IIT PALAKKAD

CS3100: Paradigms Of Programming

Under the guidance of Dr. Mrinal Kanti Das

Imperative Paradigm Mini Project

Date of submission : Sept 13 2017

Submitted by

Jayaprakash A

Siddhardha SST

Vinay Ande

Aditya M

Chapter 1

Imperative Paradigm

1.1 Problem Statement

Compute distance between two documents.

1.2 Data Structures and algorithms

Some basic data structures like lists and dictionaries were used. Term frequency–inverse document frequency (Tf-Idf), is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. This method has been used to give weightage to more important words rather than just the word frequency.

1.3 Methodology

The entire methodology is divide into 4 stages. These include formatting text, creating word frequency vector, computing tf-idf weights and calculation of distance.

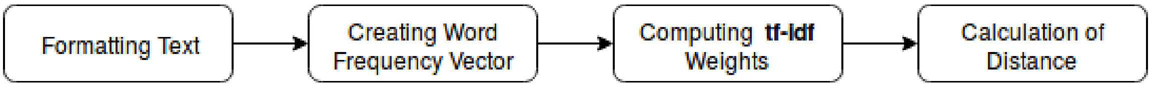

Figure 1.1: The flow chart depicting stages of algorithm

- **Formatting text :** Read the file and extract words from it. Words are obtained by splitting the text by various delimiters (variable `delimiters`).
- **Word frequency vector :** After extracting the words, create a vector that contains the frequency of each and every word. This is done for each and every file in the corpus.
- **Computing tf-idf weights :** Tf-Idf weights corresponding to the above word frequency vector is calculated as per the given formula.

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

$$IDF(t) = \ln\left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}\right)$$

Tf-idf weight(t) = TF(t) * IDF(t)

- **Calculation of distance :** The distance between each and every pair of files is done by calculating the cosine distance and later sorted.

1.4 Observations and results

Some interesting observations have been found. Distance between same files is zero.

1.5 Contribution of team member