**CE706 - Information Retrieval 2021**

**Assignment 2**

Student ID :2003667

# Test collection (Task 1)

*Include here the selected information needs and how they will be represented as a query.*

|  |  |
| --- | --- |
| **Information need** | **Query** |
| Find no more than 10 documents that describe within processed fields “pr\_title and “pr\_absract” the main symptoms that people acquire when they get infected by COVID-19.  Return information from fields “cord\_uid”, “title”, “abstract”, “publish\_time” as the result of a search. | query={  **"size"**: **"10"**,  **"query"**: {  **"multi\_match"**:{  **"query"**:**"Main symptoms of a Covid**  **disease"**,  **"fuzziness"** : **"AUTO"**,  **"fields"** : [ **"pr\_title"**,  **"pr\_abstract^5"** ],  **"type"**:**"best\_fields"**,  **"analyzer"**: **"standard"** ,  **"minimum\_should\_match"**:**"50%"** }  },  **"\_source"**: [**"cord\_uid"**,  **"title"**,**"abstract"**,**"publish\_time"**], } |
| Find no more than 10 documents that describe within processed fields “pr\_title and “pr\_absract” how having diabetes, obesity or pulmonary diseases influence on acquiring a severe form of a coronavirus disease.  Return information from fields “cord\_uid”, “title”, “abstract”, “publish\_time” as the result of a search. | query={  **"size"**: **"10"**,  **"query"**: {  **"multi\_match"**: {  **"query"**: **"Influence of diabetes, obesity, pulmonary diseases on acquiring a severe form of Coronavirus disease"**,  **"fuzziness"**:**"AUTO"**,  **"fields"**: [**"pr\_title"**,  **"pr\_abstract^3"** ],  **"analyzer"**: **"english"**,  }  },  **"\_source"**: [**"cord\_uid"**,**"title"**,**"abstract"**,**"publish\_time"**], } |
| Find no more than 10 of the latest documents dated from May 2020 till March 2021 that describe within processed fields “pr\_title and “pr\_absract” how effective is wearing of face coverings in terms of prevention of spread of COVID-19.  Return information from fields “cord\_uid”, “title”, “abstract”, “publish\_time” as the result of a search. | query={  **"size"**: **"10"**,  **"query"**: {  **"bool"**: {  **"must"**: {  **"multi\_match"**: {  **"query"**:**"The effectiveness**  **of wearing masks"**,  **"fields"**:[**"pr\_title"**,  **"pr\_abstract^5"**],  **"analyzer"**: **"english"**,  **"minimum\_should\_match"**:**"50%"** }  },  **"filter"**: {  **"range"**: {  **"publish\_time"**: {  **"gte"**: **"2020-06-01"**,  **"lte"**: **"2021-03-01"**,  **"format"**:  **"year\_month\_day"**,  }  }  }  }  },  **"\_source"**: [**"cord\_uid"**, **"title"**,**"abstract"**,  **"publish\_time"**], } |

﻿IR systems (Task 2)

*Include here the details of your two IR systems and the difference between them.*

The first Information Retrieval system is a full system from the assignment 1 that comprises such steps to perform a search:

* Data Loading
* Text Normalisation
* Text Lemmatisation
* Data Indexing
* Searching in ElasticSearch

The second Information Retrieval system is a system that, unlike the first system, neither selects keywords from sentences nor lemmatises words. Thus, its pipeline includes the following steps:

* Data Loading
* Text Normalisation
* Data Indexing
* Searching in ElasticSearch

The reason for composing the second system is to examine if an exclusion of “Keywords selection” and “Text Lemmatisation” steps from the whole sequence means a deterioration of search accuracy.

This hypothesis implies that a text of an original state is harder to be searched for due to the large amount of unnecesary words that occur in a majority of documents. Thus, the probability of obtaining irrelevant documents with the same words from a query is very high.

In contrast to it, a processed text, that contains only lemmatised keywords, delivers all essential information while having a brief form that is more convenient for a search.

In figure 1 below code commands to launch information retrieval systems 1 and 2 with their different sequences of steps to perform are shown.



Figure 1 – Code commands to run systems 1 and 2

﻿Pool method (Task 3)

*For each method retrieve the top 10 documents. Therefore for each query, you will have a maximum of 20 documents.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Query** | **# different documents** | **Id of the documents retrieved by System 1** | **Id of the documents retrieved by System 2** |
| query={  **"size"**: **"10"**,  **"query"**: {  **"multi\_match"**:{  **"query"**:**"Main symptoms of a Covid**  **disease"**,  **"fuzziness"** : **"AUTO"**,  **"fields"** : [ **"pr\_title"**,  **"pr\_abstract^5"** ],  **"type"**:**"best\_fields"**,  **"analyzer"**: **"standard"** ,  **"minimum\_should\_match"**:**"50%"** }  },  **"\_source"**: [**"cord\_uid"**,  **"title"**,**"abstract"**,**"publish\_time"**], } | *16* | **cord\_uid** | |
| ***rcc5q3rj*** | ***52t7dh4n*** |
| qcytx64m | ***x0pw0t0q*** |
| eev7ii1q | rcc5q3rj |
| ***1bf68vyq*** | ***1bf68vyq*** |
| ***tdrn8l24*** | qcytx64m |
| ***52t7dh4n*** | **43iqfoc1** |
| ***61tw26ws*** | abjnp1px |
| ***8hdok02n*** | ***6q7q8gse*** |
| 6470qlu1 | 6nerogux |
| gfvi8jvs | 3p22cl0k |
| query={  **"size"**: **"10"**,  **"query"**: {  **"multi\_match"**: {  **"query"**: **"Influence of diabetes, obesity, pulmonary diseases on acquiring a severe form of Coronavirus disease"**,  **"fuzziness"**:**"AUTO"**,  **"fields"**: [**"pr\_title"**,  **"pr\_abstract^3"** ],  **"analyzer"**: **"english"**,  }  },  **"\_source"**: [**"cord\_uid"**,**"title"**,**"abstract"**,**"publish\_time"**], } | *16* | **cord\_uid** | |
| ***3vn6yz5c*** | ***3vn6yz5c*** |
| ***mpk3m6q1*** | 0j5828ah |
| y778k3hs | ***7cmyfxu9*** |
|  |  |
| ***0j5828ah*** | ***ifnk41oq*** |
| ***7cmyfxu9*** | 3lp6vkuw |
| ***pknn1l41*** | ***yke3oqij*** |
| lkd55twg | f00758o2 |
| rcc5q3rj | ***lpuwxdik*** |
| ***w7wsftbm*** | xh0wngsr |
| jtq2enhw | ***mpk3m6q1*** |
| query={  **"size"**: **"10"**,  **"query"**: {  **"bool"**: {  **"must"**: {  **"multi\_match"**: {  **"query"**:**"The effectiveness**  **of wearing masks"**,  **"fields"**:[**"pr\_title"**,  **"pr\_abstract^5"**],  **"analyzer"**: **"english"**,  **"minimum\_should\_match"**:**"50%"** }  },  **"filter"**: {  **"range"**: {  **"publish\_time"**: {  **"gte"**: **"2020-06-01"**,  **"lte"**: **"2021-03-01"**,  **"format"**:  **"year\_month\_day"**,  }  }  }  }  },  **"\_source"**: [**"cord\_uid"**, **"title"**,**"abstract"**,  **"publish\_time"**], } | *14* | **cord\_uid** | |
| ***63izqxxl*** | i6l43agq |
| 39px06kb | ***zvgg5duf*** |
| i6l43agq | ***cn38s5tr*** |
| ***cn38s5tr*** | jcac9dwt |
| ***zvgg5duf*** | ***v9yg80jw*** |
| ***qh1osgm6*** | ***qh1osgm6*** |
| ***6ahex7xa*** | ***6ahex7xa*** |
| ceyttetj | qxcug6i8 |
| ***xeyfkjm5*** | lpuwxdik |
| ***v9yg80jw*** | i847673z |

# Relevance assessments (Task 4)

*To be consistent with all the queries, you need to define criteria to judge if a document is relevant for an information need. The same criteria should be used for all the queries. Notice that only containing the same words is not a valid criterion.*

**Relevance criteria:**

The context of documents must meet the requirements of an information need and provide all needed details to be relevant. Otherwise, if a document does contain keywords from a query but does not provide a corresponding context it can be considered irrelevant.

|  |  |
| --- | --- |
| **Query** | **ID of relevant documents** |
| query={  **"size"**: **"10"**,  **"query"**: {  **"multi\_match"**:{  **"query"**:**"Main symptoms of a Covid**  **disease"**,  **"fuzziness"** : **"AUTO"**,  **"fields"** : [ **"pr\_title"**,  **"pr\_abstract^5"** ],  **"type"**:**"best\_fields"**,  **"analyzer"**: **"standard"** ,  **"minimum\_should\_match"**:**"50%"** }  },  **"\_source"**: [**"cord\_uid"**,  **"title"**,**"abstract"**,**"publish\_time"**], } | **cord\_uid** |
| *rcc5q3rj* |
| *1bf68vyq* |
| *tdrn8l24* |
| *52t7dh4n* |
| *61tw26ws* |
| *8hdok02n* |
| *x0pw0t0q* |
| *43iqfoc1* |
| *6q7q8gse* |
| query={  **"size"**: **"10"**,  **"query"**: {  **"multi\_match"**: {  **"query"**: **"Influence of diabetes, obesity, pulmonary diseases on acquiring a severe form of Coronavirus disease"**,  **"fuzziness"**:**"AUTO"**,  **"fields"**: [**"pr\_title"**,  **"pr\_abstract^3"** ],  **"analyzer"**: **"english"**,  }  },  **"\_source"**: [**"cord\_uid"**,**"title"**,**"abstract"**,**"publish\_time"**], } | **cord\_uid** |
| *3vn6yz5c* |
| *mpk3m6q1* |
| *0j5828ah* |
| *7cmyfxu9* |
| *pknn1l41* |
| *w7wsftbm* |
| *ifnk41oq* |
| *yke3oqij* |
| *lpuwxdik* |
| query={  **"size"**: **"10"**,  **"query"**: {  **"bool"**: {  **"must"**: {  **"multi\_match"**: {  **"query"**:**"The effectiveness**  **of wearing masks"**,  **"fields"**:[**"pr\_title"**,  **"pr\_abstract^5"**],  **"analyzer"**: **"english"**,  **"minimum\_should\_match"**:**"50%"** }  },  **"filter"**: {  **"range"**: {  **"publish\_time"**: {  **"gte"**: **"2020-06-01"**,  **"lte"**: **"2021-03-01"**,  **"format"**:  **"year\_month\_day"**,  }  }  }  }  },  **"\_source"**: [**"cord\_uid"**, **"title"**,**"abstract"**,  **"publish\_time"**], } | **cord\_uid** |
| *63izqxxl* |
| *cn38s5tr* |
| *zvgg5duf* |
| *qh1osgm6* |
| *6ahex7xa* |
| *xeyfkjm5* |
| *v9yg80jw* |

# Evaluation (Task 5)

*Include here the details of how you did this step including any issue that you had and how did you face it. You may include screenshots to clarify.*

In order to conduct a convenient pair-wise comparison of documents retrieved by both systems, code commands to build and show HTML tables out of retrieved search results were written.

On figure 2 below a code to create and display HTML tables is demonstrated.



Figure 2 - Code commands that transform raw retrieved results from Elasticsearch into easy-to-read HTML tables

In the figures 3, 4, 5 below fragments of HTML tables that correspond to each of the queries are displayed

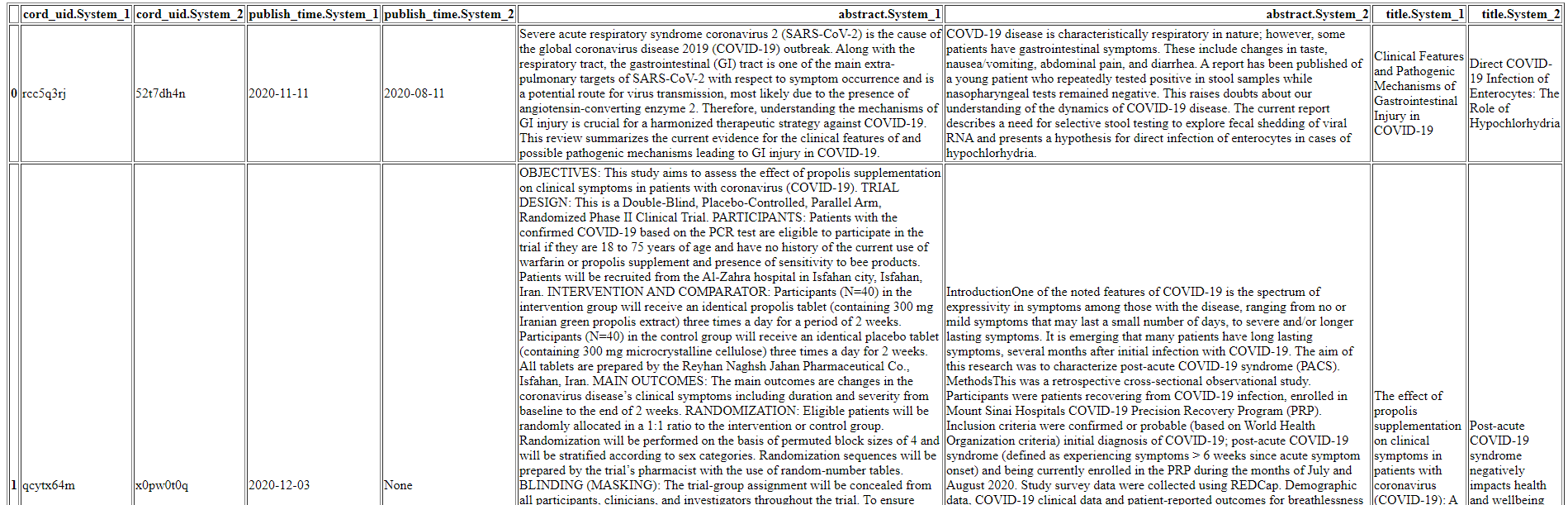


Figure 3 – HTML table of retrieved results for the query 1

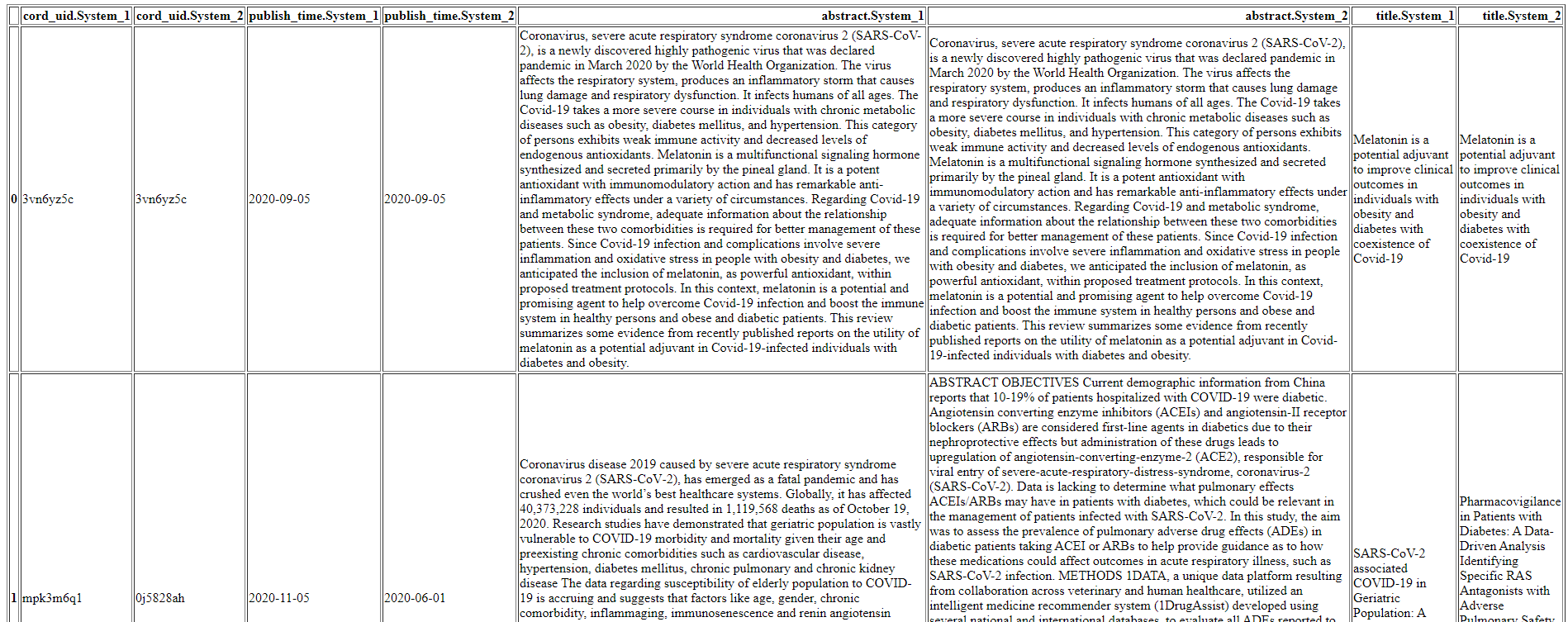


Figure 4 - HTML table of retrieved results for the query 2

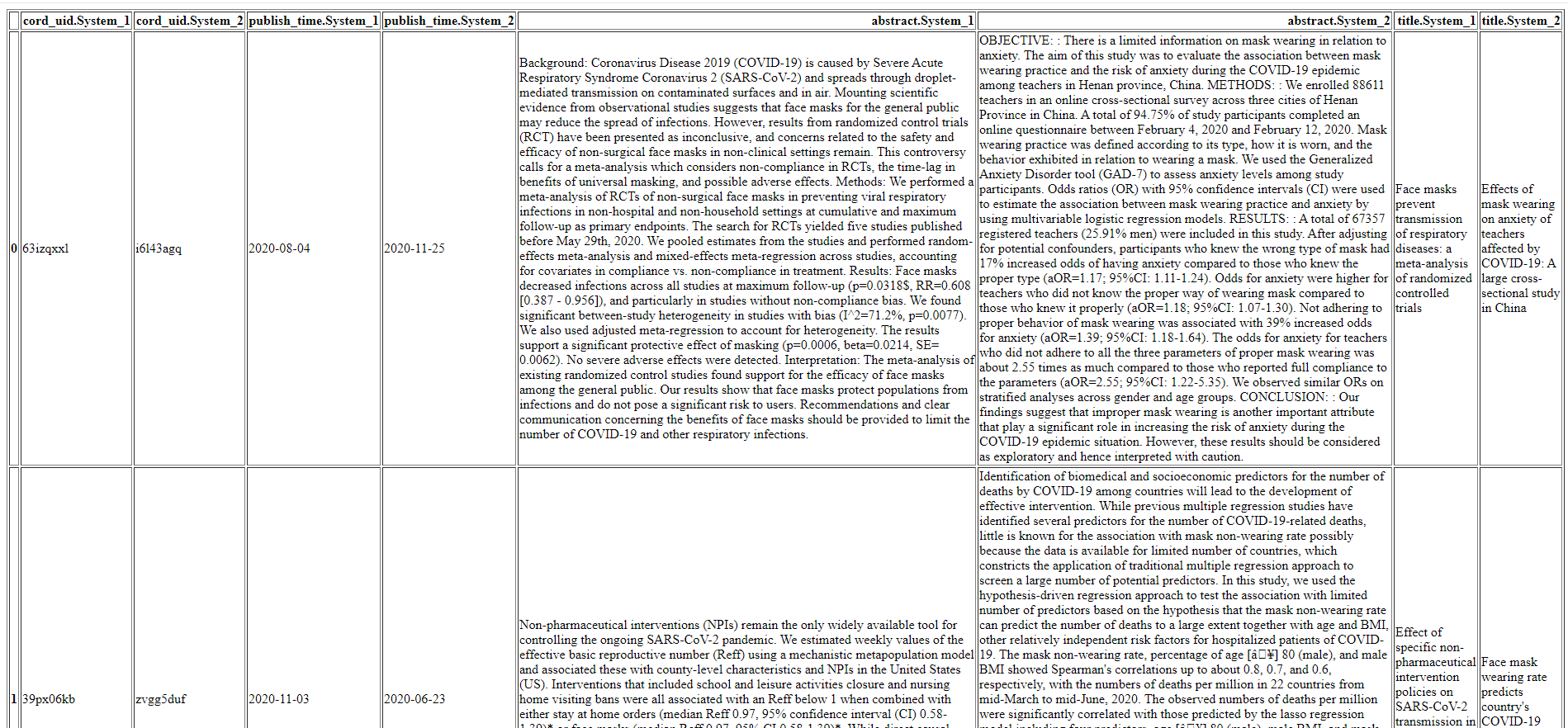


Figure 5 - HTML table of retrieved results for the query 3

In figure 6 below lists of each of the query that contain information regarding the relevancy of each document are represented. Thus, 1 – means that a document is relevant to a query, and 0 – means that a document is irrelevant.

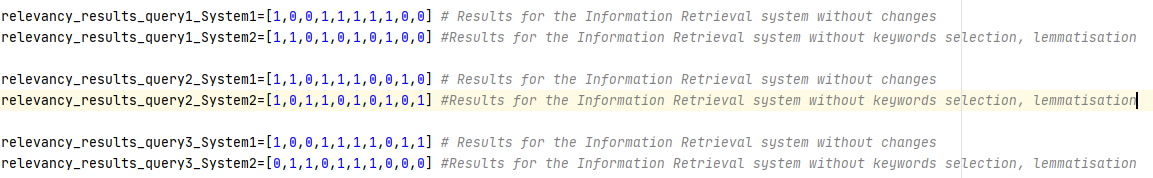


Figure 6 – Lists of relevancy for all queries

In figure 7 below, functions to calculate a precision and a recall at K are shown. Thus, these functions accepts the aforementioned lists of relevancy as an input and a K – number that means a final number of a document to evaluate a metric.

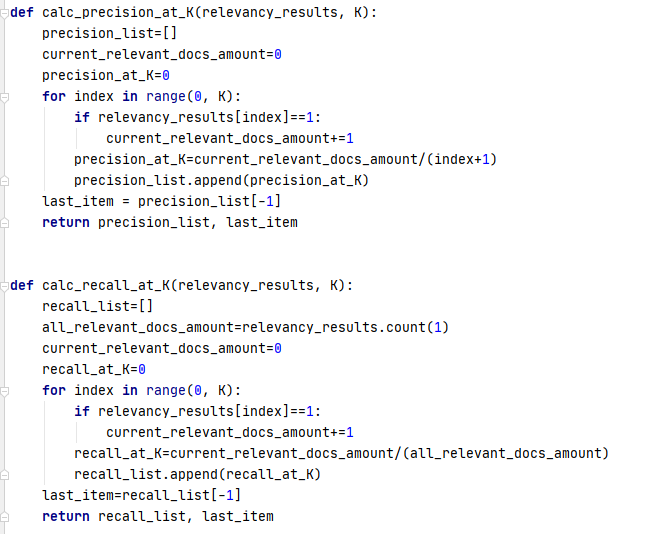


Figure 7 – Functions to evaluate such metrics as P@K and R@K

In figure 8 below, a code example to evaluate metrics P@K and R@K for the first query is shown

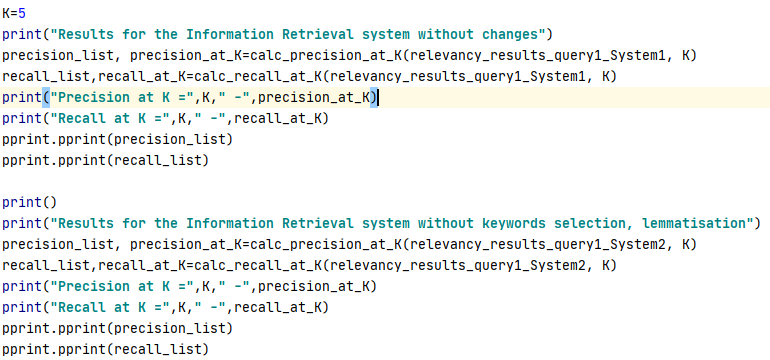


Figure 8 – a code example to evaluate metrics P@K and R@K for the first query

On figures 9,10,11, results of evaluated metrics of precision and recall for the K-number that is equal to 5, are demonstrated.

Precision at 5 and Recall at 5 values for each query along with lists of precision and recall values for each document within the scope of 5 can be also seen from these figures.

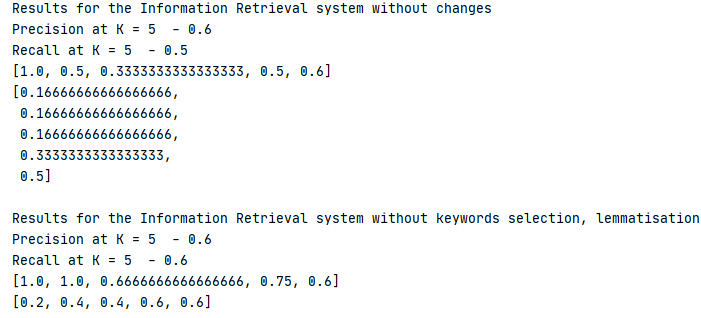


Figure 9 – Precision and Recall results at the K-value - 5 for the first query

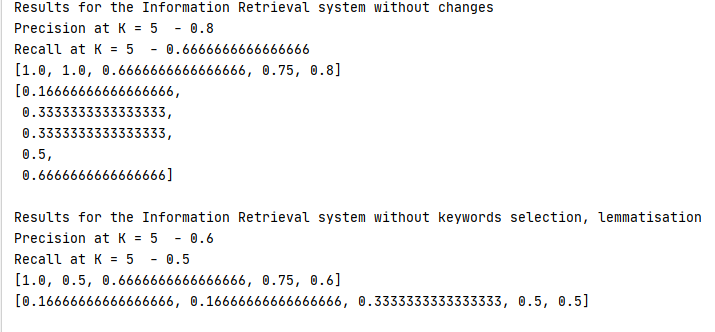


Figure 10 – Precision and Recall results at the K-value - 5 for the second query

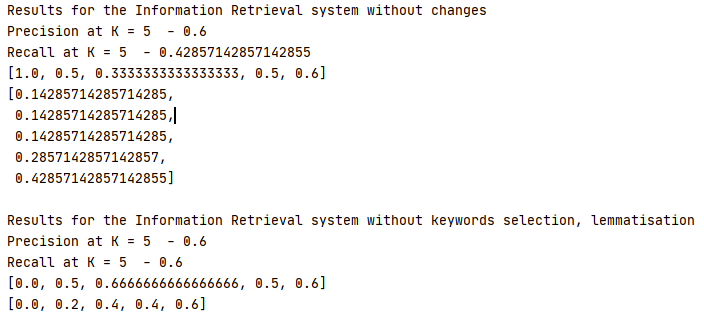


Figure 11 – Precision and Recall results at the K-value - 5 for the third query

In table 1 all precision and recall values for each of the query and each of the system are represented.

Table 1 – All retrieved results for each query and each system

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **System 1** | | **System 2** | |
|  | **P@5** | **R@5** | **P@5** | **R@5** |
| **Q1** | 0.6 | 0.5 | 0.6 | 0.6 |
| **Q2** | 0.8 | 0.667 | 0.6 | 0.5 |
| **Q3** | 0.6 | 0.429 | 0.6 | 0.6 |

According to Table 1 above, it can be noticed that the search performance of the system 2, from where steps “Keywords selection” and “Text Lemmatisation” were excluded, is quite stable for each of the query and doesn’t exhibit any sharp fluctuations.

In contrast to it, the behaviour of the first system is more unpredictable. For the second query, it is obvious that precision and recall values are higher than the System 2 ones. This fact indicates that the search performance of the first system, which includes “Keywords selection” and “Text Lemmatisation” steps, is better.

On the other hand, according to the results for the other queries, its search accuracy is slightly worse. This indication can point out that particularly the “Keywords selection” step can remove some really important words. In this case, a search efficiency may be decreased simply due to a lack of words in texts. Therefore, for some cases, this step may not work properly, since it depends on many words that were removed.

Although, considering the distance between each relevant document among the retrieved ones from the first system it is apparent that these documents are grouped more densely. This interesting detail points out that despite having a lack of words in texts, the first system puts more purpose into its returned results.

All in all, considering the obtained results, it can be stated that a comparison of these two IR systems requires more deep research that includes building more queries or/and testing other configurations of the first system to determine their real efficiency on a larger set of data.