Design Documentation

**DATA STRUCTURES USED**:

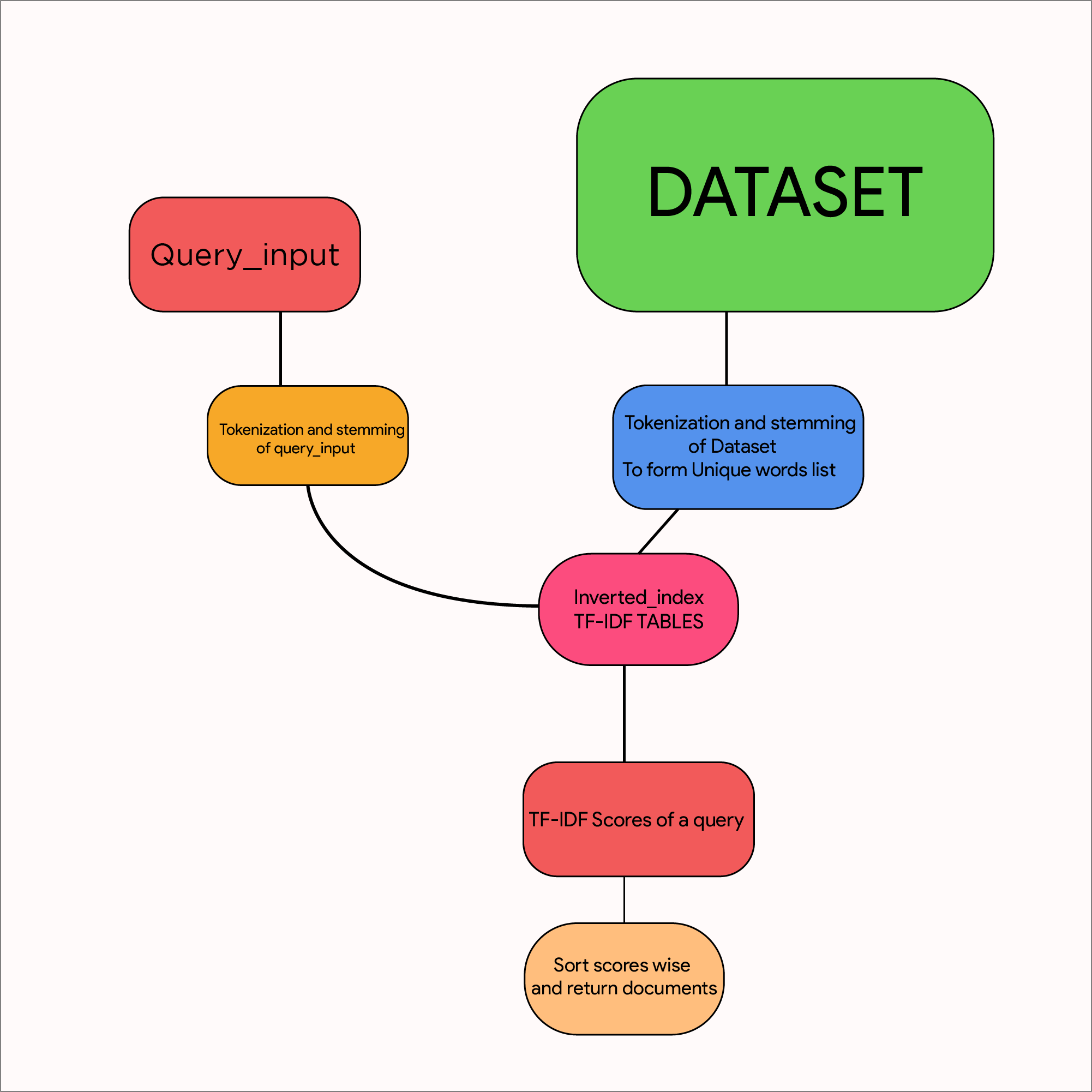
1. For Inverted Index: A dictionary called posting list is used which stores all the words occurring in the unique words list as keys and a sorted list of DocID’s in which the word occurs as their values.
2. Vector Spaced Model: A dictionary called tfd\_doc is used for tabulating words as columns, document names as their index and wtd as their corresponding values and a dictionary idf is used to tabulate the inverse document frequency of all the words.

The query is taken from the user by executing the query.py file the query is stemmed tokenized and normalized and then the tf-idf values of the query word are taken from the tf idf tables created in the inverted\_index.py file

Then the documents are ranked based on this tf-idf score of each document w.r.t to query.

**RUN TIME FOR PROCESSING**:

A corpus with 92 articles and 5000 unique words are taken from [Media news today](https://www.medicalnewstoday.com/popular) . On This Dataset, it takes about 80-90 seconds for the code to pre-process. For and average query it takes about 0.5-1second to create a csv containing the results.



Project By:

S. Preethi Sharma - 2016A7PS0001H

M.K.S Amrutha Varshini - 2016A7PS0027H

G Srinayan Reddy - 2016AAPS0171H