Ahmet Yasin Alp
2015400213

Read punctuations.txt to create translation table.
Read stopwords.txt to create stopwords list.
Read files one by one.
Split files according to ‘</REUTERS>’
Find ‘NEWID’ from first line
Find ‘<TITLE>’ , ‘</TITLE>’ for title part.
Find ‘<BODY>’ , ‘</BODY>’ for body part
Remove punctuations using translate build-in func.
Lower case
Delete stopwords using for loops
Push newid s to index dictionary
Using keys in dictionary, create bigrams dictionary.
Write these dict. as json files

Open json files
Take arguments from command-line
If it is a conjunctive query, start intersecting arrays from shortest one
Else if it is a disjunctive query, merge all of the array
Else if it is a wildcard query, intersect bigrams and postfilter then merge ids.

Number of tokens before stopword removal : 2902785
Number of tokens after stopword removal : 2227436
Number of terms after stopword removal and case-folding : 45346
Top 20 after stopword removal and case-folding : ['3', 'reuter', 'said', 'to', 'lt', 's', 'mln', 'dlrs', 'from', 'at', '1', 'year', 'pct', 'has', 'inc', 'company', '2', 'corp', 'u', '000']