CA4009 Search Technologies Lab #1

Date: 07/11/20199

Name(s): Maksims Kompanijecs - 15306971, Matthew Farelly - 15366246

Question 4 - Document Collection Statistics
Examine the occurrence count information for the terms in the list that you can see, consider the values and what they tell you about the terms and their likely usefulness for information retrieval. Enter the results of your analysis in your report file for today's laboratory

From our analysis of viewing the list of overview stats, we gathered that the list displays most frequent words in LA Times newspaper. This would mean that searching for these terms is pretty useless to gain any insightful information to a specific event. Doc Freq field displays how many documents does the term appears and Coll Freq field tells us the total occurence of the term in the collection. For example, the highest ranked term is 'time', this term appears in 360476 documents at a frequency 684393. Trying to search for an article based on 'time', will return more than half of documents published from the collection. This would prove difficult to search for information based on these keywords.

Question 5 - Interactive Searching using Lucene
Explore the relationship between the words that you entered into the query box, the contents of the snippets, the contents of the documents, the ranking function that you chose and the ranking of the documents. See if you can gauge how the term weights and document lengths might have produced the ranked output that your see. You can enter more search requests to explore this relationship further

From our analysis using BM25 ranking function. Typing in short search requests into the query box displays the list of retrieved documents.

rent - BM25 - 19/18 **+1**, 10/10 **+0**, 62/53 **+7**, 16/17 **+1**, 22/16 **+6**, 28/25 **+3**, 42/33 **+9** , 36/33 **+3**, 44/33 **+11**, 28/44 **-16**
Rent - TF-IDF - 10/10 **+0** , 9/9 **+0**, 19/18 **+1**, 9/8 **+1**, 4/4 **+0**, 17/16 **+1**, 5/4 **+1**, 22/16 **+6**, 8/7 **+1**, 7/7 **+0**