*Group members:*

* ***Alessandro Taglieri, 1890945***
* ***Guglielmo Lato, 1257406***

Part 1 – Search engine evaluation:

* **Number of indexed documents and the number of queries**
  + Cranfield\_DATASET:
    - Number of indexed documents: 1401;
    - Number of queries: 223;
  + Time\_DATASET:
    - Number of indexed documents: 424;
    - Number of queries: 84.
* **Number of queries in Ground-Truth:**
  + Cranfield : 673
  + Time\_DATASET: 322
* **MRR table for each configuration**:

Cranifeld\_DATASET Time\_DATASET

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **CONFIGURATION** | **MRR** | | **StemmingAnalyzer() TF\_IDF** | 0.35675324675324677 | | **StemmingAnalyzer() Frequency** | 0.2777525252525252 | | **StemmingAnalyzer() BM25F** | 0.4333333333333334 | | **SimpleAnalyzer() TF\_IDF** | 0.1296969696969697 | | **SimpleAnalyzer() Frequency** | 0.04773629148629149 | | **SimpleAnalyzer() BM25F** | 0.4038744588744589 | | **StandardAnalyzer() TF\_IDF** | 0.3255811360356815 | | **StandardAnalyzer() Frequency** | 0.27674242424242423 | | **StandardAnalyzer() BM25F** | 0.4460606060606062 | | **RegexAnalyzer() TF\_IDF** | 0.1296969696969697 | | **RegexAnalyzer() Frequency** | 0.04773629148629149 | | **RegexAnalyzer() BM25F** | 0.4038744588744589 | | **FancyAnalyzer() TF\_IDF** | 0.3301265905811361 | | **FancyAnalyzer() Frequency** | 0.28128787878787875 | | **FancyAnalyzer() BM25F** | 0.45363636363636384 | | **NgramAnalyzer(5) TF\_IDF** | 0.2848701298701299 | | **NgramAnalyzer(5) Frequency** | 0.25056637806637805 | | **NgramAnalyzer(5) BM25F** | 0.3315656565656567 | | **KeywordAnalyzer() TF\_IDF** | 0.10984848484848485 | | **KeywordAnalyzer() Frequency** | 0.03838203463203464 | | **KeywordAnalyzer() BM25F** | 0.390995670995671 | | **LanguageAnalyzer() TF\_IDF** | 0.36357142857142866 | | **LanguageAnalyzer() Frequency** | 0.30906204906204904 | | **LanguageAnalyzer() BM25F** | 0.4416017316017317 | | |  |  | | --- | --- | | **CONFIGURATION** | **MRR** | | **StemmingAnalyzer() TF\_IDF** | 0.43479166666666663 | | **StemmingAnalyzer() Frequency** | 0.3785416666666666 | | **StemmingAnalyzer() BM25F** | 0.6770833333333334 | | **SimpleAnalyzer() TF\_IDF** | 0.20897321428571428 | | **SimpleAnalyzer() Frequency** | 0.12666666666666665 | | **SimpleAnalyzer() BM25F** | 0.6264583333333333 | | **StandardAnalyzer() TF\_IDF** | 0.45979166666666665 | | **StandardAnalyzer() Frequency** | 0.4025 | | **StandardAnalyzer() BM25F** | 0.6243749999999999 | | **RegexAnalyzer() TF\_IDF** | 0.0 | | **RegexAnalyzer() Frequency** | 0.0 | | **RegexAnalyzer() BM25F** | 0.0 | | **FancyAnalyzer() TF\_IDF** | 0.45979166666666665 | | **FancyAnalyzer() Frequency** | 0.4025 | | **FancyAnalyzer() BM25F** | 0.6302083333333333 | | **NgramAnalyzer(5) TF\_IDF** | 0.380625 | | **NgramAnalyzer(5) Frequency** | 0.30124999999999996 | | **NgramAnalyzer(5) BM25F** | 0.5583333333333333 | | **KeywordAnalyzer() TF\_IDF** | 0.0 | | **KeywordAnalyzer() Frequency** | 0.0 | | **KeywordAnalyzer() BM25F** | 0.0 | | **LanguageAnalyzer() TF\_IDF** | 0.49312500000000004 | | **LanguageAnalyzer() Frequency** | 0.40145833333333336 | | **LanguageAnalyzer() BM25F** | 0.6900000000000001 | |

* **A schematic description of all configurations:**

We created 24 different search engine configurations. We changed these configurations based on different analyzer and different scoring function: We have used eight analyzer (Stemming Analyzer, Simple Analyzer, Standard Analyzer, Regex Analyzer, Fancy Analyzer, Ngram Analyzer, Keyword Analyzer and Language Analyzer). We have combined each analyzer with three different scoring function, that are: Tf\_Idf, Bm25f, and Frequency. In the previous table we can see every configuration with its respective mrr score.

* **R-precision distribution table for top five search configurations:**

*Cranfield\_DATASET*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SE conf** | **mean** | **min** | **1quartile** | **median** | **3quartile** | **max** |
| StandardAnalyzer() BM25F | 522 | 0.0 | 186 | 0.5 | 878 | 1.0 |
| SimpleAnalyzer() BM25F | 534 | 0.0 | 0.2 | 0.5 | 889 | 1.0 |
| FancyAnalyzer() BM25F | 529 | 0.0 | 0.2 | 0.5 | 889 | 1.0 |
| StemmingAnalyzer() BM25F | 536 | 0.0 | 0.2 | 0.5 | 865 | 1.0 |
| LanguageAnalyzer() BM25F | 543 | 0.0 | 0.2 | 517 | 885 | 1.0 |

*Time\_DATASET*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SE conf** | **mean** | **min** | **1quartile** | **median** | **3quartile** | **max** |
| SimpleAnalyzer() BM25F | 237 | 0.0 | 0.0 | 218 | 362 | 1.0 |
| StemmingAnalyzer() BM25F | 252 | 0.0 | 0.0 | 244 | 0.44 | 1.0 |
| LanguageAnalyzer() BM25F | 263 | 0.0 | 0.0 | 0.25 | 0.5 | 1.0 |
| StandardAnalyzer() BM25F | 256 | 0.0 | 0.0 | 0.25 | 421 | 1.0 |
| FancyAnalyzer() BM25F | 256 | 0.0 | 0.0 | 0.25 | 421 | 1.0 |

* **p@K plot:**

*Cranfield\_DATASET Time\_DATASET*

Immagine che contiene testo, mappa

Descrizione generata automaticamenteImmagine che contiene testo, mappa

Descrizione generata automaticamente

* **nCDG@K plot:**

*Cranfield\_DATASET Time\_DATASET*

Immagine che contiene testo, mappa

Descrizione generata automaticamenteImmagine che contiene testo, mappa

Descrizione generata automaticamente

Part 2 – Near-Duplicates-Detection:

* **Number of row and the number of bands that you chose**:
  + R = 5
  + B = 30
* **The probability to have false-Negatives, in the set of candidate pairs, for the following Jaccard values: 0.89, 0.9, 0.95 and 1:**

*P(False negatives) = (1-J\*\*r)\*\*b*

|  |  |
| --- | --- |
| **JACCARD VALUE** | **PRIOBABILITY TO HAVE FN** |
| **0.89** | 2.2423145959674837e^(-11) |
| **0.9** | 2.333110689493024e^(-12) |
| **0.95** | 4.323903046637328e^(-20) |
| **1** | 0 |

* **The probability to have false-Positives, in the set of candidate pairs, for the following Jaccard values: 0.85, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55 and 0.5.**

*P(False positives) = 1-(1-J\*\*r)\*\*b*

|  |  |
| --- | --- |
| **JACCARD VALUE** | **PROBABILITY TO HAVE FP** |
| **0.85** | 0.9999999771362545 |
| **0.8** | 0.9999932813669531 |
| **0.75** | 0.9997045171242237 |
| **0.7** | 0.9959949903519956 |
| **0.65** | 0.9752748118175822 |
| **0.6** | 0.9118304460022172 |
| **0.55** | 0.7875762833291795 |
| **0.5** | 0.6142095565052361 |

* **How did you reduce the probability to have False-Negatives?**

It is sure that we will have false positives in our candidates results, based on the values of r and b that we chose. We decided to emphasize on limiting the false negatives for this homework. We could try to increase r and minimize b in order to have smaller false positives, but when we would have bigger false negatives, due to the tradeoff between them. For this specific homework, our conception was to minimize the false negatives. So, to reduce the probability to have FN, we reduced r.

* **The execution time of the Near-Duplicates-Detection tool.**
  + Execution of time: 1 min 33 sec
* **The number of Near-Duplicates couples you found**.
  + Number of near-duplicates: 25276
* **The number of Near-Duplicates couples you found with an approximated Jaccard value at most 0.89, 0.90, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 1.**

|  |  |
| --- | --- |
| **JACCARD VALUE** | **NUMBER OF NEAR DUPLICATES** |
| **0.89** | 25276 |
| **0.90** | 24899 |
| **0.91** | 24122 |
| **0.92** | 23848 |
| **0.93** | 23111 |
| **0.94** | 22761 |
| **0.95** | 22164 |
| **0.96** | 21858 |
| **0.97** | 21252 |
| **0.98** | 21060 |
| **0.99** | 20574 |
| **1** | 20478 |