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.	`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": {`

Return information from fields “cord_uid”, “title”, “abstract”, “publish_time” as the result of a search.	"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:

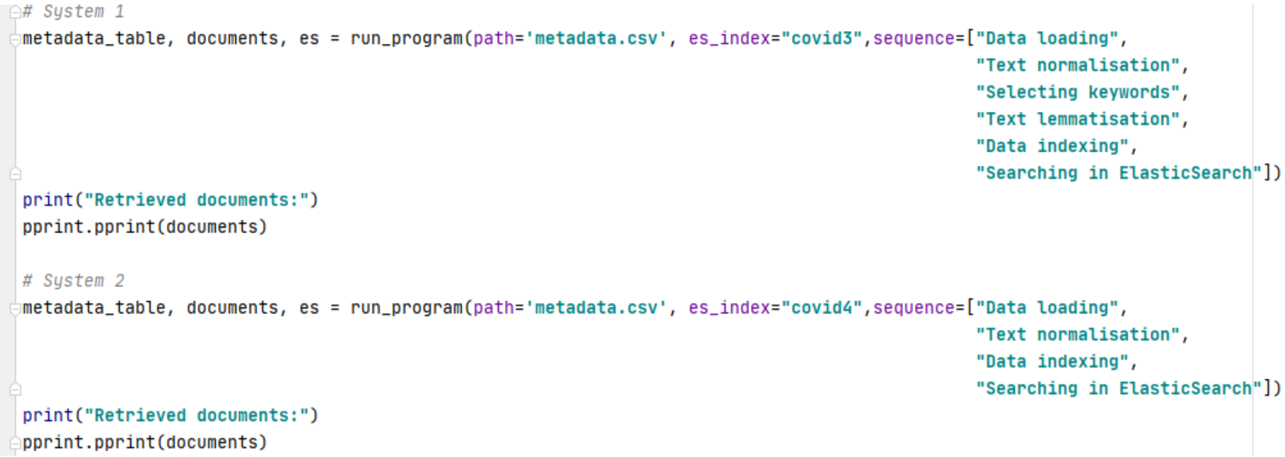
- Data Loading
- Text Normalisation
- Data Indexing
- Searching in ElasticSearch

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.



```
# System 1
metadata_table, documents, es = run_program(path='metadata.csv', es_index="covid3",sequence=["Data loading",
                                                                                           "Text normalisation",
                                                                                           "Selecting keywords",
                                                                                           "Text lemmatisation",
                                                                                           "Data indexing",
                                                                                           "Searching in ElasticSearch"])
print("Retrieved documents:")
pprint.pprint(documents)

# System 2
metadata_table, documents, es = run_program(path='metadata.csv', es_index="covid4",sequence=["Data loading",
                                                                                           "Text normalisation",
                                                                                           "Data indexing",
                                                                                           "Searching in ElasticSearch"])
print("Retrieved documents:")
pprint.pprint(documents)
```

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
	cord_uid
	63izqxxl

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"],
}
	cn38s5tr

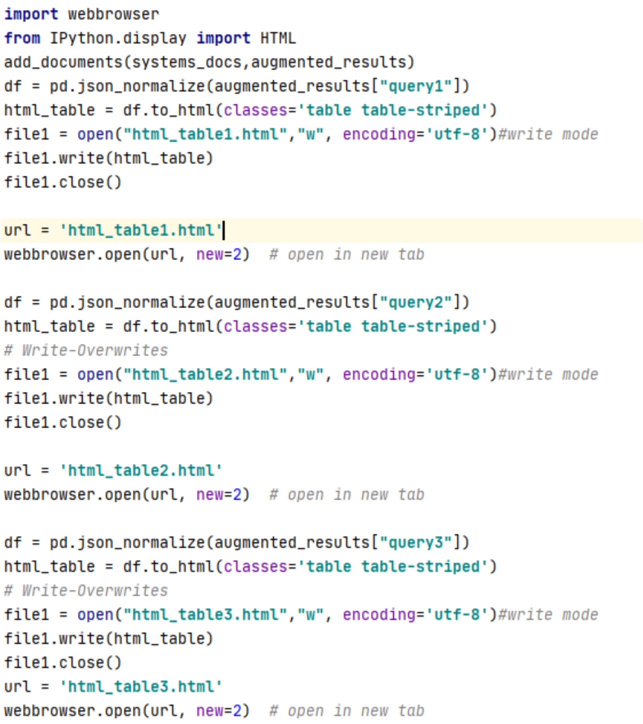
	zvgg5duf
	qh1osgm6
	6ahex7xa
	xeyfkjm5
	v9yg80jw

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.

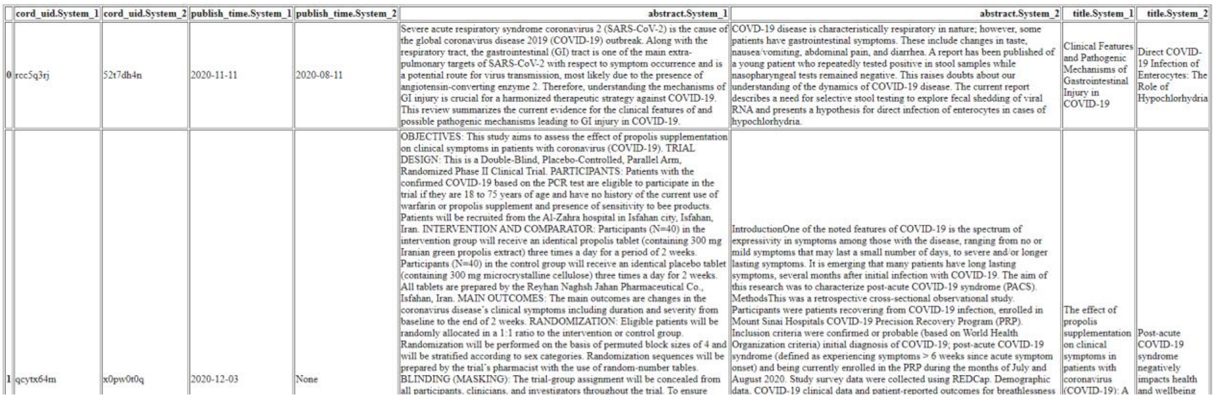
On figure 2 below a code to create and display HTML tables is demonstrated.



```
import webbrowser
from IPython.display import HTML
add_documents(systems_docs,augmented_results)
df = pd.json_normalize(augmented_results["query1"])
html_table = df.to_html(classes='table table-striped')
file1 = open("html_table1.html","w", encoding='utf-8')#write mode
file1.write(html_table)
file1.close()

url = 'html_table1.html'
webbrowser.open(url, new=2)  # open in new tab

df = pd.json_normalize(augmented_results["query2"])
html_table = df.to_html(classes='table table-striped')
# Write-Overwrites
file1 = open("html_table2.html","w", encoding='utf-8')#write mode
file1.write(html_table)
file1.close()

url = 'html_table2.html'
webbrowser.open(url, new=2)  # open in new tab

df = pd.json_normalize(augmented_results["query3"])
html_table = df.to_html(classes='table table-striped')
# Write-Overwrites
file1 = open("html_table3.html","w", encoding='utf-8')#write mode
file1.write(html_table)
file1.close()
url = 'html_table3.html'
webbrowser.open(url, new=2)  # open in new tab
```

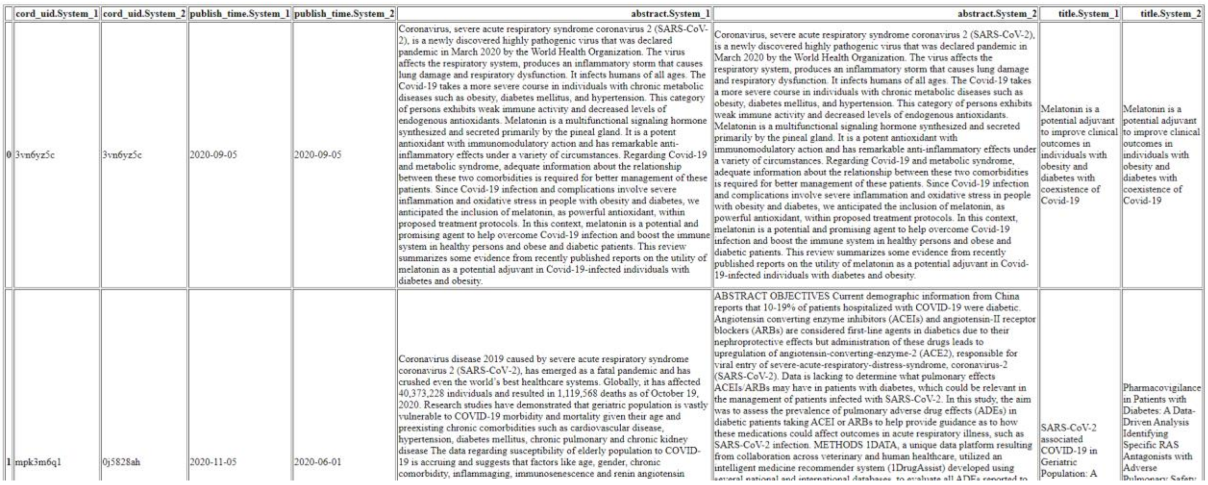
Figure 2 - Code commands that transform raw retrieved results from Elasticsearch into easy-to-read HTML tables

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



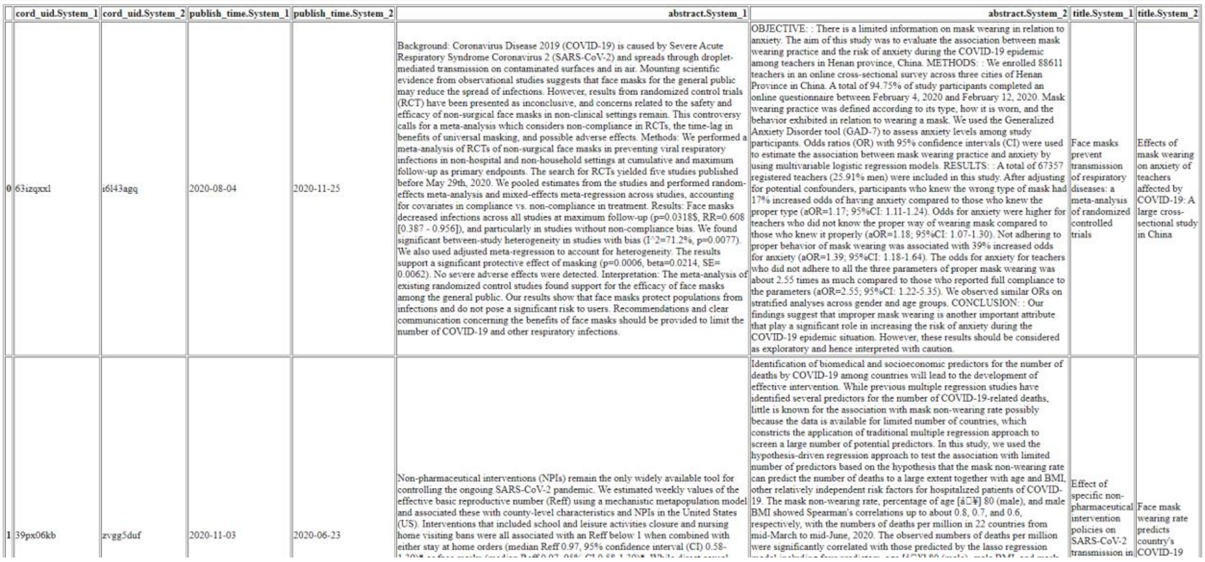
	cord_uid.System_1	cord_uid.System_2	publish_time.System_1	publish_time.System_2	abstract.System_1	abstract.System_2	title.System_1	title.System_2
0	rcc5q3rj	52t7dh4n	2020-11-11	2020-08-11	Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the cause of the global coronavirus disease 2019 (COVID-19) outbreak. Along with the respiratory tract, the gastrointestinal (GI) tract is one of the main extra-pulmonary targets of SARS-CoV-2 with respect to symptom occurrence and is a potential route for virus transmission, most likely due to the presence of angiotensin-converting enzyme 2. Therefore, understanding the mechanisms of GI injury is crucial for a harmonized therapeutic strategy against COVID-19. This review summarizes the current evidence for the clinical features of and possible pathogenic mechanisms leading to GI injury in COVID-19.	COVD-19 disease is characteristically respiratory in nature; however, some patients have gastrointestinal symptoms. These include changes in taste, nausea/vomiting, abdominal pain, and diarrhea. A report has been published of a young patient who repeatedly tested positive in stool samples while nasopharyngeal tests remained negative. This raises doubts about our understanding of the dynamics of COVID-19 disease. The current report describes a need for selective stool testing to explore fecal shedding of viral RNA and presents a hypothesis for direct infection of enterocytes in cases of hypochlorhydria.	Clinical Features and Pathogenic Mechanisms of Gastrointestinal Injury in COVID-19	Direct COVID-19 Infection of Enterocytes: The Role of Hypochlorhydria
1	qcytx64m	x0pw0t0q	2020-12-03	None	OBJECTIVES: This study aims to assess the effect of propolis supplementation on clinical symptoms in patients with coronavirus (COVID-19). TRIAL DESIGN: This is a Double-Blind, Placebo-Controlled, Parallel Arm, Randomized Phase II Clinical Trial. PARTICIPANTS: Patients with the confirmed COVID-19 based on the PCR test are eligible to participate in the trial if they are 18 to 75 years of age and have no history of the current use of warfarin or propolis supplement and presence of sensitivity to bee products. Patients will be recruited from the Al-Zahra hospital in Isfahan city, Isfahan, Iran. INTERVENTION AND COMPARATOR: Participants (N=40) in the intervention group will receive an identical propolis tablet (containing 300 mg Iranian green propolis extract) three times a day for a period of 2 weeks. Participants (N=40) in the control group will receive an identical placebo tablet (containing 300 mg microcrystalline cellulose) three times a day for 2 weeks. All tablets are prepared by the Reyhan Naghsh Jahan Pharmaceutical Co., Isfahan, Iran. MAIN OUTCOMES: The main outcomes are changes in the coronavirus disease's clinical symptoms including duration and severity from baseline to the end of 2 weeks. RANDOMIZATION: Eligible patients will be randomly allocated in a 1:1 ratio to the intervention or control group. Randomization will be performed on the basis of permuted block sizes of 4 and will be stratified according to sex categories. Randomization sequences will be prepared by the trial's pharmacist with the use of random-number tables. BLINDING (MASKING): The trial-group assignment will be concealed from all participants, clinicians, and investigators throughout the trial. To ensure	IntroductionOne of the noted features of COVID-19 is the spectrum of expressivity in symptoms among those with the disease, ranging from no or mild symptoms that may last a small number of days, to severe and/or longer lasting symptoms. It is emerging that many patients have long lasting symptoms, several months after initial infection with COVID-19. The aim of this research was to characterize post-acute COVID-19 syndrome (PACS). MethodsThis was a retrospective cross-sectional observational study. Participants were patients recovering from COVID-19 infection, enrolled in Mount Sinai Hospitals COVID-19 Precision Recovery Program (PRP). Inclusion criteria were confirmed or probable (based on World Health Organization criteria) initial diagnosis of COVID-19; post-acute COVID-19 syndrome (defined as experiencing symptoms > 6 weeks since acute symptom onset) and being currently enrolled in the PRP during the months of July and August 2020. Study survey data were collected using REDCap. Demographic data, COVID-19 clinical data and patient-reported outcomes for breathlessness	The effect of propolis supplementation on clinical symptoms in patients with coronavirus (COVID-19): A	Post-acute COVID-19 syndrome negatively impacts health and wellbeing

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



	cord_uid.System_1	cord_uid.System_2	publish_time.System_1	publish_time.System_2	abstract.System_1	abstract.System_2	title.System_1	title.System_2
0	3vn6yz5c	3vn6yz5c	2020-09-05	2020-09-05	Coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a newly discovered highly pathogenic virus that was declared pandemic in March 2020 by the World Health Organization. The virus affects the respiratory system, produces an inflammatory storm that causes lung damage and respiratory dysfunction. It infects humans of all ages. The Covid-19 takes a more severe course in individuals with chronic metabolic diseases such as obesity, diabetes mellitus, and hypertension. This category of persons exhibits weak immune activity and decreased levels of endogenous antioxidants. Melatonin is a multifunctional signaling hormone synthesized and secreted primarily by the pineal gland. It is a potent antioxidant with immunomodulatory action and has remarkable anti-inflammatory effects under a variety of circumstances. Regarding Covid-19 and metabolic syndrome, adequate information about the relationship between these two comorbidities is required for better management of these patients. Since Covid-19 infection and complications involve severe inflammation and oxidative stress in people with obesity and diabetes, we anticipated the inclusion of melatonin, as powerful antioxidant, within proposed treatment protocols. In this context, melatonin is a potential and promising agent to help overcome Covid-19 infection and boost the immune system in healthy persons and obese and diabetic patients. This review summarizes some evidence from recently published reports on the utility of melatonin as a potential adjuvant in Covid-19-infected individuals with diabetes and obesity.	Coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a newly discovered highly pathogenic virus that was declared pandemic in March 2020 by the World Health Organization. The virus affects the respiratory system, produces an inflammatory storm that causes lung damage and respiratory dysfunction. It infects humans of all ages. The Covid-19 takes a more severe course in individuals with chronic metabolic diseases such as obesity, diabetes mellitus, and hypertension. This category of persons exhibits weak immune activity and decreased levels of endogenous antioxidants. Melatonin is a multifunctional signaling hormone synthesized and secreted primarily by the pineal gland. It is a potent antioxidant with immunomodulatory action and has remarkable anti-inflammatory effects under a variety of circumstances. Regarding Covid-19 and metabolic syndrome, adequate information about the relationship between these two comorbidities is required for better management of these patients. Since Covid-19 infection and complications involve severe inflammation and oxidative stress in people with obesity and diabetes, we anticipated the inclusion of melatonin, as powerful antioxidant, within proposed treatment protocols. In this context, melatonin is a potential and promising agent to help overcome Covid-19 infection and boost the immune system in healthy persons and obese and diabetic patients. This review summarizes some evidence from recently published reports on the utility of melatonin as a potential adjuvant in Covid-19-infected individuals with diabetes and obesity.	Melatonin is a potential adjuvant to improve clinical outcomes in individuals with obesity and diabetes with coexistence of Covid-19	Melatonin is a potential adjuvant to improve clinical outcomes in individuals with obesity and diabetes with coexistence of Covid-19
1	mpk3m6q1	0j5828ah	2020-11-05	2020-06-01	Coronavirus disease 2019 caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has emerged as a fatal pandemic and has crushed even the world's best healthcare systems. Globally, it has affected 40,373,228 individuals and resulted in 1,119,568 deaths as of October 19, 2020. Research studies have demonstrated that geriatric population is vastly vulnerable to COVID-19 morbidity and mortality given their age and preexisting chronic comorbidities such as cardiovascular disease, hypertension, diabetes mellitus, chronic pulmonary and chronic kidney disease The data regarding susceptibility of elderly population to COVID-19 is accruing and suggests that factors like age, gender, chronic comorbidity, inflammaging, immunosenescence and renin angiotensin	ABSTRACT OBJECTIVES Current demographic information from China reports that 10-19% of patients hospitalized with COVID-19 were diabetic. Angiotensin converting enzyme inhibitors (ACEIs) and angiotensin-II receptor blockers (ARBs) are considered first-line agents in diabetics due to their nephroprotective effects but administration of these drugs leads to upregulation of angiotensin-converting-enzyme-2 (ACE2), responsible for viral entry of severe-acute-respiratory-distress-syndrome, coronavirus-2 (SARS-CoV-2). Data is lacking to determine what pulmonary effects ACEIs/ARBs may have in patients with diabetes, which could be relevant in the management of patients infected with SARS-CoV-2. In this study, the aim was to assess the prevalence of pulmonary adverse drug effects (ADEs) in diabetic patients taking ACEI or ARBs to help provide guidance as to how these medications could affect outcomes in acute respiratory illness, such as SARS-CoV-2 infection. METHODS 1DATA, a unique data platform resulting from collaboration across veterinary and human healthcare, utilized an intelligent medicine recommender system (1DrugAssist) developed using several national and international databases, to evaluate all ADEs reported to	SARS-CoV-2 associated COVID-19 in Geriatric Population: A	Pharmacovigilance in Patients with Diabetes: A Data-Driven Analysis Identifying Specific RAS Antagonists with Adverse Pulmonary Safety

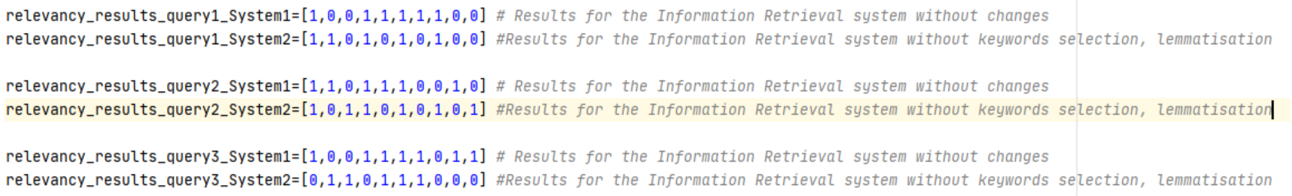
Figure 4 - HTML table of retrieved results for the query 2



	cord_uid.System_1	cord_uid.System_2	publish_time.System_1	publish_time.System_2	abstract.System_1	abstract.System_2	title.System_1	title.System_2
0	63izqxxl	i6l43agq	2020-08-04	2020-11-25	Background: Coronavirus Disease 2019 (COVID-19) is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and spreads through droplet-mediated transmission on contaminated surfaces and in air. Mounting scientific evidence from observational studies suggests that face masks for the general public may reduce the spread of infections. However, results from randomized control trials (RCT) have been presented as inconclusive, and concerns related to the safety and efficacy of non-surgical face masks in non-clinical settings remain. This controversy calls for a meta-analysis which considers non-compliance in RCTs, the time-lag in benefits of universal masking, and possible adverse effects. Methods: We performed a meta-analysis of RCTs of non-surgical face masks in preventing viral respiratory infections in non-hospital and non-household settings at cumulative and maximum follow-up as primary endpoints. The search for RCTs yielded five studies published before May 29th, 2020. We pooled estimates from the studies and performed random-effects meta-analysis and mixed-effects meta-regression across studies, accounting for covariates in compliance vs. non-compliance in treatment. Results: Face masks decreased infections across all studies at maximum follow-up (p=0.0318$, RR=0.608 [0.387 - 0.956]), and particularly in studies without non-compliance bias. We found significant between-study heterogeneity in studies with bias (I^2=71.2%, p=0.0077). We also used adjusted meta-regression to account for heterogeneity. The results support a significant protective effect of masking (p=0.0006, beta=0.0214, SE= 0.0062). No severe adverse effects were detected. Interpretation: The meta-analysis of existing randomized control studies found support for the efficacy of face masks among the general public. Our results show that face masks protect populations from infections and do not pose a significant risk to users. Recommendations and clear communication concerning the benefits of face masks should be provided to limit the number of COVID-19 and other respiratory infections.	OBJECTIVE: : There is a limited information on mask wearing in relation to anxiety. The aim of this study was to evaluate the association between mask wearing practice and the risk of anxiety during the COVID-19 epidemic among teachers in Henan province, China. METHODS: : We enrolled 88611 teachers in an online cross-sectional survey across three cities of Henan Province in China. A total of 94.75% of study participants completed an online questionnaire between February 4, 2020 and February 12, 2020. Mask wearing practice was defined according to its type, how it is worn, and the behavior exhibited in relation to wearing a mask. We used the Generalized Anxiety Disorder tool (GAD-7) to assess anxiety levels among study participants. Odds ratios (OR) with 95% confidence intervals (CI) were used to estimate the association between mask wearing practice and anxiety by using multivariable logistic regression models. RESULTS: : A total of 67357 registered teachers (25.91% men) were included in this study. After adjusting for potential confounders, participants who knew the wrong type of mask had 17% increased odds of having anxiety compared to those who knew the proper type (aOR=1.17; 95%CI: 1.11-1.24). Odds for anxiety were higher for teachers who did not know the proper way of wearing mask compared to those who knew it properly (aOR=1.18; 95%CI: 1.07-1.30). Not adhering to proper behavior of mask wearing was associated with 39% increased odds for anxiety (aOR=1.39; 95%CI: 1.18-1.64). The odds for anxiety for teachers who did not adhere to all the three parameters of proper mask wearing was about 2.55 times as much compared to those who reported full compliance to the parameters (aOR=2.55; 95%CI: 1.22-5.35). We observed similar ORs on stratified analyses across gender and age groups. CONCLUSION: : Our findings suggest that improper mask wearing is another important attribute that play a significant role in increasing the risk of anxiety during the COVID-19 epidemic situation. However, these results should be considered as exploratory and hence interpreted with caution.	Face masks prevent transmission of respiratory diseases: a meta-analysis of randomized controlled trials	Effects of mask wearing on anxiety of teachers affected by COVID-19: A large cross-sectional study in China
1	39px06kb	zvgg5duf	2020-11-03	2020-06-23	Non-pharmaceutical interventions (NPIs) remain the only widely available tool for controlling the ongoing SARS-CoV-2 pandemic. We estimated weekly values of the effective basic reproductive number (Reff) using a mechanistic metapopulation model and associated these with county-level characteristics and NPIs in the United States (US). Interventions that included school and leisure activities closure and nursing home visiting bans were all associated with an Reff below 1 when combined with either stay at home orders (median Reff 0.97, 95% confidence interval (CI) 0.58-	Identification of biomedical and socioeconomic predictors for the number of deaths by COVID-19 among countries will lead to the development of effective intervention. While previous multiple regression studies have identified several predictors for the number of COVID-19-related deaths, little is known for the association with mask non-wearing rate possibly because the data is available for limited number of countries, which constricts the application of traditional multiple regression approach to screen a large number of potential predictors. In this study, we used the hypothesis-driven regression approach to test the association with limited number of predictors based on the hypothesis that the mask non-wearing rate can predict the number of deaths to a large extent together with age and BMI, other relatively independent risk factors for hospitalized patients of COVID-19. The mask non-wearing rate, percentage of age [â□¥] 80 (male), and male BMI showed Spearman's correlations up to about 0.8, 0.7, and 0.6, respectively, with the numbers of deaths per million in 22 countries from mid-March to mid-June, 2020. The observed numbers of deaths per million were significantly correlated with those predicted by the lasso regression	Effect of specific non-pharmaceutical intervention policies on SARS-CoV-2 transmission in	Face mask wearing rate predicts country's COVID-19

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

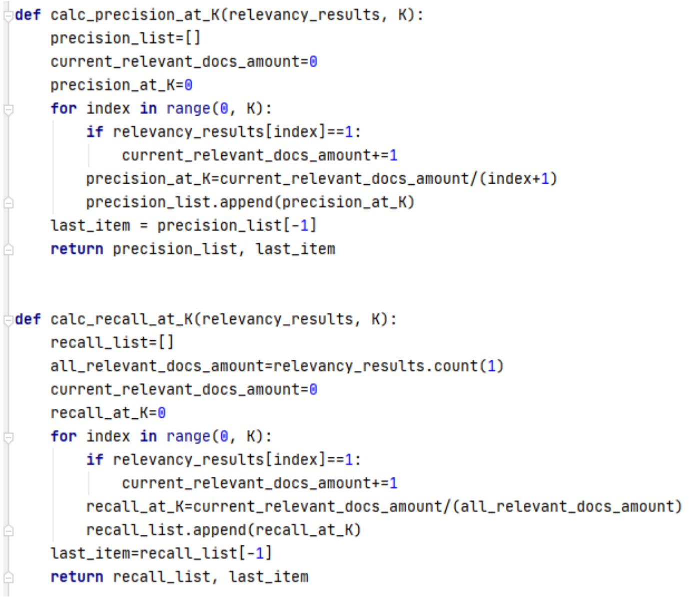
In figure 6 below lists of each of the query that contain information regarding the relevancy of each document are represented. Thus, 1 – means that a document is relevant to a query, and 0 – means that a document is irrelevant.



```
relevancy_results_query1_System1=[1,0,0,1,1,1,1,1,0,0] # Results for the Information Retrieval system without changes
relevancy_results_query1_System2=[1,1,0,1,0,1,0,1,0,0] #Results for the Information Retrieval system without keywords selection, lemmatisation

relevancy_results_query2_System1=[1,1,0,1,1,1,0,0,1,0] # Results for the Information Retrieval system without changes
relevancy_results_query2_System2=[1,0,1,1,0,1,0,1,0,1] #Results for the Information Retrieval system without keywords selection, lemmatisation

relevancy_results_query3_System1=[1,0,0,1,1,1,1,0,1,1] # Results for the Information Retrieval system without changes
relevancy_results_query3_System2=[0,1,1,0,1,1,1,0,0,0] #Results for the Information Retrieval system without keywords selection, lemmatisation
```

Figure 6 – Lists of relevancy for all queries

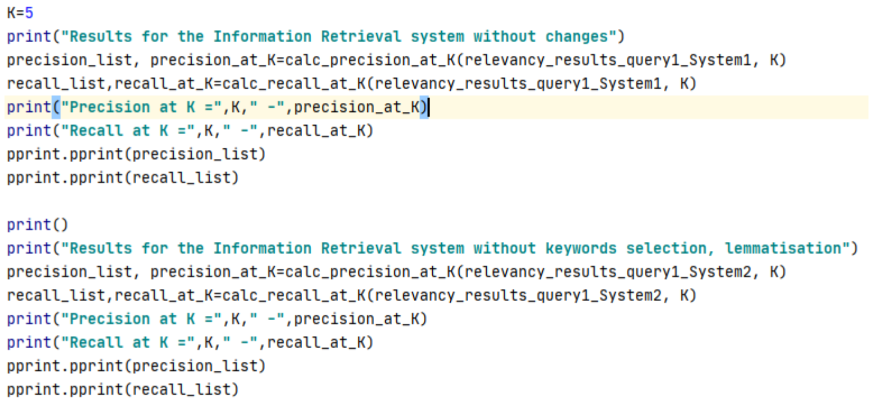
In figure 7 below, functions to calculate a precision and a recall at K are shown. Thus, these functions accepts the aforementioned lists of relevancy as an input and a K – number that means a final number of a document to evaluate a metric.



```
def calc_precision_at_K(relevancy_results, K):
    precision_list=[]
    current_relevant_docs_amount=0
    precision_at_K=0
    for index in range(0, K):
        if relevancy_results[index]==1:
            current_relevant_docs_amount+=1
        precision_at_K=current_relevant_docs_amount/(index+1)
        precision_list.append(precision_at_K)
    last_item = precision_list[-1]
    return precision_list, last_item


def calc_recall_at_K(relevancy_results, K):
    recall_list=[]
    all_relevant_docs_amount=relevancy_results.count(1)
    current_relevant_docs_amount=0
    recall_at_K=0
    for index in range(0, K):
        if relevancy_results[index]==1:
            current_relevant_docs_amount+=1
        recall_at_K=current_relevant_docs_amount/(all_relevant_docs_amount)
        recall_list.append(recall_at_K)
    last_item=recall_list[-1]
    return recall_list, last_item
```

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

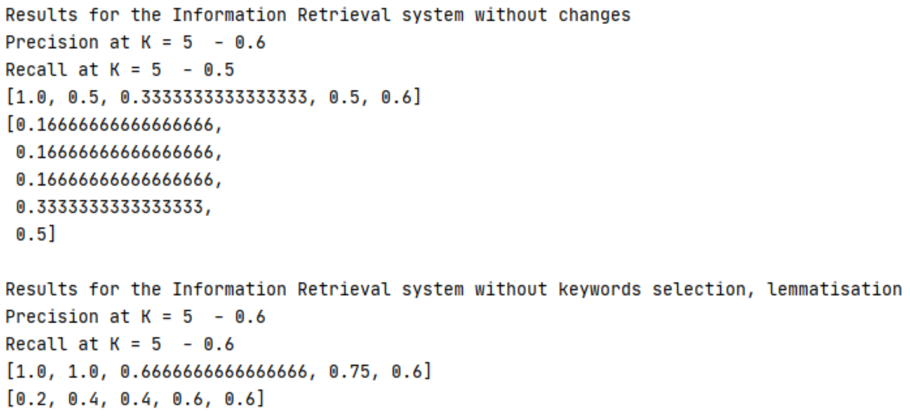


```
K=5
print("Results for the Information Retrieval system without changes")
precision_list, precision_at_K=calc_precision_at_K(relevancy_results_query1_System1, K)
recall_list,recall_at_K=calc_recall_at_K(relevancy_results_query1_System1, K)
print("Precision at K =",K," -",precision_at_K)
print("Recall at K =",K," -",recall_at_K)
pprint.pprint(precision_list)
pprint.pprint(recall_list)

print()
print("Results for the Information Retrieval system without keywords selection, lemmatisation")
precision_list, precision_at_K=calc_precision_at_K(relevancy_results_query1_System2, K)
recall_list,recall_at_K=calc_recall_at_K(relevancy_results_query1_System2, K)
print("Precision at K =",K," -",precision_at_K)
print("Recall at K =",K," -",recall_at_K)
pprint.pprint(precision_list)
pprint.pprint(recall_list)
```

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

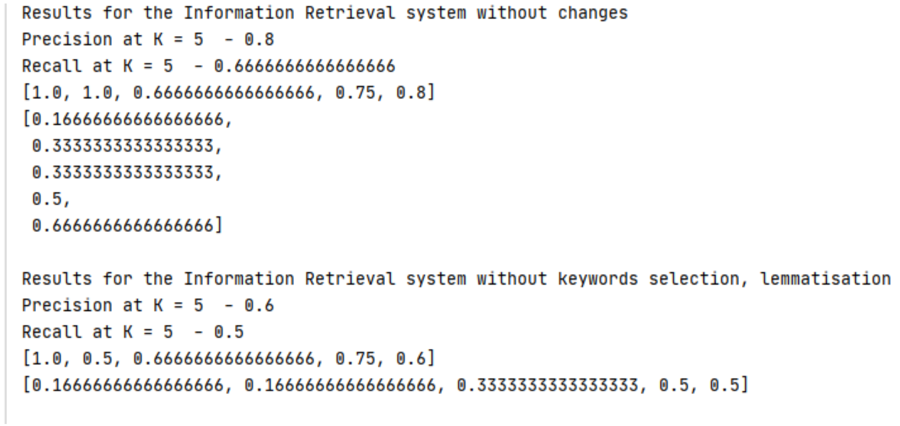
On figures 9,10,11, results of evaluated metrics of precision and recall for the K-number that is equal to 5, are demonstrated.

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.



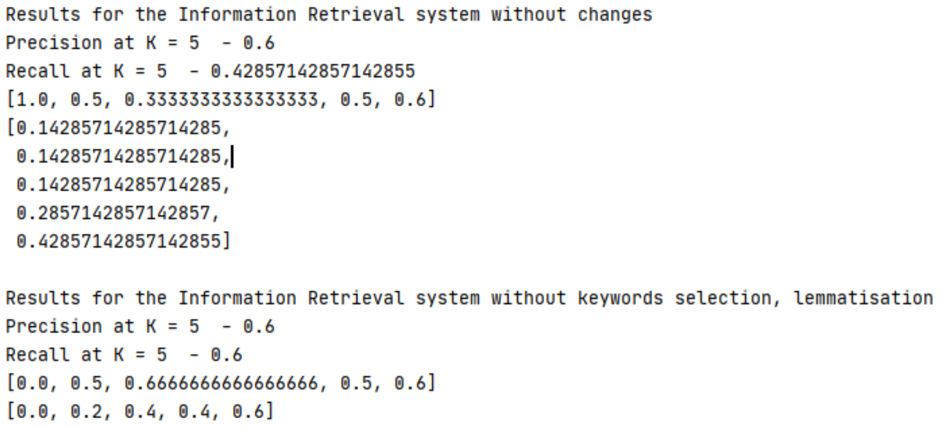
```
Results for the Information Retrieval system without changes
Precision at K = 5  - 0.6
Recall at K = 5  - 0.5
[1.0, 0.5, 0.3333333333333333, 0.5, 0.6]
[0.16666666666666666,
 0.16666666666666666,
 0.16666666666666666,
 0.3333333333333333,
 0.5]

Results for the Information Retrieval system without keywords selection, lemmatisation
Precision at K = 5  - 0.6
Recall at K = 5  - 0.6
[1.0, 1.0, 0.6666666666666666, 0.75, 0.6]
[0.2, 0.4, 0.4, 0.6, 0.6]
```

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



```
Results for the Information Retrieval system without changes
Precision at K = 5  - 0.8
Recall at K = 5  - 0.6666666666666666
[1.0, 1.0, 0.6666666666666666, 0.75, 0.8]
[0.16666666666666666,
 0.3333333333333333,
 0.3333333333333333,
 0.5,
 0.6666666666666666]

Results for the Information Retrieval system without keywords selection, lemmatisation
Precision at K = 5  - 0.6
Recall at K = 5  - 0.5
[1.0, 0.5, 0.6666666666666666, 0.75, 0.6]
[0.16666666666666666, 0.16666666666666666, 0.3333333333333333, 0.5, 0.5]
```

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



```
Results for the Information Retrieval system without changes
Precision at K = 5  - 0.6
Recall at K = 5  - 0.42857142857142855
[1.0, 0.5, 0.3333333333333333, 0.5, 0.6]
[0.14285714285714285,
 0.14285714285714285,
 0.14285714285714285,
 0.2857142857142857,
 0.42857142857142855]

Results for the Information Retrieval system without keywords selection, lemmatisation
Precision at K = 5  - 0.6
Recall at K = 5  - 0.6
[0.0, 0.5, 0.6666666666666666, 0.5, 0.6]
[0.0, 0.2, 0.4, 0.4, 0.6]
```

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.