**DESIGN DOCUMENT**

CS- F469

ASSIGNMENT-1

Domain Specific Search Engine (Sports)



Piyush Mishra – 2016B2AA0633H

Abhishek Singh – 2016A7PS0093H

Nimit Kanani – 2016B3A70464H

1. Introduction

In this project we retrieve the top results for queries pertaining to the sports domain, using ranked retrieval from a corpus consisting of various news articles related to the same.

1. Language Used : Python
2. Frameworks Used : PyQt5 (For GUI)
3. Libraries Used : NLTK for text processing

2. Dataset & Pre-processing

The dataset used in the project was obtained from Kaggle.

For pre-processing we used a regular expression (see the code) to remove all punctuations such as full stops, commas, exclamation points, semi colons, etc. This was followed by removal of stopwords and all the text was also converted into lowercase.

3. TF-IDF Score Generation

We use TF-IDF score to rank the queries based on relevance. The score is calculated using the following formulas :

TF refers to Term frequency = no of times a term appears in a document

To get better results we use

Weight of term t in document d :*Wt,d = 1+log10(tf)t,d*

Given query q and document d,

*Score(q,d) = Σt belongs in q and d Wt,d*

IDF refers to Inverse Document Frequency, it measures the rarity of a word in a document

*dft = document freque ncy of t(no of documents that contain t)*

*idft = log10(No of documents/dft)*

Therefore, TF-IDF score for a term t given document d =

*TF-IDFt,d = Wt,d\*idft*

For a query q and document d, the TF-IDF score is sum of TF-IDF scores of all terms t and document d.

We use an array of dictionaries to store TF scores of terms in each document.

We also maintain a dictionary containing all unique words in the documents corpus for the purpose of calculating IDF scores.

4. Presenting Results

User inputs a query on the user interface which then retrieves the documents in order of their tf-idf score. The retrieved documents are also displayed on UI.

**Special Feature** :

Whenever user makes a spelling mistake in their query and it does not return any results, he can select auto correct option, the results will be returned corresponding to the terms having least edit distance from the query.