Team: Voltron
Topic: Book Recommendation System

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

Name	***NetIds***	***Captain***
Adeeti Kaushal	*Adeetik2*	*Yes*
Vivek Bansal	*Vivekb3*	

2. What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

   Topic: Book recommendation system.

   Description: In the course, we discussed about Recommendation systems and Search. There are two main recommendation system approaches that were discussed further, Content-based filtering and Collaborative based filtering. In this project, we are focusing on simple content-based book recommendation system. The content will be downloaded from public sites (Project Gutenberg).

   Tasks involved: The tasks involve loading the content of book into python after downloading it. Search for relevant content/words in the loaded data. Tokenize the corpus and perform stemming on the tokenized corpus. Next step would involve building bag of words model and find stop words, build term frequency-inverted document frequent model and show the results of tf-idf model. Compute the distance between texts. Look into search criteria and find similar books matching the content using Cosine Similarity.

   Important and Interesting: This topic is important and interesting because big corporations like Facebook, Amazon, Apple, Netflix, Google (FAANG) and others big companies use recommendation system to show/target the content based on similarities in content. This is important for business to present or build the resources that users are searching and show them similarities between other books.

   Planned Approach: The approach that we plan to take involves collecting the books from Project Gutenberg which offers free books and that will be used as our dataset. We will find a topic/search term or title that will be used to create the model and then then find similarities in other books. Python libraries will be used to perform various tasks like stemming, tokenizing etc. The outcome would involve books with similarities.

   Tools, Systems, Dataset:

   Python Libraries that can be used for NLP, tokenization, stemming, parsing, classification etc. which are important tasks. Jupyter Notebook will be used to read the

dataset and execute the python program. From Project Gutenberg, free books would be downloaded to build the dataset.

Expected outcome: Being able to retrieve/recommend similar books based on the content.

Evaluation: Results retrieved from the program should give recommendation of books that are matching or similar in some way or form. For example, a crime mystery novel should not return any match with children book.

3. Which programming language do you plan to use?

   *We are planning to use **Python**.*

4. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

   We are planning to cover the following topics in the project.

Tasks	**Estimated Time (hrs**
Learning Python	8*
Research/Build dataset	5
Code for tokenization, Stemming on corpus	4
Build bag of words, find stop words	2
Build tf-idf	4
Build Similarity matrix	2
Full end to end integration, tuning	6
Testing	5
Visual representation of results (matplotlib)	6
Other use case – Content based Similarity	6
Report	4
Total	**52**

N=2
20*2 hours = 40 hrs project

Estimated hours = 52

* Since, both need to learn Python, so we have added 8 hrs for learning but without that the efforts required for project are above the required hours.