18/01/2020

Genre can be predicted by using cover image or title as well. If we want to use the literature the best way to do this will be by using the summary of the books. Dataset for the same is yet to be found.

While predicting genre of a book various things need to be kept in mind. I referred a thesis by Emily Jordan on AUTOMATED GENRE CLASSIFICATION IN LITERATURE.

Link: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.864.9862&rep=rep1&type=pdf>

A book can belong to multiple genres, and same words can belong to different genres in different contexts for example: “time stood still” could be from a romance novel, a science fiction or a fantasy.

Nouns are not required while finding genre, for example: “Harry Potter”, Harry has no meaning when identifying the genre of the book.

BSB Sir’s guideline:

1. Extract important terms (Key words/phrases)

N gram +TF-IDF based term extraction

1. Train RNN with terms extracted in 1 to predict the genre

19/01/2020

TF-IDF representation of text:

TF and DF values for all unique words in the stories, based on the story itself. Currently I have not included the title weight of the word.

<https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089>

Ranking:

Matching based

Cosine similarity

Neither implemented yet, will do so after cleaning and working on our dataset.

24,25,26/01/2020

Dataset is not available for book texts and their genres, but found a dataset containing the summaries of the books and their genres. Multiple genres have been given in the file, so cleaning will be required. We will define some major genres and classify the summaries into that.

Dataset:

<http://www.cs.cmu.edu/~dbamman/booksummaries.html>

<https://www.kaggle.com/ymaricar/cmu-book-summary-dataset>

CSV

The [csv](https://docs.python.org/3/library/csv.html#module-csv) module implements classes to read and write tabular data in CSV format. It allows programmers to say, “write this data in the format preferred by Excel,” or “read data from this file which was generated by Excel,” without knowing the precise details of the CSV format used by Excel. Programmers can also describe the CSV formats understood by other applications or define their own special-purpose CSV formats.

<https://docs.python.org/3/library/csv.html>

discourse:

df based on file or sentence

compare between LDA and word2vec and tfidf

02/02/2020

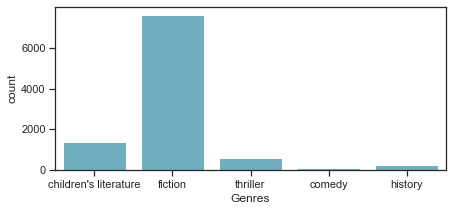
WordCloud- visual representation of words in each genre, gives better understanding of the genre

SeaBorn-graph of our dataset

<https://seaborn.pydata.org/tutorial/categorical.html>

dataset:whole

1 genre/dataset after cleaning



Scikit Learn:

<https://scikit-learn.org/stable/tutorial/basic/tutorial.html#machine-learning-the-problem-setting>

Multinomial Naive Bayes

Linear support Vector machine—better result on first 100 summaries

<https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>