**Information Retrieval**

****

|  |  |
| --- | --- |
| ***Section*** | BCS-7C |
| ***Assignment No*** | 2 |
| ***Due Date*** | 28-Nov-2021 |
| ***Submitted To*** | Mr. Saad Farooq |
| **Group Members** | |
| ***Muhammad Hassan Arshad*** | 18L-2149 |
| ***Haris Ali*** | 18L-1247 |

**IR Assignment 2 – Report**

The entire observation and execution is performed on this corpus consisting a total of 3465 documents. There was a total of 10 queries that were executed on the entire corpus with the usage of three techniques: TF – IDF, OkapiBM25, Jelinek-Mercer. Each method TF-IDF, OkapiBM25 or the Jelinek-Mercer, took separate times for execution with TF-IDF being the most while BM25 being the least time-consuming. Each technique had a huge impact on the result as each method has its own technique.

|  |  |  |  |
| --- | --- | --- | --- |
| **Query - ID** | **TF - IDF** | **Okapi BM 25** | **Jelinek Mercer** |
| **202** | 0.058823529411764705 | 0.01694915254237288 | 0.01694915254237288 |
| **214** | 0.5490477981881747 | 0.5620641869177754 | 0.02879127151063576 |
| **216** | 0.3805247328172806 | 0.4630680444075006 | 0.012883121335351156 |
| **221** | 0.352000530485123 | 0.4125820016768598 | 0.014023691441608654 |
| **227** | 0.1840243413459332 | 0.4463294902366337 | 0.00453385283822771 |
| **230** | 0.2237753419118438 | 0.26396220227913636 | 0.011988384475519517 |
| **234** | 0.5499767183221833 | 0.6567890099479397 | 0.02079328589145774 |
| **243** | 0.49663047322173476 | 0.4707236410614154 | 0.007520923811215719 |
| **246** | 0.23190520353766644 | 0.24516908237083432 | 0.01001988041663499 |
| **250** | 0.06146417484664943 | 0.46450906302951656 | 0.0026714014275887824 |
| **Average** | 0.30881728440883544 | 0.4002145874469985 | 0.013017496569061291 |

## **Observation:**

From the table above we can get the information:

* TF-IDF and OkapiBM25 performed almost similarly, however, Jelinek-Mercer performed very poorly.
* The OkapiBM25 performed the best and also it had the best performance overall. This is due to the fact that it only was processing the words that were in the queries while ignoring the rest. It was surprisingly very good even better than the TF-IDF method which was performing in a brute-force like manner; as it avoids this by having an idf score of the words.
* Initially, Jelinek-Mercer was of high expectations that it will perform quite accurately as it is a smoothing technique. However, smoothing took a lot of probabilities from the doc vectors and that is why the document relevance was reduced immensely. One reason could be because the queries length was quite short and that’s why smoothing took quite a probability from each result.
* Although, there was not much of a difference. Query no 9 is the hardest of all, as we can see, on the average, it’s GAP score turned out to be not so good.