[Eoin Quinn](mailto:eoin.quinn55@mail.dcu.ie)-21356173

Ronan Kelly-21374881

**CA4009 Lab 1**

| **Term** | **Doc Freq** | **Coll Freq** |
| --- | --- | --- |
| time | 360476 | 684393 |
| page | 347260 | 383440 |
| year | 266010 | 827942 |
| financi | 247956 | 344316 |
| countri | 229732 | 417625 |

This is a list of the top 5 most common terms in the newspaper. As we can see the LA Times

is a financial newspaper as the terms are clearly relating to finance and industry. The list is sorted based on the number of different documents the term shows up in.

Terms with high document frequency can be useful for finding unique documents, filtering less relevant information while conducting information retrieval.

As for terms with high collection frequency, this can be useful when conducting information retrieval, when trying to find patterns between documents this will be much preferred.

As we can see the word ‘time’ shows up in the most amount of documents, however, this isn't the most useful for information retrieval as it is just a stemmed version of the newspaper title.

The term ‘year’ shows up the most amount of times between the documents, this could be useful for finding trends or patterns between the different news articles.

A term like ‘financial’ which shows up a lot of times in lots of documents could be useful for information retrieval if we are looking for articles specifically related to finance, whereas a term like ‘countries’ is a bit broad when looking at information retrieval.

**Section 5**

I entered “Sports”, as a short search query.. This returned me a list of documents which had a high frequency count of the word sport. As the lab instructions described, the ID within each TREC collection for each document were displayed. And a list of the most frequent features (words) in the document, along with the number of times that word occurred.

The documents returned included letters from LA times’ readers to editors, daily reports etc. The word sport appears relatively frequently in all the documents returned. For example, the full document for the top document returned was:

*“LA030489-0046 26201 March 4, 1989, Saturday, Home Edition Sports; Part 3; Page 3; Column 1; Sports Desk 41 words THE SPORTS NEWS GETS LOST AMID THE GLOSS If Scott Ostler really needs to keep tabs on sports news, I suggest he skip Sports Illustrated and invest in a magazine without glossy pages and lots of advertisements. You get more sports and less skin that way. CHRIS KRECHTER Santa Ana Letter to the Editor.”*

The snippet for this document that appeared upon searching “Sports” was displayed as the following:

*“*[*LA030489-0046*](https://clueweb.adaptcentre.ie/IRModelGenerator/DocumentViewer?id=478161)*Term Freq Vector:  
sport:6, 15.521645 new:2, 3.8393402 page:2, 1.5209209 advertis:1, 46.43121 amid:1, 65.99463 ana:1, 97.915276 chri:1, 57.949856 column:1, 3.864511 desk:1, 4.628067 edit:1, 4.1606994 editor:1, 23.728771 gloss:1, 916.9358 glossi:1, 992.7726 home:1, 4.517582 illustr:1, 83.29207 invest:1, 7.955939 krechter:1, 528155.0 letter:1, 14.380173 lost:1, 14.927645 lot:1, 14.825403 magazin:1, 53.543694 march:1, 7.452974 ostler:1, 8518.629 part:1, 2.3533697 santa:1, 30.543314 saturdai:1, 14.082255 scott:1, 77.58998 skin:1, 161.6636 skip:1, 415.54288 suggest:1, 11.045801 tab:1, 677.12177 word:1, 3.476968  
LA030489-0046 26201 March 4, 1989, Saturday, Home Edition* ***Sports****; Part 3; Page 3; Column 1;* ***Sports*** *Desk 41 words THE* ***SPORTS*** *NEWS GETS LOST AMID THE GLOSS If Scott Ostler really needs to keep tabs on* ***sports*** *news, I suggest he skip* ***Sports*** *Illustrated and invest in a magazine without glossy pages and lots of advertisements. You get more* ***sports*** *and less skin that way. CHRIS KRECHTER Santa Ana Letter”*

This complaint letter is related to the search query as the complaint is regarding sport.and the word sport occurs frequently. My search query also is related to the content in the snippets as the word sports occurs frequently in the snippets of the documents returned. As for the lower ranked documents returned from the search query. These tend to be longer documents, likely due to the way TF-IDF operates. Since the documents are longer and contain more words,, the weight on the word “Sport” each time it occurs is lower. For example the following snippet was returned ranked as the 17th document.

*“*[*LA121389-0036*](https://clueweb.adaptcentre.ie/IRModelGenerator/DocumentViewer?id=521428)

*Term Freq Vector:*

*sport:12, 15.521645 athlet:7, 63.4878 east:4, 9.873901 offici:4, 4.8586082 feder:3, 7.0993347 germani:3, 16.240929 leadership:3, 26.693369 resign:3, 33.95622 attent:2, 20.71196 call:2, 5.006873 column:2, 3.864511 critic:2, 13.606982 modrow:2, 8658.278 paid:2, 16.3177 presid:2, 5.118029 public:2, 4.7937393 quit:2, 104.378456 aid:1, 18.29932 basketbal:1, 88.75063 berlin:1, 100.79294 camp:1, 53.001003 champion:1, 73.62071 chang:1, 7.0286655 charg:1, 9.539682 commerci:1, 13.564696 competit:1, 13.754395 decemb:1, 11.316556 deem:1, 89.11001 deleg:1, 24.309813 democrat:1, 14.2563505 desk:1, 4.628067 dtsb:1, 176051.67 e.german:1, 528155.0 edit:1, 4.1606994 eichler:1, 3799.6763 elit:1, 200.28632 embattl:1, 625.77606 emerg:1, 14.6877 endang:1, 112.541016 enjoi:1, 31.595776 entir:1, 26.41964 excess:1, 36.69272 facil:1, 15.926513 favorit:1, 102.01951 fund:1, 7.965538 german:1, 17.98403 govern:1, 3.3375988 gymnast:1, 1168.4845 han:1, 140.9541 hockei:1, 464.92517 home:1, 4.517582 includ:1, 3.5955327 increas:1, 5.3094244 kienbaum:1, 75450.71 klau:1, 236.20528 lead:1, 7.0955195 leader:1, 8.47312 local:1, 8.321071 maintain:1, 12.542866 met:1, 18.802242 mondai:1, 11.293085 mount:1, 56.493206 newswir:1, 193.46336 olymp:1, 133.40616 organ:1, 9.121691 page:1, 1.5209209 part:1, 2.3533697 peac:1, 15.639306 place:1, 6.585556 post:1, 13.686317 premier:1, 46.374134 pressur:1, 13.202555 previous:1, 26.593906 privileg:1, 117.133514 promot:1, 16.983023 rapid:1, 49.489788 respect:1, 16.096888 respons:1, 7.309497 revolt:1, 204.79062 secondari:1, 84.31593 secretariat:1, 167.7215 session:1, 23.747976 shotput:1, 4126.211 sign:1, 8.550071 societi:1, 21.18977 sponsor:1, 43.739544 star:1, 33.410614 state:1, 2.766687 statement:1, 11.338665 statu:1, 23.928734 stress:1, 20.094164 support:1, 5.7452493 tenni:1, 163.61679 time:1, 1.46516 timmermann:1, 40627.31 top:1, 11.849746 train:1, 16.756187 tuesdai:1, 10.767905 ulf:1, 3912.2593 unabl:1, 40.658585 urg:1, 24.178493 util:1, 37.32808 vice:1, 20.18401 want:1, 19.252542 wednesdai:1, 11.208485 word:1, 3.476968*

*LA121389-0036 147672 December 13, 1989, Wednesday, Home Edition* ***Sports****; Part C; Page 7; Column 1;* ***Sports*** *Desk 224 words NEWSWIRE: E.GERMAN* ***SPORTS*** *OFFICIALS QUIT The leadership of East Germany's embattled* ***sports*** *federation resigned Tuesday in response to mounting public criticism of excessive secretariat quit the German Gymnastics and* ***Sports*** *Federation (DTSB) in an emergency session were looking out for their own posts and urged that more attention be paid to local* ***sports****. Since East Germany's peaceful”*

Trying a different term such as ‘Ireland’ we get another long list of documents containing the term.

[*FT931-4691*](https://clueweb.adaptcentre.ie/IRModelGenerator/DocumentViewer?id=69237)*Term Freq Vector:  
bank:3, 7.489117 ireland:3, 70.98858 appoint:2, 18.164637 \_an:1, 2.512583 commerci:1, 13.564696 compani:1, 3.3356385 countri:1, 2.299005 court:1, 13.315392 dcjctaceft:1, 528155.0 director:1, 6.5301065 ec:1, 5.1114907 execut:1, 8.0267935 financi:1, 2.1300352 ft:1, 2.4916968 iez:1, 319.70642 industri:1, 2.6442788 london:1, 2.4773097 macsharri:1, 2870.4077 mar:1, 11.091033 page:1, 1.5209209 peop:1, 27.695595 peopl:1, 3.824605 rai:1, 85.92078 time:1, 1.46516 type:1, 3.20984  
has been appointed to the court of directors of Bank of* ***Ireland****. Companies:- Bank of* ***Ireland****. Countries:- IEZ* ***Ireland****, EC. Industries:- P602 Commercial Banks. Types:- PEOP Appointments. The Financial Times*

This was the highest ranked result from the search query, as we can see the portion in green shows the weights and frequency of each word within the document and lists them based on their importance when ranking. The part in black is a brief snippet from the document itself containing the terms that I searched for in the query. The article itself is very short which is likely why it is ranked so high.

*“*[*FT934-2050*](https://clueweb.adaptcentre.ie/IRModelGenerator/DocumentViewer?id=205209)*Term Freq Vector:  
ireland:7, 70.98858 govern:5, 3.3375988 decid:2, 10.467427 declar:2, 20.444973 ec:2, 5.1114907 establish:2, 8.158598 irish:2, 101.179115 london:2, 2.4773097 north:2, 10.549808 peopl:2, 3.824605 polit:2, 6.0952682 unit:2, 2.9128659 violenc:2, 49.341835 \_an:1, 2.512583 affair:1, 8.156456 affirm:1, 101.4707 agre:1, 9.018424 arrang:1, 23.924398 britain:1, 21.19147 chang:1, 7.0286655 claim:1, 11.167484 commun:1, 7.398994 consent:1, 113.09529 constitut:1, 19.298267 countri:1, 2.299005 dec:1, 20.59887 down:1, 103.498924 drop:1, 16.252424 dublin:1, 207.28218 econom:1, 5.4726553 effect:1, 6.451142 encompass:1, 218.78831 end:1, 4.1984706 event:1, 16.246923 fein:1, 561.2699 financi:1, 2.1300352 forum:1, 72.34009 framework:1, 45.88662 ft:1, 2.4916968 futur:1, 8.108745 gbz:1, 6.895964 give:1, 8.017168 govt:1, 45.467888 greater:1, 17.895674 iez:1, 319.70642 industri:1, 2.6442788 interest:1, 4.877633 intern:1, 3.8523903 ira:1, 233.9039 island:1, 39.30017 kingdom:1, 6.507417 legisl:1, 20.273108 main:1, 10.240127 major:1, 7.2614594 nation:1, 3.5148706 negoti:1, 12.680488 new:1, 3.8393402 number:1, 5.8547926 offer:1, 7.7248397 page:1, 1.5209209 peac:1, 15.639306 place:1, 6.585556 point:1, 5.6887507 recommend:1, 20.204859 reconcili:1, 146.75049 renounc:1, 378.60574 secur:1, 6.0898566 selfish:1, 666.86237 settlement:1, 31.835745 sinn:1, 584.24225 south:1, 8.453997 statu:1, 23.928734 strateg:1, 36.026943 street:1, 11.605762 tabl:1, 26.706867 time:1, 1.46516 type:1, 3.209849 ulster:1, 319.513 wai:1, 27.27228  
Both governments affirm that the status of N* ***Ireland*** *can be changed only with the consent of 'a greater number of its people'. But the future of* ***Ireland*** *to be decided by people of north and south alone. Britain declares it has no selfish, strategic or economic interest in N* ***Ireland****. It will give legislative effect to a united* ***Ireland*** *if a majority in the north so decides. London and Dublin to establish political framework encompassing arrangements within N* ***Ireland****, for the whole island”*

When looking at this second piece it is lower ranked by the search engine, we see it is a bit longer than the first one we had, this is likely the cause of the lower ranking as the relevance of my singular term within the article is lower because of the length.

The shorter the length of the document and the higher the relative frequency leads to a higher ranking for the document. This is because with the TFIDF model it multiplies the count of the word within the document, multiplies by the weight and then divides by the total number of words in the document.

Now moving on to the BM-25 model, using it with the “Ireland” term so we can compare the two models. This returns us a significantly different list of results.

“[*LA103089-0034*](https://clueweb.adaptcentre.ie/IRModelGenerator/DocumentViewer?id=456931)*Term Freq Vector:  
ireland:11, 70.98858 northern:6, 27.303299 worm:5, 1017.6397 agricultur:4, 21.98539 zealand:4, 138.29668 column:2, 3.864511 long:2, 5.58575 metro:2, 13.23929 nativ:2, 85.076515 popul:2, 21.926973 recent:2, 5.781474 report:2, 2.316825 ago:1, 8.942231 arriv:1, 19.630365 artiopositalthia:1, 528155.0 attack:1, 17.65165 border:1, 25.780006 british:1, 12.523534 call:1, 5.006873 capsul:1, 851.8629 damag:1, 22.620026 deplet:1, 317.0198 desk:1, 4.628067 dip:1, 105.08456 earth:1, 74.127014 earthworm:1, 17037.258 edit:1, 4.1606994 egg:1, 163.92148 end:1, 4.1984706 enter:1, 18.57281 flatworm:1, 264077.5 found:1, 9.5647335 home:1, 4.517582 inch:1, 77.34002 invad:1, 223.51036 invas:1, 127.60449 irish:1, 101.179115 journal:1, 119.817375 lowli:1, 1833.8716 medicin:1, 82.524216 mondai:1, 11.293085 mysteri:1, 130.63443 octob:1, 12.604831 page:1, 1.5209209 part:1, 2.3533697 plagu:1, 191.56874 product:1, 5.8490868 prolifer:1, 152.60185 provinc:1, 38.24439 republ:1, 15.220166 research:1, 15.522557 respect:1, 16.096888 root:1, 61.9464 scienc:1, 34.671764 scientist:1, 89.95998 shrub:1, 983.5289 side:1, 10.646355 slowli:1, 93.13261 southern:1, 19.842024 spread:1, 34.59455 staff:1, 7.3542805 suffer:1, 18.019617 term:1, 6.7086897 thought:1, 17.126759 threaten:1, 24.181814 time:1, 1.46516 trade:1, 5.983403 triangulata:1, 528155.0 wipe:1, 166.191 wire:1, 16.863726 word:1, 3.4769685 year:1, 1.985470  
Desk 154 words IN BRIEF: SCIENCE / MEDICINE;* ***IRELAND*** *UNDER ATTACK BY NEW ZEALAND WORMS From Times staff and wire reports* ***Ireland*** *has suffered a mysterious invasion from the other end of the Earth that threatens to do serious long-term damage. The invader, which has spread out across Northern* ***Ireland*** *and is now entering the Republic of* ***Ireland****, is a lowly New Zealand worm called Artiopositalthia New Zealand worm got to* ***Ireland*** *nobody knows. But it is thought that it probably arrived in Northern”*

This was the first result returned by the search and as you can see this is a much different result than the other model, the document although still relatively short it’s significantly longer than the number document returned by the TF-IDF model.

The reason for this is because the BM-25 model addresses some of the limitations of the TF-IDF model, it takes into consideration a saturation function, which diminishes the impact of high term frequencies as for, it also normalises the length of the documents penalising the longer documents, and also uses parameter tuning through the K and B values, this improves the overall relevance of the articles when doing a search.

The K value affects the sensitivity of term frequency, whereas the B value affects the impact of document length, a higher K value means that the frequency of terms is very important, whereas with a higher B value means that the longer the document the more it’s penalised by our ranking model.

Now switch the search query to “Economics” with the BM25 model with parameters k=1.2 and b=0.75. The documents retrieved seem to be longer in comparison to using TF-IDF. The model still takes weights of words into account, however the documents retrieved have much higher total occurrences than those returned using TF-IDF. For example the first snippet returned was:

“[*FBIS3-27640*](https://clueweb.adaptcentre.ie/IRModelGenerator/DocumentViewer?id=327350)*Term Freq Vector:  
econom:22, 5.4726553 zone:10, 36.199795 corridor:9, 199.68053 meet:7, 4.9425874 provinci:5, 81.73244 construct:4, 13.55808 hunan:4, 1200.3523 provinc:4, 38.24439 changsha:3, 2917.9834 citi:3, 7.320134 develop:3, 4.813267 area:2, 6.401103 belt:2, 172.31812 build:2, 8.077741 coastal:2, 129.07013 directli:2, 28.057533 feb:2, 12.911747 gnp:2, 353.0448 govern:2, 3.3375988 increas:2, 5.3094244 industri:2, 2.6442788 leader:2, 8.47312 made:2, 3.9773703 make:2, 3.423398 mandarin:2, 231.54538 network:2, 12.110869 open:2, 6.1597447 parti:2, 6.676801 peopl:2, 3.824605 percent:2, 17.624554 radio:2, 13.676041 reform:2, 15.490687 region:2, 8.286994 requir:2, 6.344128 set:2, 5.3337135 tech:2, 31.374302 year:2, 1.9854704 accord:1, 15.480245 achiev:1, 12.236007 ahead:1, 16.223469 amount:1, 11.909062 appli:1, 14.6222315 articl:1, 6.381227 ask:1, 12.040191 attend:1, 20.540388 author:1, 6.9454784 averag:1, 13.976052 bangzhu:1, 14670.973 billion:1, 23.914648 bo:1, 470.72638 breakthrough:1, 137.68379 calcul:1, 39.142887 central:1, 9.204995 chen:1, 269.05502 chi:1, 32.515854 chu:1, 489.0324 committe:1, 10.909814 compar:1, 11.799446 countri:1, 2.299005 demand:1, 9.4546385 depart:1, 7.959656 divid:1, 40.03297 document:1, 5.88028 end:1, 4.1984706 experi:1, 17.154573 explicitli:1, 300.25867 export:1, 18.825699 express:1, 12.860813 fbi:1, 4.515535 februari:1, 10.864704 forward:1, 22.083752 function:1, 30.149275 gmt:1, 6.22494 governor:1, 35.86304 guiyu:1, 132038.75 high:1, 5.5923147 higher:1, 11.763966 hold:1, 7.6489882 hope:1, 8.490556 import:1, 6.48035 kei:1, 18.703035 languag:1, 7.2994957 larg:1, 7.562357 lead:1, 7.0955195 locat:1, 19.911592 major:1, 7.2614594 mark:1, 12.305569 move:1, 6.667782 orient:1, 52.563198 pan:1, 108.629166 part:1, 2.3533697 plai:1, 10.0333395 plan:1, 4.645122 point:1, 5.6887507 polici:1, 6.303694 prefectur:1, 305.11554 preferenti:1, 244.85628 previou:1, 16.443182 price:1, 6.662315 program:1, 7.6774526 proport:1, 47.63732 rank:1, 26.779991 rapidli:1, 55.30998 relev:1, 43.69612 report:1, 2.316825 respect:1, 16.096888 role:1, 13.720806 south:1, 8.453997 special:1, 9.780286 specif:1, 13.822067 strong:1, 11.500131 success:1, 11.361592 support:1, 5.7452493 tang:1, 962.03094 target:1, 20.238926 ten:1, 43.312695 text:1, 4.375694 thoroughfar:1, 1450.9752 top:1, 11.849746 transplant:1, 452.1875 turn:1, 8.733588 type:bfn:1, 9.673877 type:daili:1, 10.209642 vital:1, 55.73019 wang:1, 212.28094 work:1, 4.2602763 yesterdai:1, 5.5204186 yuan:1, 138.00758 zhixiang:1, 33009.68  
Region Hunan Holds Meeting on Building* ***Economic*** *Zones HK2402054294 Changsha Hunan People's Radio: Mandarin Article Type:BFN [Text] A provincial meeting on building five* ***economic*** *zones and one* ***economic*** *of these five* ***economic*** *zones and one* ***economic*** *corridor will play an important role in making Hunan rank among the country's top ten provinces. These five* ***economic*** *zones and one* ***economic*** *corridor will become the province's hi-tech industrial thoroughfare and export-oriented* ***economic*** *belt. Provincial party”*

Following this, switching the parameters k=1.3 and b=0.8 the order of the articles retrieved changed, this meant the document snippet above was the second snippet returned. The order of documents retrieved altered for the same reasons listed above.