**Community Questions Answering - Question Comment Similarity with the use of Unsupervised ranking model**

4 study points project work on Text-Based Information Retrieval

KU Leuven, Belgium.

by,

Marimuthu Ananthavelu (r0652832) Monika Filipčiková – (r0683254)

**Summary**

This project work demonstrates the use of the Unsupervised model for ranking the comments with respect to a Question in the Community Questions Answering forums. The model which is used in this task is a Vector Space Model. The text pre-processing of the questions and comments were done with the help of available libraries for stop words, punctuations marks. The relevancy between each question and all the comments were referred as independent Vectors and their similarities were calculated with the help of Cosine scoring. The evaluation of model is done on the development set and compared with the baseline results and they are discussed. The output prediction file is generated for the ‘test\_input.xml’ is submitted for evaluation.

Introduction to the Project task:

Community Questions Answering forums are helpful to know the information a user wants to know. The comments in this task are the different answers for a specific question which is presented the in the file with the .xml format. The objective of this task is to rank the very relevant answers to a question as higher and rank less to the least relevant answers.

Approach:

The Vector Space model considers all the information needs and the relevant documents in the Vector space. An example shall be below:

Term-2

Term-3

Sentence-3

Sentence-2

Sentence-1

Term-1

In our task, we presented the model where the Question and all the comments were represented as Vectors. These vectors are weighted and normalized to each question and document pair and the dot product of those are used for calculating the Cosine score as follows:

1. For each question, the following are calculated:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Wordset | Question | | | | | |
| Term-Frequency | Term- weight frequency | Document- Frequency | Inverse document  frequency | Weight | Normalized  weights |
| Term-1 |  |  |  |  |  |  |
| Term-2 |  |  |  |  |  |  |
| … |  |  |  |  |  |  |
| … |  |  |  |  |  |  |
| … |  |  |  |  |  |  |

*Whereas,*

***Question =*** Mix of Question Category, Subject, and Body

***Wordset =*** Words which all are part of a single Thread (Question + all the comments in a Thread)

***Term Frequency (tf-raw)=*** Number of times a term appears in a Question

***Term weight frequency (tf-weight) =*** Converting the raw number of terms into weights in logarithmic scale as below:

tf-weight= 1+ log(tf-raw)

***Document-Frequency (df)=*** Number of documents a term appears

***Inverse document frequency (idf) =*** Converted values of Document frequency using the total number of documents.

idf = log (N / log (df))

*Whereas,*

N – Total number of documents