### Abstract

One of my favorite things to do at my spare time is to read books. In this project, I wanted to build a book recommendation engine. I have previously implemented a recommendation system, but it was basic; it used collaborative filtering (looking at users who have liked similar items and recommend their other likes to the target user) and SVD matrix factorization. For this project, I wanted to implement content-based filtering, and use natural language processing, by considering the descriptions/summaries of the books.

### Design

The initial design of the project was based on book descriptions obtained through web scrapping from Goodreads.com. However, the final product uses book plot summaries instead of short descriptions, as many of them turned out to be quite vague. It uses state-of-the-art natural language processing transformer, BERT, to generate embeddings for the book summaries, as well as cosine similarity as a way to find most similar books in the vector space. In addition, other experiments which have been conducted, that include an unsupervised machine learning approaches, can be found in the notebook.

### Data

The experiment started off with a dataset collected through web scraping (request, BeautifulSoap, and Selenium) and included book descriptions, and contained 30,000 rows and 7 columns. This was later replaced by the [CMU Book Summary Dataset] 16,559 rows and 7 columns, of which only the book title and summary columns are relevant.

### Algorithms

###### Feature Engineering

* Data preprocessing:
  + Removal of non-alphanumeric characters except for the space: using a regular expression.
  + Converting the entire text to lowercase for unification purposes.
  + Stopword removal using NLTK.
  + Creation of bigrams using Gensim
  + Lemmatization using NLTK WordNet lemmatizer
  + Stemming using NLTK Snowball stemmer.
* TF-IDF for vectorizing the book summaries.
* SequenceTransformer for vectorizing the book summaries (embedding)

###### Modeling

* DBSCAN for clustering.

###### Similarity metric

* Cosine Similarity

### Tools

* [request]
* [BeautifulSoap]
* [Selenium]
* [NLTK]
* [Gensim]
* [SequenceTransformer]
* [Scikit-learn]

### Communication

Presentation slides can be found in the repository in both PowerPoint and PDF formats.

[CMU Book Summary Dataset]: ( <https://www.cs.cmu.edu/~dbamman/booksummaries.html>)

[SequenceTransformer]: (<https://www.sbert.net>)

[request]: (<https://docs.python-requests.org/en/latest/>)

[BeautifulSoap]: (<https://beautiful-soup-4.readthedocs.io/en/latest/>)

[Selenium]: (<https://selenium-python.readthedocs.io>)

[NLTK]: (<http://www.nltk.org>)

[Gensim]: (<https://radimrehurek.com/gensim/>)

[Scikit-learn]: (<https://scikit-learn.org/stable/>)