Book Recommender

using Natural Language Processing

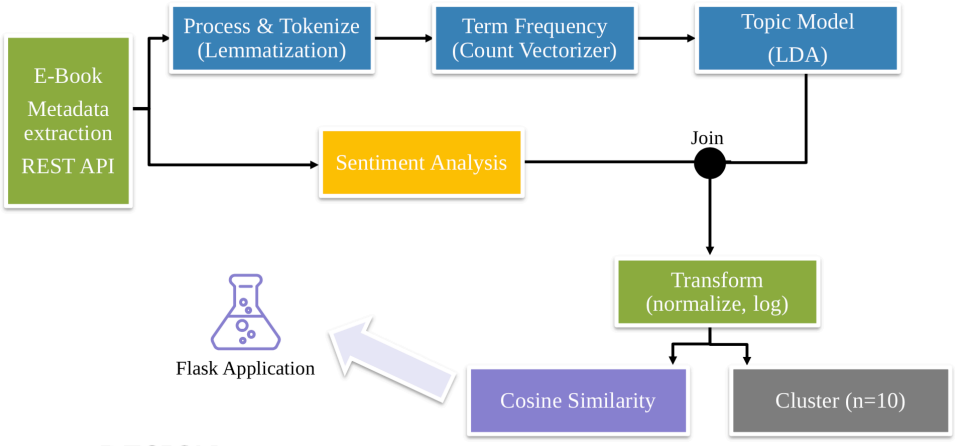
Anjali Sebastian - November 18, 2018

Motivation

I love to read books and I am a huge bookworm. I am always excited to try a new book and look for suggestions from family and friends on what I should read next. Users are rarely given recommendations for free books by commercial book recommender systems (Amazon, Google books, etc …) since it is not profitable for them to make such suggestions (they want to sell you a book not give you a free book !). On the other hand, free book projects such as gutenberg.org have very basic interfaces for users to interact with and get recommendations. There is a large corpus of copyright-free books, that is available for free to readers today and my project would like to make use of these free resources and provide free book suggestions to users. My key motivation behind this project is to create a book recommendation system that works by taking an excerpt of a book that a user likes in order to give a suggestion of a similar book that is freely available.

Design

For the purpose of the project I have used gutenberg.org to download full .txt versions of the books. The current version uses 3000 ebooks of around 150 authors. The flow chart below outlines my design process.



DESIGN

I carried out sentiment analysis along with creating a bag of words model and topic modeling. At each stage I stored the resulting vectors and pipelines as either compressed csv files or dill objects. I used dill objects over ordinary pickles since dill objects store not only the pipeline but any functions that are used by the pipeline.

Sentiment Analysis: Used the TextbBlob sentiment analyzer to determine the polarity and subjectivity of each book. Additionally I calculated the following features - word count, sentence count, average sentence length

Making bag-of-words model: I processed and tokenized all 3000 books using multiple methods … with 1 ngram and 2 ngrams, with the snowball stemmer, porter stemmer and the wordnet lemmatizer. My final model uses a 1 ngram with a word net lemmatizer. I chose this combination because using 2 ngrams was simply too computationally intensive and I didn't feel that the benefit in terms of more context was enough to justify using ngrams that were greater than one. I used the lemmatizer because it seemed to retain more meaning of the words when chopping the words to their roots.

Topic Modeling:
I tried topic modeling with both LDA (used Term Frequency) and NMF algorithms (used Term Frequency Inverse Document Frequency) to generate topics. I looked at 10, 15 and 20 topics and used 20 topics as generated by LDA for my final model.

Developing Features
I used the topics generated from LDA + sentiment analysis outputs to create a set of features that I then used to cluster the books using the K-Means (10 clusters) unsupervised learning algorithm.

Finding a similar Book: After developing the features I used cosine similarity to compare each book in the corpus to a book excerpt provided by the user. I created a flask application that suggests the closest book in the the corpus to the user.

Data

I used 3000 full books (english) from project gutenberg.org and I am trying to extend this to all 30,000 available English Books at gutenberg.org . My features were developed using Natural Language Processing. I generated 20 topic features and 5 other features - word_count, sentence_count, sentence_length, polarity, subjectivity for a total of 25 features.

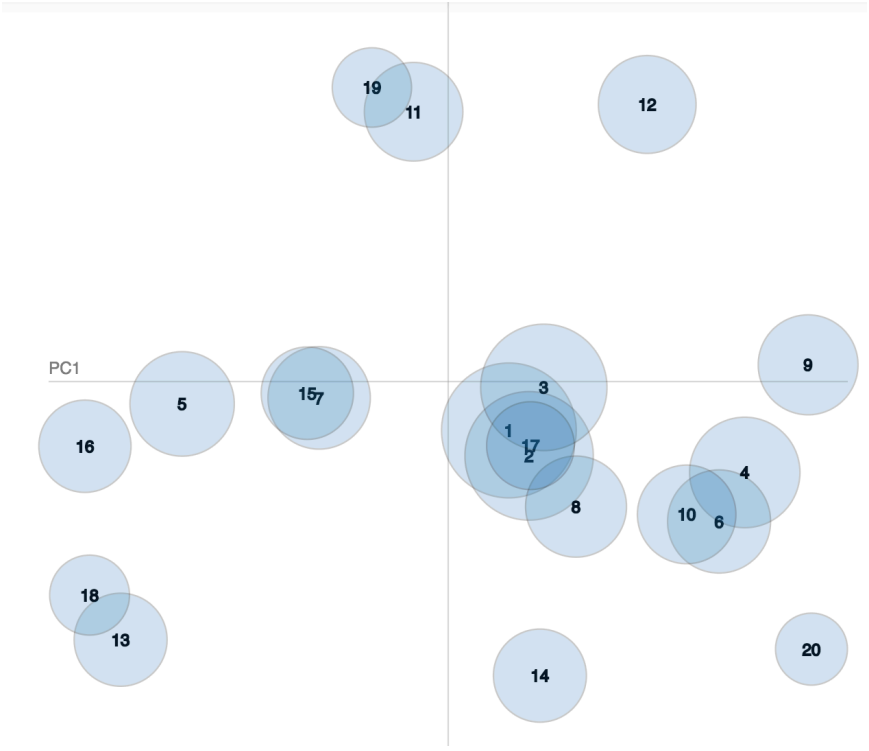
Train-test-split: I performed an 80-20 train test split before creating my bag of word model and carrying out sentiment analysis. I did this so that in the future I can used the test data set for carrying out supervised learning.

Tools

IDE	Jupyter Notebooks	Atom	
Data handling and cleaning	Numpy	Pandas	
Data Visualization	Matplotlib	Seaborn	pyLDAvis
API development	Flask	Javascript	HTML
Modeling	NLTK	Sklearn	Textblob
Documentation	Powerpoint	Typora	Pages
Version Management	Github		

Results

Both LDA and NMF generated topics that made sense. I used the topics generated by LDA. The perplexity was close to 5000 for both the train and the test set.



Topic: 1 - Dramatic tragedies in older English	Topic: 2 - Walter Scott	Topic: 3 - Harold Bindloss
thou , king , thee , thy , sir	sir , lord , letter , lady , person	girl , don , sat , horse , replied
Comic Tragedies by Louisa May Alcott	Trial of Duncan Terig by Sir Walter Scott	The Greater Power by Harold Bindloss
Cleopatra Henry Rider Haggard	The Heart of Mid-Lothian by Sir Walter Scott	Winston of the Prairie Harold by Bindloss
Idylls of the King by Lord Tennyson	Rob Roy by Sir Walter Scott	Alton of Somasco by Harold Bindloss
Topic: 4 - Seafaring adventures	**Topic: 5 - Poetry**	**Topic: 6 - Alfred Henty - Kingly adventures with a religious tone**
sea , boat , ship , captain , wind , island	thy , thou , poem , poet , soul , nature	king , ship , father , town , prince , army
Youth by Joseph Conrad	The Daemon of the World by P B Shelley	Saint George for England George by Alfred Henty
The Life of a Ship by R M Ballantyne	Fugitive Pieces by Lord Byron	The Dragon and the Raven George by Alfred Henty
South Sea Tales by Jack London	The Complete Poetical Works of P B Shelley	Wulf the Saxon George by Alfred Henty
Topic: 7 - My favorites - classic victorian novels	**Topic: 8 - The Rover Boys - not a good topic since it latches on to the names of the rover boys**	**Topic: 9 - William Wymark Jacobs**
sir , lady , gentleman , dear , cried , answered , miss , table	tom , dick , boy , sam , answered , cried , yes , fellow	don , mr , miss , yes , sir , boy , didn , ain , girl , won
Nicholas Nickleby - Charles Dickens	The Rover Boys in New York by Edward Stratemeyer	Bob's Redemption, Captains All, Book 5 W
The Works of Robert Louis Stevenson	The Rover Boys in Alaska by Edward Stratemeyer	Odd Charges, Odd Craft, Part 13 by William Wymark Jacobs
The New Magdalen by Wilkie Collins	The Rover Boys at Big Horn Ranch by Edward Stratemeyer	The Nest Egg, Captains All, Book 3 by William Wymark Jacobs
Topic: 10 - Children's Books	**Topic: 11 - Economics and Society**	**Topic: 12 - Government and Politics**
girl , boy , mother , don , peter , tree	law , class , money , general , person , power , fact	government , general , power , law , war , king , public
The Tale Of Benjamin Bunny Beatrix Potter	Principles Of Political Economy by John Stuart Mill	The Emancipation Proclamation by Abraham Lincoln

The Wonderful Wizard of Oz by Lyman Frank Baum	Nature and Progress of Rent Thomas by Robert Malthus	The works of Edmund Burke
The Tailor of Gloucester by Beatrix Potter	Socialism by John Stuart Mill	
Topic: 13 - America / New world	**Topic: 14 - War**	**Topic: 15 - Science and Evolution**
indian , chief , tree , river , returned , horse , party , savage	officer , general , captain , army , enemy , war	specie , form , plant , animal , fact , nature , body
The Prairie Chief by R M Ballantyne	Through Three Campaigns by George Alfred Henty	The origin of species by Charles Darwin
The Adventures of Captain by Bonneville Washington Irving	Personal Memoirs of U. S. By Grant Ulysses Grant	Experimental Researches in Electricity, Volume 1 by Michael Faraday
The Last of the Mohicans by James Fenimore Cooper	The Great Boer War by Sir Arthur Conan Doyle	On Some Fossil Remains of Man by Thomas Henry Huxley
Topic: 16 - Talking about other works. Critics	**Topic: 17 - Anthony Trollop**	**Topic: 18 - Architecture and Art**
act , english , story , sense , art , history	mr , lady , miss , don , father , dear , mother	city , mountain , stone , king , wall , church
Some Anomalies of the Short Story by William Dean Howells	The Belton Estate by Anthony Trollope	The Stones of Venice
George Bernard Shaw by G K Chesterton	Is He Popenjoy? By Anthony Trollope	The Poetry of Architecture John Ruskin
Topic: 19 - Religion	**Topic: 20 - Henry James**	
god , soul , spirit , father , lord , heaven	don , mr , really , yes , girl , sense	
Sermons for the Times by Charles Kingsley	The Jolly Corner by Henry James	
The Pharisee And The Publican by John Bunyan	The Beast in the Jungle by Henry James	

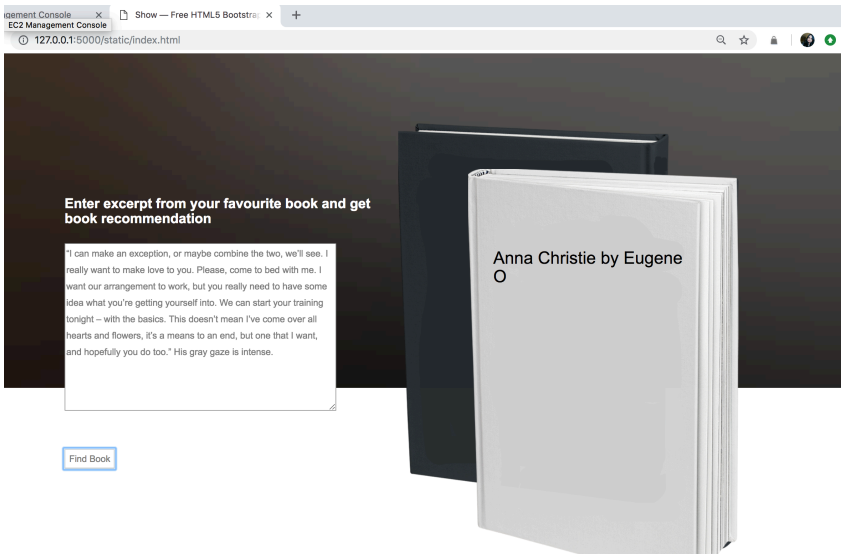
Real World Use

Sentiment Analysis on children's : While carrying out sentiment analysis for each book I grouped them by author and noticed that Beatrix Potter's books ('Peter Rabbit Series') was very negative. On closer observation and reading the text I agreed with the sentiment analyzer that the books very quite negative and at least in todays modern context completely

unsuitable for children. Here are some quotes form the books that were quite inappropriate for very young kids.
“When Benjamin Bunny grew up, he married his Cousin Flopsy.”
“As there was not always quite enough to eat,—Benjamin used to borrow cabbages from Flopsy’s brother, Peter Rabbit, who kept a nursery garden.”
“Sometimes Peter Rabbit had no cabbages to spare.”

Based on this observation I would suggest using sentiment analyzer’s and Natural Language Processing in general to determine if a book is appropriate for children.

I developed a flask application which can take an excerpt provided by the user and returns a book title that is similar to the excerpt entered.



What I would do differently

I felt that with this project there weren’t any metrics I could use to make concrete decisions along the pipeline. I would spend less time creating so many multiple pipelines. I think with all the options I tried out between stemmers, frequency counters, topic models I felt more and more confused and had no time to do word2vec or make a better recommender system. Next time will try to get to the MVP faster just choose one path and get a result before going back and trying every possible combination.

Future Work

- Bring more context using Neural Networks
- Enhance the corpus and add more books
- Remember user history and adapt to user choice
- Create checker for books whether they are appropriate for children