



PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100 Ft. Road, BSK III Stage, Bengaluru – 560 085
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SESSION: AUG-DEC 2020

Course Title: Algorithms for Information Retrieval		
Course code: UE17CS412		
Semester : VII sem	Section:G	Team Id:3
SRN:PES1201701581	Name:RAVICHANDRA GIDD	
SRN:PES1201701635	Name:AMBARISH D Y	
SRN:PES1201701626	Name:VENKATESH K	
SRN:PES1201700239	Name:PAVAN MITRA	

ASSIGNMENT REPORT

Problem Statement

Search Engine based on a Vector-Space Model and a comparative evaluation alongside ElasticSearch Engine.

Description

Our 4 stage comparative study includes:

1. Pre-processing:

- **We use the famed NLTK module to implement the following pre-processing techniques**

- **Tokenization**

It is the transformation of a bunch of meaningful data into a string of abstract characters that can be mapped to, later on.

1. This helps in bringing in the abstraction of **term-id** which can later be used **not only** for more efficient querying **but also**
2. Correction measures such as edit distance measurement for the sake of error correction.

- **Lemmatization**

The process of reducing a term to its “root” word

Ex: standardise -> standard

Another option could have been stemming, where the same outcome is done using certain rules such as “remove -ish” etc. but sometimes, especially when there are a lot of proper nouns involved, it may lead to semantic loss.

Ex: “fetish” -> “fet”

We prefer lemmatization over stemming here since stemming may lead to loss of meaningful queries, owing to journalism usually dealing with tons of “**people, places and things!**”

We use NLTK’s built-in **WordNet lemmatizer** that is sufficiently large and tailored to the English language

- **Case Folding**

Since case (lower or upper) does not make differ semantically for searching, we “fold” all root query words to a single type (here, all lower case)

2. B-tree implementation:

We use OOBTree to build tree-based dictionaries and posting lists that are mapped to each entry

- We use a unique <term-frequency, doc-id> structure to index the inverted indices
- This structure can be put to use to find similarity measures later on

3. <Op>Closest word:

We used edit distance(Levenshtein distance) to obtain the closest word in the vocabulary.

It makes use of Dynamic Programming, a paradigm that invests in **overlapping subproblems** with repeatable measurements/operations that build on each other.

$$\text{lev}(a, b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \text{lev}(\text{tail}(a), \text{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \text{lev}(\text{tail}(a), b) \\ \text{lev}(a, \text{tail}(b)) \\ \text{lev}(\text{tail}(a), \text{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

4. Ranking process

After applying similar pre-processing measures to the query as well, we begin the 4 stage ranking process

- **Create a Term frequency vector**
Refer to the term frequency using the “files” OOBTree and keep track of each query term’s tf score.
- **Created a inverse frequency vector**
Refer to the posting lists and use a tf-idf matrix to similarly obtain idf scores.
- **Create a tf-idf vector**
Using the “**document-at-a-time**” strategy, we obtain tf-idf score vector by multiplication.
- **Finally cosine similarity for retrieving top k results**

We look at a simple Euclidean distance to calculate similarity between document vectors and the query vector, but this leads to some blatant dissimilarities.

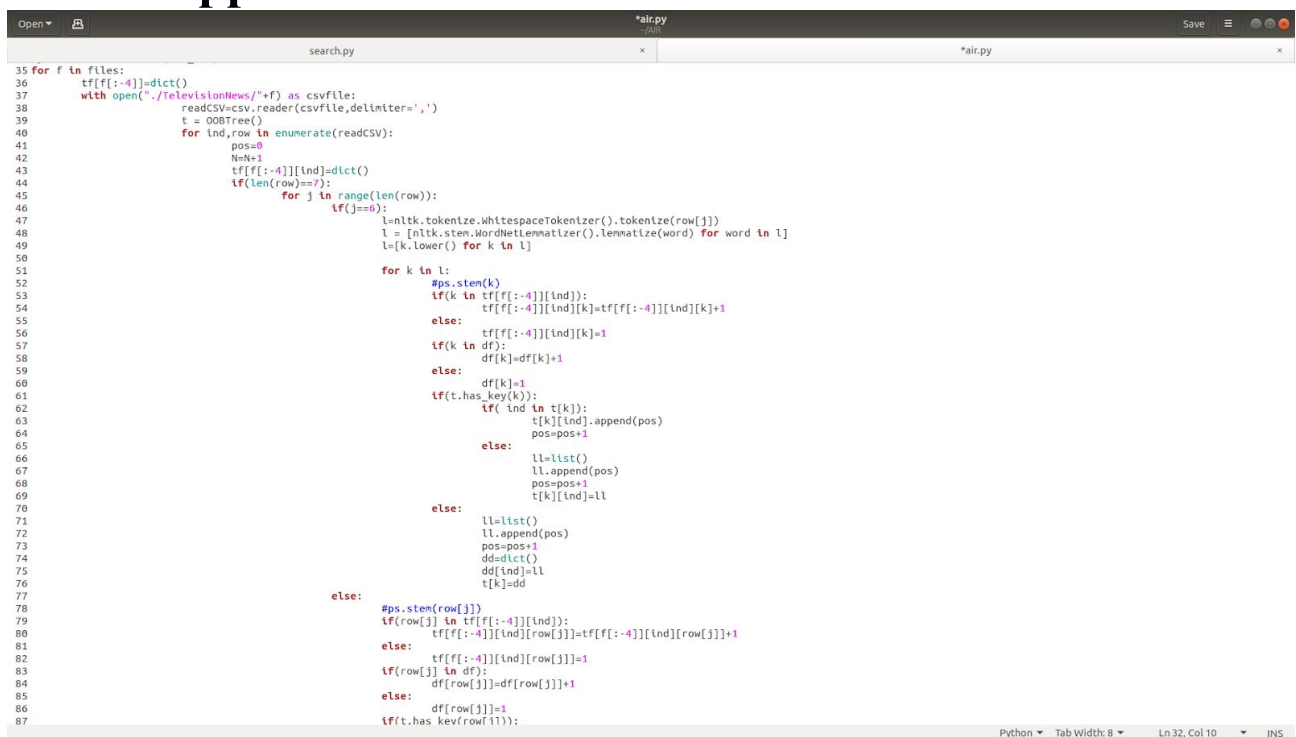
Since the document tf-idf vectors exist in the same dimensions of the query tf-idf vectors, a simple dot product helps obtain a vector-based similarity measure for the same.

Now, a normalised dot-product of these vectors is equivalent to a cosine similarity(since cosine theta decreases with theta, and the closer the query to the document, the lower the angle, and hence larger the cosine measure)

$$\text{similarity}(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

This gives a convenient metric between [0,1] that can be ranked for all documents and returned by orders of top-k i.e. top-10, top-20 etc.

Code Snippets



```

35 for f in files:
36     tfidf[-4]=dict()
37     with open("./TelevisionNews/"+f) as csvfile:
38         readCSV=csv.reader(csvfile,delimiter=',')
39         t = OOBTree()
40         for ind,row in enumerate(readCSV):
41             pos=0
42             N=N+1
43             tfidf[-4][ind]=dict()
44             if(len(row)==7):
45                 for j in range(len(row)):
46                     if(j==0):
47                         l=nlTK.tokenize(whitespaceTokenizer().tokenize(row[j]))
48                         l = [nlTK.stem.WordNetLemmatizer().lemmatize(word) for word in l]
49                         l=[k.lower() for k in l]
50
51                     for k in l:
52                         #ps.stem(k)
53                         if(k in tfidf[-4][ind]):
54                             tfidf[-4][ind][k]=tfidf[-4][ind][k]+1
55                         else:
56                             tfidf[-4][ind][k]=1
57                         if(k in df):
58                             df[k]=df[k]+1
59                         else:
60                             df[k]=1
61             if(t.has_key(k)):
62                 if( ind in t[k]):
63                     t[k][ind].append(pos)
64                     pos=pos+1
65                 else:
66                     ll=lst()
67                     ll.append(pos)
68                     pos=pos+1
69                     t[k][ind]=ll
70             else:
71                 ll=lst()
72                 ll.append(pos)
73                 pos=pos+1
74                 dd=dict()
75                 dd[ind]=ll
76                 t[k]=dd
77
78     else:
79         #ps.stem(row[j])
80         if(row[j] in tfidf[-4][ind]):
81             tfidf[-4][ind][row[j]]=tfidf[-4][ind][row[j]]+1
82         else:
83             tfidf[-4][ind][row[j]]=1
84         if(row[j] in df):
85             df[row[j]]=df[row[j]]+1
86         else:
87             df[row[j]]=1
88         if(t.has_key(row[j])):

```

```
Open  search.py  *air.py
74                                     dd=dict()
75                                     dd[ind]=ll
76                                     t[k]=dd
77
78                                     else:
79                                     #ps.sten(row[j])
80                                     if(row[j] in tf[f[:4]][ind]):
81                                     tf[f[:4]][ind][row[j]]=tf[f[:4]][ind][row[j]]+1
82
83                                     else:
84                                     tf[f[:4]][ind][row[j]]=1
85                                     if(row[j] in df):
86                                     df[row[j]]=df[row[j]]+1
87
88                                     else:
89                                     df[row[j]]=1
90                                     if(t.has_key(row[j])):
91                                     if ind in t[row[j]]:
92                                     t[row[j]][ind].append(pos)
93                                     pos=pos+1
94
95                                     else:
96                                     ll=list()
97                                     ll.append(pos)
98                                     pos=pos+1
99                                     t[row[j]][ind]=ll
100
101                                     else:
102                                     ll=list()
103                                     ll.append(pos)
104                                     pos=pos+1
105                                     dd=dict()
106                                     dd[ind]=ll
107                                     t[row[j]]=dd
108
109                                     afile = open(r"./"+f[:4]+".pkl", 'wb')
110                                     pickle.dump(t, afile)
111                                     afile.close()
112
113                                     df = open(r"./"+d+"df.pkl", 'wb')
114                                     pickle.dump(df, afile)
115                                     df.close()
116
117                                     end=time.time()
118                                     df["N"]=N
119                                     afile = open(r"./"+t+"tf.pkl", 'wb')
120                                     pickle.dump(tf, afile)
121                                     afile.close()
122
123                                     df = open(r"./"+d+"df.pkl", 'wb')
124                                     pickle.dump(df, afile)
125                                     df.close()
126
127                                     file2 = open(r"df.pkl", 'rb')
128                                     new_d = pickle.load(file2)
129                                     file2.close()
130
131                                     file1 = open(r"df.pkl", 'rb')
132                                     df = pickle.load(file1)
133                                     file1.close()
134
135                                     file2 = open(r"tf.pkl", 'rb')
136                                     tf = pickle.load(file2)
137                                     file2.close()
138
139                                     Python  Tab Width: 8  Ln 32, Col 10  INS
```

```
Open  search.py  *air.py
128
129 def editDistDP(str1):
130     if(str1 in df):
131         return str1
132     maxi=1000000000
133     res=str1
134     m=len(str1)
135     for str2 in df:
136         n=len(str2)
137         dp = [[0 for x in range(n + 1)] for x in range(m + 1)]
138         for i in range(n + 1):
139             for j in range(m + 1):
140                 if(i == 0):
141                     dp[i][j] = j
142                 elif(j==0):
143                     dp[i][j] = i
144                 elif(str1[i-1] == str2[j-1]):
145                     dp[i][j] = dp[i-1][j-1]
146                 else:
147                     dp[i][j] = 1 + min(dp[i][j-1],dp[i-1][j],dp[i-1][j-1])
148             if(dp[n][n]<maxi):
149                 res=str2
150                 maxi=dp[n][n]
151             if(maxi==0):
152                 return res;
153     return res
154
155 l=list()
156 #print(new_d["warning"],new_d["greenhouse"],new_d["gas"])
157 #query="greenhouses gasses and global warming "
158 query=input("ENTER YOUR QUERY\n")
159 start=time.time()
160 query=nltk.tokenize.WhitespaceTokenizer().tokenize(query)
161 query=[nltk.stem.WordNetLemmatizer().lemmatize(word) for word in query]
162
163 query=[i.lower() for i in query]
164 """for i,q in enumerate(query):
165     query[i]=editDistDP(q)"""
166 ql=[query.count(i)/len(query) for i in query]
167 for i in range(len(ql)):
168     if(query[i] in df):
169         ql[i]=ql[i]*math.log(df["N"]/df[query[i]])
170     else:
171         ql[i]=0
172
173 a=0
174 for i in ql:
175     a=a+i*i
176 a=pow(a,0.5)
177
178 for f in files:
179     for i,row in enumerate(tf[f[:4]]):
180         s=ntf(tf[f[:4]][i])
181         #print(s)
182         ll=[0]*len(ql)
183         for ni,q in enumerate(query):
184             ll[ni]=ql[ni]*s
185         #print(ll)
186         #print(s)
187         #print(ll)
188
189     Python  Tab Width: 8  Ln 32, Col 10  INS
```

```
Open search.py x *air.py Save
164 query[l]=editedLSDP(q)""
165 ql=query.count(l)/len(query) for l in query
166 for i in range(len(ql)):
167     if(query[l] in df):
168         ql[i]=ql[i]*math.log(df["N"]/df[query[l]])
169     else:
170         ql[i]=0
171 a=0
172 for i in ql:
173     a=a+i*i
174 a=pow(a,0.5)
175 for f in files:
176     for l,row in enumerate(tf[::4]):
177         s=ntf(tf[::4])[l]
178         #print(s)
179         ll=[0]*len(ql)
180         for ql,q in enumerate(query):
181             if(q in tf[::4]):
182                 ll[q]=tf[::4][l][q]/s
183                 ll[q]*=math.log(df["N"]/df[q])
184
185         b=0
186         for j in ll:
187             b=b+j*j
188         b=pow(b,0.5)
189         if(b!=0):
190             s=0
191             for j in range(len(ql)):
192                 s+=ql[j]*ll[j]
193             s=s/a
194             s=round(s,10)
195             res=llst()
196             res.append(f)
197             res.append(l)
198             res.append(s)
199             l.append(res)
200 l.sort(key=lambda x:x[2])
201 l.reverse()
202
203 for i in range(10):
204     if(i>=len(l)):
205         print("no entries")
206         break
207     with open("./televisionNews/"+l[i][0]) as csvfile:
208         readCSV=csv.reader(csvfile,delimiter=',')
209         row=[l for l in readCSV]
210         print(l[i])
211         print("doc:",l[i][0],"trows:",l[i][1],"tURL:",row[l[i][1]][0],"MatchDateTime:",row[l[i][1]][1],"Station:",row[l[i][1]][2],"tShow:",row[l[i][1]][3],"tIASHowID:",row[l[i][1]][4],"tIAPreviewThumb:",row[l[i][1]][5],"tSnippet:",row[l[i][1]][6],"n\n")
212
213 end=time.time()
214 print("Time taken: ",end-start)
215
```

Output Screenshots

```
amb@ambi-Inspiron-5567: ~/AIR
File Edit View Search Terminal Help
amb@ambi-Inspiron-5567:~/AIR$ python3 air.py
class 'Itst'> 417
greENTER YOUR QUERY
greenhouse gas and global warming
doc: BBCNEWS.201908.csv row: 379 URL: https://archive.org/details/BBCNEWS_20190816_033000_HARDtalk#start/82/end/117 MatchDateTme : 8/16/2019 3:31:37 Station : BBCNEWS Show :
HARDtalk IAShowID: BBCNEWS_20190816_033000_HARDtalk IAPreviewThumb: https://archive.org/download/BBCNEWS_20190816_033000_HARDtalk/BBCNEWS_20190816_033000
HARDtalk.000059.jpg Snippet: global greenhouse gas emissions are still rising. the data suggest the planet is warming at an alarming rate. what to do about it: well, my guest today is roger hallan, co
-founder of extinction rebellion, a movement dedicated to mass resistance and civil disobedience.
doc: BBCNEWS.201908.csv row: 70 URL: https://archive.org/details/BBCNEWS_20190808_170000_BBC_News_at_Six#start/203/end/238 MatchDateTme : 8/8/2019 17:03:38 Station : BBCNEWS S
how: BBC News at Six IAShowID: BBCNEWS_20190808_170000_BBC_News_at_Six IAPreviewThumb: https://archive.org/download/BBCNEWS_20190808_170000_BBC_News_at_Six
thumbs/BBCNEWS_20190808_170000_BBC_News_at_Six.000178.jpg Snippet: so what we choose to put an our plate helps define what the carbon footprint on the level of greenhouse gases in the atmosphere, so
that choice between broccoli and ribs on your plate actually has a real link to the level of global warming that we're likely to see.
doc: FOXNEWS.201901.csv row: 80 URL: https://archive.org/details/FOXNEWSW_20190115_050000_Tucker_Carson_Tonight#start/3296/end/3331 MatchDateTme : 1/15/2019 5:55:11 Station : FO
XNEWS Show: Tucker Carlson Tonight IAShowID: FOXNEWSW_20190115_050000_Tucker_Carson_Tonight IAPreviewThumb: https://archive.org/download/FOXNEWSW_20190115_050000_Tucker_Carson_Tonight/FOXNEWS
W_20190115_050000_Tucker_Carson_Tonight.thumbs/FOXNEWSW_20190115_050000_Tucker_Carson_Tonight.003297.jpg Snippet: admitted that protesters have destroyed about 60% of the speed cameras in all of fr
ance. you might think global warming and speed cameras are not related, but they are. both our weapons are leaders use to bully the population. left to impose gas taxes on everyone else while flying
doc: BBCNEWS.201906.csv row: 5 URL: https://archive.org/details/BBCNEWS_20190629_050000_Breakfast#start/2939/end/2974 MatchDateTme : 6/29/2019 5:49:14 Station : BBCNEWS Show: Breakf
ast IAShowID: BBCNEWS_20190629_050000_Breakfast IAPreviewThumb: https://archive.org/download/BBCNEWS_20190629_050000_Breakfast/BBCNEWS_20190629_050000_Brea
kfast.002937.jpg Snippet: because we were emitting carbon dioxide and methane into the air at a runaway rate, but what i hadn't fully understood is this - simply reducing greenhouse gas emissions w
ill not bring global warming under control.
doc: BBCNEWS.201906.csv row: 4 URL: https://archive.org/details/BBCNEWS_20190630_033000_Click#start/119/end/154 MatchDateTme : 6/30/2019 3:32:14 Station : BBCNEWS Show: Click IASH
owID: BBCNEWS_20190630_033000_Click IAPreviewThumb: https://archive.org/download/BBCNEWS_20190630_033000_Click/BBCNEWS_20190630_033000_Click.thumbs/BBCNEWS_20190630_033000_Click.000118.jpg Snp
pet: dioxide and methane into the air at a runaway rate, but what i hadn't fully understood is this - simply reducing greenhouse gas emissions will not bring global warming under control.
doc: FOXNEWS.201203.csv row: 70 URL: https://archive.org/details/FOXNEWSW_20120314_010000_Hannity#start/1786/end/1821 MatchDateTme : 3/14/2012 1:30:01 Station : FOXNEWS Show :
Hannity IAShowID: FOXNEWSW_20120314_010000_Hannity IAPreviewThumb: https://archive.org/download/FOXNEWSW_20120314_010000_Hannity/FOXNEWSW_20120314_010000_Hannity.thumbs/FOXNEWSW_20120314_0100
00_Hannity.001765.jpg Snippet: fuels and get the gasoline prices up to $9 dis. it hark back to those days? i think it does. we have moved on since then. global warning is not the headline that it was. i
instead, we want a rational policy of getting gas prices under control, a rational energy
doc: BBCNEWS.201912.csv row: 619 URL: https://archive.org/details/BBCNEWS_20191224_003000_Review_2019#start/717/end/752 MatchDateTme : 12/24/2019 8:42:12 Station : BBCNEWS Show :
Review 2019 IAShowID: BBCNEWS_20191224_003000_Review_2019 IAPreviewThumb: https://archive.org/download/BBCNEWS_20191224_003000_Review_2019.thumbs/BBCNEWS_20191224
_003000_Review_2019.000718.jpg Snippet: that even sharp reductions in emissions won't, on their own, be enough to halt irreversible damage. so researchers here at cambridge university are looking at idea
s, however crazy, to try and take global warming gases out of the atmosphere
doc: BBCNEWS.201912.csv row: 409 URL: https://archive.org/details/BBCNEWS_20191231_143000_Review_2019#start/771/end/886 MatchDateTme : 12/31/2019 14:43:06 Station : BBCNEWS Show :
Review 2019 IAShowID: BBCNEWS_20191231_143000_Review_2019 IAPreviewThumb: https://archive.org/download/BBCNEWS_20191231_143000_Review_2019.thumbs/BBCNEWS_20191231
_143000_Review_2019.000779.jpg Snippet: global warming gases out of the atmosphere and to actually repair the earth's climate.
doc: BBCNEWS.201912.csv row: 123 URL: https://archive.org/details/BBCNEWS_20191208_040000_BBC_News#start/1263/end/1298 MatchDateTme : 12/8/2019 4:21:18 Station : BBCNEWS Show :
BBC News IAShowID: BBCNEWS_20191208_040000_BBC_News IAPreviewThumb: https://archive.org/download/BBCNEWS_20191208_040000_BBC_News/BBCNEWS_20191208_040000_BBC_News.thumbs/BBCNEWS_20191208_040000
_BBC_News.001258.jpg Snippet: greenhouse gas emissions. the findings suggest larger fish are being affected both by global warming and chemical pollution. tens of thousands of people are taking part in
the world's big sleep out to raise awareness about homelessness. an estimated 100 million people across the globe don't have a home.
doc: BBCNEWS.201912.csv row: 92 URL: https://archive.org/details/BBCNEWS_20191208_050000_BBC_News#start/351/end/386 MatchDateTme : 12/8/2019 5:06:06 Station : BBCNEWS Show :
BBC News IAShowID: BBCNEWS_20191208_050000_BBC_News IAPreviewThumb: https://archive.org/download/BBCNEWS_20191208_050000_BBC_News/BBCNEWS_20191208_050000_BBC_News.thumbs/BBCNEWS_20191208_050000
_BBC_News.000327.jpg Snippet: which shows the earth's oceans are becoming starved of oxygen is further evidence that urgent action is needed to cut greenhouse gas emissions. the findings suggest larger
fish are being affected both by global warming and chemical pollution. a state funeral has been
time taken: 3.6258487701410010
amb@ambi-Inspiron-5567:~/AIR$
amb@ambi-Inspiron-5567:~/AIR$
```



```
File Edit View Search Terminal Help
ambig@ambi-inspiron-5567: ~/AIR$
ambig@ambi-inspiron-5567:~/AIR$ python3 air.py
<class 'list'> 417
ENTER YOUR QUERY
barack obama
doc: MSNBC.201409.csv row: 203 URL: https://archive.org/details/MSNBCW_20140916_030000_All_In_With_Chris_Hayes#start/3359/end/3394 MatchDateTme : 9/16/2014 3:56:14 Station : MSNBC S
how: All In With Chris Hayes IAShowID: MSNBCW_20140916_030000_All_In_With_Chris_Hayes IAPreviewThumb: https://archive.org/download/MSNBCW_20140916_030000_All_In_With_Chris_Hayes/thumbs/MSNBCW_20140916_030000_All_In_With_Chris_Hayes.003345.jpg Snippet: country in mexico. what were the politics like there. were you getting beaten up in the san e ways that barack obama on the epa is getting beaten up? in that sense, mexico is suffering a lot from climate change. right now today the hurricane in
doc: MSNBC.201409.csv row: 128 URL: https://archive.org/details/MSNBCW_20140922_230000_Hardball_With_Chris_Matthews#start/3147/end/3182 MatchDateTme : 9/22/2014 23:52:42 Station : MSNBC S
how: Hardball With Chris Matthews IAShowID: MSNBCW_20140922_230000_Hardball_With_Chris_Matthews IAPreviewThumb: https://archive.org/download/MSNBCW_20140922_230000_Hardball_With_Chris_Matthews/MSNBCW_20140922_230000_Hardball_With_Chris_Matthews.thumbs/MSNBCW_20140922_230000_Hardball_With_Chris_Matthews.003135.jpg Snippet: climate change was not a partisan issue. and then, all of the sudden, in th e barack obama age, it became if you're a republican you're against it, if you're a democrat, you're for it.
doc: MSNBC.201802.csv row: 30 URL: https://archive.org/details/MSNBCW_20180223_010000_All_In_With_Chris_Hayes#start/3368/end/3403 MatchDateTme : 2/23/2018 1:56:23 Station : MSNBC S
how: All In With Chris Hayes IAShowID: MSNBCW_20180223_010000_All_In_With_Chris_Hayes IAPreviewThumb: https://archive.org/download/MSNBCW_20180223_010000_All_In_With_Chris_Hayes/thumbs/MSNBCW_20180223_010000_All_In_With_Chris_Hayes.003357.jpg Snippet: and that way even if you don't own a gun, even if you don't like guns, if you think of your self as a conservative this is one of those hot button issues you must believe in. just like you believe in climate change denial, just like you believe barack obama was not born in the united states so
doc: FOXNEWS.201707.csv row: 110 URL: https://archive.org/details/FOXNEWSW_20170703_090000_Fox_and_Friends_First#start/3111/end/3146 MatchDateTme : 7/3/2017 9:52:00 Station : FOX NEWS S
Show: Fox and Friends First IAShowID: FOXNEWSW_20170703_090000_Fox_and_Friends_First IAPreviewThumb: https://archive.org/download/FOXNEWSW_20170703_090000_Fox_and_Friends_First/thumbs/FOXNEWSW_20170703_090000_Fox_and_Friends_First.003087.jpg Snippet: democratic strategists, good morning to you both. carrie, we will start with you, w hat do you make of this? the paris climate change agreement, barack obama did not have authority to get into it.
doc: CNN.201703.csv row: 145 URL: https://archive.org/details/CNNW_20170309_070000_CNN_Tonight_With_Don_Lemon#start/1220/end/1255 MatchDateTme : 3/9/2017 7:20:35 Station : CNN Show S
Show: CNN Tonight With Don Lemon IAShowID: CNNW_20170309_070000_CNN_Tonight_With_Don_Lemon IAPreviewThumb: https://archive.org/download/CNNW_20170309_070000_CNN_Tonight_With_Don_Lemon/thumbs/CNNW_20170309_070000_CNN_Tonight_With_Don_Lemon.thumbs/CNNW_20170309_070000_CNN_Tonight_With_Don_Lemon.001197.jpg Snippet: president to have a good relationship with their predecessor? it's helpful, because barack obama knows a lot on what is going on right now, the war on terror, the middle east, climate change, donald trump could tap him as a resource, i mean after
doc: CNN.201703.csv row: 140 URL: https://archive.org/details/CNNW_20170328_180000_CNN_Newsroom_With_Brooke_Baldwin#start/2408/end/2443 MatchDateTme : 3/28/2017 18:40:23 Station : CNN Show S
Show: CNN Newsroom With Brooke Baldwin IAShowID: CNNW_20170328_180000_CNN_Newsroom_With_Brooke_Baldwin IAPreviewThumb: https://archive.org/download/CNNW_20170328_180000_CNN_Newsroom_With_Brooke_Baldwin/thumbs/CNNW_20170328_180000_CNN_Newsroom_With_Brooke_Baldwin.002397.jpg Snippet: what president trump has done with his own executiv e orders, we are going to see about half a dozen of these executive orders signed by president barack obama that directed the federal government to prepare for climate change. that's now gone.
doc: FOXNEWS.201906.csv row: 269 URL: https://archive.org/details/FOXNEWSW_20190628_100000_Fox_Friends#start/8265/end/8300 MatchDateTme : 6/28/2019 12:18:00 Station : FOXNEWS S
Show: FOX Friends IAShowID: FOXNEWSW_20190628_100000_Fox_Friends IAPreviewThumb: https://archive.org/download/FOXNEWSW_20190628_100000_Fox_Friends/thumbs/FOXNEWSW_20190628_100000_Fox_Friends.thu mbs/FOXNEWSW_20190628_100000_Fox_Friends.008248.jpg Snippet: what we did. build on the obama-biden situation. so under estimated what barack obama did, first man to bring together entire world, 196 na tions to commit to a deal with climate change. brian: wow, joe Biden says he doesn't want help from
doc: FOXNEWS.201304.csv row: 37 URL: https://archive.org/details/FOXNEWSW_20130407_030000_The_Journal_Editorial_Report#start/1115/end/1150 MatchDateTme : 4/7/2013 3:18:50 Station : FOXNEWS S
Show: The Journal Editorial Report IAShowID: FOXNEWSW_20130407_030000_The_Journal_Editorial_Report IAPreviewThumb: https://archive.org/download/FOXNEWSW_20130407_030000_The_Jo urnal_Editorial_Report/FOXNEWSW_20130407_030000_The_Journal_Editorial_Report.thumbs/FOXNEWSW_20130407_030000_The_Journal_Editorial_Report.001095.jpg Snippet: ending with, greg, get out of my bedroom. b ut, anyway, the stories are their teachers talk about global warming as if it's fact and talk about barack obama and they talk about how mitt romney is a joke
doc: FOXNEWS.201304.csv row: 36 URL: https://archive.org/details/FOXNEWSW_20130407_060000_Red_Eye#start/1113/end/1148 MatchDateTme : 4/7/2013 6:18:48 Station : FOXNEWS S
Show: Red Eye IAShowID: FOXNEWSW_20130407_060000_Red_Eye IAPreviewThumb: https://archive.org/download/FOXNEWSW_20130407_060000_Red_Eye/FOXNEWSW_20130407_060000_Red_Eye.thumbs/FOXNEWSW_20130407_0600 00_Red_Eye.001095.jpg Snippet: like stories that are often ending with, greg, get out of my bedroom. but, anyway, the stories are their teachers talk about global warming as if it's fact and talk about barack obama and they talk
doc: CNN.201207.csv row: 53 URL: https://archive.org/details/CNNW_20120702_130000_CNN_Newsroom#start/4757/end/4792 MatchDateTme : 7/2/2012 14:19:32 Station : CNN Show: CNN Newsroom I AShowID: CNNW_20120702_130000_CNN_Newsroom IAPreviewThumb: https://archive.org/download/CNNW_20120702_130000_CNN_Newsroom/thumbs/CNNW_20120702_130000_CNN_Newsroom.00 4740.jpg Snippet: and energy distribution and storage? wouldn't you want that, but the people who are politicizing this issue they seem to be winning because not much is being done on the issue of climate change even though president barack obama promised that, you know, back in the day, 2008.
time taken: 3.037285804748535
```

Interpretation of efficiency

Matches in the query	Time taken by custom search engine	Time taken by elastic search engine	Number of matches in top k	NAL score
greenhouse gas and global warming	3.6258	0.0109 s	12/18	1.13
barack obama	3.03728	0.0110 s	29/39	2.47
carbon dioxide	3.10845	0.0571 s	19/40	1.58

We observe that F scores are in the ranges of **[0.09, 0.15]**
This is insufficient as a comparative measure

Custom Metric:

We notice that as word count increases, the proportion of top-ks decreases, since ElasticSearch has complex context-based implementations, hence we penalise the lower word error more.

Normalised average log-depreciation top-k comparison (NAL score):

$$\text{NAL} = \frac{n(\text{Custom U Es})}{\log(N)}$$

, where N is the number of terms in query

It is also observed that the performance of professional IR systems is substantially higher for proper nouns (of the same length) which further leads us to believe they use complex cotext0based algorithms.

This leads to an idea of finding global reference index of a specific term and incorporating it into our measure as well (G-NAL score)

Learning Outcome

- Appreciating the nuances of a complex search engine
- Implementing tf.idf scoring, vector space ranking from scratch helps keep track of various data structures involved and their interconnectedness.
- Observed the importance of metrics and how specific measures can be used to seed insights from and tell stories about IR systems.

Name and Signature of the Faculty