Flow of Control

Cleaning text using cleaner.py (for word count), cooccurrence.py(for cooccurence)

→ **WordCount** →

data_cleaned.txt is the files cleaned to remove special character and stopwords

→ **Mapper** →

Co-occurence →

data_cooccur.txt has preprocessed data which also finds pairs of words. Removes special characters and stop words

→ **Mapper** →

Data is cleaned now it is put in Mapper which is in WordCount.java then put through reducer to get final count

→ **Data for plots** →

Plotting for analysis