Program:

Write a (Perl or Python) program that generates the inverted index of a set of already preprocessed files. The files are stored in a directory which is given as an input parameter to the program. Use the

files preprocessed in the previous assignment(s) as test data. Use raw term frequency (tf) in the document without normalizing it. Think about saving the generated index, including the document frequency (df), in a file so that you can retrieve it later Remove the following during the preprocessing:

Solution: Please Run the Assignment5.py File for output. Other files of .py are external functions.

Name: Sumanth Reddy Ganta

UID: U00906099

Subject: Information Retrieval/ Web Search

Date: 10/12/2023 (mm/dd/yyyy)

Executive Summary:

The main purpose of the program is to preprocess the data from the text files with the conditions provided such as removing uppercases, Morphological changes etc.

By making this we can get the list of Vocabulary in the document along with the term and inverte4d document frequency.

Methodology:

To solve the problem, I've broken the program into different functions and modules with their necessaryfunctionalities. I've used NLTK library in python for stemming and lemmatization if the words in the output.

Functionality:

For a user to use this program, we are going to use the output of the previous assignment as input to this program. Now for the vocabulary present in those files we are going to calculate the term frequency and the inverted document frequency for all the vocabulary.

Results:

The result or output of the program is in Text Document .For more information about the result refer to

the output text files.

Discussion:

The result is the output file with the term and inverted document frequencies for each term containing

We need to improve in the Stemming using porter Stemmer as it is not providing the Accurate output for some of the Vocabulary. So Improvement is needed in this area.

References:

NLTK library in python for Lemmatization and Stemming