Final Report

Humaira Halim (hbh4bv@virginia.edu) DS 5001 Spring 2023

Code for report worked in collaboration with Nikita Amanna and Nicholas Kalinowski -- report and conclusions are all mine

Introduction.

According to the Encyclopedia Brittanica, the genre of science fiction is defined as "a form of fiction that deals principally with the impact of actual or imagined science upon society or individuals" (2023). Science fiction is a relatively modern subject in written literature; forming alongside the socioeconomic changes of the Industrial Revolution, the West is accredited for its origins as writers began contemplating the consequences of technological developments. Combining text analysis with 19th-20th century Science Fiction offers an anthropological glimpse into what the visionary's imagined as the consequences of post-Industrial Revolution.

For my project, I have sourced 6 popular scifi novels of the 19th-20th century from American authors. I was interested in what these authors imagined for the future of science, humored by my reality as a data scientist two centuries later. The text models used in this report to characterize my corpora are Principal Components (PCA), Topic Models (LDA), Word Embeddings (word2vec), and Sentiment Analysis.

Source Data.

The six American scifi books I have selected for analysis are: *Looking Backward* (1887) by Edward Bellamy, *The Iron Heel* (1998) by Jack London, *A Voyage to the Moon* (1827) by George Tucker, *The Variable Man* (1953) by Philip K Dick, *The Brick Moon* (1869) by Edward Everett Hale, and *Youth* (1952) by Isaac Asimov.

The novels in my corpus were all sourced from Project Gutenberg, an online archive dedicated to digitizing and preserving older works. The link to Project Guntenberg can be found here. Additionally, access to the raw .txt files of my corpus via UVA box can be found here.

Below is a basic summary chart of these observations:

Corpus	Avg # Words per Book:	Avg # Chapters per Book:	Most Common POS Count:
American	50192.0	18.67	57230

The average length of the books within my corpora are about 19 chapters, with a mean of 50,200 words.

Data Model.

PCA

For the PCA, I needed to generate a bag of words and tfidf. The TFIDF table is an extension of an implied DOC table, where each doc is an observation (particularly a chapter in this case). PCA results in two tables: loadings (language model) and docs and components (which replaces the original document-term matrix with a reduced version).

Part of the loadings table is shown below:

term_str	PC0	PC1	PC2	PC3	PC4	PC5	PC6
year	-0.053656064999109555	0.007769112154987246	-0.02878947352014494	-0.03178593362306816	0.005842871129240983	-0.001034956693584325	-0.0
use	-0.04235574949630194	0.003950406839817844	-0.004963310600787502	-0.027388008882604634	0.006797976434846156	0.003436229283808301	-0.0
cause	-0.03327047644698494	0.01241832318761675	0.0020547234795388336	0.03208373509780746	0.020771388115637945	0.0011205488581936133	-0.0
morning	-0.04928309722131909	0.01183616159508366	0.0010564573333013971	-0.034885849100616984	0.05696105899176856	-0.030138512678318366	0.0
hope	-0.033962139641387654	-0.0018585980091765375	0.01586112871718469	-0.002018148545843118	0.0007440462074681797	-0.013197783706713446	-0.0
return	-0.04448720776803042	-0.007500068732421816	-0.004626207785665526	-0.007941501438849086	0.0325203701932573	0.006219727638008061	-0.0
words	-0.03898372888340651	0.0027600538988942214	0.014251684711202528	-0.034586975138353904	0.020473615291198205	0.018979768559578294	0.0
oh	-0.03533760547154869	0.01687633567082296	0.04893292854061487	-0.03917377229713326	0.0004186188346710362	0.041294227561037684	0.0
sense	-0.037288446910479	0.013376054071256478	0.0055738162428412424	-0.043036305431505884	0.038337202580614016	0.01784499687572556	0.0
word	-0.03885264912945833	0.01865350545733442	0.00403003789095973	0.00303106277219993	-0.012205549093648238	0.030748293682964667	0.0

Part of the documents and components table is shown below:

pc_id	eig_val	year	use	cause	morning	hope	return
PC0	0.12240732155130885	-0.053656064999109555	-0.04235574949630194	-0.03327047644698494	-0.04928309722131909	-0.033962139641387654	-0.04448720776803042
PC1	0.04720190529366207	0.007769112154987246	0.003950406839817844	0.01241832318761675	0.01183616159508366	-0.0018585980091765375	-0.007500068732421816
PC2	0.03254496984098815	-0.02878947352014494	-0.004963310600787502	0.0020547234795388336	0.0010564573333013971	0.01586112871718469	-0.004626207785665526
PC3	0.028144997490662817	-0.03178593362306816	-0.027388008882604634	0.03208373509780746	-0.034885849100616984	-0.002018148545843118	-0.007941501438849086
PC4	0.022753192309548545	0.005842871129240983	0.006797976434846156	0.020771388115637945	0.05696105899176856	0.0007440462074681797	0.0325203701932573
PC5	0.019776502417202267	-0.001034956693584325	0.003436229283808301	0.0011205488581936133	-0.030138512678318366	-0.013197783706713446	0.006219727638008061
PC6	0.018078172793520673	-0.008955590330203607	-0.02566859395917397	-0.018980853246527624	0.04461308557610259	-0.0011635449598358522	-0.008879968797189886
PC7	0.01738926635447485	0.0011833371823402996	0.019252088520819246	-0.03336674502026897	-0.034747048833704455	0.00140073858543531	0.013252652718220494
PC8	0.01670276917729117	0.01665640511185262	0.02293204648345435	-0.0007557122128168253	0.012475764639582866	-0.01228022480512192	-0.009548403104679554
PC9	0.015982039974866184	0.00695975931402156	-0.025466779883817924	0.012550735495139138	0.023852175162067057	0.0010386605184319608	-0.012405445596236776

LDA

For topic models, I uses Scikit-Learn's CountVectorizer function to convert the F1 corpus into a document-term (aka DTM) vector space of word counts. From there, I could use Scikit-Learn's LatentDirichletAllocation algorithm to extract the THETA (doctopic where counts get normalized into PDFs) and PHI (topicword where counts get normalized into PDFs) tables. I also created a TOPICS table:

topic_id	0	1	2	3	4	5	6	7	8	label	doc_weight_sum_amer	term_freq
T00	time	machines	war	bomb	thing	arm	profits	strength	power	T00 time, machines, war, bomb, thing, arm, profits, strength, power	449.17471007016456	0.08499057
T01	man	yes	moon	face	right	course	cart	time	street	T01 man, yes, moon, face, right, course, cart, time, street	453.8507121494376	0.08704988
T02	men	way	unions	business	city	police	field	end	time	T02 men, way, unions, business, city, police, field, end, time	382.7191615173024	0.078139211
T03	time	man	character	ship	earth	people	animals	view	purpose	T03 time, man, character, ship, earth, people, animals, view, purpose	416.67780103041076	0.10683434
T04	said	way	eyes	moment	face	house	door	room	hand	T04 said, way, eyes, moment, face, house, door, room, hand	548.9678580305524	0.09749330
T05	men	life	mind	time	world	women	man	century	society	T05 men, life, mind, time, world, women, man, century, society	471.378629056762	0.122914802
T06	earth	air	night	moon	time	day	course	man	machine	T06 earth, air, night, moon, time, day, course, man, machine	377.03764345126405	0.09173043
T07	life	world	time	surplus	country	years	man	things	day	T07 life, world, time, surplus, country, years, man, things, day	453.45356679384804	0.095354312
T08	labor	class	men	nation	day	time	people	course	army	T08 labor, class, men, nation, day, time, people, course, army	545.0433751072758	0.16442362
T09	music	father	room	dont	children	day	course	meat	dollars	T09 music, father, room, dont, children, day, course, meat, dollars	395.69654279301164	0.07106949

this table includes a list of the topics (0-8), most associated words by each topic, document weight and term frequency.

Word Embeddings (word2vec)

For word embeddings, it was necessary to import our TOKENS tables and generate a DOCS table for Gensim. Using Genim's word2vec function and the tsne engine, it was possible to vectorize terms into a semantic space and generate coordinates. As we've learned in lecture, the tsne function

embodies a clustering method applied to high dimensional vectors, comparing pairwise similarities into probabilities. I set the parameters to vector_size=256, window=2, and min_count=50.

Sentiment Analysis

For sentiment analysis, I needed to generate a vocab sentiment table. I imported the Syuzhet.csv file which provided a table of words and their "sentiments". Our generated table incorporates dfidf, ifidf, syu_sentiment and weighted_sentiment.

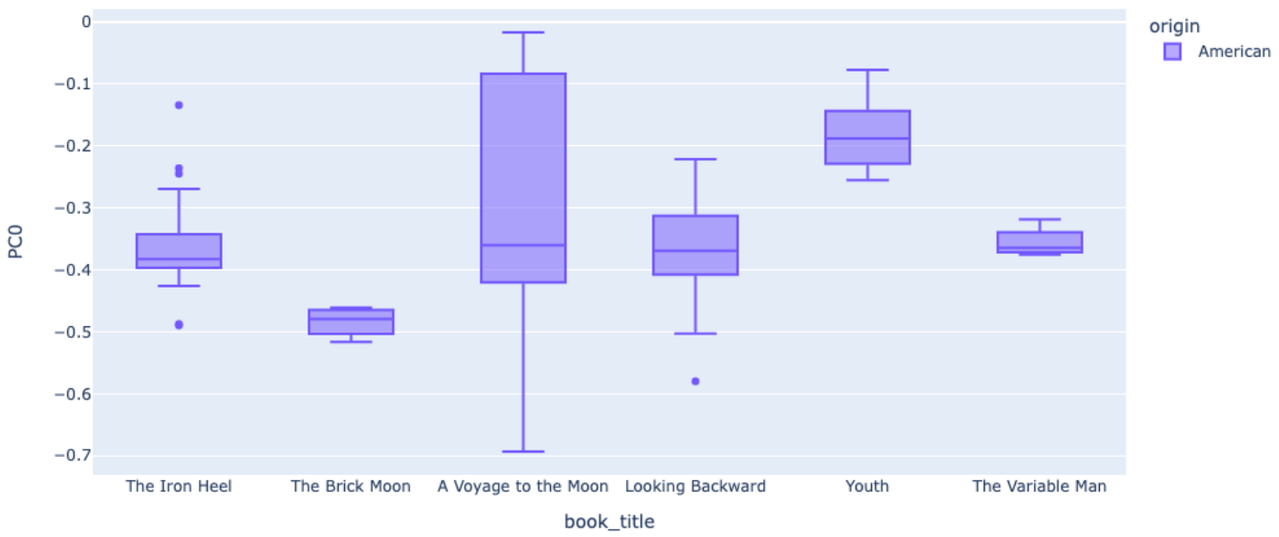
term_str	n	tfidf_x	dfidf	syu_sentiment	weighted_sentiment
abandon	1.0	0.047939119169419744	6.807354922057604	-0.75	-0.03595433937706481
abandoned	1.2857142857142858	0.016160395594536105	28.0	-0.5	-0.008080197797268053
aberration	1.0	0.022540910337939084	6.807354922057604	-0.8	-0.018032728270351267
abhorred	1.0	0.04862396372898288	6.807354922057604	-1.0	-0.04862396372898288
abhorrent	1.0	0.03137029917998896	6.807354922057604	-0.5	-0.01568514958999448
abilities	1.0	0.02076060673174118	25.33435452801869	0.6	0.012456364039044708
ability	1.5833333333333333	0.018728311356214188	38.66870905603738	0.5	0.009364155678107094
abnormal	1.0	0.021140853795209948	6.807354922057604	-0.5	-0.010570426897604974

Exploration.

PCA

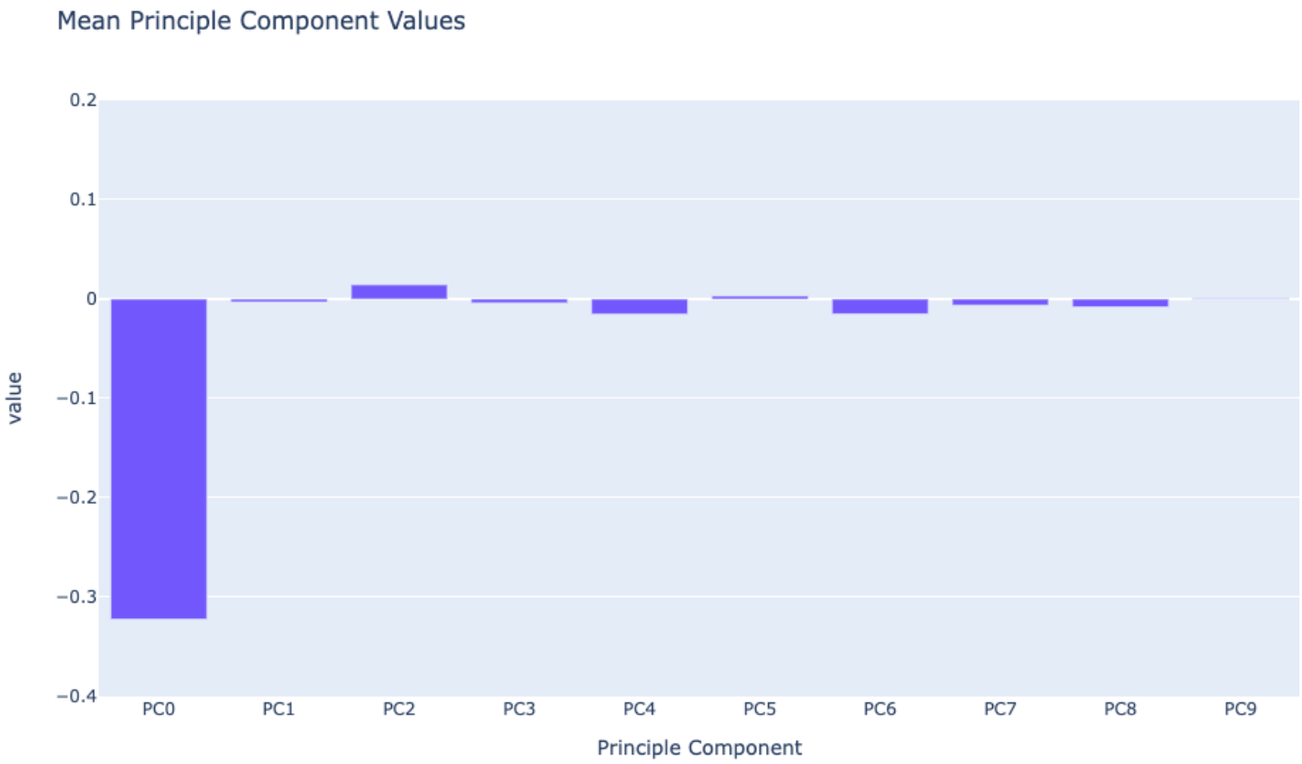
Principle Component Analysis, or PCA, identifies and combines features with max variance. Firstly, I looked at PC0 Dispersion by book to see if there was clear differences in the spread:

PCA Dispersion Plot



A Voyage to the Moon has a much larger range than the other books, whereas The Brick Moon and the Variable Man have a very small spread. Given outliers have a significant impact on the results of a PCA dispersion, it is suggestable that A Voyage to the Moon has a significant impact on our PCA results.

Next for PCA, I wanted to look at PC0-8 to see the differences in correlation by PC:

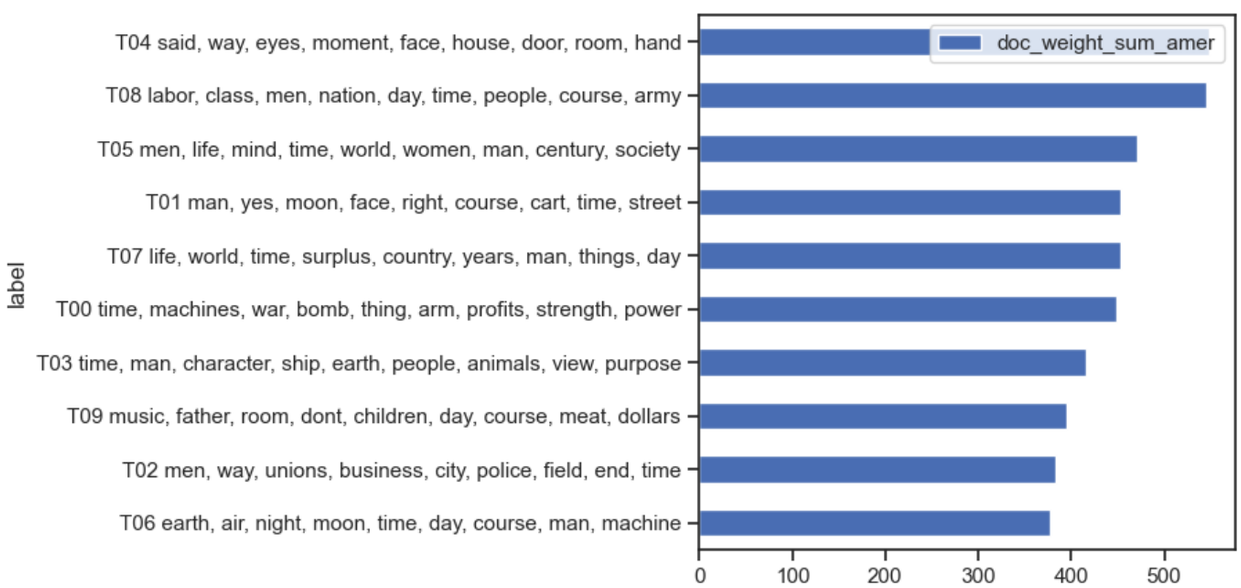


PC0 had a much stronger correlation than the PC's 1-8 which was pretty much what I expected to see.

LDA

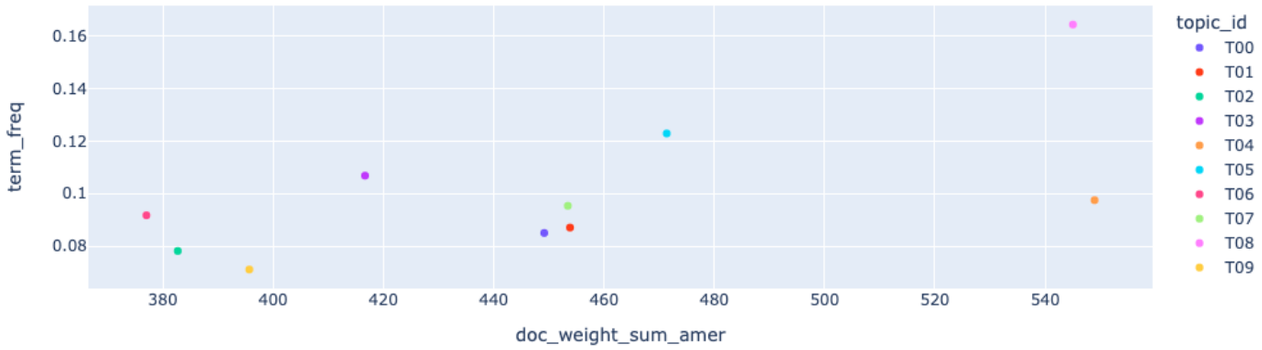
The LDA model attempts to estimate probability distributions for topics in documents plus words in topics. LDA gives us THETA (distribution of topics over documents) and PHI (distribution of words over topics) tables. Firstly, we can see how our model grouped topics by weighting each document:

TOPICS by Doc Weight



By looking at the doc weight of each topic, we can determine which topics are most strongly represented in the document. I noted that Topic 4 (said, ways, eyes, moment, face, house, door, room, hand) & Topic 8 (labor, class, men, nation, day, time, people, course, army) had the greatest document weight. This strong weight highlights the importance of these two topics in scifi. Now that we are able to understand the core terms of each topic, I wanted to see their term frequency by document weight:

TOPIC Doc Weight vs Term Freq



Analyzing term frequency allowed me to estimate the likelihood of the words within my topic to be important; plotting term frequency and document weight together gives a glimpse into the most important topics and words in my corpus. While Topic 4 wasn't as important for term frequency, Topic 8 definitively excelled in both metrics.

THETA Sample

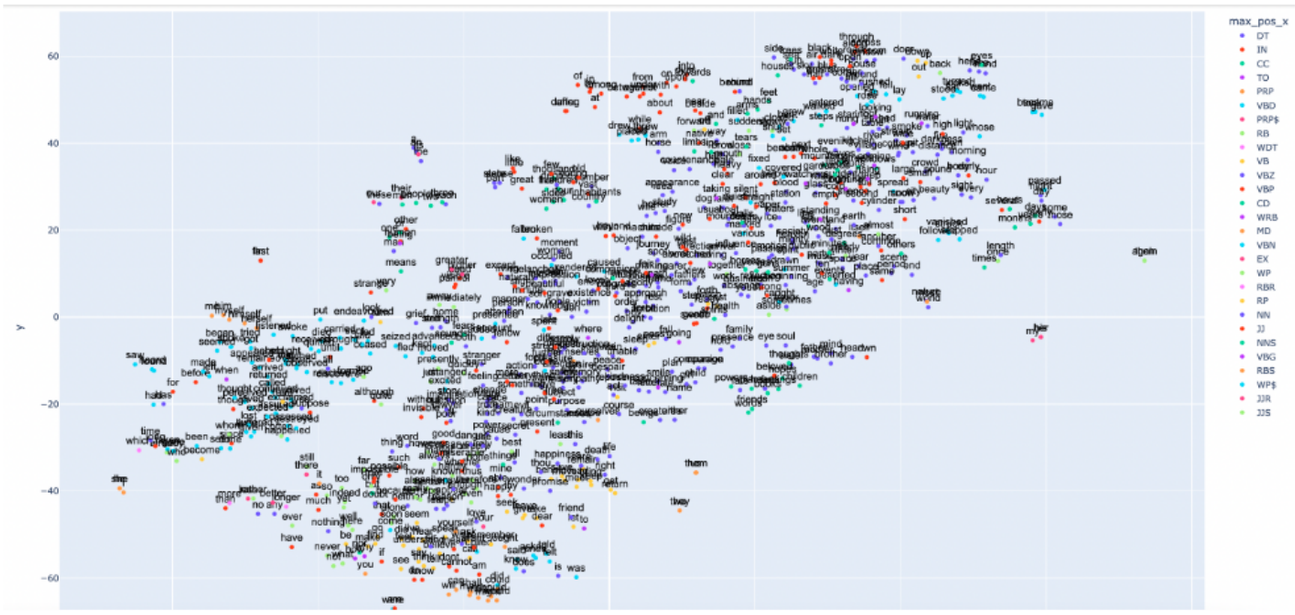
Another method of LDA is to observe theta, my table which explores the relationships between documents and topics. Here, I took a sample of 20 of my OHCO paragraphs and illustrated a heatmap with THETA values:

			T00	T01	T02	T03	T04	T05	T06	T07	T08	T09
book_id	chap_id	para_num										
1164	10	80	0.827621	0.006668	0.006667	0.006668	0.119038	0.006667	0.006668	0.006667	0.006669	0.006668
10005	34	134	0.011115	0.011113	0.011114	0.011114	0.011111	0.011112	0.011112	0.011113	0.899981	0.011115
25439	30	25	0.003704	0.003704	0.003704	0.003704	0.003704	0.003704	0.003704	0.003704	0.793809	0.176558
10005	32	9	0.005264	0.005264	0.005264	0.005265	0.005264	0.005264	0.005264	0.952622	0.005264	0.005263
25439	21	8	0.020000	0.020002	0.020000	0.020001	0.020001	0.020008	0.020000	0.020003	0.020003	0.819982
32154	2	14	0.025005	0.025006	0.774962	0.025003	0.025003	0.025006	0.025004	0.025003	0.025005	0.025003
25439	30	4	0.006250	0.006252	0.190816	0.006250	0.318278	0.006252	0.006250	0.447149	0.006251	0.006251
1164	14	6	0.005263	0.292662	0.005264	0.005264	0.325858	0.344632	0.005264	0.005265	0.005264	0.005264
	12	34	0.033333	0.033337	0.033335	0.033333	0.366661	0.366660	0.033339	0.033333	0.033334	0.033333
	25	43	0.004349	0.004349	0.960860	0.004349	0.004349	0.004349	0.004349	0.004349	0.004348	0.004349
32154	4	342	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000	0.100000
31547	9	8	0.033333	0.033333	0.033335	0.699982	0.033333	0.033333	0.033339	0.033341	0.033336	0.033334
25439	30	3	0.020001	0.020013	0.020000	0.819983	0.020001	0.020000	0.020000	0.020001	0.020000	0.020000
	9	2	0.033334	0.033337	0.033336	0.033334	0.033333	0.033334	0.033338	0.033342	0.699979	0.033333
1633	2	32	0.009092	0.140147	0.009092	0.009092	0.467646	0.009092	0.009093	0.009092	0.328561	0.009094
25439	9	4	0.001563	0.001563	0.001563	0.190515	0.001563	0.417933	0.001563	0.033569	0.348607	0.001563
31547	1	30	0.020005	0.020001	0.020001	0.455167	0.020003	0.384817	0.020002	0.020004	0.020000	0.020001
1633	2	15	0.007143	0.007143	0.007143	0.007144	0.935706	0.007143	0.007145	0.007144	0.007144	0.007145
		31	0.025001	0.485527	0.025000	0.025007	0.025004	0.025007	0.025000	0.025002	0.025001	0.314450
32154	2	116	0.207543	0.306746	0.007693	0.007694	0.431857	0.007693	0.007693	0.007694	0.007693	0.007693

I was not surprised to see that Topic 8 had some stronger/more polar probabilities, whereas Topic 4 had a larger number of "lukewarm" 0.3 values than the other topics.

Word Embeddings

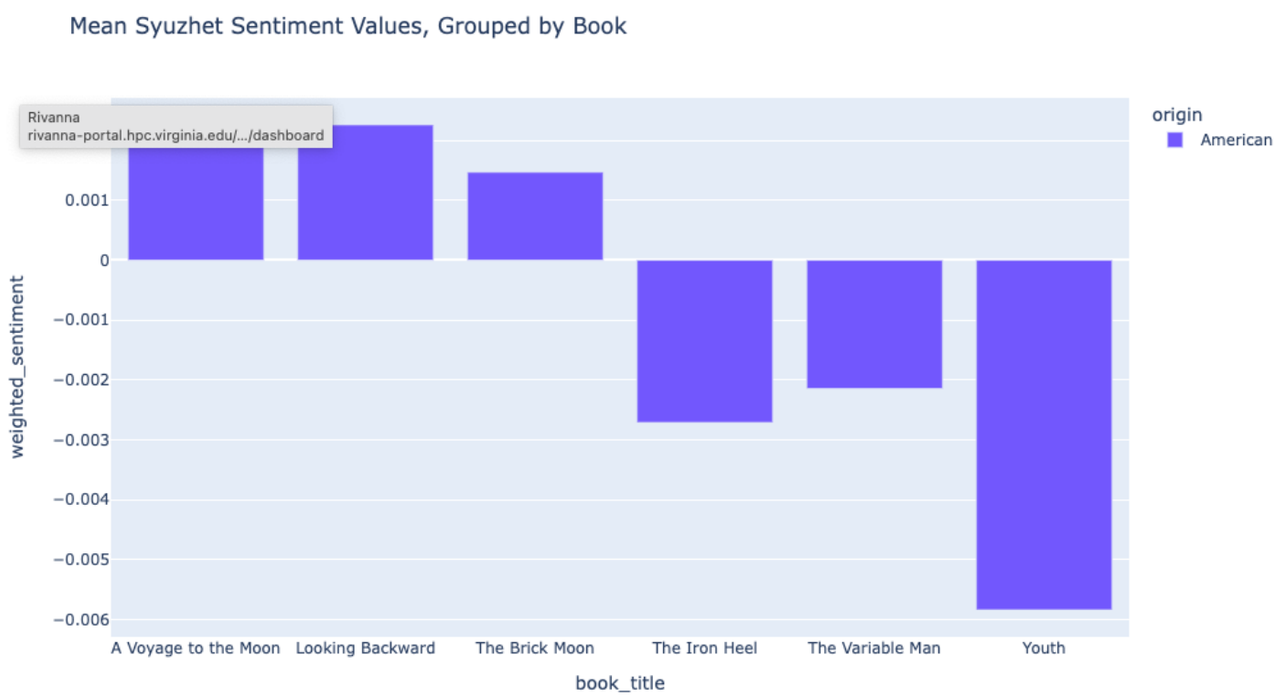
The process of word embeddings allow us to examine meanings independent of documents. As seen in lecture, word vectors are generated by a class of unsupervised algorithms in order to 'learn' the meanings of words from their embedded contexts. Using word2vec and my generated x/y coordinates, I was able to map vectors onto a semantic space, colored by POS tags:



When mapping words, clustering and closeness between points suggests similar meaning.

Sentiment Analysis

Sentiment Analysis allows us to computationally observe "opinion, sentiment and subjectivity in text". Essentially, we can detect emotions and attitudes. However, Syuzhet sentiment analysis is unable to capture as much of the complexity in emotion. A sentiment score assigns "positive" or "negative" implications to terms, so I was able to plot the average sentiment of each of my novels:



My corpus was evenly split between averaged positive and negative sentiments by book.

Interpretation.

By performing the four different text analytics methods we learned in class, I was able to computationally analyze the 19th-20th century scifi genre as well as compare the books within it. PCA gave me some interesting results to compare the books in my corpus, and it singled out George Tucker's *A Voyage to the Moon* (1827). *A Voyage to the Moon* is regarded as earlier proto-scifi and social satire, so it is understandable why it's range is so much more wider and differing than that of the other novels in my corpus.

Given that the LDA model outputted Topic 8 (labor, class, men, nation, day, time, people, course, army) to have the largest term frequency and document weight, my results for topic analysis suggest that these are the subjects which visionary scifi authors were thinking most about for society's future. These terms are strongly instutional which makes sense. These authors were very astute to consider how society, particularly nationhood, would maintain its structure. Coming out of the Industrial Revolution, it also makes sense that conversations surrounding class and labor were on their minds. The word "army" in this topic was also striking and suggests how power and nationhood is maintained by power and violence.

The Word2vec was a little more difficult to apply meaningful interpretations to for the sorts of questions I wanted to ask of my corpus, but it was interesting to see how words were clustered together. I could see elements of the body ('hands', 'eyes', etc) clustered together. One fascinating extention of Word Embedding is to capture analogies between words which would be a good idea for future work.

Lastly, I wondered why certain books were overall positive or negative in sentiment, and tie that to the synopsis of each book. According to Goodreads, In *A Voyage to the Moon* the narrator is shipwrecked off the coast of Burma, where he befriends a brahmin, who starts telling him of his trips to the moon using a space ship of his own invention. I can already imagine the whimsy albeit absurdity of this book, so it makes sense to me this one would have a positive overall tone. Next, Encyclopedia Brittanica describes *Looking Backward* as "an indictment of the capitalistic system and an imaginative picturing of a utopia achieved by a collectivist society in the year 2000". Achieving utopia is enough of a description to call this book a fairly optimistic one in the genre of scifi. Lastly for the positive books, *The Brick Moon* describes a retelling of a group of college friends who engineer the idea of a mechanism for more accurately telling longitude. Although there are setback and failures for these friends, the book frames their experience with joy, curiousity and fulfillment in their work. I understand why it's slightly less positive than the other two positive books.

For the negative books, *The Iron Heel* tells the story of a "20th century America that falls to a dictatorial oligarchy". This tied into our most important topic within our topic model: the words "nation" and "army" specifically in mind. *The Variable Man* discusses a similar idea of societal downfall. In the novel, "the human race has achieved space travel and begun to spread out from Earth, but is limited by an old and corrupt Centauran Empire, ruled from Proxima Centauri." This narrative is very similar to *the Iron Heel* in the sense of highlight corruption in governance and the crumbling of humankind as a result. Lastly with the most negative weighted sentiment, *Youth* narrates two boys who find two strange animals and capture them in attempt to put them in a circus; meanwhile, two professionals are tasked with the decision to open up their world to interstellar trade. *Youth* was the one book that surprised me for a negative sentiment. My guess is that some of the animals not surviving/succeeding in the book, coupled by the anxiety of the working professionals about the consequences of their decisions, gave this book a negative sentiment.

Resources.

Dropbox link: https://www.dropbox.com/scl/fo/vo4jx7bw8d0ybyiy9bhdm/h?dl=0&rlkey=u49a6y1hb67ic1nnypw554ahk

All code resourced from class labs and lecture: https://github.com/ontoligent/DS5001-2023-01-R

“Science Fiction.” Encyclopædia Britannica, Encyclopædia Britannica, Inc., 21 Apr. 2023, https://www.britannica.com/art/science-fiction.