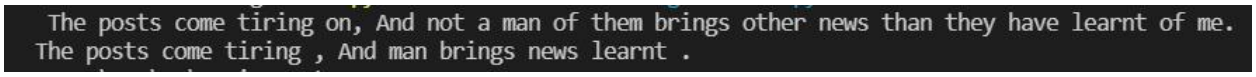
Boolean Retrieval System

Our retrieval system returns the names of documents from the given dataset that satisfy the boolean query entered. The boolean queries supported are AND, OR, and NOT and their different combinations. The system is also capable of handling different wildcard queries.
Our code performs 4 different operations:

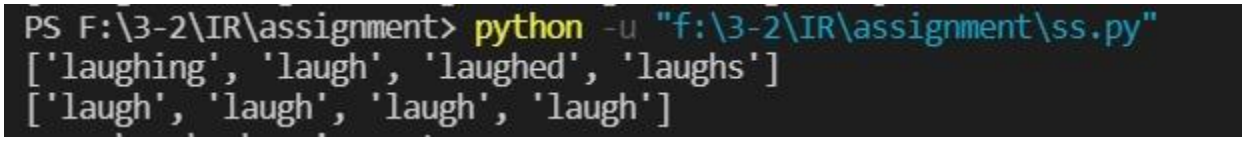
- Preprocessing (Stopword removal and Stemming)
- Index Building (Inverted Index and Permuterm)
- Query Handling
- Spelling Correction (Edit Distance Method)

For preprocessing, we run through the dataset to tokenize all the documents. This is followed by removing all the stopwords.



In this example, our code removes words like ‘on’, ‘not’, ‘a’, ‘of’, ’them’, ‘other’, ‘they’, ‘have’, ‘of’, and ‘me’.

The remaining tokens are then stemmed using Porter Stemmer to give the final set of tokens.



In this example, the words ‘laughing’, ‘laugh’, ‘laughed’, and ‘laughs’ are all stemmed to ‘laugh’.

We then create an inverted index for each stemmed word. This gives us a list of all document names in which each word is present.
The structure of our inverted index dictionary is:
‘antonio’ -> [list of documents which contain word ‘antonio’]
‘brutus’ -> [list of documents which contain word ‘brutus’]
And so on.

For creating a permuterm index, we append a ‘$’ sign to each word and generate different permutations for it.
The structure of our permuterm index dictionary is:
‘bat$’ -> [bat]
‘$bat’ -> [bat]
‘t$ba’ -> [bat]
‘at$b’ -> [bat]
And so on for other words.

When the code is run, the user enters a query which is either a boolean query or a wildcard query.
If it is a boolean query, we call the function bool_query which returns the retrieved documents.

```
Enter Query : bum AND NOT brutus
DOCUMENTS RETRIEVED
 ['a-midsummer-nights-dream_TXT_FolgerShakespeare.txt', 'measure-for-measure_TXT_FolgerShakespeare.txt', 'timon-of-athens
_TXT_FolgerShakespeare.txt']
```

If it is a wildcard query, we first process it and then call the processQuery function which prints the retrieved document list.

```
Enter Query : anton*
DOCUMENTS RETRIEVED
['antony-and-cleopatra_TXT_FolgerShakespeare.txt', 'henry-v_TXT_FolgerShakespeare.txt', 'julius-caesar_TXT_FolgerShakespeare.txt', 'macbeth_TXT_FolgerShakespeare.txt', 'all
s-well-that-ends-well_TXT_FolgerShakespeare.txt', 'much-ado-about-nothing_TXT_FolgerShakespeare.txt', 'the-merchant-of-venice_TXT_FolgerShakespeare.txt', 'the-taming-of-the
-shrew_TXT_FolgerShakespeare.txt', 'the-tempest_TXT_FolgerShakespeare.txt', 'the-two-gentlemen-of-verona_TXT_FolgerShakespeare.txt', 'twelfth-night_TXT_FolgerShakespeare.tx
t', 'a-midsummer-nights-dream_TXT_FolgerShakespeare.txt', 'coriolanus_TXT_FolgerShakespeare.txt', 'cymbeline_TXT_FolgerShakespeare.txt', 'hamlet_TXT_FolgerShakespeare.txt',
 'henry-iv-part-1_TXT_FolgerShakespeare.txt', 'henry-iv-part-2_TXT_FolgerShakespeare.txt', 'henry-vi-part-1_TXT_FolgerShakespeare.txt', 'henry-vi-part-3_TXT_FolgerShakespea
re.txt', 'henry-viii_TXT_FolgerShakespeare.txt', 'king-john_TXT_FolgerShakespeare.txt', 'king-lear_TXT_FolgerShakespeare.txt', 'loves-labors-lost_TXT_FolgerShakespeare.txt'
, 'lucrece_TXT_FolgerShakespeare.txt', 'measure-for-measure_TXT_FolgerShakespeare.txt', 'othello_TXT_FolgerShakespeare.txt', 'richard-iii_TXT_FolgerShakespeare.txt', 'richa
rd-ii_TXT_FolgerShakespeare.txt', 'romeo-and-juliet_TXT_FolgerShakespeare.txt', 'shakespeares-sonnets_TXT_FolgerShakespeare.txt', 'the-merry-wives-of-windsor_TXT_FolgerShak
espeare.txt', 'the-two-noble-kinsmen_TXT_FolgerShakespeare.txt', 'the-winters-tale_TXT_FolgerShakespeare.txt', 'titus-andronicus_TXT_FolgerShakespeare.txt', 'troilus-and-cr
essida_TXT_FolgerShakespeare.txt', 'venus-and-adonis_TXT_FolgerShakespeare.txt']
The execution time is : 0.010098788823618164
```

For boolean queries, in case of spelling errors, we also check for edit distance to get the closest possible words.

```
Enter Query : brutu
DOCUMENTS RETRIEVED
 ['antony-and-cleopatra_TXT_FolgerShakespeare.txt', 'coriolanus_TXT_FolgerShakespeare.txt', 'hamlet_TXT_FolgerShakespeare
.txt', 'henry-vi-part-2_TXT_FolgerShakespeare.txt', 'henry-v_TXT_FolgerShakespeare.txt', 'julius-caesar_TXT_FolgerShakesp
eare.txt', 'lucrece_TXT_FolgerShakespeare.txt', 'the-merchant-of-venice_TXT_FolgerShakespeare.txt', 'titus-andronicus_TXT
_FolgerShakespeare.txt']
```