Text Analytics

Using SAS Enterprise Miner and R Studio

By Sazee S. (Maninderpreet Singh Puri)

Table of Contents

1. Introduction- Case Study A

The Case Study A is a data analysis assignment on the dataset *'news_cleaned_new.csv'* which consists of 17 variables with 12273 observations. The goal is to explore the impact of article properties and to investigate what properties of the article correlate with the high number of comments of the article on social media. Also, we intend to use SAS Enterprise Miner for keyword analysis. Out of the 17 variables *'headline', 'abstract', 'keyword'* and *'url'* are nominal variables, whereas *'pub_date'* is a nominal variable with dates. All other variables are numerical variables. I just used *'n_comments', 'n_words_content', 'n_words_headline', 'n_words_abstract', 'n_keywords',* and *'is_weekend'* parameters for analysis in R studio. For text analysis I used the *'abstract'* parameter and perform parsing, filtering, and clustering in the SAS Enterprise miner.

2. Task 1

After splitting the data into five article domains, I filtered top 20% articles in each domain with the highest number of comments. I investigated the correlation of number of comments with other variables in each domain by plotting correlation matrices and linear plots**.**

1. **Entertainment domain:**

Figure 1 Correlation matrix of Entertainment domain

As we can see in the correlation matrix that the number of comments (*'n_comments'*) is positively correlated to *'n_keywords'* (0.36), *'n_words_abstract'* (0.34), *'n_words_headline'* (0.33) and *'n_words_content'* (0.28). However, we can see that *'is_weekend'* (-0.18) is negatively correlated to *'n_comments'*. This means that if number of keywords or number of words in abstract/headline/ content are high, the number of comments on the articles will be high. This also shows that people at weekends don't like commenting about entertainment articles.

2. **Lifestyle domain:**

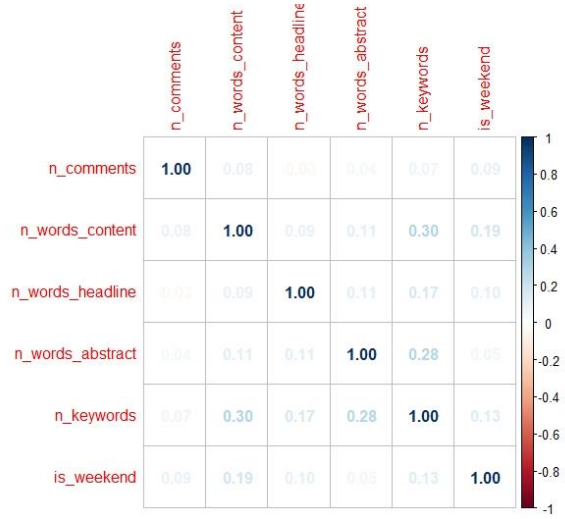

Figure 2 Correlation matrix of Lifestyle domain

As we can see in the correlation matrix that the number of comments (*'n_comments'*) is positively correlated to *'is_weekend'* (0.09), *'n_words_content'* (0.08), *'n_keywords'* (0.07) and 'n_words_abstract' (0.04), and. However, we can see that *'n_words_headline'* (-0.03) is negatively correlated to *'n_comments'*. This means that if number of keywords or number of words in abstract/ content are high or if it is a weekend, the number of comments on the articles will be high. This also shows that if number of words are high in headline the number of comments will be low. We can also notice that the impact of variables on number of comments is not very huge as all correlation values are not large.

3. Scitech domain:

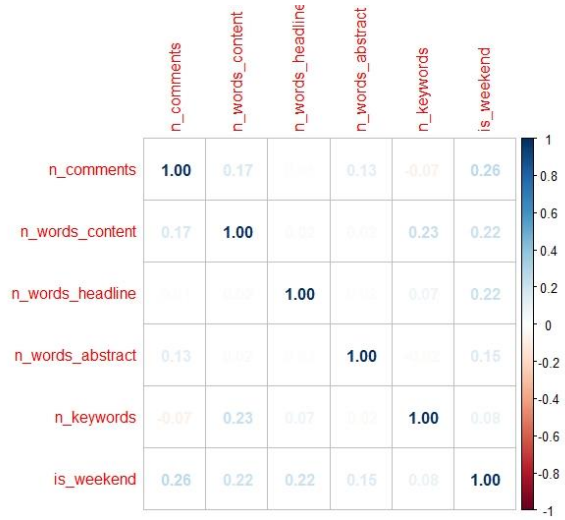

Figure 3 Correlation matrix of Scitech domain

As we can see in the correlation matrix that the number of comments (*'n_comments'*) is positively correlated to *'is_weekend'* (0.26), *'n_words_content'* (0.17) and *'n_words_abstract'* (0.13). However, we can see that *'n_keywords'* (-0.07) parameter is negatively correlated to *'n_comments'*. Also, we can notice that *'n_words_headline'* has no correlation with number of comments. This means that if number of words in abstract/ content are high or if it is a weekend, the number of comments on the science and technology articles will be high. This also shows that if number of keywords are high, the number of comments will be low.

4. Sports domain:

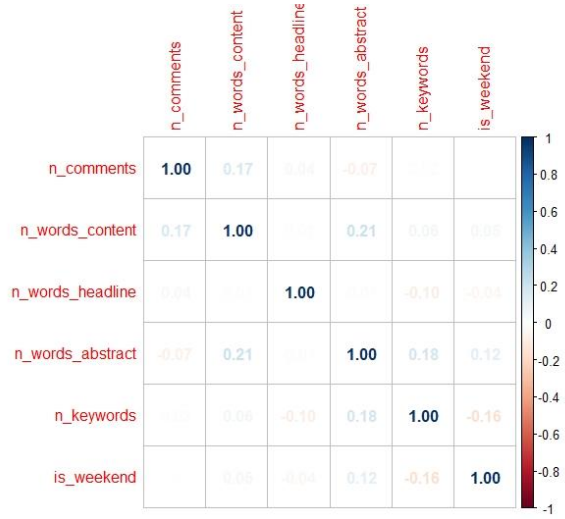

Figure 4 Correlation matrix of Sports domain

As we can see in the correlation matrix that the number of comments (*'n_comments'*) is positively correlated to *'n_words_content'* (0.17). However, we can see that *'n_words_abstract'* (0.07) and *'n_words_headline'* (0.04) parameters are negatively correlated to *'n_comments'*. Also, we can notice that *'n_keywords'* and *'is_weekend'* parameters have no correlation with number of comments. This means that if number of words in content are high, the number of comments on the Sports articles will be high. This also shows that if number of words in abstract are high, the number of comments will be low. The parameters will no correlations show that they don't impact the number of comments on the articles. We can also notice that the impact of variables on number of comments is not very huge as all correlation values are not large.

5. **World domain:**

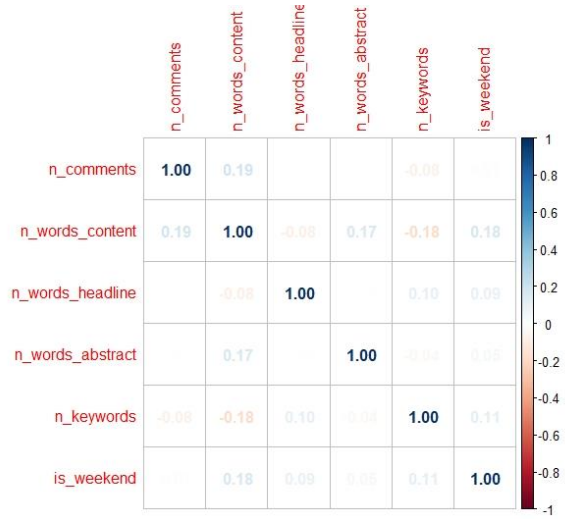

Figure 5 Correlation matrix of World domain

As we can see in the correlation matrix that the number of comments (*'n_comments'*) is positively correlated to *'n_words_content'* (0.19). However, we can see that *'n_keywords'* (-0.08) parameter is negatively correlated to *'n_comments'*. Also, we can notice that *'n_words_headline'*, *'n_words_abstract'* and *'is_weekend'* has no correlation with number of comments. This means that if number of words in content are high, the number of comments on the articles related to world will be high. This also shows that if number of keywords are high, the number of comments will be low. The parameters will no correlations show that they don't impact the number of comments on the articles.

I also plotted linear graphs for all variables against the number of comments in each domain to visualize the correlation and draw a line of best fit. All plots confirmed the results we got in correlation matrices are true.

3. Task 2

I split the *'news_cleaned_new.csv'* into five subsets to perform analysis in SAS Enterprise Miner. In SAS Enterprise miner I created five diagrams and used *'Text Parsing'* for tokenization, *'Text Filter'* for filtering out keywords in each article domain.

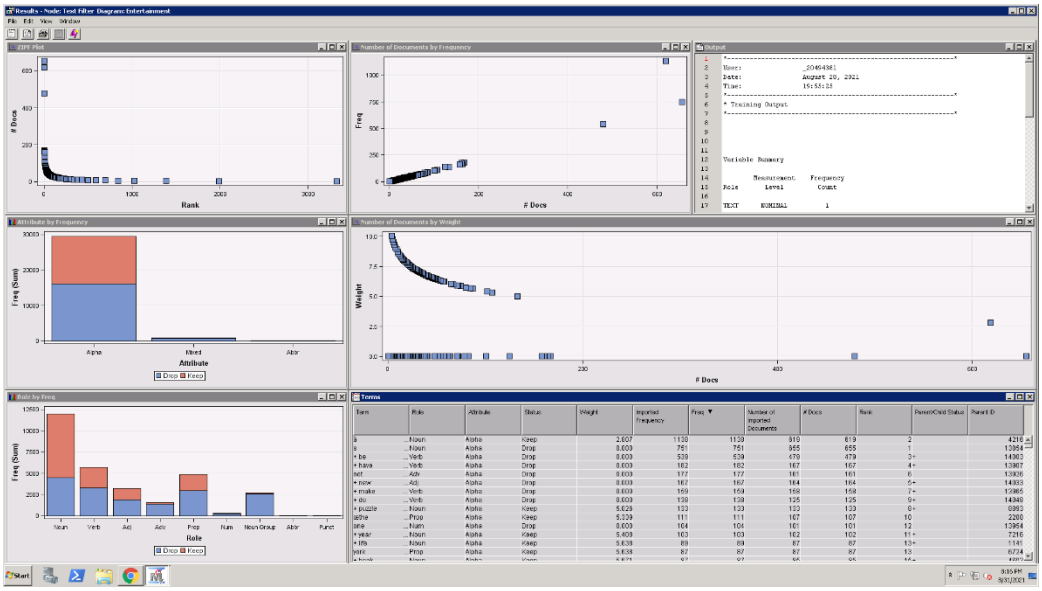

Figure 6 Results window for text filtering for Entertainment domain.

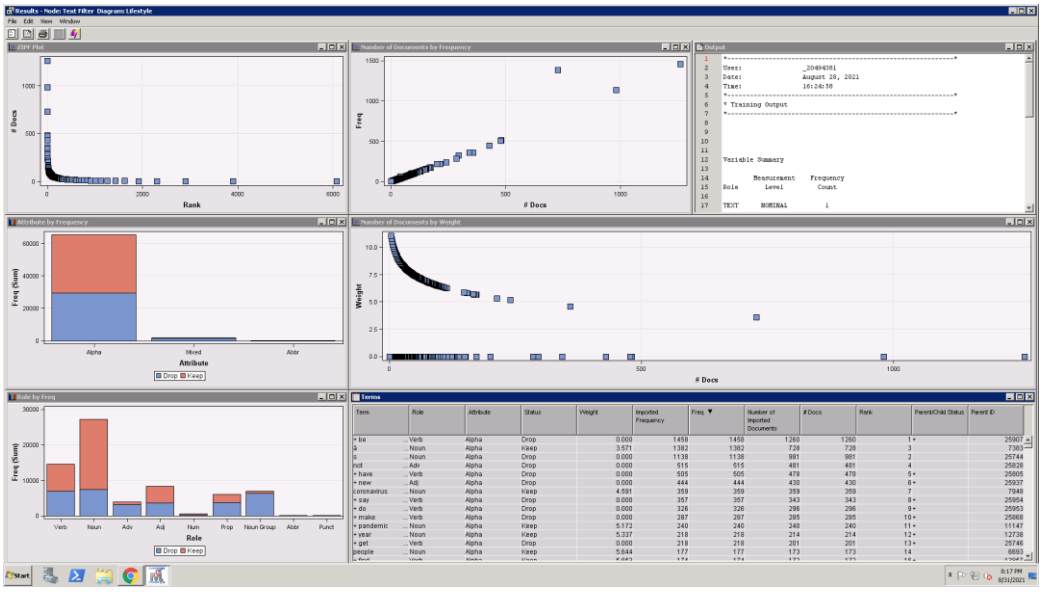

Figure 7 Results window for text filtering for Lifestyle domain.

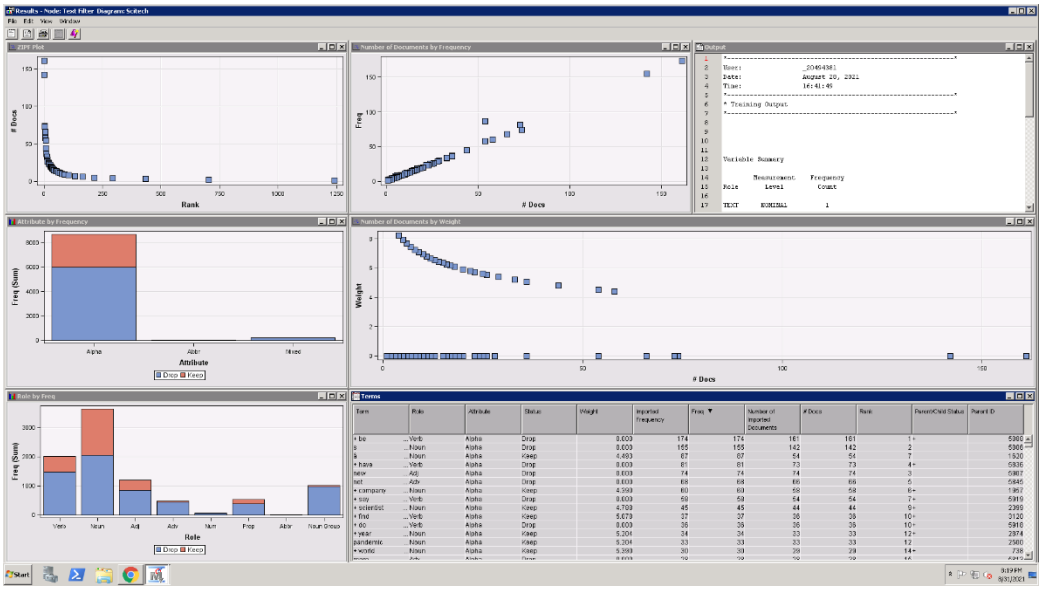

Figure 8 Results window for text filtering for Scitech domain.

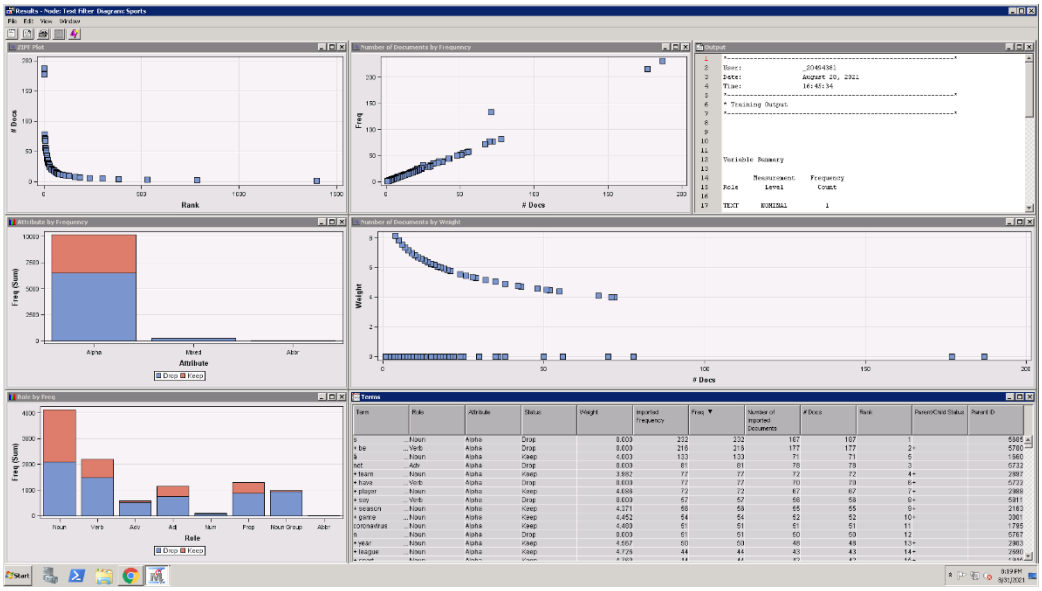

Figure 9 Results window for text filtering for Sports domain.

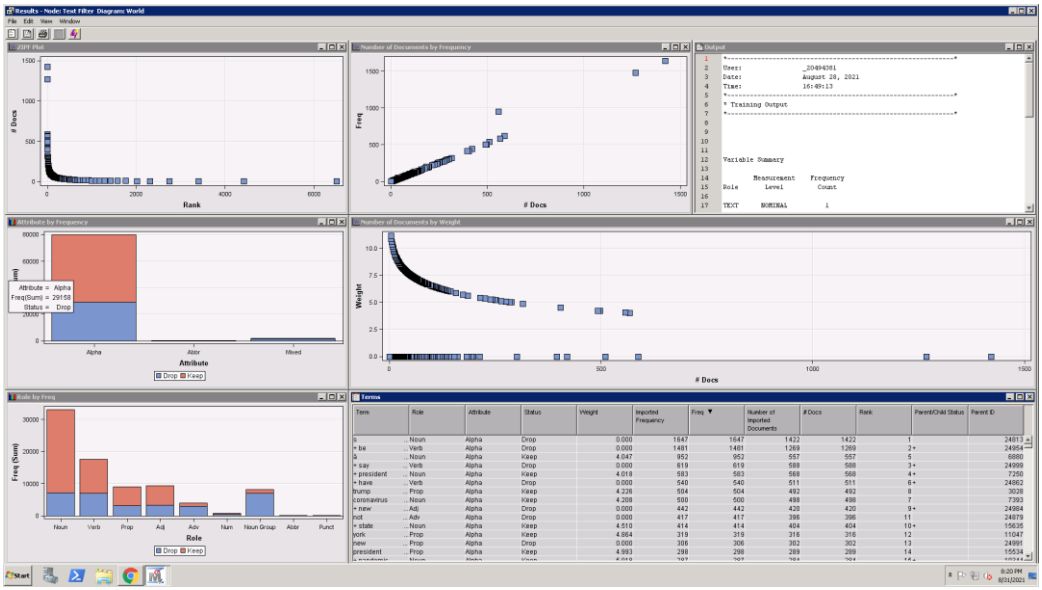

Figure 10 Results window for text filtering for World domain.

As we can see in the results window each results window shows a many different panels. All panels show plots for different measures like number of documents by frequency, output, number of documents by weight, attribute by frequency, role by freq and terms. To finds the top frequent keywords we can see in the '*Freq'* column in the *Terms* table. Some words are considered in analysis and while other words which are not considered as a topic are dropped in analysis. This can be checked in the '*Status'* column in *Terms* table.

To find the top five topics in each category I used '*Text topic'* function in the SAS Enterprise miner. According to the SAS Enterprise help manual (SAS, n.d.), the function finds the collection of terms that describe and characterize a main theme or idea as. I added this function in all the diagrams to get the top five topics occurring in each domain.

The results window shows multiple panels with information about number of documents by topics, topics terms, number of terms by topics, terms, topics, and outputs. The *'Topics'* column shows the top five topics which are listed below in the table:

Table 1 Top five topics in each domain

Top five topics list	Entertainment	Lifestyle	Scitech	Sports	World
1	â,œthe,+life,+book,+novel	+coronavirus outbreak,stock market,business news,stock,+outbreak	â,+world,œthe,+people,+scientist	â,œi,œit,+want,+pass	trump,president,+president,administration,impeachment
2	+week,+puzzle,+entry,+look,+solver	â,œthe,trump,+president,+joke	+company,pandemic,tech,internet,google	coronavirus,pandemic,coronavirus pandemic,+sport,+play	york,city,+reader tale,metropolitan,diary
3	times,york,debut,crossword,â	+risk,people,coronavirus,+patient,health	+find,+study,species,+suggest,+researcher	+team,+season,+win,+chief,+quarterback	â,+president,œi,œthe,+want
4	+columnist,ethicist,magazine's ethicist,+magazine,â	+find,+look,+event,+tune,online	+scientist,+turn,+year,+virus,+help	+player,+league,+season,+team,union	coronavirus,+pandemic,+case,people,+city
5	emily,henry,cox,rathvon,+puzzle	+year,+pandemic,+time,york,+work	social,media,social media,+network,+app.	+year,+game,+cheat,houston,astros	+state,biden,+president,democratic,joe

I split the whole dataset into top 20 % and bottom 20% datasets according to the number of comments using R-studio. I imported the subsets to SAS Enterprise Miner and conducted the cluster analysis to find common topics which span across data channels.

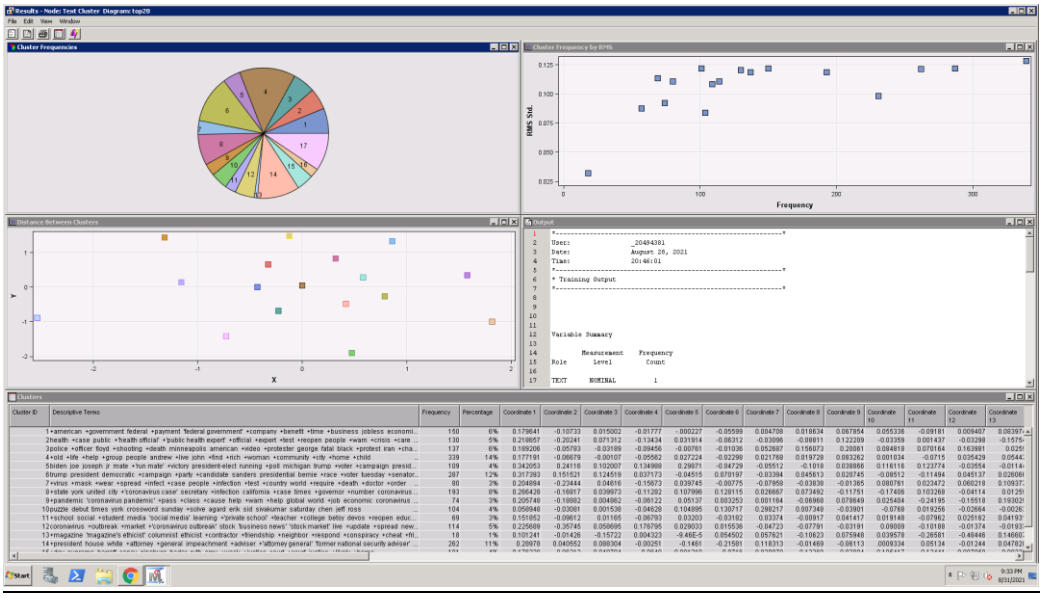

Figure 11 Results window showing clusters in top 20% of the whole data by number comment.

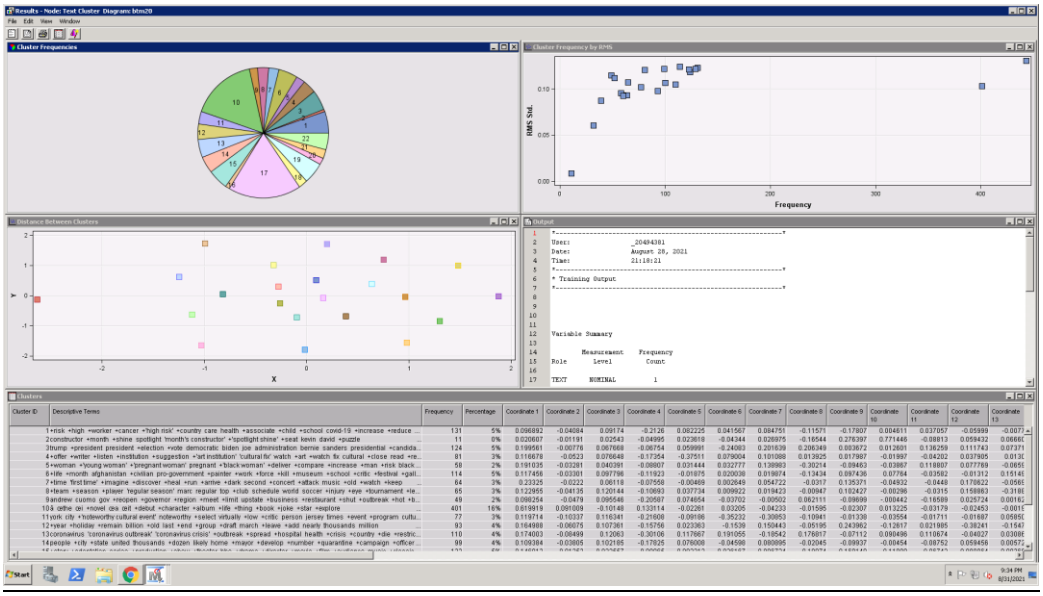

Figure 12 Results window showing clusters in bottom 20% of the whole data by number comment.

We can see in the results window that there are seventeen clusters of common topics across domains in top 20% of the whole dataset. The scatter plot in *Distance between Clusters* panels in the results window shows that the clusters are different to each other and are not closely related. It also shows the distribution of clusters using *'Cluster Frequencies'* plot. Cluster four is occurring more frequently in the dataset whereas cluster thirteen is the least frequent one. Similarly, we can analyze the results in the bottom 20% of the data set. We can see there are twenty-two clusters in the subset. The scatter plot in *Distance between Clusters* panels shows that some clusters are close to each other which are related. The pie chart shows cluster seventeen and ten are occurring most frequently whereas cluster two and sixteen are occurring least frequently occurring in the bottom 20% dataset.

4. Conclusion

We can see from the analysis that number of comments on the article domains are impacted differently in each domain. Some parameters affect the number of comments positively and some affect negatively. In the text analysis on '*abstract*' parameter we found clusters in top and bottom 20% which shows relation between topics across domains.

5. References

1. SAS. (n.d.). *SAS® Text Miner 14.1 Reference Help* [Press release].

6. Appendices

1. Screen shots of results window of cluster analysis in SAS Enterprise Miner.

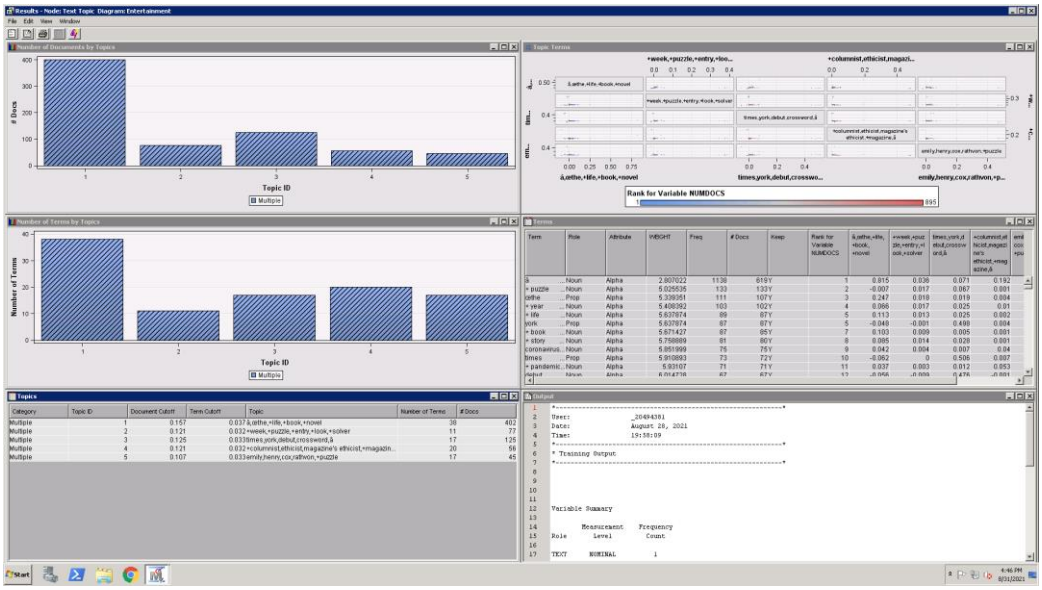

Figure 13 Top 5 topics in Entertainment domain.

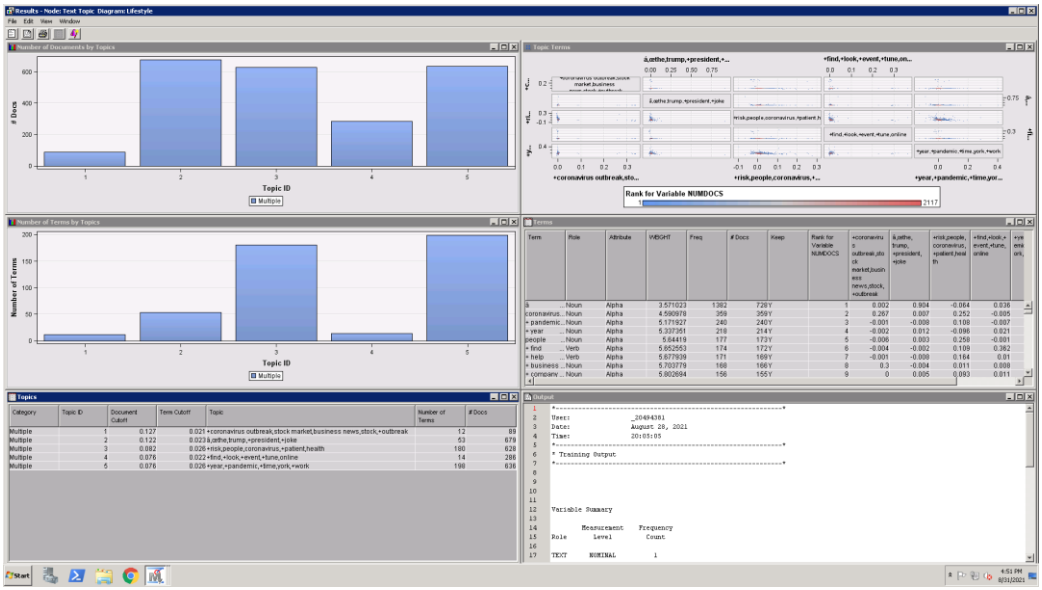

Figure 14 Top 5 topics in Lifestyle domain.

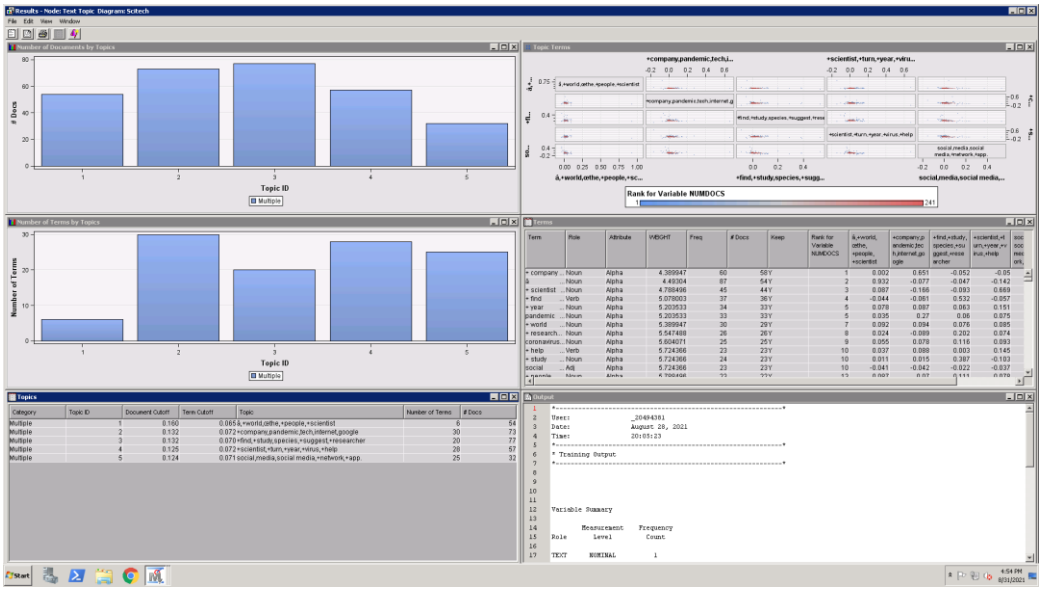

Figure 15 Top 5 topics in Scitech domain.

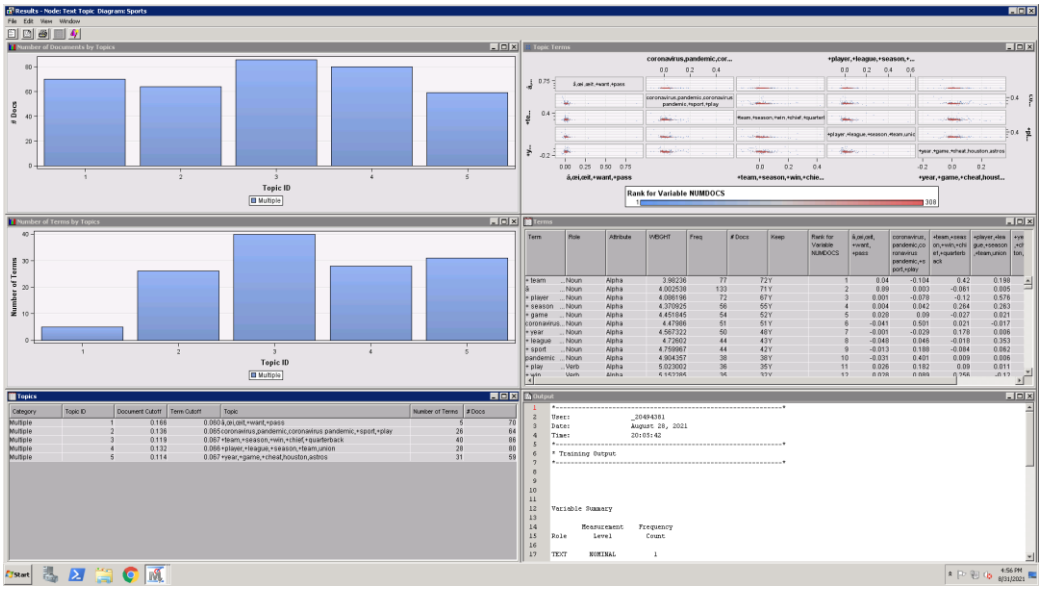

Figure 16 Top 5 topics in Sports domain.

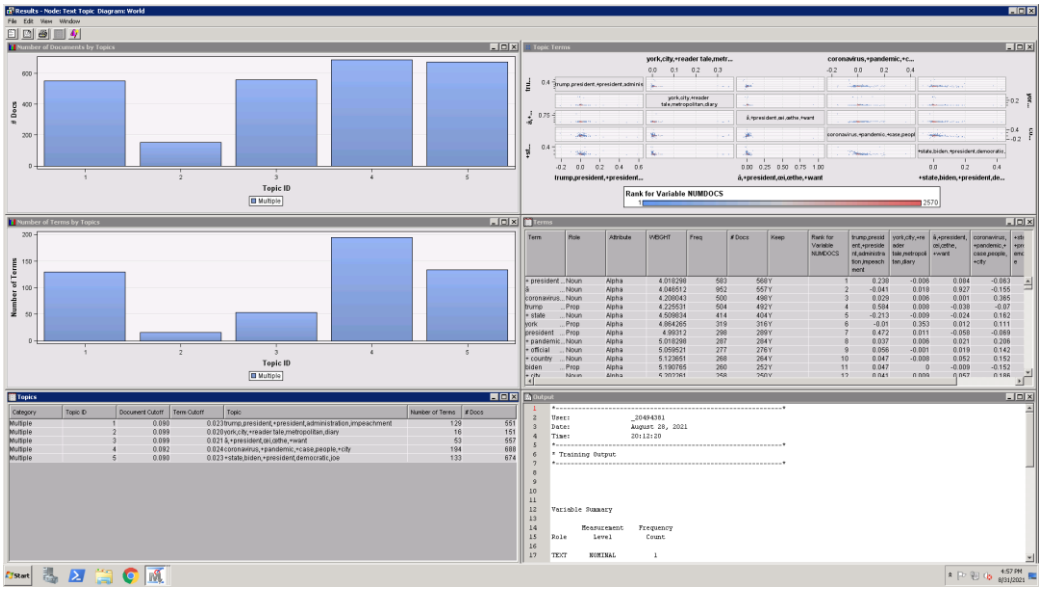

Figure 17 Top 5 topics in World domain.