CAPP 30123 Project Proposal

- Group name: SSSP
- Members:
    - Yilun Dai
    - Ling Dai
    - Jie Heng
    - Fangfang Wan

- Data:
    - Source:
      Word frequencies in English-language literature, 1700-1922; Genre-specific word counts for 178,381 volumes from the HathiTrust Digital Library
    - Size:
      There are around 30 files in the dataset, and each of them is 100-500MB.
    - Link:
      https://wiki.htrc.illinois.edu/display/COM/Word+Frequencies+in+English-Language+Literature%2C+1700-1922
    -
- Hypothesis:
    - We plan to investigate the frequencies of n-grams in those English language literature from 1700-1922, and use topic modeling to analyze the involving trend of topic in English language literature(fiction, poetry and drama), and explain such change with historical events in corresponding time periods.
- Algorithms
    - Topic Modeling
    - K mean clustering