CS F469 INFORMATION RETRIEVAL

First Semester 2019-2020

Team Members:

Jaju Vedant Vinod – 2016A3PS0303H

Tulika Manek – 2017A7PS0147hH

Tanuja - 2017A7PS0044H

Pavan Chakradhar Reddy - 2017A7PS0189H

Aim:

The aim of the project is to build a domain specific information retrieval model.

Dataset used are news articles from BBC news with different topics.

This document describes the implementation details of the model built.

Design :

To retreive documents most relevant to the query, documents -term 2D matrix is generated, assigning scores .

The scores

The implementation is based on vector space model. The weights are calculated by tf-idf-cf method.

The model organises and treats each document and query as a vector .Weights are generated based on the tf-idf scores

and class-document-frequency score . The logarithm value is taken to moderate the extreme values .

Class document frequency score tells the how many documents contain the particular term in a particular class, to increase the particular weight of the class.

The corpus is processed to remove stop words, and punctuations and whitespaces etc. Cosine distance is used to calculate the angle between two vectors (query and document). Finally, the top ten documents are retreived.

Data structures used are Dataframe and dictionary . The dictionary of dictionary is dataframe.