CS410 Project proposal

Team members:

First name	Last Name	Net ID	Comments
Venkata	Gottiparthi	vg24@illinois.edu	Captain
Zhao	Li	zhaol4@illinois.edu	Team member
Hannah	ke	yuhanke2@illinois.edu	Team member
Qi	Zhou	qizhou8@illinois.edu	Team member
Yixin	Xu	yx33@illinois.edu	Team member

Topic selection and relevance to the theme and the class:

Free topic - Information retrieval and Sentiment analysis for news

In the era of information overload, sentiment analyzer for news is helpful to make time-sensitive decisions or pinpoint reliable information sources.
The topic directly uses the knowledge such as ranking relevant information we learnt from the course. We plan to use efficient algorithms that can quickly retrieve the information from the website, and we will use the knowledge we learnt from the course for ranking it based on relevance.

Datasets, algorithms or techniques we plan to use:
Our plan involves utilizing Flask or Streamlit to design the user interface. For our dataset, we will construct a corpus by implementing a web crawler to gather news about the stock or finance market. In terms of algorithm, we will focus on retrieval techniques, particularly the utilization of BM25 through the MeTapy toolkit.

Approach evaluation:

a. Ranking relevant documents on top of non-relevant ones, and use Precision and Recall for measuring it.
b. User feedback

The programming language:
We plan to use python for the information retrieval and build the search engine with Metapy toolkits.

The workload:

a. Crawling and scraping web pages - 15 hours
b. Cleaning data and building dataset - 20 hours

c. User interface - 15 hours

d. Information retrieval - 20 hours

   i. Constructing corpus

   ii. Indexing

   iii. Retrieval functions and algorithm

e. Data analysis: - 30 hours

   i. Ranking

   ii. Sentiment Analysis

   iii. Data Visualization: display charts or graphs based on retrieved documents and data.