NLP Literature Analysis Report

Yana Pyryalina

University of California San Diego

July 2019

Table of Contents

Problem Statement

Reading is great. And with so many amazing books out there also come great movies, reviews, and summaries. Reading those reviews and watching those films often only gives us a picture of what the book is actually like, though. With the power of data science and natural language processing, I am able to bring another dimension to how we understand literature.

For this project, I look at the following eight writings:

- **The Foundation by Isaac Asimov** - a book I am currently reading, by my favorite sci-fi writer
- **A Clockwork Orange by Anthony Burgess** - the writing behind a famous extravagant horror movie by Stanley Kubrik, a book with a unique writing style and vocabulary
- **Comments to the Society of the Spectacle by Guy Debord** - a continuation of a book I was taught in university about the influence of the capitalist media on society
- **A Brief History of Time by Stephen Hawking** - a book that excited millions about the workings of our universe
- **For Whom the Bell Tolls by Ernest Hemingway** - a book with unique writing style and themes specific to American writers
- **Carrie by Stephen King** - one of the most well-known horrors out there
- **The Hobbit by J.R.R. Tolkien** - a very long journey by very short people, one that so many people and communities hold dear to their heart
- **Slaughterhouse Five by Kurt Vonnegut** - a book highly recommended to me

I chose these writings both out of personal interest and because of their **unique subjects, settings, and writing styles.**

Process

1. Data Collection and Cleaning

1. **Finding data**
    - go to Archive.org and found .txt versions of the above-mentioned books
2. **Collecting data**
    - use data scraping using the requests and Beautiful Soup python libraries to acquire the data
3. **Cleaning the Data**
    - Corpus
        - Create a pandas data frame
        - Round 1 Cleaning - delete new lines
        - Round 2 Cleaning **-** clean up things like copyright notes
    - Document-Term Matrix
        - Round 3 Cleaning - tokenize text (i.e. lowercase, remove punctuation, remove digits)
        - Create a document-term matrix using CountVectorizer

2. Exploratory Data Analysis

1. Most common words - **Word Clouds**
    - Find top 30 words said by each author
    - Exclude words that appear in more than 50% of the books
    - Create word cloud using WordCloud and matplotlib libraries
2. **Vocabulary and Length** - Unique and Total Words
    - Find non-zero items in the document-term matrix and input the numbers into a new data frame
    - Find the total number of words that a writer uses
3. Bar plot and Scatter plot **findings**
    - Make a bar-plot of unique and total words of author using numpy and matplotlib
    - Make a scatter-plot of Book-Length vs Vocabulary (unique words) using matplotlib

3. Sentiment Analysis

1. Sentiment of books **overall**
    - Create lambda functions for polarity and subjectivity using TextBlob
    - Plot the data using matplotlib
2. Sentiment of books **over time**
    - Create a function to split each writing into 40 pieces using numpy and math
        - *40 pieces ended up a good balance between too little vs too much detail*

- Plot the data using matplotlib

4. Topic Modeling (using Latent Dirichlet Allocation)

- *I found LDA to be a good choice of tool for this project due to the interpretability of topics. Working with large and complex pieces of data like literature works, I wanted to see if I as a reader could pick up latent/hidden topics as a result of my data analysis, which was successful (as shown in the Results section)*

1. Topic modeling based on the **entire** original document-term matrix (all parts of speech)
   - Input the document-term matrix, transpose, and transform into gensim corpus required by the LDA (Latent Dirichlet Allocation)
   - Create a dictionary of all terms and their respective location in the term-document matrix
   - Specify number of topics and passes, and run the LDA model

2. **Extract parts of speech** for topic model
   - *I chose to use parts-of-speech extraction mainly to improve the results of my LDA model*

   - **Nouns only**
     - *Filtering and using only nouns for our LDA model could be an improvement simply because topics themselves (like "war" or "love") are nouns. For example, in the sentence "The war was dreadful and unforgiving, but it could not destroy their love" we could simply extract nouns "war" and "love", showing the sentence's themes much clearer.*
     - Create a function to tokenize given text and extract only the nouns
     - Input the clean data
     - Apply the noun-filtering function
     - Create a new document-term matrix using only nouns
     - Create a new gensim corpus (based on the new document-term matrix)
     - Create a new vocabulary dictionary
     - Test the LDA model, gradually increasing the number of topics
       - *Starting off with fewer topics and gradually increasing their number helped me track down at what point my LDA model starts giving redundant results (at what point and how its performance deteriorates) so that I can see what to fix*

   - **Nouns and adjectives**
     - *At the same time, adjectives could also be useful for making a better term-document matrix for our LDA model. For example, in the sentence "Her face was bloody and demonic", words like "bloody" and "demonic" hint at the themes much better than the word "face".*
     - Repeat the above process with nouns & adjectives

- **Final Model**
  - Take the most recent noun+adjective function, set the topic number to 5 and pass number to 100
    - *After gradually increasing the number of topics and number of times for the model to run, 5 topics and 100 passes ended up giving topic clusters that made sense the most in connection to the books' known themes*

5. Text Generation using Markov Chains

- *I chose Markov Chains as a starting point for text generation, as I wanted to start my learning process from simpler approaches, as well as to see the extent of the success of this approach when working with larger texts like works of literature. Surprisingly, Markov Chains resulted in text generator that preserved the writers' styles very visibly, despite of course having a limit to its capacity after couple of dozens of iterations.*

1. Select text to imitate
2. Create a **Markov Chain function** and apply it to an author
   - *For this task with Markov Chains, I chose to use the defaultdict library because it allows to efficiently create a dictionary chain of words from our corpus on the go (in a loop), avoiding KeyErrors*
   - Import defaultdict library from collections
   - Input a string of text
   - Tokenize the text by word, but include punctuation
   - Initialize a default dictionary to hold all of the words and their connections
   - Create a zipped list of all of the word pairs and put them in a dictionary format (word : list of connecting words)
   - Convert the default dictionary back into a regular dictionary and return it
3. Create a **Sentence Generator function**
   - Input the chosen author's Markov Chain and desired length of sentence
   - Capitalize the first word to create a more sentence-like formatting
   - Follow the Markov Chain text generation method by generating the second word from the value list, then third, and repeating until the end of sentence
   - End with a period to complete the sentence formatting
4. Generate sentences based on writer

Tools used

1. Data Collection and Cleaning

- Collection - requests, Beautiful Soup, pickles
- Cleaning and visualizing - pandas, lambda functions, regular expressions, CountVectorizer

2. Exploratory Data Analysis

- Word Clouds (finding most common words) - Counter, sklearn (text, CountVectorizer), WordCloud, matplotlib
- Unique and total words + Visualizing - pandas, numpy, matplotlib

3. Sentiment Analysis

- pandas, TextBlob (+ lambda functions), numpy, math

4. Topic Modeling

- LDA (Latent Dirichlet Allocation), gensim, scipi.sparse, sklearn (text, CountVectorizer)
- Extracting Parts of Speech - nltk (word_tokenize, pos_tag)

5. Text Generation

- Markov Chains (+ defaultdict), random, defaultdict

Challenges

- Finding free e-book content in a standardized web-page format
- Learning pandas, regular expressions, lambda functions, visualization tools
- Finding the optimal data and number of topics in topic modeling
- Finding the right approach to extract parts of speech
- Cleaning out each book's different tables of content, notes about the authors, copyright notes, repeating titles

Accomplishments

- Realistic text generator that preserves the style of the author
- Surprising insights into famous writing that I would not have been able to capture without data analysis
- An enjoyable learning process

Results

Data

Before cleaning

	writing
asimov	\nFOUNDATION \n\nCUE^.TIhJIIT^«I T-WCf Hi \n\n\nTHE FOUNDATION HOVELS \n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\nF ...
burgess	Being the adventures of a young man \nwhose principal interests are rape, \nultra-violence and Beethoven. \n\n\n\nA CLOCKWORK ORANGE (UK Version) ...
debord	Comments on \nthe Society of the \nSpectacle. (1) \n\n\n\nGuy Debord \n\n\n\nIn memory of Gerard Lebovici, assassinated in Paris on \n5 March 1984...
hawking	A Brief History of Time - Stephen Hawking \n\n\nA Brief History of Time \n\nStephen Hawking \n\n\nChapter 1 - Our Picture of the Universe \n\nChap...
hemingway	FOR WHOM \nTHE BELL TOLLS \n\nby \n\nERNEST HEMINGWAY \n\n\n\nJONATHAN CAPE \nTHIRTY BEDFORD SQUARE \nLONDON \n\n\n\nFIRST PUBLISHED. MARCH 1923 \...
king	\n\nSTEPHEN \n\nKING \n\nCarrie \n\n\nDOUBLEDAY \n\nNew York London Toronto Sydney Auckland \n\n\n\nCARRIE \n\n\n\nContents \n\n\nTitle Page \n\nD...
tolkien	\n\n\n\n\n\n\nTHE HOBBIT \n\n\n\nOR \n\nTHERE AND BACK \nAGAIN \n\n\n\nJ.R.R. TOLKIEN \n\n\n\nThe Hobbit is a tale of high adventure, \nundertaken...
vonnegut	SLAUGHTERHOUSE-FIVE \n\nOR THE C HIL DREN'S \nCRUSADE \n\nA Duty-dance with Death \nKURT VONNEGUT, JR. \n\n[NAL Release #21] \n\n[15 jan 2001 - OC...

Corpus (after cleaning)

	writing	full_name
asimov	The date was August 1, 1941. World War II had been raging for two years. France had fallen, the Battle of Britain had been fought, and the Soviet ...	Asimov - The Foundation
burgess	Being the adventures of a young man whose principal interests are rape, ultra-violence and Beethoven. I first published the novella A Clockwork Or...	Burgess - Clockwork Orange
debord	Comments on the Society of the Spectacle. (1) Guy Debord In memory of Gerard Lebovici, assassinated in Paris on 5 March 1984, in a trap that remai...	Debord - Comments to the Society of the Spectacle
hawking	I didn't write a foreword to the original edition of A Brief History of Time. That was done by Carl Sagan. Instead, I wrote a short piece titled "...	Hawking - A Brief History of Time
hemingway	CHAPTER ONE He lay flat on the brown, pine-needled floor of the forest, his chin on his folded arms, and high overhead the wind blew in the tops o...	Hemingway - For Whom the Bell Tolls
king	This is for Tabby, who got me into it — and then bailed me out of it. CARRIE PART ONE BLOOD SPORT News item from the Westover (Me.) weekly Enterpr...	King - Carrie
tolkien	The Hobbit was first published in September 1937. Its 1951 second edition (fifth impression) contains a significantly revised portion of Chapter V...	Tolkien - The Hobbit
vonnegut	The cattle are lowing, The Baby awakes, But the little Lord Jesus No crying He makes. One All this happened, more or less. The war parts, anyway, ...	Vonnegut - Slaughterhouse Five

Clean data (input for Document-Term Matrix)

	writing
asimov	the date was august world war ii had been raging for two years france had fallen the battle of britain had been fought and the soviet union had ...
burgess	being the adventures of a young man whose principal interests are rape ultraviolence and beethoven i first published the novella a clockwork orang...
debord	comments on the society of the spectacle guy debord in memory of gerard lebovici assassinated in paris on march in a trap that remains mysterio...
hawking	i didnt write a foreword to the original edition of a brief history of time that was done by carl sagan instead i wrote a short piece titled ackno...
hemingway	chapter one he lay flat on the brown pineneedled floor of the forest his chin on his folded arms and high overhead the wind blew in the tops of th...
king	this is for tabby who got me into it — and then bailed me out of it carrie part one blood sport news item from the westover me weekly enterprise a...
tolkien	the hobbit was first published in september its second edition fifth impression contains a significantly revised portion of chapter v riddles in...
vonnegut	the cattle are lowing the baby awakes but the little lord jesus no crying he makes one all this happened more or less the war parts anyway are pre...

Document-term matrix

	aa	aaaaaaah	aaaaaaarhgh	aaaaah	aarguing	aat	ab	aback	abacus	aban	...	zone	zoo	zoobies	zooby	zooful	zoomed	zophar	zvono
asimov	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
burgess	0	1	1	2	0	0	0	0	0	0	...	0	5	21	2	0	0	1	
debord	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
hawking	0	0	0	0	0	0	0	0	5	0	...	0	0	0	0	0	0	0	
hemingway	1	0	0	0	0	4	1	0	0	1	...	0	0	0	0	0	0	0	
king	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	1	0	0	
tolkien	0	0	0	0	1	0	0	1	0	0	...	0	0	0	0	0	0	0	
vonnegut	0	0	0	0	0	0	0	0	0	0	...	0	16	0	0	0	1	0	

8 rows × 24231 columns

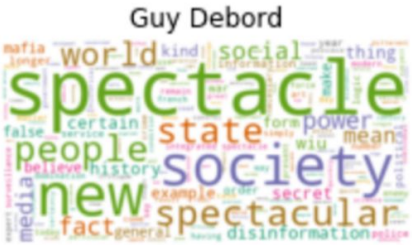
Word Clouds

Isaac Asimov

Anacreon vs Terminus - Anacreon, a smaller but rapidly developing planet, seems to be much more of a concern or interest than Terminus – Anacreon's more developed and seemingly more important neighbor

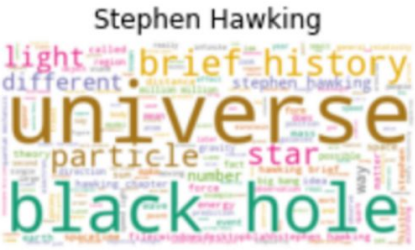
Anthony Burgess

Root of confusion - The word cloud is about as confusing as the actual book. One reason for such confusion could be because other than Brother, most words are of Russian origin, which makes it harder to grasp the important elements of the story



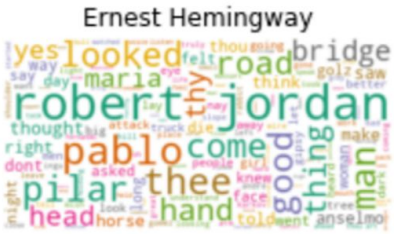
Guy Debord

Facts - A word *fact* appears a lot, which makes me wonder whether it is the author presenting his perspective as facts or it is him referring to the media's disinformation masked as "*facts*"



Stephen Hawking

Stephen Hawking seems to mention himself often enough to appear in the word cloud!

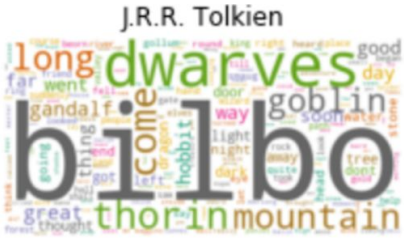


Ernest Hemingway

Hemingway seems to have the most homogeneous word cloud - other than *robert* and *pablo*, most other words seem to be pretty evenly distributed and used. What could that imply about his writing?

Stephen King

In contrast to Hemingway, Stephen King's word cloud seems to be pretty heterogeneous, with only two characters and two body parts mentioned most often



J.R.R. Tolkien

Sorry Gandalf...But the goblin seems to be more popular!

Based on the popularity of words *long*, *great*, and *far*, the journey must have definitely fit that description

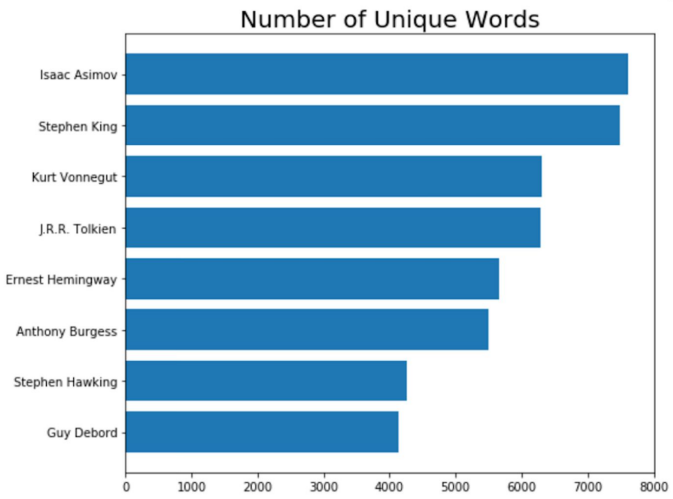
Kurt Vonnegut

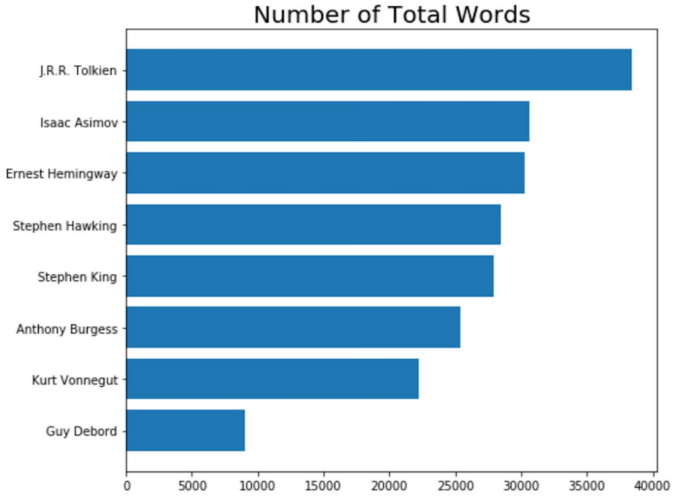
Billy, America, and War seem to be the center of the story

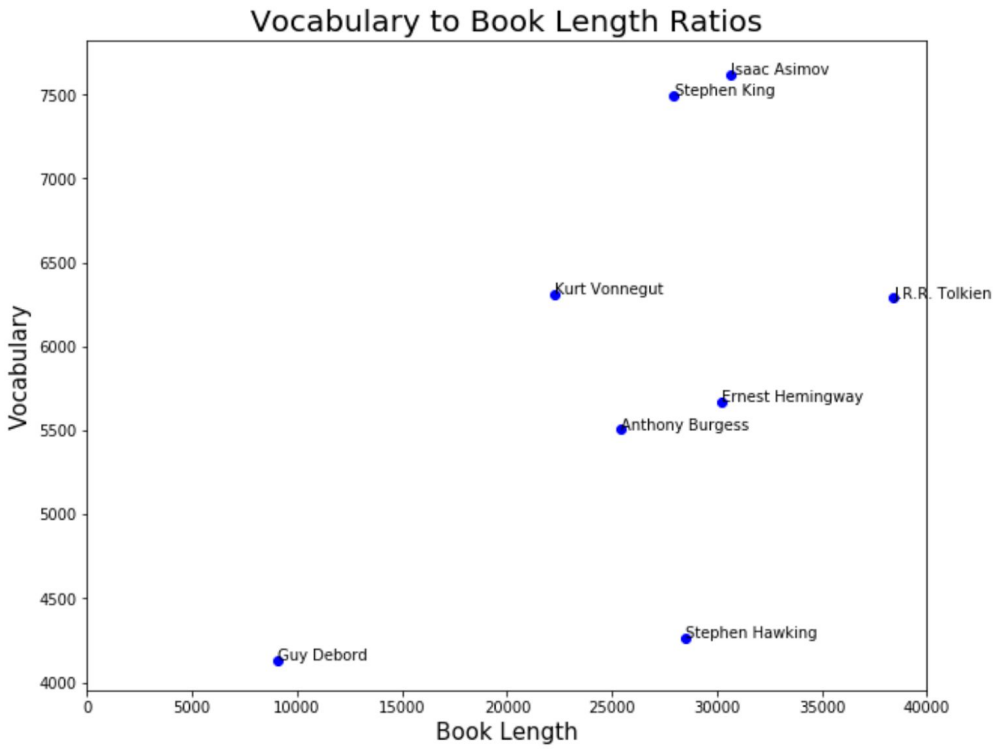
Vocabulary

As a part of Exploratory Data Analysis, I looked at the vocabulary size of each author (number of unique words), as well as its ratio to the length of the book.
Some of the findings were quite unexpected:

writer	unique_words	total_words
Guy Debord	4129	9073
Kurt Vonnegut	6309	22234
Anthony Burgess	5505	25425
Stephen King	7493	27960
Stephen Hawking	4267	28484
Ernest Hemingway	5667	30240
Isaac Asimov	7622	30640
J.R.R. Tolkien	6289	38423







Insights on Scatter Plot

- **Big universe and small vocabulary** - Surprisingly, Stephen Hawking is on the low end of vocabulary sizes. I expected his book 'A Brief History of Time' to have many technical terms, but it ended up having a low number of unique words. However, it does make sense! The purpose of this book was to explain complex astrophysics concepts in simple and understandable language, which Stephen Hawking indeed did a good job of!

- **Not enough Russian words** - It was also surprising to me to see Anthony Burgess' *A* Clockwork *Orange* on the slightly lower end of vocabulary size compared to others. With how many extravagant Russian-originated slang words there are in the book, their number still could not top Hemingway's vocabulary.

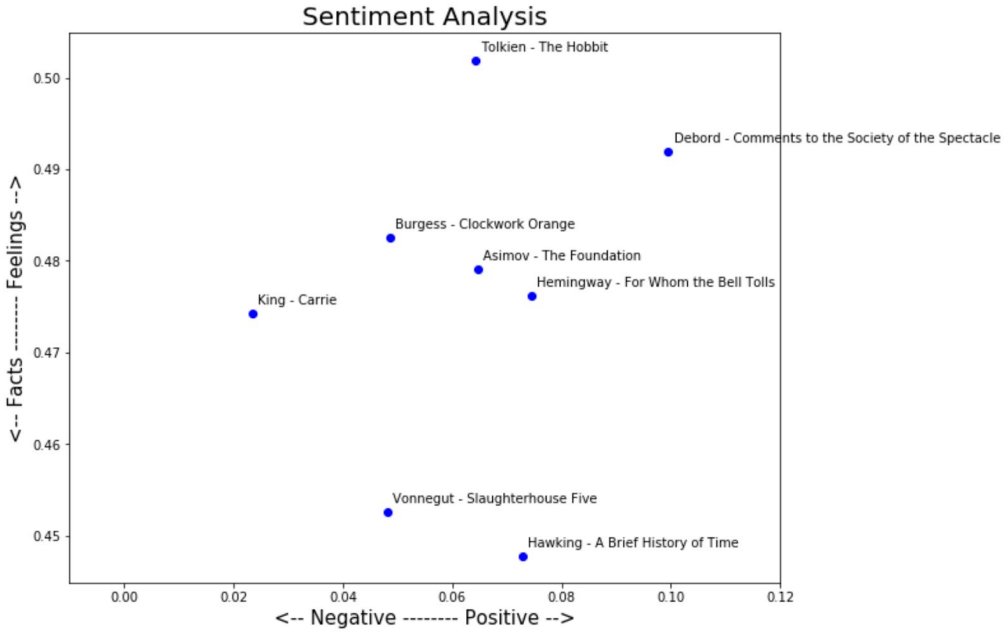
- **Horror & Sci-fi Champions** - Again surprisingly, *Carrie* (by Stephen King) and *Foundation* (by Isaac Asimov) had similarly wide vocabularies and lengths despite such different genres. Also, such a high-vocab statistic for King was a contrast to his word cloud. Looking at the word cloud (with leading words *Carrie, momma, eye, hand*), I thought his vocabulary would be pretty narrow, but, in reality, Stephen King got to the very top, almost beating Asimov's space-travel vocabulary.

- **Long journey, medium vocab** - Despite having the longest journey, The Hobbit had a relatively medium-sized vocabulary. Just like in *A* Clockwork *Orange*, I imagined that *The Hobbit*'s lexicon would be full of world-specific words, but it ended up being just a tiny bit above the average, ranking almost equal to Kurt Vonnegut's vocabulary size.

Sentiment Analysis

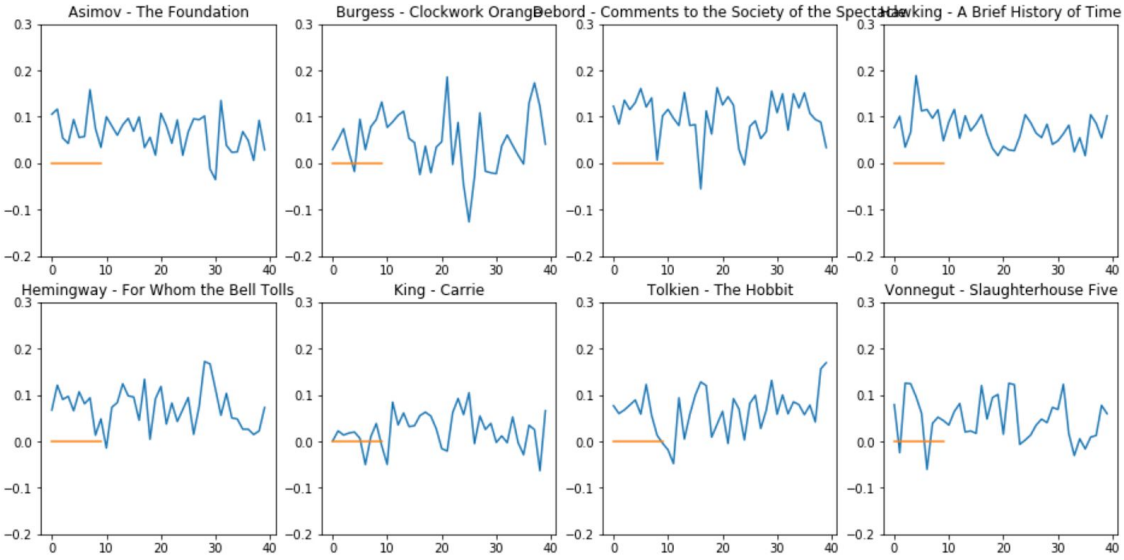
Each word in a corpus is labeled in terms of polarity and subjectivity. A corpus' sentiment is the average of these.

- **Polarity**: How positive or negative a word is. -1 is very negative. +1 is very positive. In literature this can be useful for determining whether the story's events and characters' feelings are at a low/not exciting or a high/very exciting.
- **Subjectivity**: How subjective, or opinionated a word is. 0 is fact. +1 is an opinion or feeling. In literature this could address the difference between feelings or judgements of a character or author vs actual events of the story.



- **Sentimental Hobbit** - It was surprising to see The Hobbit on the very top of subjectivity. I would think that the hobbit's journey would be filled with ____. However, this finding could hint that The Hobbit does the best job of portraying its world through the eyes of the characters.
- ***Fact*?** - In contrast to our findings of Guy Debord using the word "fact" a lot, his book was the 2nd most subjective book, especially being non-fiction
- **Fiction or Life?** - While it is not that surprising to see Stephen Hawking's writing to be most factual, it was interesting to see Kurt Vonnegut's writing to be very close on that scale. These findings could point at the book being more heavily relying on the plot rather than descriptions. At the same time, that could also be related to Vonnegut's writings being based on real-life events like WWII.

Over time



- **Positivity in outer space** - As we can see, Asimov's and Hawking's graphs look similar to each other. They both start off with a fall following a big positive spike; both generally stay on the more positive side, and both change their sentiment in smaller fluctuations. The only exception is Asimov's plot twist at 30 on x-axis. Such overall similarity is surprising due to such different Fact-VS-Feeling ratings. However, the common positivity could be influenced by the topics of the universe and its wonders, about which both authors were passionate.

- **What happened in Clockwork Orange?** - Evidently, something horrible happened in the very middle of the book. We can see a huge spike into negativity. I guess we'll have to read to find out!

- **Carrie** - It was surprising to see Carrie above the positivity line. However, the book is still overall more negative than others. This can also be seen on the overall-sentiment scatterplot, where Carrie was more negative than the others.

- **Rethinking literature?** - While learning literature in middle school and high school, I was always taught that each plot is like a pyramid: leading uphill towards one culmination in the middle, and falling downhill into a resolution. However, as we can see on these graphs, spikes of narrative happen all over the place. For example, the Hobbit's middle seems to have even and consistent spikes, creating key spikes in the beginning and end instead. Asimov, Hemingway, and Vonnegut follow a relatively similar pattern. The only writings that shift dramatically in the middle are horror writings like *Carrie* and *A Clockwork Orange*, from which the latter is arguably the most extravagant of all writings included here.

Topic Modeling

The topic modeling below was created using Latent Dirichlet Allocation (LDA), which is one of many topic modeling techniques specifically designed for text data. It is an unsupervised machine learning model. Once the topic modeling technique is applied, it is up to humans to interpret the results and see if the mix of words in each topic make sense.

Attempt #1 - All-text training data

```
[(0,
  '0.007*"bilbo" + 0.006*"jordan" + 0.006*"robert" + 0.005*"come" + 0.005*"good" + 0.004*"long" + 0.004*"dwarves" + 0.004
*"thought" + 0.004*"thee" + 0.004*"man"'),
 (1,
  '0.011*"billy" + 0.004*"foundation" + 0.004*"hardin" + 0.003*"people" + 0.003*"man" + 0.003*"war" + 0.003*"dont" + 0.003
*"mallow" + 0.003*"years" + 0.002*"way"'),
 (2,
  '0.016*"universe" + 0.008*"theory" + 0.007*"black" + 0.007*"history" + 0.007*"light" + 0.006*"brief" + 0.006*"hawking" +
0.006*"particles" + 0.004*"hole" + 0.004*"energy"'),
 (3,
  '0.006*"carrie" + 0.005*"right" + 0.004*"going" + 0.004*"brothers" + 0.004*"real" + 0.004*"went" + 0.004*"got" + 0.003*"w
hite" + 0.003*"momma" + 0.003*"dont"')]
```

- I found this model to be a decent start. However, as we can see above, the results have quite a few verbs (come, thought, dont, going, went, got) that clutter-up our word groups - the verbs don't really mean much, and if they had not been there, our results might have been more comprehensive. That is why I chose to filter out parts of speech and see if my models could be improved with it.

Attempt #2 - Nouns only

Filtered training data:

	writing
asimov	date world war ii years france battle britain ...
burgess	adventures man interests ultraviolence beethov...
debord	comments society spectacle guy debord memory g...
hawking	i foreword edition history time carl piece ack...
hemingway	chapter brown floor chin arms wind blew tops p...
king	tabby — part blood sport news item westover en...
tolkien	hobbit edition impression portion chapter v ri...
vonnegut	cattle baby awakes jesus crying war parts guy ...

Results:

```
[(0,
  '0.010*"carrie" + 0.009*"time" + 0.005*"eyes" + 0.005*"war" + 0.005*"momma" + 0.005*"school" + 0.004*"way" + 0.004*"face"
+ 0.004*"night" + 0.004*"blood"'),
 (1,
  '0.021*"time" + 0.015*"universe" + 0.009*"theory" + 0.008*"history" + 0.008*"brothers" + 0.006*"particles" + 0.006*"way"
+ 0.005*"bit" + 0.005*"hole" + 0.005*"energy"'),
 (2,
  '0.012*"jordan" + 0.010*"bilbo" + 0.010*"time" + 0.008*"dwarves" + 0.007*"man" + 0.007*"way" + 0.007*"head" + 0.006*"roa
d" + 0.006*"bridge" + 0.005*"pablo"'),
 (3,
  '0.011*"foundation" + 0.008*"hardin" + 0.007*"time" + 0.006*"man" + 0.005*"years" + 0.005*"power" + 0.005*"way" + 0.005
*"empire" + 0.005*"seldon" + 0.004*"mallow"')]
```

- After filtering nouns, this model definitely seemed like a big improvement. We got to see a group with more “creepy” words (*carrie, momma, school, face, night, blood*), more science-y words, hobbit-y words, and power-related words (*foundation, man, power, empire*). This was a huge encouragement, however, the topics and words did not seem very inclusive of all books and possible topics. That is why I wanted to see what happens when we incorporate adjectives.

Attempt #3 - Nouns and adjectives

Filtered training data:

	writing
asimov	date august world war ii years france battle b...
burgess	adventures young man principal interests rape ...
debord	comments society spectacle guy debord memory g...
hawking	i foreword original edition brief history time...
hemingway	chapter flat brown floor forest chin folded ar...
king	tabby — part blood sport news item westover we...
tolkien	hobbit second edition fifth impression portion...
vonnegut	cattle baby awakes little lord jesus crying wa...

Results:

```
[(0,
  '0.011*"universe" + 0.006*"carrie" + 0.006*"theory" + 0.005*"foundation" + 0.004*"brief" + 0.004*"particles" + 0.004*"har
din" + 0.003*"momma" + 0.003*"energy" + 0.003*"mallow"'),
 (1,
  '0.014*"jordan" + 0.011*"robert" + 0.008*"brothers" + 0.007*"thee" + 0.007*"pilar" + 0.007*"pablo" + 0.006*"thou" + 0.006
*"bridge" + 0.006*"road" + 0.005*"thy"'),
 (2,
  '0.013*"bilbo" + 0.009*"dwarves" + 0.006*"thorin" + 0.005*"gandalf" + 0.005*"goblins" + 0.005*"mountain" + 0.005*"door" +
0.004*"hobbit" + 0.003*"river" + 0.003*"mountains"')]
```

- This model initially seemed like loss of improvement from the last model. As we can see, at 3 topics (and 10 passes), the results seem to be preoccupied with names of main characters (*carrie, hardin, mallow, jordan, robert, pablo, bilbo, thorin, gandalf*) which is not necessarily very descriptive of the topics. However, I still had hope that with more topics the results would be more inclusive of other words, and that with more passes the model would be better trained.

Attempt #4 - Final model - Nouns and adjectives, 5 topics, 100 passes at a time

```
[(0,
  '0.009*"carrie" + 0.006*"brothers" + 0.004*"momma" + 0.004*"veck" + 0.004*"door" + 0.004*"school" + 0.003*"dim" + 0.003
*"bed" + 0.003*"horrorshow" + 0.003*"malenky"'),
 (1,
  '0.013*"jordan" + 0.012*"bilbo" + 0.010*"robert" + 0.009*"dwarves" + 0.006*"pilar" + 0.006*"road" + 0.006*"thee" + 0.006
*"pablo" + 0.006*"bridge" + 0.006*"thorin"'),
 (2,
  '0.010*"spectacle" + 0.006*"spectacular" + 0.003*"social" + 0.003*"media" + 0.003*"disinformation" + 0.003*"false" + 0.00
2*"mafia" + 0.002*"services" + 0.002*"information" + 0.002*"example"'),
 (3,
  '0.018*"universe" + 0.009*"theory" + 0.008*"foundation" + 0.007*"brief" + 0.006*"particles" + 0.006*"hardin" + 0.005*"ene
rgy" + 0.005*"mallow" + 0.004*"seldon" + 0.004*"star"'),
 (4,
  '0.000*"carrie" + 0.000*"jordan" + 0.000*"robert" + 0.000*"road" + 0.000*"thee" + 0.000*"momma" + 0.000*"universe" + 0.00
0*"bilbo" + 0.000*"door" + 0.000*"brothers"')]
```

- Indeed, the final model with 5 topics and 100 passes yielded much better results than the mini-noun-adjective model and even the noun-only model. As we can see above, the results include more comprehensive words (including adjectives), take words from more books, and incorporate a much wider variety of topics.

After a long time of iterating the final LDA model, I ended up with these five topics:

- **Topic 1: horror:** carrie, brothers, momma, school, horrorshow
- **Topic 2: hobbit's world:** bilbo, dwarves, three, goblin, mountain, hobbit
- **Topic 3: media:** spectacle, disinformation, false, mafia, services
- **Topic 4: outer space:** universe, theory, foundation, particles, hardin, energy, star
- **Topic 5: the way somewhere:** bridge, road, door

Overall, the topics seem to make sense, as they are mainly divided by genres like sci-fi, horror, fantasy, non-fiction. However, in addition to books overlapping in their genres, one major visible overlap between all books seems to be “the way somewhere”. I find this interesting because while for humans concept of journey in stories makes sense, a computer may not have a general assumption that books (even non-fiction or scientific reports) are overall descriptions of some kind of journey (*road*) through obstacles (*bridge, door*) rather than random collections of facts or events. Through such relatively simple data analysis, a program can have a chance to gain a deeper understanding of literature - its essence and (in the future) possibly even its appeal.

Text Generation - writing style preservation

Here are some of my favorite examples of generated sentences:

Stephen King:

> 'Comment. Nobody was overcast and sanitary napkins, chanting, laughing, shrieking of course, forbade her heart..'

Guy Debord:

> 'Erasure of the century language of goods is insignificant or in this mystery is genuine.'

Stephen Hawking:

> 'Suggests, cosmic censorship hypothesis tells us the inner edge of one ignored the very close.'

J.R.R. Tolkien:

> 'Loosened his door, and giving his ring of the balance as if he had baked.'
>
> Grandmother, teaching his legs with them hoping against the chill flame, beating wigs bear to.
>
> Saucer, and Bilbo had lost the foot of the foot right away.
>
> "Try," said a gathering together! There was not forbear to sniff.
>
> What about the hobbit? He slashed the current market value"

Isaac Asimov:

> 'Sake, as to admit you speak, Hardin. And two; after Emperor would one of Imperial.'

Ernest Hemingway:

> 'Material and headed toward Segovia at the cup out and the gipsy’s voice thickening.'

Kurt Vonnegut:

> 'Architect. The atmosphere now, big kiss. She had been stolen from the opinion of his.'

Credits

Thank you so much to Alice Zhao who made such an amazing PyOhio 2018 conference workshop that inspired this project.

Check out her workshop here:

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