

CA4009: Search Technologies

Laboratory Session 1

26th October 2021

Joanna Talvo - 18342523

Chloe Ward - 18302716

4: Examining Document Collection Statistics

Each term in the list has a corresponding document frequency and collection frequency. Document Frequency is the number of documents that contain a term that we have searched for. Document frequency eliminates unimportant words that exist in the analysis of the dataset in relation to words that we are searching for.

This document collection was taken from news articles which are taken mainly from the LA Time newspaper. Since this data came from newspaper articles, the top 20 terms are mainly related to words that we see in news articles. The ranking of the terms are based on their document frequency and are shown in descending order.

From the notes, Hans Luhn said that if a word occurs more often than expected in the document, the word is more likely to be related to this document and therefore more likely to be relevant to a query that contains this word. “**time**” is the most used term in this collection having the highest document frequency value. So if this collection consists of about 500,000 and “**time**” having 360,476, this means that “time” appears to be very related to this document collection. The rest of the list such as page, year, financial are also part of the top 20 most searched terms in news articles which are basically common words that we also find in news articles.

However, it is also said in the notes that the terms present in less documents were said to have greater discriminative power. Schultz, who introduced the Gaussian bell curve, interpreted that frequent terms lack discriminative power and rare terms are not sufficient to be shared across documents with similar meaning. So to find the better keywords in this document collection, we should not be looking for the most frequent or rarest words, but rather the ones that occur in average amount of times which can be found in the middle part of the top terms in the list.

Num top terms:

Term	Doc Freq	Coll Freq
time	360476	684393
page	347260	383440
year	266010	827942
financi	247956	344316
countri	229732	417625
report	227965	428085
94	227231	406769
part	224425	354664
london	213197	298164
ft	211966	242930
_an	210204	210215
1	202111	499593
industri	199735	382327
state	190898	550368
unit	181318	323331
1994	173764	272099
2	171375	404964
type	164542	186382
compani	158337	525676
govern	158244	449893

figure 1. Top 20 terms

In figure 1, we noticed that some terms have a lower document frequency but a higher collection frequency than the ones above them. For example, “year” has the highest collection frequency value among the terms. From the notes, the key concept of Collection Frequency is that the terms that occur in less documents are more valuable than the ones that occur in many documents. So if we are taking that concept here in this collection, terms such as “year”, “time” and “state” are the good keywords.

There are also terms that are integers in the list like single digits, double digits and years like 1994. These terms might not be very useful for information retrieval as they are very generic terms and unlikely to relate specific parts in the documents. Some terms were also stemmed down like “compani”, “financi” and stemming of words helps increase the accuracy of information retrieval.

5: Interactive Searching using Lucene

Using TF-IDF

At the start, we used Lucene , with the default values of $k = 1.2$ and $b = 0.75$ for BM25. We entered the query “**Medicine**” into the query box, and documents with our chosen search term appeared.

error:1, 430.44418 expertis:1, 107.78474 fact:1, 10.156046 fridai:1, 9.775217 govern:1, 3.3375988 grant:1, 17.2425 home:1, 4.517582 involv:1, 8.572136 medic:1, 29.337055 metro:1, 13.23929 nov:1, 21.428774 page:1, 1.5209209 part:1, 2.3533697 patient:1, 68.54705 power:1, 7.7429595 recogn:1, 40.577366 record:1, 8.502173 report:1, 2.316825 specialist:1, 40.4159 subspecialti:1, 40627.31 tension:1, 59.644833 trauma:1, 454.52237 usc:1, 175.70027 word:1, 3.476968

words FOR THE RECORD Emergency **medicine** -- A Nov. 19 article about tensions between two specialties that emergency **medicine** was recognized with its own specialty board examinations only a year ago. In fact, the American Board of Emergency **Medicine** has been empowered to certify doctors as emergency **medicine** in subspecialties of emergency **medicine**. Correction

• [FBIS4-51121](#)
Term Freq Vector:
buddhist:10, 785.94495 medicin:9, 82.524216 symposium:4, 605.6823 close:3, 6.3912654 fujian:3, 823.9548 provinc:3, 38.24439 region:3, 8.286994 research:3, 15.522557 xinhua:3, 62.48137 beiji:2, 30.195814 china:2, 20.070492 confer:2, 13.39509 cure:2, 212.11044 east:2, 9.873901 english:2, 8.705804 establish:2, 8.158598 municip:2, 54.38729 wuyishan:2, 88025.836 academ:1, 78.61789 affair:1, 8.156456 announ:1, 7.383272 associ:1, 13.690929 attend:1, 20.540388 autonom:1, 132.03876 believ:1, 11.696231 bfn:1, 8.5514555 cancer:1, 128.75548 case:1, 7.778768 chi:1, 32.515854 chines:1, 26.945309 citi:1, 7.320134 clinic:1, 111.89725 combin:1, 19.962777 commun:1, 7.399894 complic:1, 57.9435 countri:1, 2.299005 dai:1, 4.1854615 dalli:1, 6.8694325 develop:1, 4.813267 difficult:1, 14.271374 dozen:1, 44.068 effect:1, 6.451142 enhanc:1, 41.329918 exchang:1, 9.574616 expert:1, 27.413837 famou:1, 76.37817 fbi:1, 4.515535 field:1, 15.254015 fuzhou:1, 1778.2997 gmt:1, 6.22494 grass:1, 145.49724 great:1, 10.910265 gynecolog:1, 37.19.4014 heard:1, 31.00047 held:1, 8.504089 hospi:1, 31.922333 human:1, 20.181696 illi:1, 47.770893 import:1, 6.48035 institut:1, 10.178162 made:1, 3.9773703 medic:1, 29.337055 meet:1, 4.9425874 natur:1, 12.473549 ophthalmolog:1, 8951.779 organ:1, 9.121691 oversea:1, 38.50649 paper:1, 20.223427 part:1, 2.3533697 particip:1, 12.969452 past:1, 8.204732 progress:1, 18.391079 prove:1, 20.4624 receiv:1, 6.547349 relat:1, 6.7995496 report:1, 2.316825 root:1, 61.9464 scholar:1, 245.42519 scienc:1, 34.671764 scienti:1, 50.439785 societi:1, 21.18977 strengthen:1, 26.816704 stress:1, 20.094164 studi:1, 14.69342 surgeri:1, 153.35512 teach:1, 71.94592 templ:1, 235.8888 ten:1, 43.312695 text:1, 4.375694 toda:1, 5.9891024 tradit:1, 22.989248 year:1, 1.985470

Region First Buddhist **Medicine** Symposium Closes in Fujian First Buddhist **Medicine** Symposium Closes BFN [Text] Fuzhou, May 24 (XINHUA) -- The first China Buddhist **medicine** symposium came to a close of Buddhist **Medicine**. Participants in the symposium held that Buddhist **medicine** is a combination of natural science and humane studies, and is an important part of traditional Chinese **medicine**. They believed that Buddhist **medicine** has also proved effective in curing some difficult and complicated

• [FBIS4-57406](#)
Term Freq Vector:
medicin:25, 82.524216 teja:16, 13203.875 plan:10, 4.645122 add:9, 22.221264 distribut:9, 21.568792 medic:7, 29.337055 million:7, 9.764555 peopl:6, 3.824605 product:6, 5.8490868 cuba:5, 130.11949 import:5, 6.48035 ensur:4, 16.864801 health:4, 15.862416 pharmaci:4, 740.75037 program:4, 7.6774528 resilleaz:4, 88025.836 state:4, 2.766687 avail:3, 74.56657 bui:3, 12.793213 card:3, 52.662777 control:3, 6.977778 doctor:3, 57.47062 explain:3, 17.46371 famili:3, 12.962768 figur:3, 10.2250595 goal:3, 24.1796 includ:3, 3.5955327 increas:3, 5.3094244 level:3, 6.999788 minist:3, 6.0379205 popul:3, 21.926973 price:3, 6.662315 process:3, 8.970024 produc:3, 9.346057 stress:3, 20.094164 acquir:2, 25.555475 agenda:2, 46.41081 approxim:2, 44.83489 began:2, 15.665748 certifi:2, 44.755104 chronic:2, 233.49028 commun:2, 7.398994 conclud:2, 23.727705 condit:2, 9.777389 entitl:2, 51.11837 essenti:2, 32.58004 expect:2, 5.123341 havana:2, 189.77902 higher:2, 11.763966 intensifi:2, 92.46411 issu:2, 4.2502656 made:2, 3.9773703 main:2, 10.240127 make:2, 3.423398 materi:2, 15.561891 mechan:2, 29.974745 meet:2, 4.9425874 monitor:2, 32.896606 nationwid:2, 77.53303 neighborhood:2, 78.46605 open:2, 6.1597447 packag:2, 26.139816 pad:2, 293.09378 point:2, 5.6887507 previou:2, 16.443182 public:2, 4.7937393 record:2, 8.502173 recoveri:2, 24.010319 replac:2, 18.026999 report:2, 2.316825 review:2, 10.82773 rigid:2, 246.91678 sale:2, 8.253453 sanitari:2, 582.9525 shortag:2, 61.91735 sold:2, 19.257456 stabl:2, 55.327362 system:2, 6.752001 vitamin:2, 962.03094 year:2, 1.9854704 abroad:1, 41.890465 ailment:1, 788.291 alloc:1, 47.99664 allow:1, 13.308346 altern:1, 22.02665 amount:1, 11.909062 annual:1, 10.915225 answer:1, 22.403181 antonio:1, 147.11838 appear:1, 12.562856 asthma:1, 1200.3523 attain:1, 103.60043 avoid:1, 21.536251 bee:1, 815.054 bfn:1, 8.5514555 book:1, 21.051258 boost:1, 29.499273 build:1, 8.077741 cardiovascular:1, 1272.6626 caribbean:1, 114.567245 carri:1, 10.239928 categori:1, 47.14407 caus:1, 19.695518 chang:1, 7.0286655 charg:1, 9.539682 comment:1, 6.2439265 committe:1, 10.909814 communist:1, 37.228096 comparison:1, 73.202354 complex:1, 27.320246 compli:1, 46.76421 compris:1, 76.90084 concept:1, 44.995316 confus:1, 54.38169 consumpt:1, 66.568565 contribut:1, 14.917526 cost:1, 6.4599004 council:1, 10.228822 creat:1, 10.400027 cuban:1, 172.2619 dalli:1, 6.8684325 deriv:1, 68.06121 develop:1, 4.813267 diabet:1, 933.13605 dire:1, 443.45508 document:1, 5.88028 draft:1, 78.130761 editor:1, 40.085077 educ:1, 18.782183 effect:1, 6.454149 effect:1, 9.033701 elimin:1, 31.865867 emerg:1, 14.6877 experi:1, 17.154573 event:1, 18.826600 express:1, 12.860813 fale:1

We then kept playing around with both of the k and b values, and we came to the conclusion that by changing the b value it ultimately changes the normalisation for the documents that are retrieved from our search term. We also realised that when we change the k value, the documents that were retrieved were much less related to the search term “medicine”.

Using the search term “olympic”, the summary lengths of the contents are quite small. The doc IDs start with FT and LA.

- [FR940919-2-00118](#)
Term Freq Vector:
time:10, 1.46516 task:9, 24.566492 hyph:4, 13.092913 number:4, 5.8547926 repres:4, 8.015343 dai:3, 4.1854615 laps:3, 367.54 shown:3, 32.96135 assumpt:2, 104.21369 bill:2, 9.10722 code:2, 17.213278 action:1, 7.730606 calendar:1, 30.357225 common:1, 17.252073 complet:1, 9.054448 employe:1, 20.076595 entr:1, 45.79908 estim:1, 13.339605 extrem:1, 28.202862 high:1, 5.5923147 kei:1, 18.703035 list:1, 11.034033 low:1, 12.092014 midd:1, 21.468843 minut:1, 20.530807 model:1, 29.430235 process:1, 8.970024 redesign:1, 224.5557 team:1, 14.797159 work:1, 4.260276
- FR940919-2-00118 FR940919-2-00046 Assumptions, Task Times and Lapse Times Listed below are key assumptions, task times and lapse times that the Team used to model the redesigned process. The task times are shown in minutes and represent the estimated time it will take an employee to complete the described task. For each task time entry, three task time numbers are shown. The middle number represents the most common task time, while the first and last number represent the low and high extremes
- [FT911-1466](#)
Term Freq Vector:
financ:3, 2.1300352 time:3, 1.46516 an:1, 2.512583 bank:1, 7.489117 beeaadzt:1, 528155.0 ft:1, 2.4916968 holidai:1, 47.7493 london:1, 2.4773097 mondai:1, 11.293085 new:1, 3.83893402 page:1, 1.5209209 public:1, 4.7937393 publish:1, 11.970604 resum:1, 50.89179 world:1, 5.191222
- FT911-1466 _AN-BEEAAADZFT 910504 FT 04 MAY 91 / World News In Brief: Financial Times The Financial Times will not be published on the bank holiday, Monday May 6. Publication will resume on May 7. The Financial Times London Page 1
- [FT923-5675](#)
Term Freq Vector:
financ:3, 2.1300352 time:3, 1.46516 an:1, 2.512583 aug:1, 26.331388 bank:1, 7.489117 cibbtaaaft:1, 528155.0 ft:1, 2.4916968 holidai:1, 47.7493 london:1, 2.4773097 mondai:1, 11.293085 new:1, 3.83893402 page:1, 1.5209209 publish:1, 11.970604 world:1, 5.191222
- FT923-5675 _AN-CIBBTAAAF 920829 FT 29 AUG 92 / World News In Brief: Financial Times The Financial Times will not be published on Monday because of the bank holiday. The Financial Times London

Short snippets of the term “olympics”

We then moved onto TF-IDF and When we searched the term “drug” in the query box, we noticed that between the next occurrence of “**drug**” that there was at least 5 or less word difference between the next occurrence of “**drug**” this highlighted to us that the data was generated using Luhn’s keyword cluster. In Luhn’s keyword cluster he determined that two words are significantly related if they are not separated from more than two five insignificant words which can be seen when we search the term “drug”. This made us aware that it was probably scored using Luhn’s keyword cluster.

We noticed that with TF-IDF, the summarisation of the snippets were small, whereas with the BM25 the snippets were longer. Another thing we noticed was that the term frequency vectors list for the top documents using BM25 were longer than TF-IDF.

- [FBIS3-22105](#)

Term Freq Vector:

drug:16, 25.26332 addict:11, 227.65302 cina:3, 24007.045 neatkariņa:3, 27797.63 riga:3, 556.53845 article:2, 6.381227 batti:2, 29.698324 document:2, 5.88028 februari:2, 10.864704 institut:2, 10.178162 latvia:2, 274.65158 latvian:2, 467.3938 medic:2, 29.337055 movement:2, 21.23407 number:2, 5.8547926 parent:2, 33.01794 part:2, 2.3533697 people:2, 3.824605 regist:2, 20.461607 user:2, 58.80149 year:2, 1.9854704 ag:1, 20.624609 arrest:1, 32.18887 bureau:1, 46.83886 center:1, 12.6289425 central:1, 9.204995 club:1, 35.432377 confidenti:1, 118.7933 continu:1, 4.803068 dealer:1, 35.3092 den:1, 424.5619 establish:1, 8.158598 eurasia:1, 71.73095 feb:1, 12.911747 gather:1, 39.323578 health:1, 15.862416 iceberg:1, 1184.204 inform:1, 6.424696 jpr:1, 31.357536 languag:1, 7.2994957 nation:1, 3.5148706 percent:1, 17.624554 pm:1, 167.66826 remind:1, 63.36593 renci:1, 88025.836 report:1, 2.316825 scandinavia:1, 52815.5 specialist:1, 40.4159 state:1, 2.766687 statist:1, 33.87129 street:1, 11.605762 strelnieku:1, 176051.67 tdi:1, 149.40735 ten:1, 43.312695 text:1, 4.375694 thousand:1, 28.913067 title.narcot:1, 472.8335 treat:1, 33.0862 treatment:1, 33.72422 type.cso:1, 78.39617 type.jpr:1, 55.00469 unit:1, 2.9128659 viestur:1, 52815.5 visibl:1, 95.026085 work:1, 4.2602763 written:1, 21.201677 ziemeļblāzma:1, 264077.

February 1994 CENTRAL EURASIA LATVIA Statistics On Number Of Drug Addicts Reported 94WD0271A Riga Type:CSO [Article by Viesturs Rencis: "Still Drug Addicts"] [Text] According to the State Drug Treatment and Health Center medical institutions have registered 794 drug addicts and 358 drug users. In 1993, 125 drug addicts were registered and 50 drug users. Last year 383 drug addicts were treated in various medical institutions. For its part in Riga alone there were 2,536 drug addicts, people arrested

- [FBIS3-24331](#)

Term Freq Vector:

drug:16, 25.26332 addict:11, 227.65302 cina:3, 24007.045 latvia:3, 274.65158 neatkariņa:3, 27797.63 riga:3, 556.53845 article:2, 6.381227 batti:2, 29.698324 document:2, 5.88028 feb:2, 4.515535 februari:2, 10.864704 institut:2, 10.178162 latvian:2, 467.3938 medic:2, 29.337055 movement:2, 21.23407 number:2, 5.8547926 parent:2, 33.01794 part:2, 2.3533697 people:2, 3.824605 regist:2, 20.461607 state:2, 2.766687 user:2, 58.80149 year:2, 1.9854704 ag:1, 20.624609 arrest:1, 32.18887 baltic:1, 135.77249 bureau:1, 46.83886 center:1, 12.6289425 central:1, 9.204995 club:1, 35.432377 confidenti:1, 118.7933 continu:1, 4.803068 dealer:1, 35.3092 den:1, 424.5619 establish:1, 8.158598 eurasia:1, 71.73095 feb:1, 12.911747 gather:1, 39.323578 health:1, 15.862416 iceberg:1, 1184.204 inform:1, 6.424696 languag:1, 7.2994957 nation:1, 3.5148706 percent:1, 17.624554 pm:1, 167.66826 remind:1, 63.36593 renci:1, 88025.836 report:1, 2.316825 scandinavia:1, 52815.5 specialist:1, 40.4159 statist:1, 33.87129 street:1, 11.605762 strelnieku:1, 176051.67 ten:1, 43.312695 text:1, 4.375694 thousand:1, 28.913067 tit:1, 23.295475 treat:1, 33.0862 treatment:1, 33.72422 type.cso:1, 78.39617 type.jpr:1, 55.00469 unit:1, 2.9128659 usr:1, 163.16188 viestur:1, 52815.5 visibl:1, 95.026085 work:1, 4.2602763 written:1, 21.201677 ziemeļblāzma:1, 264077.

: Central Eurasia 25 February 1994 BALTIC STATES LATVIA Statistics On Number Of Drug Addicts In Latvia: Latvian Article Type:CSO [Article by Viesturs Rencis: "Still Drug Addicts"] [Text] According to the State Drug Treatment and Health Center medical institutions have registered 794 drug addicts and 358 drug users. In 1993 125 drug addicts were registered and 50 drug users. Last year 383 drug addicts were treated in various medical institutions. For its part in Riga alone there were 2,536 drug addicts