

CE706 - Information Retrieval 2021

Assignment 2

2003930

Test collection (Task 1)

Information need	Query
What can we learn about coronaviruses by studying genes?	"query": { "match_phrase": { "title": "gene" } }
How do coronaviruses and other infectious diseases cause outbreaks?	"query": { "match_phrase_prefix": { "title": "outbreak" } }
What is the effects of coronaviruses and infectious diseases other viruses in America?	"query": { "match": { "title": "america usa" } }

IR systems (Task 2)

The original system used is based on that used in the first assignment and then I made modifications from that system to create the second system. The two systems used have 5 differences between them. Some are much more impactful than the others. The first difference is that in system 1 I decided to implement lemmatization using the wordnet lemmatizer. In system 2 I decided to change this and implement stemming instead. This is done using a porter stemmer which is also from NLTK. This means that system 2 will look for the stem of a word and use that for indexing, this will likely make more errors than the method implemented in the first system. The second difference is that in the in system 1 each document gets split up into tokens of its sentences, whereas in system 2 the documents are split up into word tokens which are indexed. This means that system 2 will produce more tokens of a shorter length. The third difference is a minor one. This is how the stop words are generated. The first system uses the NLTK stopword corpus whereas the second uses the Gensim stopword corpus. These corpus' have a few very minor differences, so I do not expect this change to have a massive affect on the systems performance. The next change I made is that while system 1 converts all text to lowercase during the preprocessing steps, the second system does not do this. If the text in the document is uppcase, then system 2 will index it as so. The final difference between the two systems is that in system 1 the original document gets indexed alongside the tokens of the processed document. This does not happen in system 2. System 2 will only index the tokens generated from the preprocessing steps. I predict that system 1 will perform much better than the second system. Examples of the abstract and title fields from a document can be seen below to illustrate the differences.

System 1

t abstract	<p>BACKGROUND: There was a pandemic influenza around the world in 2009 including South Korea since last pandemic occurred four decades ago. We aimed to evaluate the epidemiological and clinical characteristics of this infection in childhood. METHODS: We evaluated the epidemiologic characteristics of all the subjects infected with the 2009 H1N1 influenza A virus (2,971 patients, ≤ 15 years of age), and the clinical and laboratory findings of the inpatients (217 patients, 80 had pneumonia) between 1 September 2009 and 31 January 2010 in a single hospital throughout the epidemic. RESULTS: The age distribution of all the subjects was relatively even. Over 90% of cases occurred during a two-month period. Two hundred and five patients (94.5%) received oseltamivir within 48 h of fever onset, and 97% of inpatients defervesced within 48 h of medication. The group with pneumonia included more males than females, and had higher leukocytes counts with lower lymphocyte differentials than the group without pneumonia. The white blood cell count and lymphocyte differential were associated with the severity of pneumonia. Corticosteroid treatment for severe pneumonia patients was highly effective in preventing disease progression. CONCLUSION: Children of all ages affected with even rates of infection, but males were predominant in pneumonia patients. Pneumonia patients showed lymphopenia and its severity was associated with the severity of illness. Our results suggest that the mechanism of lung injury in 2009 H1N1 virus infection may be associated with the host immune response., background pandemic influenza around world 2009 including south korea since last pandemic occurred four decades ago, aimed evaluate epidemiological clinical characteristics infection childhood, methods evaluated epidemiologic characteristics subjects infected 2009 h1n1 influenza virus 2,971 patients ≤ 15 years age clinical laboratory findings inpatients 217 patients 80 pneumonia 1 september 2009 31 january 2010 single hospital throughout epidemic, results age distribution subjects relatively even, 90 cases occurred two-month period, two hundred five patients 94.5 received oseltamivir within 48 h fever onset 97 inpatients defervesced within 48 h medication, group pneumonia included males females higher leukocytes counts lower lymphocyte differentials group without pneumonia, white blood cell count lymphocyte differential associated severity pneumonia, corticosteroid treatment severe pneumonia patients highly effective preventing disease progression, conclusion children ages affected even rates infection males predominant pneumonia patients, pneumonia patients showed lymphopenia severity associated severity illness, results suggest mechanism lung injury 2009 h1n1 virus infection may associated host immune response</p>
t title	Epidemiological and clinical characteristics of childhood pandemic 2009 H1N1 virus infection: an observational cohort study, epidemiological clinical characteristics childhood pandemic 2009 h1n1 virus infection observational cohort study

System 2

t abstract	<p>background, there, pandem, influenza, world, 2009, includ, south, korea, pandem, occur, decad, ago, We, aim, evalu, epidemiolog, clinic, characterist, infect, childhood, method, We, evalu, epidemiolog, characterist, subject, infect, 2009, h1n1, influenza, A, viru, 2,971, patient, ≤, 15, year, age, clinic, laborator, find, inpati, 217, patient, 80, pneumonia, 1, septemb, 2009, 31, januari, 2010, single, hospit, epidem, result, the, age, distribut, subject, rel, even, over, 90, case, occur, two-month, period, two, patient, 94.5, received, oseltamivir, 48, h, fever, onset, 97, inpati, defervesc, 48, h, medic, the, group, pneumonia, includ, male, femal, higher, leukocyte, count, lower, lymphocyt, differenti, group, pneumonia, the, white, blood, cell, count, lymphocyt, differenti, associ, sever, pneumonia, corticosteroid, treatment, sever, pneumonia, patient, highli, effect, prevent, diseas, progress, conclus, children, age, affect, rate, infect, male, predomin, pneumonia, patient, pneumonia, patient, show, lymphopenia, sever, associ, sever, ill, our, result, suggest, st, mechan, lung, injuri, 2009, h1n1, viru, infect, associ, host, immun, respons</p>
t title	epidemiolog, clinic, characterist, childhood, pandem, 2009, h1n1, viru, infect, observ, cohort, studi

Pool method (Task 3)

Query	# different documents	Id of the documents retrieve by System 1	Id of the documents retrieve by System 2
"query": { "match_phrase": { "title": "gene" } }	16	bjjft7ut x53t9i4k sbnnh2mm tloaa3v1 jzwcy7dr k1hwh640 nhb4o6ty hkc4vbmj dcjwfes7 baugu1gh	x53t9i4k bjjft7ut nhb4o6ti mxjtj5c0 t6l692zu 6d9x0xbj 3lq7fnd sbnnh2mm dcjwfes7 hkc4vbmj
"query": { "match_phrase_prefix": { "title": "outbreak" } }	11	ttyo4z6f t7004uw2 2mfbqs8i 3lq7fnd crjwej14 5b936n3g fite9vs8 h3yxymh3 9fr0m92p 36dhfptw	t7004uw2 ttyo4z6f 3lq7fnd crjwej14 2mfbqs8i 5b936n3g h3yxymh3 36dhfptw fite9vs8 2ad1tu4

"query": { "match": { "title": "america usa" } }	4	7vhcf929 hwlvk68z ntx35a8s v9k7vpi8	7vhcf929 hwlvk68z ntx35a8 v9k7vpi8
---	---	--	---

Relevance assessments (Task 4)

Relevance criteria: Documents give some detail which can be used to answer the question set out by the information needs. Does not have to answer the question. But must outline some example of information which could be used to answer the question.

Query	ID of relevant documents
"query": { "match_phrase": { "title": "gene" } }	x53t9i4k k1hwh640 nhb4o6ty hkc4vbmj dcjwfes7 baugulgh mxjtj5c0 t6l692zu 31q7ftnd
"query": { "match_phrase_prefix": { "title": "outbreak" } }	ttyo4z6f t7004uw2 2mfbqs8i 31q7ftnd crjwej14 5b936n3g fite9vs8 h3yxymh3 9fr0m92p 36dhfptw
"query": { "match": { "title": "america usa" } }	hwlvk68z ntx35a8s v9k7vpi8

Evaluation (Task 5)

To calculate the precision and recall I created functions which took a list of the ids of the relevant documents in each query and then a second argument of the results provided by a system.

Pseudocode for P@K:

```

k_pred = the first k values in prediction list
actual = relevant documents
correct_values = documents which appear in k_pred AND actual
pk = correct_values/k
return pk

```

Pseudocode for R@K:

```

correct_values = documents which appear in the first k values of prediction list AND items in the
relevant document list
if correct_values is 0
return 0
else
return correct_values / relevant document list length

```

	System 1		System 2	
	P@5	R@5	P@5	R@5
Q1	0.2	0.11	0.6	0.33
Q2	1.0	0.5	1.0	0.5
Q3	0.6	0.75	0.6	0.75

Discussion: In my initial discussion of the two systems I predicted that system 1 would perform much better than system 2. However, the results above show a slightly different story. Both query 2 and 3 provide us the exact same results. However, in the first query you can see that system 2 is more accurate than the first. However, it is worth noting that the sample size of the data used is small which is likely part of the reason for both query 2 and 3 yielding the same results across the two systems. Especially in the case of query 3 where each system returned the same 4 results – potentially the query was too specific for this and the information needed to be wider.

In the first query I was looking to solve the information need of “What can we learn about coronaviruses by studying genes?”. The simplest way to try and solve this was by looking for the phrase of gene. By using the match_phrase parameter it allowed me to include words such as “genetic” which would also be applicable. However, it also includes words such as “generally” in the scope. This is not what we want to include. The reason for the score of system 1 being so low is that it included more of these sorts of words than system 2 and therefore had less relevant documents. This is likely since system 1 used lemmatization as opposed to stemming which return different values.