Twitter Search Engine

-Group 20

Group Members:

- Logu R (S20190010111)
- Harshith Jupuru (S20190010072)
- Sree Nitish B (S20190010166)
- Koushik Bhukya (S20190010)

Tasks:

- Text Preprocessing
  - Stopword removal / Stemming
  - Creating Tokens
  - Parsing Corpus
  - Corpus Counter
- Creating Frequency Index
  - Inverse document frequency
  - Getting unique tokens
- Creating TF-IDF Index
  - Finding max frequency of any term in corpus
  - Computing term frequency index
  - Computing idf value
  - Computing tf-idf weight
- Query Refinement using
  - Jaccard Distance
    - Find N-grams
  - Edit Distance
  - Thesaurus
- Retrieving Relevant Documents
- Ranking with Cosine Similarity
  - Generate document and query vectors
- Integrating the IR System in Django

Features :

- Spell corrections using jaccard coefficients and edit distance from correctly spelled words taken from nltk.corpus.words
- Replacing some query terms with synonyms using wordnet thesaurus to improve recall
- Setting threshold for cosine similarity and maximum documents to retrieve as results