Information Retrieval

Assignment 1

Dibyendu Roy Chaudhuri
MT19034

Q1) Inverted Index Table

Methodology:

1. For creating an inverted index, I have used a dictionary of list. In the list first index is used to show the size posting list and the second index is the posting list where indexes of every document are stored in ascending order.
2. I have also maintained another list where document names are stored in ascending order.
3. An inverted index algorithm is used to perform logical operations between words.

Preprocessing Step:

1. Tokenize is done on following delimiters-
   ("\s", "-", ".", "@", "t", "\n", "[0-9]", ‘"’, ">", ",", "?", ":", "{", "(", "[", ")", "}", "]", "<", "_", "!", "/", "|", "\", "*", "=", "^")
2. Convert whole text into lower case.
3. Lemmatization is used.
4. Pickle library is used to store the inverted index.

Assumption:

1. I have taken stop words as valid words. For that, while entering a query Word should start with a capital letter and the operator should be in lowercase.
2. I have assumed there is no alphanumeric word. So I just removed numeric digit during processing.

Q2) Positional Index Table

Methodology:

1. For creating an inverted index, I have used a dictionary of list. In the list first index is used to show the size posting list and the second index is the posting list where indexes of every document are stored in ascending order.
2. For creating the positional index, a dictionary of dictionary is used.
3. I have also maintained another list where document names are stored in ascending order.
4. For the searching phase, a simple positional index algorithm has been used.

Preprocessing Step:

1. Tokenize is done on following delimiters-
   ("\s", "-", ".", "@", "t", "\n", "[0-9]", ‘"’, ">", ",", "?", ":", "{", "(", "[", ")", "}", "]", "<", "_", "!", "/", "|", "\", "*", "=", "^")
2. Convert whole text into lower case.
3. Pickle library is used to store the inverted index and positional index.

Assumption:

1. I have taken stop words as valid words. For that, while entering a query Word should start with a capital letter and the operator should be in lowercase.
2. I have assumed there is no alphanumeric word. So I just removed numeric digit during processing.