CE706 - Information Retrieval 2021

Assignment 2

2003258

Test collection

Information need	Query
how do we track and deal with outbreak of deceases in real-time?	Health surveillance system in real-time
how multi-versatile microarray is better than PCR methods in simultaneous detection of many viruses and their expression profiles?	multi-versatile microarray vs PCR methods in simultaneous detection of viruses
how to assess preparedness of health departments for high-impact public health emergencies?	Assess preparedness of health departments for public health emergencies

IR systems

I am using python elastic search client to index and retrieve documents in this project. And leveraging the standard python input interface to accept queries from the users.

The two IR systems that are being compared are vector space models with different type of vector representations, one is sparse vector representation while other being dense vector representation.

Vector space models fetch the relevant documents by measuring the similarity between query vector and document vector.

Here systems being compared have a different way of representing TF-IDF vector and more importantly the length of vectors each system support.

System 1 (Sparse Vector implementation):

In sparse vector implementation, initially, most informative words in the corpus are identified by removing all the words which have document frequency more than certain threshold (40% document in this case) like shown below

```
tfidf_vectorizer=TfidfVectorizer(max_df=0.4,use_idf=True)
fitted_vectorizer=tfidf_vectorizer.fit(docs_1)
tfidf_vectorizer_vectors=fitted_vectorizer.transform(docs_1)
```

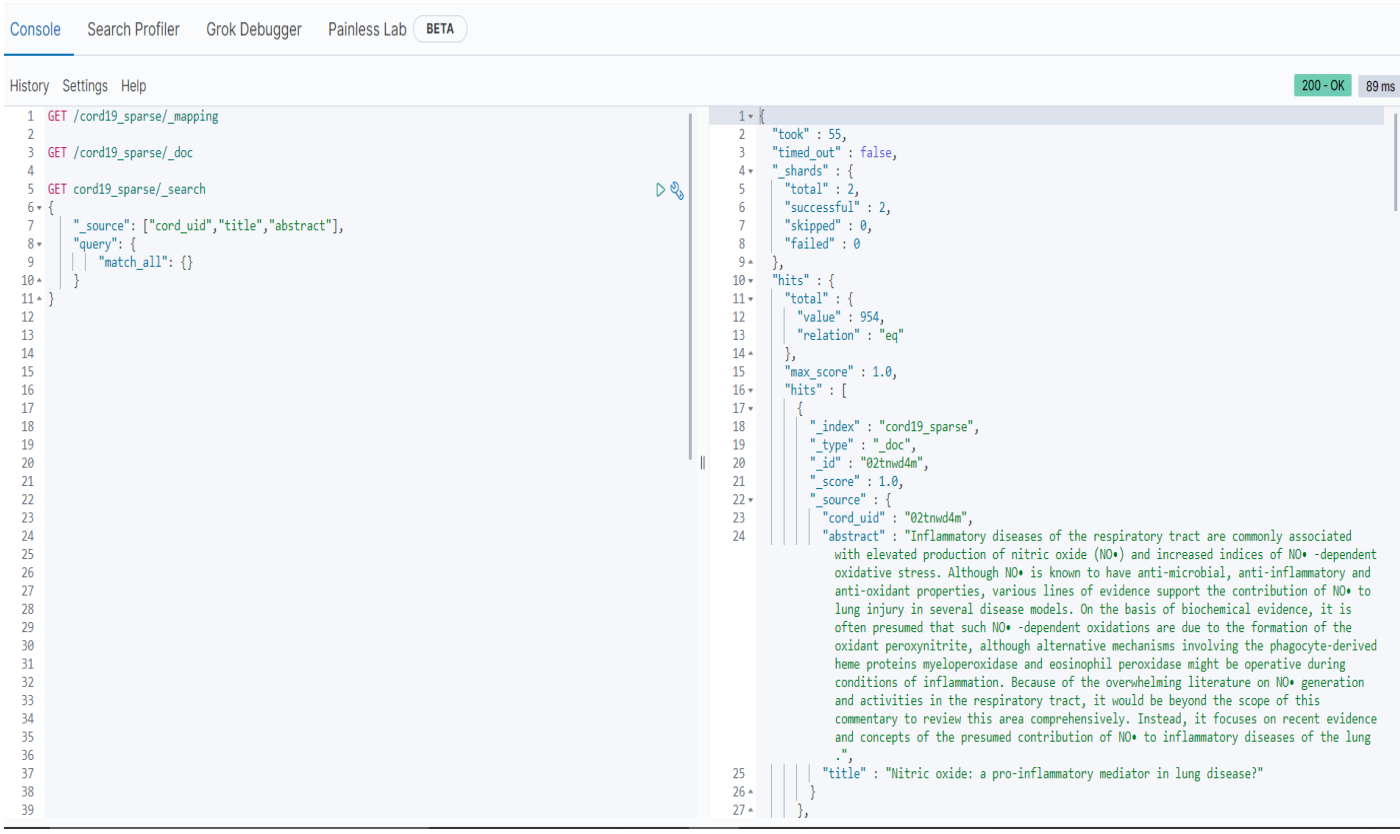
Then for each document sparse vector is calculated using “TFidfVectorizer’ as shown in the above snap and stored as an index vector for a document while indexing to elastic search.

Sparse vector is a dictionary with informative terms as keys and corresponding TF-IDF values as values.

Mapping:

```
mapping = {
    "settings": {
        "number_of_shards": 2,
        "number_of_replicas": 1
    },
    "mappings": {
        "properties": {
                    "cord_uid": {
                        "type": "keyword"
                        },
                    "source_x": {
                        "type": "keyword",
                    },
                    "authors": {
                        "type": "text",
                    },
                    "journal": {
                        "type" :"text"
                    },
                    "url": {
                        "type" :"keyword"
                    },
                    "tfidf_vector": {
                      "type": "sparse_vector"
                    },
                        "license":{
                            "type":"keyword"
                    }
                }
    }
}
```

Indexed Docs (Kibana):



When user enters a query, it goes through the same pre-processing steps all documents processed through and the sparse vector is made for query and it is passed to "cosineSimilaritySparse" function elastic search natively provides to calculate cosine similarity to rank the documents and bring top 10 results.

Below is the query used to fetch the relevant documents

Query:

```
req_get_all = {
   "size":10,
   "query":{
      "script_score":{
         "query":{
           "wildcard":{
               field:
               {
                "value": "*"+value+"*",
                "case_insensitive":True
               }
           }
         },
         "script":{
            "source":"cosineSimilaritySparse(params.query_vector, 'tfidf_vector') + 1.0",
            "params":{
                "query_vector":vec
                }
             }
          }
       }
    }
```

The maximum length of the sparse vector is 1024, since vector contains only informative terms that exist in the documents, retrieval model can accommodate good number of important features. In this project, around 9000 of important features from corpus are chosen based on the document frequency.

System 2 (Dense vector implementation):

Here, dense vector is a list with TF-IDF values of most important features in the corpus, the short coming of dense vector in our implementation is , it has to contain all TF-IDF values of features in the system not just the features that exist in the document like sparse vector implementation. And maximum length of dense vector is 2048, that puts a limitation of selecting a maximum of only 2048 important features, which would considerably affect the performance compared to sparse vector implementation.

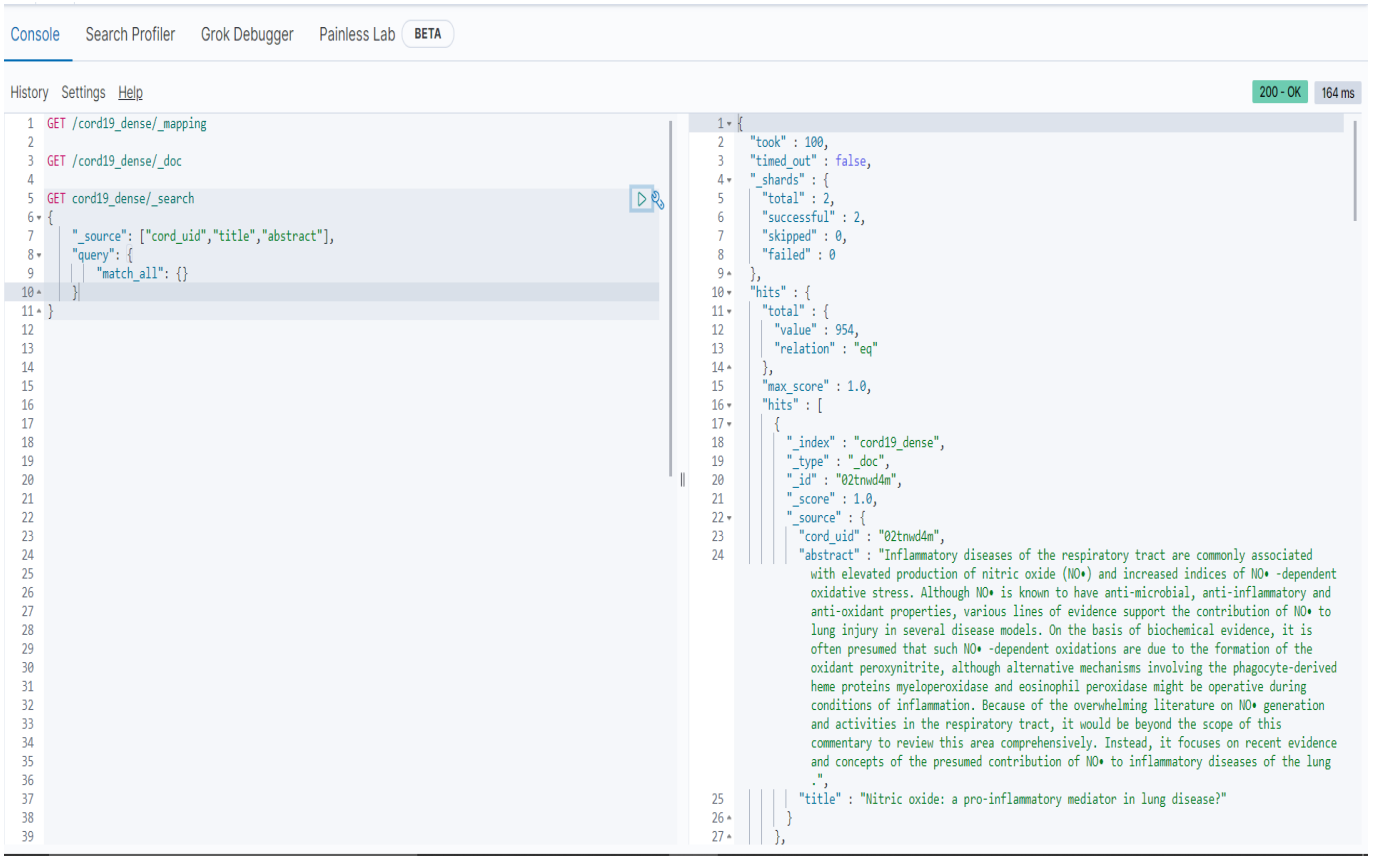
```
#Considers 2048 most informative terms based on IDF scores and buids TF-IDF vecotor for every document
#and fitted tfidf_vectorizer can be used in future to transform user query
tfidf_vectorizer=TfidfVectorizer(max_features=2048,use_idf=True)
fitted_vectorizer=tfidf_vectorizer.fit(docs_1)
tfidf_vectorizer_vectors=fitted_vectorizer.transform(docs_1)
```

Then for each document dense vector is calculated using “TFidfVectorizer’ as shown in the above snap and stored as an index vector for a document while indexing to elastic search.

Mapping:

```
mapping = {
    "settings": {
        "number_of_shards": 2,
        "number_of_replicas": 1
    },
    "mappings": {
         "properties": {
                    "cord_uid": {
                        "type": "keyword"
                        },
                    "source_x": {
                        "type": "keyword",
                    },
                    "title": {
                        "type": "text"
                    },
                    "abstract": {
                        "type": "text",
                    },
                    "url": {
                        "type" :"keyword"
                    },
                    "tfidf_vector": {
                      "type": "dense_vector",
                      "dims": 2048
                    },
                        "license":{
                            "type":"keyword"
                    }
                }
    }
}
```

Indexed Docs (Kibana):



When user enters a query ,it goes through the same pre-processing steps all documents processed through and the dense vector is made for query and it is passed to “cosineSimilarity” function elastic search natively provides to calculate cosine similarity to rank the documents and bring top 10 results

Query:

```
try:
    req_get_all = {
        "size":10,
        "query":{
            "script_score":{
                "query":{
                  "wildcard":{
                      field:
                      {
                       "value": "*"+value+"*",
                       "case_insensitive":True
                      }
                  }
                },
                "script":{
                    "source":"cosineSimilarity(params.query_vector, 'tfidf_vector') + 1.0",
                    "params":{
                        "query_vector":vec
                        }
                    }
                }
            }
        }
```

Pool method

Query	# different documents	Id of the documents retrieve by System 1	Id of the documents retrieve by System 2
Health surveillance system in real-time	05ppugs7 085v7n6k 5hwqdnx1 emnln2ix fite9vs8 fvfjz7al ge5iri3v jg13scgo m1xf62bf nge1itgk qfb7074e tw6wusxe w5fxen70 wnuqe66q yz2wbpuu zzc7n84w	emnln2ix zzc7n84w nge1itgk jg13scgo fite9vs8 05ppugs7 fvfjz7al yz2wbpuu w5fxen70 tw6wusxe	emnln2ix jg13scgo fite9vs8 ge5iri3v yz2wbpuu 5hwqdnx1 qfb7074e 085v7n6k wnuqe66q m1xf62bf
multi-versatile microarray vs PCR methods in simultaneous detection of viruses	3tt99oax 5gglmx9d 6f8vyziv 6gow0x1v 7s5b3lpn 8ctsa9sd 9r62ffew cj4zlr3c cpjlzutk cxzlmfst d3cko4j2 ngn9x3lc ps3mnpzq qfb7074e rzzsmuoc t35n7bk9 zgapjjjw	t35n7bk9 ngn9x3lc 6gow0x1v 6f8vyziv ps3mnpzq qfb7074e zgapjjjw 9r62ffew cxzlmfst 7s5b3lpn	qfb7074e ps3mnpzq zgapjjjw d3cko4j2 8ctsa9sd cpjlzutk 3tt99oax cj4zlr3c 5gglmx9d rzzsmuoc

Assess preparedness of health departments public health emergencies	05ppugs7 0jj9svwj dzzlpz4o fite9vs8 fvfjz7al ge5iri3v h8xxpbr6 hf3nytb2 jg13scgo m1xf62bf p34ezktf rsnblu4d tw6wusxe wgxt36jv wnuqe66q yz2wbpuu zzc7n84w	fvfjz7al 05ppugs7 wgxt36jv p34ezktf tw6wusxe ge5iri3v zzc7n84w rsnblu4d fite9vs8 h8xxpbr6	ge5iri3v wgxt36jv fite9vs8 dzzlpz4o jg13scgo wnuqe66q 0jj9svwj hf3nytb2 m1xf62bf yz2wbpuu

----------Question 1----------

1) how do we track and deal with outbreak of deceases in real-time?
   Query: Health surveillance system in real-time

Input Query 1

```
Do you want to search by query?(y/n)
y

Please enter the query!!
Health surveillance system in real-time
```

System 1 (Sparse vector implementation) output

	score	cord_uid	title
0	1.378537	emnln2ix	Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan
1	1.294169	zzc7n84w	Strengthening systems for communicable disease surveillance: creating a laboratory network in Rwanda
2	1.281083	nge1itgk	Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends
3	1.251538	jg13scgo	Technical Description of RODS: A Real-time Public Health Surveillance System
4	1.247661	fite9vs8	Avian influenza outbreak in Turkey through health personnel's views: a qualitative study
5	1.213475	05ppugs7	A multidimensional classification of public health activity in Australia
6	1.185136	fvfjz7al	Public health preparedness in Alberta: a systems-level study
7	1.166522	yz2wbpuu	Globalization and Health
8	1.164496	w5fxen70	Reliability of case definitions for public health surveillance assessed by Round-Robin test methodology
9	1.156728	tw6wusxe	Local public health workers' perceptions toward responding to an influenza pandemic

System 2 (Dense vector implementation) output

	score	cord_uid	title
0	1.505887	emnln2ix	Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan
1	1.411795	jg13scgo	Technical Description of RODS: A Real-time Public Health Surveillance System
2	1.279736	fite9vs8	Avian influenza outbreak in Turkey through health personnel's views: a qualitative study
3	1.219024	ge5iri3v	Global Public Health Security
4	1.205416	yz2wbpuu	Globalization and Health
5	1.187347	5hwqdnx1	Enhancing Time-Series Detection Algorithms for Automated Biosurveillance
6	1.184918	qfb7074e	Development of TaqMan(®)MGB fluorescent real-time PCR assay for the detection of anatid herpesvirus 1
7	1.170037	085v7n6k	Pilot Evaluation of RT-PCR/Electrospray Ionization Mass Spectrometry (PLEX-ID/Flu assay) on Influenza-Positive Specimens
8	1.166481	wnuqe66q	Global public goods and the global health agenda: problems, priorities and potential
9	1.163274	m1xf62bf	Community responses to communication campaigns for influenza A (H1N1): a focus group study

Unique documents(Pooling) from both the systems are:

'05ppugs7','085v7n6k', '5hwqdnx1', 'emnln2ix','fite9vs8', 'fvfjz7al','ge5iri3v','jg13scgo', 'm1xf62bf','nge1itgk', 'qfb7074e', 'tw6wusxe', 'w5fxen70', 'wnuqe66q', 'yz2wbpuu', 'zzc7n84w'

----------Question2 ----------

2) how multi-versatile microarray is better than PCR methods in simultaneous detection of many viruses and their expression profiles?
   Query: multi-versatile microarray vs PCR methods in simultaneous detection of viruses

Input Query 2

```
Do you want to search by query?(y/n)
y

Please enter the query!!
multi-versatile microarray vs PCR methods in s
```

System 1 (Sparse vector implementation) output

	score	cord_uid	title
0	1.293545	t35n7bk9	Multi-faceted, multi-versatile microarray: simultaneous detection of many viruses and their expression profiles
1	1.184133	ngn9x3lc	The Microbial Detection Array Combined with Random Phi29-Amplification Used as a Diagnostic Tool for Virus Detection in Clinical Samples
2	1.133153	6gow0x1v	Virus Identification in Unknown Tropical Febrile Illness Cases Using Deep Sequencing
3	1.128034	6f8vyziv	Development of Real-Time PCR Array for Simultaneous Detection of Eight Human Blood-Borne Viral Pathogens
4	1.120492	ps3mnpzq	ProCAT: a data analysis approach for protein microarrays
5	1.119242	qfb7074e	Development of TaqMan(®)MGB fluorescent real-time PCR assay for the detection of anatid herpesvirus 1
6	1.108473	zgapjjjw	Usefulness of Published PCR Primers in Detecting Human Rhinovirus Infection
7	1.104239	9r62ffew	Differentially profiling the low-expression transcriptomes of human hepatoma using a novel SSH/microarray approach
8	1.099112	cxzlmfst	Automated identification of multiple micro-organisms from resequencing DNA microarrays
9	1.095267	7s5b3lpn	A Microarray Based Approach for the Identification of Common Foodborne Viruses

System 2 (Dense vector implementation) output

	score	cord_uid	title
0	1.226891	qfb7074e	Development of TaqMan(®)MGB fluorescent real-time PCR assay for the detection of anatid herpesvirus 1
1	1.207257	ps3mnpzq	ProCAT: a data analysis approach for protein microarrays
2	1.191949	zgapjjjw	Usefulness of Published PCR Primers in Detecting Human Rhinovirus Infection
3	1.162975	d3cko4j2	From Functional Genomics to Functional Immunomics: New Challenges, Old Problems, Big Rewards
4	1.151251	8ctsa9sd	RNA and DNA Bacteriophages as Molecular Diagnosis Controls in Clinical Virology: A Comprehensive Study of More than 45,000 Routine PCR Tests
5	1.145013	cpjlzutk	Detection of Mycobacterium ulcerans by the Loop Mediated Isothermal Amplification Method
6	1.137929	3tt99oax	Polyomaviruses KI and WU in Immunocompromised Patients with Respiratory Disease
7	1.132318	cj4zlr3c	Primate-to-Human Retroviral Transmission in Asia
8	1.126702	5gglmx9d	Evaluation of the Seeplex® Meningitis ACE Detection Kit for the Detection of 12 Common Bacterial and Viral Pathogens of Acute Meningitis
9	1.122781	rzzsmuoc	DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data

Unique documents (Pooling) from both the systems are:

't35n7bk9','ngn9x3lc','6f8vyziv','9r62ffew','cxzlmfst','7s5b3lpn','d3cko4j2','rzzsmuoc','6gow0x1v','ps3mnpzq','qfb7074e','zgapjjjw','8ctsa9sd','cpjlzutk','3tt99oax','cj4zlr3c','5gglmx9d'

----------Question 3----------

3) how to assess preparedness of health departments for high-impact public health emergencies?

Query: Assess preparedness of health departments for public health emergencies

```
Do you want to search by query?(y/n)
y

Please enter the query!!
Assess preparedness of public health departmen
```

System 1 (Sparse vector implementation) output

	score	cord_uid	title
0	1.474076	fvfjz7al	Public health preparedness in Alberta: a systems-level study
1	1.417010	05ppugs7	A multidimensional classification of public health activity in Australia
2	1.323412	wgxt36jv	Designing and conducting tabletop exercises to assess public health preparedness for manmade and naturally occurring biological threats
3	1.318034	p34ezktf	Australian public health policy in 2003 – 2004
4	1.315986	tw6wusxe	Local public health workers' perceptions toward responding to an influenza pandemic
5	1.266282	ge5iri3v	Global Public Health Security
6	1.261727	zzc7n84w	Strengthening systems for communicable disease surveillance: creating a laboratory network in Rwanda
7	1.250893	rsnblu4d	Measuring healthcare preparedness: an all-hazards approach
8	1.229302	fite9vs8	Avian influenza outbreak in Turkey through health personnel's views: a qualitative study
9	1.213453	h8xxpbr6	Missing and accounted for: gaps and areas of wealth in the public health review literature

System 2 (Dense vector implementation) output

	score	cord_uid	title
0	1.520514	ge5iri3v	Global Public Health Security
1	1.392174	wgxt36jv	Designing and conducting tabletop exercises to assess public health preparedness for manmade and naturally occurring biological threats
2	1.311986	fite9vs8	Avian influenza outbreak in Turkey through health personnel's views: a qualitative study
3	1.285337	dzzlpz4o	We should not be complacent about our population-based public health response to the first influenza pandemic of the 21(st)century
4	1.282423	jg13scgo	Technical Description of RODS: A Real-time Public Health Surveillance System
5	1.275622	wnuqe66q	Global public goods and the global health agenda: problems, priorities and potential
6	1.269383	0jj9svwj	Globalization and emerging governance modalities
7	1.256540	hf3nytb2	The influenza pandemic preparedness planning tool InfluSim
8	1.255321	m1xf62bf	Community responses to communication campaigns for influenza A (H1N1): a focus group study
9	1.247969	yz2wbpuu	Globalization and Health

Unique documents (Pooling) from both the systems are:

'fvfjz7al','wgxt36jv','05ppugs7','tw6wusxe','ge5iri3v','rsnblu4d','dzzlpz4o','0jj9svwj','hf3nytb2','p34ezktf','zzc7n84w','fite9vs8','h8xxpbr6','jg13scgo','wnuqe66q','m1xf62bf','yz2wbpuu'

Relevance assessments

Relevance criteria: Documents are relevant if they are discussing same methodologies or techniques as needed or trying to tackle similar problems or discussing about the information needed.

All the retrieved documents of different IR systems are briefly gone through and hand picked the documents as relevant if they seem to be catering the information needed.

Query	**ID of relevant documents**
Health surveillance system in real-time	emnln2ix zzc7n84w nge1itgk jg13scgo fvfjz7al w5fxen70 qfb7074e wnuqe66q

multi-versatile microarray vs PCR methods in simultaneous detection of viruses	t35n7bk9 ngn9x3lc 6f8vyziv 9r62ffew cxzlmfst 7s5b3lpn d3cko4j2 rzzsmuoc
Assess public health preparedness for high-impact public health emergencies	fvfjz7al wgxt36jv 05ppugs7 tw6wusxe ge5iri3v rsnblu4d dzzlpz4o 0jj9svwj hf3nytb2

find Relevant/ Irrelevant document from pool

In [87]:
```
#below documents are marked as irrelevant for queries after going through them
q1_irrelevent_docs = ['fite9vs8','05ppugs7','yz2wbpuu','tw6wusxe','ge5iri3v','5hwqdnx1','085v7n6k','m1xf62bf']
q2_irrelevent_docs = ['6gow0x1v','ps3mnpzq','qfb7074e','zgapjjjw','8ctsa9sd','cpjlzutk','3tt99oax','cj4zlr3c','5gglmx9d']
q3_irrelevent_docs = ['p34ezktf','zzc7n84w','fite9vs8','h8xxpbr6','jg13scgo','wnuqe66q','m1xf62bf','yz2wbpuu']

if question == 1:
    irrelevant = q1_irrelevent_docs
elif question == 2:
     irrelevant = q2_irrelevent_docs
else  question == 3:
     irrelevant = q3_irrelevent_docs

       # pooled docs - irrelevant gives relevant documents
pool=set(pool)-set(irrelevant)
```

Evaluation

	System 1		**System 2**	
	P@5	**R@5**	**P@5**	**R@5**
Q1	0.8	0.5	0.4	0.25
Q2	0.6	0.38	0.2	0.12
Q3	0.8	0.44	0.6	0.33

I am using the “ExampleData” class suggested in the week 23 lab with slight modifications to evaluate the performance of the both systems.
As can been seen in the below snippets, I am invoking method “pk_table” in class “ExampleData” with relevant, irrelevant and system output lists and the rank at which system needs to be evaluated. This method would display the precision and recall performance of the both the systems in the tabular format as shown in the following pages.

```
## Precision recall for 1nd system
predict = sparse_predicted
ex = ExampleData()
ex.pk_table(relevant,predict,irrelevant,5)
```

```
## Precision recall for 2nd system
predict = dense_predicted
ex = ExampleData()
ex.pk_table(relevant,predict,irrelevant,5)
```

Query 1: Health surveillance system in real-time

System 2 (Dense vector implementation) output

```
Relevant: ['zzc7n84w', 'emnln2ix', 'wnuqe66q', 'w5fxen70', 'nge1itgk', 'fvfjz7al', 'qfb7074e', 'jg13scgo']

Irrelevant: ['fite9vs8', '05ppugs7', 'yz2wbpuu', 'tw6wusxe', 'ge5iri3v', '5hwqdnx1', '085v7n6k', 'm1xf62bf']

Predicted: ['emnln2ix', 'jg13scgo', 'fite9vs8', 'ge5iri3v', 'yz2wbpuu', '5hwqdnx1', 'qfb7074e', '085v7n6k', 'wnuqe66q', 'm1x
f62bf']
```

k	Result	R@k	P@k
1	emnln2ix	0.12	1.0
2	jg13scgo	0.25	1.0
3	fite9vs8	0.25	0.67
4	ge5iri3v	0.25	0.5
5	yz2wbpuu	0.25	0.4

System 1 (Sparse vector implementation) output

```
Relevant: {'zzc7n84w', 'emnln2ix', 'wnuqe66q', 'w5fxen70', 'nge1itgk', 'fvfjz7al', 'qfb7074e', 'jg13scgo'}

Irrelevant: ['fite9vs8', '05ppugs7', 'yz2wbpuu', 'tw6wusxe', 'ge5iri3v', '5hwqdnx1', '085v7n6k', 'm1xf62bf']

Predicted: ['emnln2ix', 'zzc7n84w', 'nge1itgk', 'jg13scgo', 'fite9vs8', '05ppugs7', 'fvfjz7al', 'yz2wbpuu', 'w5fxen70', 'tw6
wusxe']
```

k	Result	R@k	P@k
1	emnln2ix	0.12	1.0
2	zzc7n84w	0.25	1.0
3	nge1itgk	0.38	1.0
4	jg13scgo	0.5	1.0
5	fite9vs8	0.5	0.8

Query 2: Multi-versatile microarray vs PCR methods in simultaneous detection of viruses

System 2 (Dense vector implementation) output

```
Relevant: ['9r62ffew', 'd3cko4j2', 'ngn9x3lc', 't35n7bk9', '6f8vyziv', '7s5b3lpn', 'cxzlmfst', 'rzzsmuoc']

Irrelevant: ['6gow0x1v', 'ps3mnpzq', 'qfb7074e', 'zgapjjjw', '8ctsa9sd', 'cpjlzutk', '3tt99oax', 'cj4zlr3c', '5gglmx9d']

Predicted: ['qfb7074e', 'ps3mnpzq', 'zgapjjjw', 'd3cko4j2', '8ctsa9sd', 'cpjlzutk', '3tt99oax', 'cj4zlr3c', '5gglmx9d', 'rzz
smuoc']
```

k	Result	R@k	P@k
1	qfb7074e	0	0.0
2	ps3mnpzq	0	0.0
3	zgapjjjw	0	0.0
4	d3cko4j2	0.12	0.25
5	8ctsa9sd	0.12	0.2

System 1 (Sparse vector implementation) output

```
Relevant: {'9r62ffew', 'd3cko4j2', 'ngn9x3lc', 't35n7bk9', '6f8vyziv', '7s5b3lpn', 'cxzlmfst', 'rzzsmuoc'}

Irrelevant: ['6gow0x1v', 'ps3mnpzq', 'qfb7074e', 'zgapjjjw', '8ctsa9sd', 'cpjlzutk', '3tt99oax', 'cj4zlr3c', '5gglmx9d']

Predicted: ['t35n7bk9', 'ngn9x3lc', '6gow0x1v', '6f8vyziv', 'ps3mnpzq', 'qfb7074e', 'zgapjjjw', '9r62ffew', 'cxzlmfst', '7s5
b3lpn']
```

k	Result	R@k	P@k
1	t35n7bk9	0.12	1.0
2	ngn9x3lc	0.25	1.0
3	6gow0x1v	0.25	0.67
4	6f8vyziv	0.38	0.75
5	ps3mnpzq	0.38	0.6

Query 3: Assess preparedness of health departments for public health emergencies

System 2 (Dense vector implementation) output

```
Relevant: ['0jj9svwj', 'tw6wusxe', 'ge5iri3v', 'rsnblu4d', 'fvfjz7al', 'hf3nytb2', '05ppugs7', 'dzzlpz4o', 'wgxt36jv']

Irrelevant: ['p34ezktf', 'zzc7n84w', 'fite9vs8', 'h8xxpbr6', 'jg13scgo', 'wnuqe66q', 'm1xf62bf', 'yz2wbpuu']

Predicted: ['ge5iri3v', 'wgxt36jv', 'fite9vs8', 'dzzlpz4o', 'jg13scgo', 'wnuqe66q', '0jj9svwj', 'hf3nytb2', 'm1xf62bf', 'yz2
wbpuu']
```

k	Result	R@k	P@k
1	ge5iri3v	0.11	1.0
2	wgxt36jv	0.22	1.0
3	fite9vs8	0.22	0.67
4	dzzlpz4o	0.33	0.75
5	jg13scgo	0.33	0.6

System 1 (Sparse vector implementation) output

```
Relevant: {'0jj9svwj', 'tw6wusxe', 'ge5iri3v', 'rsnblu4d', 'fvfjz7al', 'hf3nytb2', '05ppugs7', 'dzzlpz4o', 'wgxt36jv'}

Irrelevant: ['p34ezktf', 'zzc7n84w', 'fite9vs8', 'h8xxpbr6', 'jg13scgo', 'wnuqe66q', 'm1xf62bf', 'yz2wbpuu']

Predicted: ['fvfjz7al', '05ppugs7', 'wgxt36jv', 'p34ezktf', 'tw6wusxe', 'ge5iri3v', 'zzc7n84w', 'rsnblu4d', 'fite9vs8', 'h8x
xpbr6']
```

k	Result	R@k	P@k
1	fvfjz7al	0.11	1.0
2	05ppugs7	0.22	1.0
3	wgxt36jv	0.33	1.0
4	p34ezktf	0.33	0.75
5	tw6wusxe	0.44	0.8

Here, In Sparse vector implementation (system 1), even though it is of a max length 1024 , because of its sparse implementation (it would have TF-IDF values of only terms in the document) it is capable of handling a large number of features in corpus , while dense vector implementation (system 2) can only handle maximum of 2048 important features owning to vector size limit. As the system corpus gets bigger, feature size also grows, and sparse vector implementation tends to work better than dense implementation.
In the similar lines of our expectation, system 1(Sparse implementation) outperformed system 2 (dense implementation) for all the 3 queries.
Here, in system 1 around 9000 features are selected based on document frequency while only 2048 are chosen from system 2. This makes both the system different in implementation and their performance is compared using precision and recall at rank 5.