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.

Describe the nature of your corpus and the question(s) you've asked of the data.

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.

Provide a description of all relativant source files and describe the following features for each source file:

Provenance: Where did they come from? Describe the website or other source and provide relevant URLs. Location: Provide a link to the source files in UVA Box. Description: What is the general subject matter of the corpus? How many observations are there? What is the average document length? Format: A description of both the file formats of the source files, e.g., plaintext, XML, CSV, etc., and the internal structure where applicable. For - example, if XML then specify document type (e.g., TEI or XHTML).

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
- *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:

1.png

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

Data Model.

Describe the analytical tables you generated in the process of tokenization, annotation, and analysis of your corpus. You provide a list of tables with field names and their definition, along with URLs to each associated CSV file.

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).

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.png

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.

Exploration.

Describe each of your explorations, such as PCA and topic models. For each, include the relevant parameters and hyperparemeters used to generate each model and visualization. For your visualizations, you should use at least three (but likely more) of the following visualization types:

Hierarchical cluster diagrams

Heatmaps showing correlations

Scatter plots

KDE plots

Dispersion plots

t-SNE plots

PCA

PCA Dispersion Plot

pca1.png pca2.png

LDA

TOPICS by Doc Weight

LDA1.png

TOPIC Doc Weight vs Term Freq

LDA2.png

THETA Sample

LDA3.png

The LDA model attempts to estimate probability distributions for topics in documents plus words in 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, we can map vectors onto a semantic space.

Sentiment Analysis

SA1.png

Sentiment Analysis allows us to computationally observe "opinion, sentiment and subjectivity in text". Essentially, we can detect emotions and attitudes.

Interpretation.

Provide your interpretation of the results of exploration, and any conclusion if you are comfortable making them.

```
In [ ]:
```

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.