352	200	317	745	411	355	446	191
13	68	140	64	204	364	226	126	234	97
14	24	57	48	79	156	64	56	46	13
15	15	26	2	15	48	25	11	16	2
16	4	5	0	11	17	27	4	11	2

Another interesting thing to note is the increase of usage in may **indefinite pronouns** which according to the research is apparent in Alzheimers patients. In this example, the word "something" has a significant increase in usage amongst authors Christie and Murdoch, but is barely used in P.D. James' novels. Other indefinite pronouns also demonstrate an increase in usage.



Last noteable result is the decrease in unique vocabulary usage over time. All three authors have a signicant decrease.

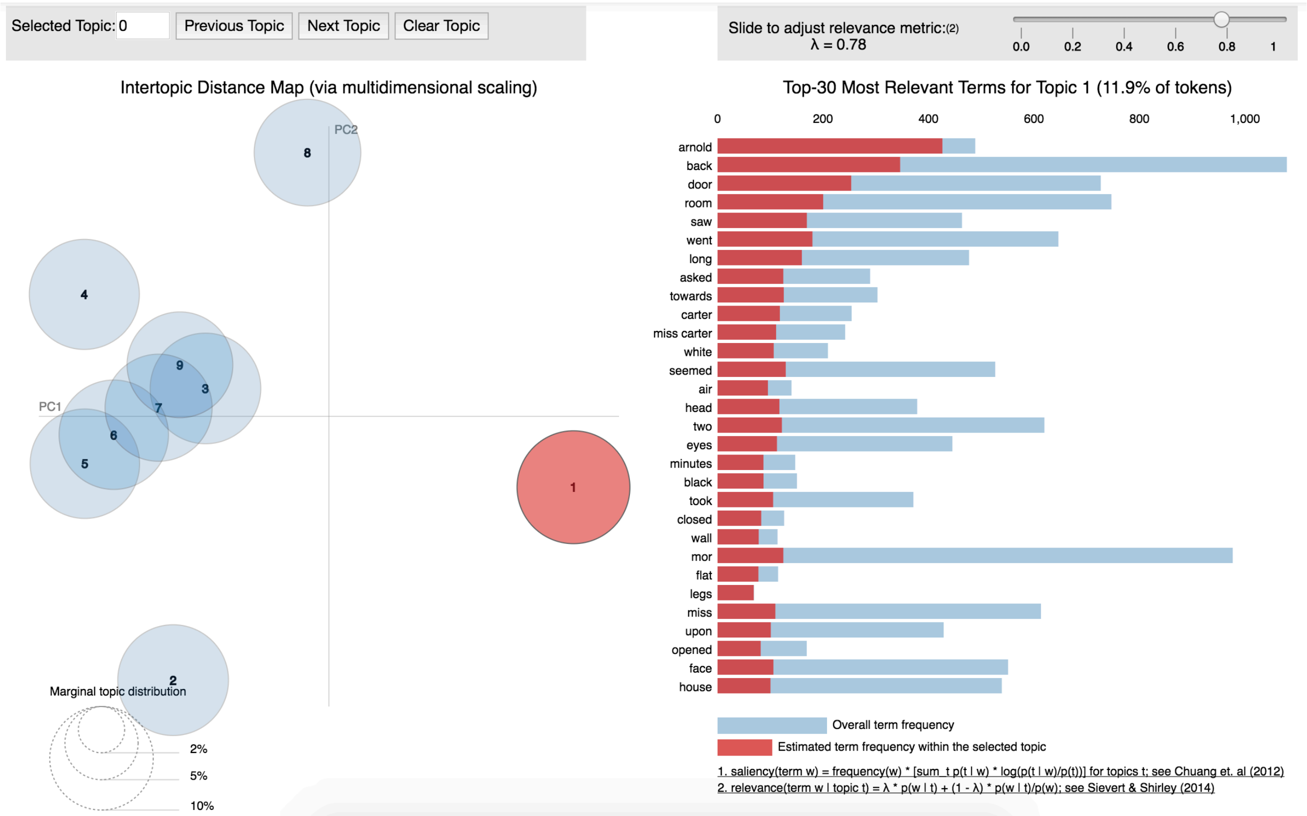
```
Cristie vocabulary rate of change: 0.27394372651452664
James vocabulary rate of change: 0.19186046511627908
Murdoch vocabulary rate of change: 0.272967994473866
```

Topic Modeling and Clustering

I chose to use LDA mostly out of curiousity to see how the algorithm would cluster the 9 books. In preparation for the LDA I removed the stop words and used 2-3 length "n-grams" as part of the tokens. Each sentence across all books was used as a "document" for the model. Below is a screen shot of the results.

The top 20 terms of the "topics" are listed below.

```
['like', 'time', 'love', 'would', 'one', 'know', 'could', 'perhaps', 'people',
'much', 'really', 'let', 'go', 'come', 'something', 'thought', 'way', 'think',
'back', 'never']
['would', 'see', 'think', 'felt', 'want', 'could', 'get', 'julian', 'must',
'mrs', 'like', 'go', 'going', 'make', 'need', 'leave', 'sort', 'come', 'may',
'soon']
['yes', 'oh', 'good', 'would', 'know', 'mrs', 'elizabeth', 'come', 'one', 'told',
'edward', 'back', 'man', 'knew', 'made', 'quite', 'thought', 'room', 'miss',
'came']
['could', 'priscilla', 'one', 'see', 'thought', 'donald', 'go', 'sorry', 'would',
'door', 'rosalind', 'felt', 'back', 'us', 'much', 'first', 'moment', 'even',
'mr', 'pain']
['know', 'would', 'right', 'course', 'one', 'please', 'last', 'night', 'mr',
'could', 'arms', 'take', 'time', 'never', 'sir', 'oh', 'wickham', 'got', 'find',
'dark']
['away', 'go', 'well', 'must', 'one', 'began', 'god', 'put', 'got', 'tim',
'might', 'think', 'little', 'hand', 'mind', 'mrs', 'could', 'mor', 'mean', 'oh']
['one', 'could', 'would', 'looked', 'even', 'mor', 'well', 'look', 'never',
'see', 'al', 'done', 'thought', 'know', 'time', 'back', 'rain', 'quite', 'got',
'might']
['like', 'could', 'say', 'something', 'man', 'nothing', 'door', 'stephen',
'reached', 'francis', 'julian', 'one', 'mor', 'saw', 'side', 'road', 'would',
'little', 'came', 'began']
['arnold', 'back', 'door', 'room', 'went', 'saw', 'long', 'seemed', 'towards',
'mor', 'asked', 'two', 'carter', 'head', 'time', 'one', 'eyes', 'miss carter',
'miss', 'white']
```



Selected Topic: 0

Slide to adjust relevance metric:(2)
λ = 0.78

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Given more time...

If there were more time, it would be interesting to explore some other avenues of the dataset.

- Incorporating more books from each of the authors.
- Find "documents" such as speeches, transcripts or writings of other well known people who have been known to have dementia or Alzheimers.
- Do more statistical anaylsis of the differences of the counts/frequencies.
- Gain access to the "Dementia DataBase" that is currently onle accessable to researchers that qualify.