Text clustering model for categorizing books

Project Proposal by: Himani Kaushik

- **Question/need:**
  - What is the question behind your analysis or model and what practical impact will your work have?
    - The purpose of the model is to use text clustering techniques to assign books into similar groups or clusters. This unsupervised learning model will apply various transformation and clustering techniques to choose the best algorithm to assign the correct labels to the books.
  - Who is your client and how will that client benefits from exploring this question or building this model/system?
    - A local library wants to implement a categorizing system based on the similarity of the genres and the authors. This text clustering model will enable the librarian to assign the books into groups and subgroups based on their labels. This will ensure an effective and streamlined process to catalogue books in the library.

- **Data Description:**
  - What dataset(s) do you plan to use, and how will you obtain the data?
    - The data will be obtained from the Project Gutenberg website - https://www.gutenberg.org .
    - I plan to extract five books by different authors and from different genres from the website. The books will be randomly partitioned into 200 partitions and each partition will consist of 150 words.
  - What is an individual sample/unit of analysis in this project?
    - The individual sample will include index, author, title, label and partition details.

- **Tools:**
  - How do you intend to meet the tools requirement of the project?
    - Following Python tools would be used:
      - Requests: For extracting books using their urls.
      - Pandas and Numpy: For cleaning data and preprocessing.
      - NLTK, spaCy, gensim, scikit-learn: For processing text, topic modeling, clustering and evaluation
      - Matplotlib and Seaborn: For visualizing the data.

- **MVP Goal:**
  - What would a minimum viable product (MVP) look like for this project?
    - MVP for the project would include results from data transformations using BOW, TF-IDF, LDA and implementation of initial K-means clustering algorithm.