Data visualization can be useful tool to help understand or view a large corpus, especially those that are primarily composed of text- associations or trends that otherwise may have been missed can be found and explored further, and viewed in a way which is easier for us (as humans) to comprehend than a string of text by itself. Through the utilization of Voyant, I was able to help get a better understanding of the trends and topics of both pieces of texts from my corpus.

Text

Description automatically generated

(A word cloud of the first part of my corpus)  
Text, whiteboard

Description automatically generated

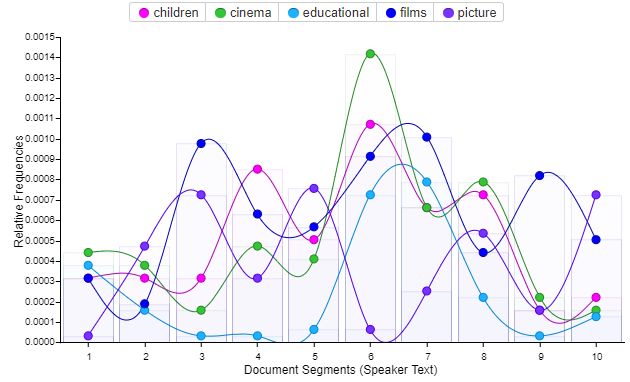
(A word cloud of the second part of my corpus)

I immediately turned to the word clouds created from the texts as a starting point, to get a quick idea of which topics were discussed the most.

A more detailed breakdown of the total frequency of the most commonly occurring words:

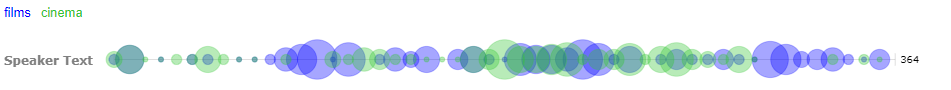
|  |  |
| --- | --- |
| **Frequent Word Count Pt. 1** | **Frequent Word Count Pt. 2** |
| *Films* - 202 | *Films* - 728 |
| *Children* - 163 | *Cinema* - 664 |
| *Cinema* - 162 | *Children* - 643 |
| *Picture* - 128 | *Think* - 629 |
| *Educational* - 81 | *Say* -417 |

It is worth noting the similarities between the two texts- They both share the same 3 most frequently used words, and almost in the same order. Additionally, the frequent occurrence of “children” would suggest that they are a primary focus of the discussion and concerns of the committee.



(Charting the most frequently appearing words in Part 1)

This assumption is further strengthened when taking a closer look at the overall distribution of words in the first part of the corpus- “children” and “cinema” have very similar distributions, especially in where they rise and fall in frequency. A correlation analysis between the two terms revealed the two terms have a Correlation of 0.8660, and a Significance of 0.0011- suggesting the two terms are commonly (and frequently) used in conjunction with each other in some capacity.

Chart, line chart

Description automatically generated

The high correlation, the similar distribution, and almost identical word count between cinema and children had me wondering why there was not a similar trend between children and films (since cinema and films are similar on a conceptual level), so I took a closer look at the relationship between “children” and “films”- with a Correlation of 0.296 and a Significance of 0.405, it appeared cinema and films were being used in entirely separate contexts. A closer look at their distributions shows that the two terms have low correlation or relation with each other; in segments 3-4 and 8-9, the frequency of one increases while the other decreases, and vice versa. While this is not the trend for the entire document (segments 4-5 are fairly similar in distribution), it definitely suggests there is a clear distinction between the two terms usage.

I wanted to focus on another significant term that had relevance to the relation between cinema and children- “education”. Cinema and education have a Correlation value of 0.7905, and a Significance of 0.00648, while children and education have respective values of 0.676 and 0.0317.

Chart, line chart

Description automatically generated

The distribution of education does seem to match, albeit in lower frequency, the overall distribution of children and cinema. When re-introducing the “Heading” category, it reveals the noticeable spike in all three terms- segment 6- is from a section entitled “The Cinema in its Relation to the Education of Children”, which explains the noticeable increase in frequency for these terms. However, they maintain a similar distribution throughout the entire text, suggesting these three terms are omnipresent throughout it, and these concerns are the most significant when moral or ethical concerns in relation to film are brought up. Ultimately, children were a key consideration for the council.

Chart, line chart

Description automatically generated

Moving onto other terms, a bubble map revealed that “picture”, the fourth most frequent word, had a high frequency occurrence with the term “house”- i.e. a “picture house”, or an old name for movie theatres. The above chart confirms this association, as the two have very similar distributions.

Chart

Description automatically generated

I eventually turned my attention to “trade”, a fairly frequent word (the 10th) which had an interesting correlation with “censor\*”- with respective Correlation/Significance values of .8809/0.0007. “Censorship” had appeared in the word cloud, so I figured it would be worth taking a closer look at, especially in relation to film. Interestingly, the frequency of censorship has a sharp increase towards the end of the text, something that “children” does not do. The correlation between the two is well below 0.25, which suggests any talk about censorship did not directly have to do with children, which can end up serving as the inspiration to censor things in the first place. “Trade”, “country”, and “public” all had similar Correlation scores of ~0.72, and public was another frequently occurring word (the 9th). I also compared the frequency of public with censorship and noticed a more similar distribution of frequency. I also noticed protect had a fairly strong correlations core, and decided to include it as well (under the assumption they might be referring to censorship to “protect” the public), and noticed its noticeably low frequency did occur in spikes of public and censorship. Something else that was interesting and worth noting was trade’s surprising lack of association with terms like “country”, “foreign”, or “international”- my first assumption was it was referencing international films exchanges of some sort, but it appears to be functioning under a different definition, one closely associated with censorship.  
Chart, line chart

Description automatically generated

I decided to more properly graph the frequency distributions of these terms, to better visualize their relationship. The overall distribution seemed to align with what I previously suggested. Censorship immediately stood out again, as no other terms, with the exception of “house” and “picture”, drastically increase in frequency near the end like that. Upon returning to the correlations page to explore this further, I noticed another word with a strong association with censorship- “posters” (as in movie posters) has respective values of 0.818/0.0038. During the censorship spike at segments 8-9, there is discussion over the censorship of specifically the posters for films, or “billposters”, for potentially offensive or perceived as obscene content. This was also discussed in conjunction with the censorship of films, and how films are reviewed for their content, and their posters should accurately reflect both their content without themselves depicting any explicit content.

Chart

Description automatically generated

This graph also shows that even with billposters low relative frequency, it still aligns with the distribution of censorship, protect, and public- two peaks at segment 3 and 5, and an increase from segments 9-10. Considerations other than just the films themselves were taken heavily into consideration- a notable portion of the discussion of censorship was in relation to movie posters, and unsurprisingly, the considerations of the public.

After my initial poking and prodding, I decided to try to collect the names of prominent individuals in this text.

|  |  |  |
| --- | --- | --- |
| **NAME** | **FREQUENCY** | **TOPIC** |
| F.R. Goodwin | 9 | List of Witnesses, Moral and Social Aspects, Educational Aspects |
| N. Bishop Harman | 8 | Educational Aspects, Music |
| A.E. Newbould | 8 | Moral and Social Aspects |
| Leon Gaster | 6 | List of Witnesses, Lighting and Eye Strain |
| J.G. Legge | 5 | Educational Aspects, List of Witnesses |
| C.W. Crook | 4 | Moral and Social Aspects, The Cinema |
| A.P. Graves | 4 | List of Witnesses, Educational Aspects |
| Dr. C.W. Kimmins | 4 | Educational Aspects |

Voyant was giving me difficulty when it came to analyzing the names, but I was able to extract some of the most commonly referenced people by name, and which heading category they appear/are mentioned in. Ideally, I would have liked to do some exploratory analysis on the names of the individuals listed above and the word with the highest correlation frequency with their name, and see where that would lead, but neither the name recognition or correlation frequency was functioning properly at this point, which provided some time to reflect on the corpus as a whole.

While I may have uncovered some interesting observations, such as the focus of censorship of movie posters, and the association of cinema with the role of education with children, it certainly would provide a far more interesting analysis if more time allowed and a deeper dive on the text could be made. Associations of what the speaker(s) think or say with the text would have yielded interesting findings, and I would have liked to break down into the individual heading categories to see any sort of correlation or relationship between them. Having had the opportunity to really play with and get a feel with the text, I noticed there were several references to film genre(s)- had time allowed, I would like to have gone back to the cleaning/mark-up phase, and try to tag/mark all the references made to movie genres- I believe this would have yielded really interesting results, as films are often divided into genre when they are being discussed, and it would paint a better picture on what the council was against, based on which genre(s) were viewed as less favorable. I noticed reference to “American ‘crime’ films” in the text, so a sweep for genre to properly analyze it would hopefully yield interesting results. Doing another mark-up for the names of council/committees, to better understand the people and ideology behind their thoughts and reasonings. A list of council names would be useful in further textual analysis and visualizations, to see who belonged to what groups, what they are trying to do/accomplish, and their relevance to this text. Finally, it would be nice to have the time to better clean and organize the speaker files, so I could easily call upon only text that is a direct quote from the various speakers, and perform textual analysis on it to see positive/negative association with what the speaker has to say.