# Information Retrieval: Assignment I

Group No. 42.
Harsh Bandhey, 2017234
Md Talib, 2018245

## Preprocessing

First we removed all special characters and digits, only latin alphabetical characters remain.
Stop words were removed using the NLTK packages. We tokenize the strings into corresponding words. Every word was converted into lower case. All words with less than two characters and digits were removed. We also lemmatize words to their closest meaningful stem word.

## Methodology

We use a hash and a variable list to create an inverted index, the hash contains words as key and posting as values. Practically for our case we use a dictionary of lists in python, we create the dictionary with unique words stems in preprocessed files as keys and document name as postings in a sorted list. We pickle and store this hash structure as a pkl file to save runtime so the next time when we run the code, we don’t need to do the same preprocessing again. We process the query similarly to stems to feed it for fetching documents.
For merging boolean queries, we make computationally appropriate merging algorithms taking advantage of the sorted positionals.

## Assumptions

We assume as per google classroom clarification that queries are considered from left to right, thus no optimization has to be done.
NOT operation is assumed to have 0 comparisons. It is O(n), but we remove words using python implementation.

## Steps of Use

All code has been written and tested in ipynb notebook, the pkl file for the hash has also been provided. A cli version of the same is also given as per input specifications, it uses preconstructed inverted index hash.