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: 20 Get Overview Stats

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

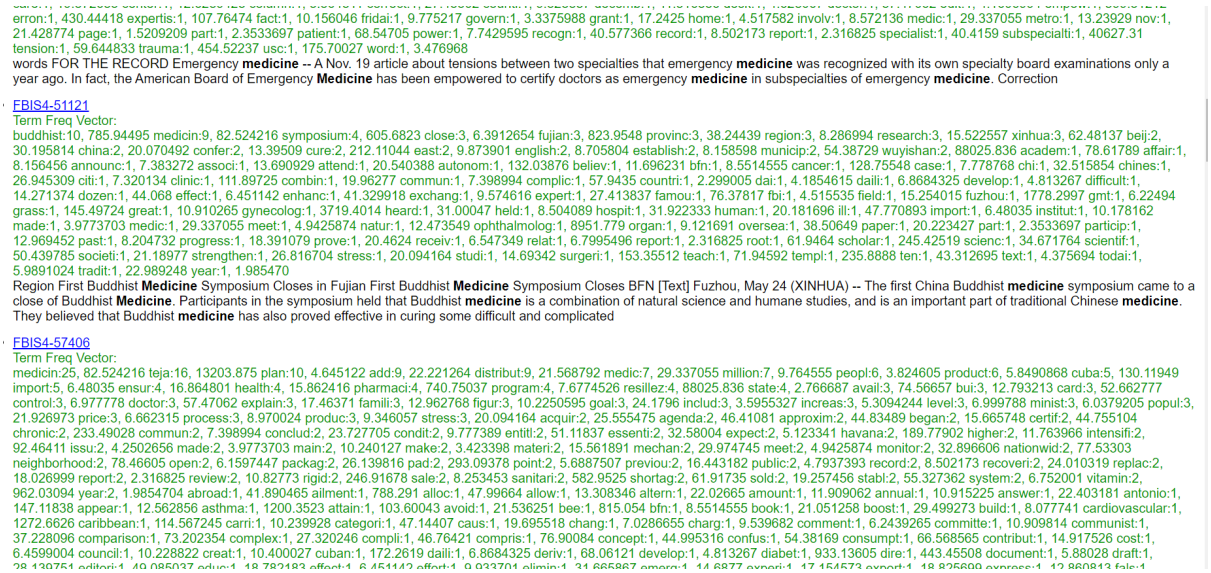
In figure 1, we noticed that some terms have a lower document frequency but a higher collection frequency than the ones above them. For example, “year” has the highest collection frequency value among the terms. From the notes, the key concept of Collection Frequency is that the terms that occur in less 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.



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.

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 each occurence 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 neatkariga:3, 27797.63 riga:3, 556.53845 articl:2, 6.381227 battl: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 peopl: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 scandanavia:1, 52815.5 specialist:1, 40.4159 state:1, 2.766687 statist:1, 33.87129 street:1, 11.605762 strelnieku:1, 176051.67 tdd: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 ziemelblazma: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 neatkariga:3, 27797.63 riga:3, 556.53845 articl:2, 6.381227 battl:2, 29.698324 document:2, 5.88028 fbi: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 peopl: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 scandanavia: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 titl: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 ziemelblazma: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