REPORT ON INFORMATION RETRIEVAL SYSTEM

- **UID: 180128022**

We have implemented an IR System that works on different Term Weighting schemes i.e., Binary, Term Frequency and TFIDF.

Code Implementation:

We have created a function **candidatePicker** which returns a set of candidate documents that contain at least one term from the query.

We have passed this function as an argument to our functions for the different Term Weighting (TW) schemes. We have also created three dictionaries called **termsInDocDict**, **docstermsDict** and **tfidfcalcDict** for our TW functions, **binary**, **termFrequency** and **tfIdf** respectively.

These dictionaries store the size of the document vectors for each TW scheme, after iterating through the index file. We have used the respective dictionary in each of our TW functions in order to generate the top 10 computations of how well the occurrences of each term correlate in query and document for each TW scheme.

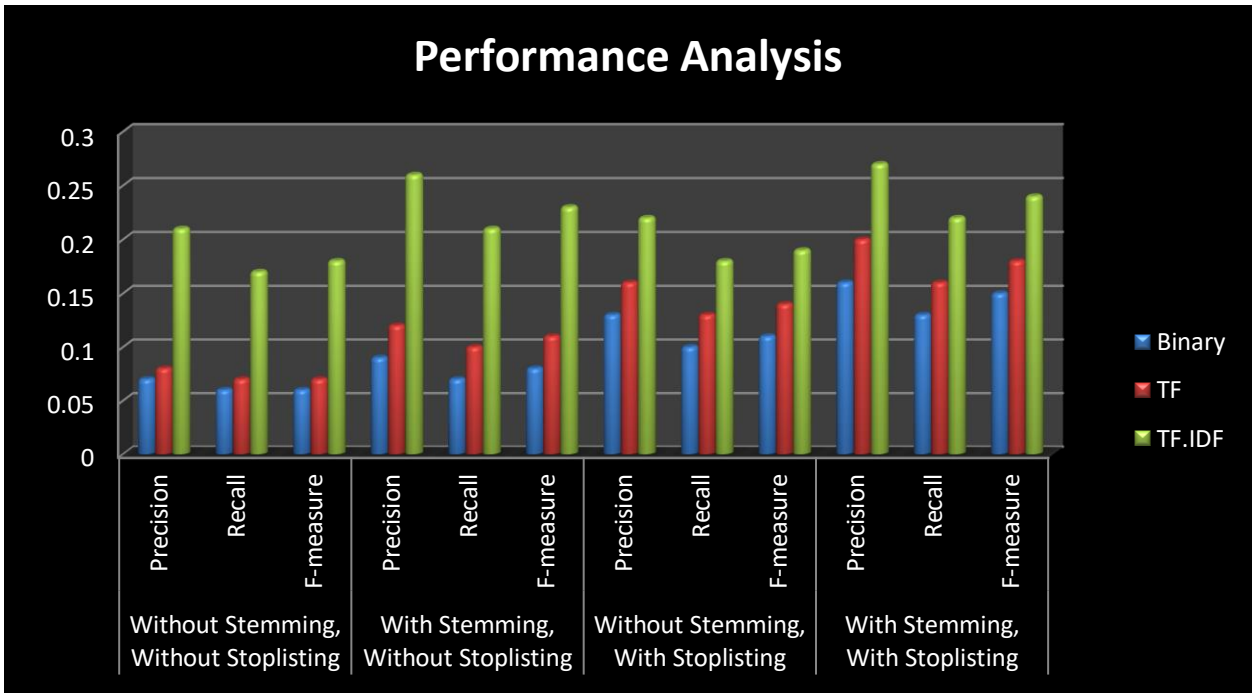
We call the respective function for each scheme through the return statements which are conditioned on the TW schemes, in ***forQuery*** method.

Below is the table of the observations:

Configuration	Prec/Rec/F	Binary	TF	TFIDF
Without Stemming, Without Stoplisting	Precision	0.07	0.08	0.21
	Recall	0.06	0.07	0.17
	F-measure	0.06	0.07	0.18
With Stemming, Without Stoplisting	Precision	0.09	0.12	0.26
	Recall	0.07	0.10	0.21
	F-measure	0.08	0.11	0.23
Without Stemming, With Stoplisting	Precision	0.13	0.16	0.22
	Recall	0.10	0.13	0.18
	F-measure	0.11	0.14	0.19
With Stemming, With Stoplisting	Precision	0.16	0.20	0.27
	Recall	0.13	0.16	0.22
	F-measure	0.15	0.18	0.24

We have taken the below colour-coding scheme to indicate the range of values:

Lowest Values  Highest Values



Observations:

While Precision and Recall address the relation between the retrieved and relevant sets of documents, there is always a trade-off between them. F-measure combines precision and recall in to a single figure and gives them equal weight to both. [*Ref: COM3110/lectures/lecture_IR4.pdf*]
This can be noticed in our observation.

We can see that the **Precision, Recall** and **F-measure** is the **lowest (0.07, 0.6** and **0.6** respectively**)** for the Configuration – **Without Stemming, Without Stoplisting** of **Binary** TW Scheme, whereas, it is the **highest (0.27, 0.22** and **0.24** respectively**)** for the Configuration – **With Stemming, With Stoplisting** of **TFIDF** TW Scheme.

Inference:

We can infer that our implementation for the **TFIDF system** for the Configuration – **With Stemming, With Stoplisting** works the best as it has the **highest efficiency** as given by its F-measure.