Week 5

Description:

For this week's fifth assignment according to the task I have found one of the data sets with content of almost 20 Usenet and 20,000 messages from 20 newsgroups.

For this analysis, I focused on a subset of 2,992 messages from these three groups to compare technology discussions across different Usenet Newsgroups in the early 1990s (because after reading 4000 messages my R studio was not working and stopped between the process that’s why I have selected 2,992 messages subset).

This comparison is an amazing opportunity to explore how technology was discussed in various contexts like graphics, space science, and general discussions during the early days of the internet.

Following the instructions for text mining, I used certain techniques for example TF-IDF analysis, topic modeling, sentiment analysis, and n-gram analysis. I would try to understand the patterns in language use, prevalent topics, and emotional tones across these newsgroups.

This analysis works on these main points:

1. Analysis of words within newsgroups using TF-IDF metric: This metric measures the importance of a word within a document to a collection of documents, highlighting words that are frequent in a specific document but rare across the entire dataset.
2. Topic modeling analysis: It's used to identify hidden topics(hidden) within a collection of documents by grouping words that frequently appear together.
3. Sentiment Analysis by word: It assesses the emotional tone of the text by analyzing individual word's sentiments.
4. Sentiment analysis by the message: This method evaluates the overall sentiment of an entire message based on the combined sentiment of its words.
5. N-gram analysis: Identifies recurring sequences of words within a text to detect patterns and context in language.

Graph explanation:

1. Top 10 most Distinctive terms by newsgroup (TF-IDF analysis):

A graph with multiple colored bars

Description automatically generated with medium confidence

This graph explains the most common word phrases used in three different newsgroups like comp. Graphics, sci. Space and talk.politics.misc.

According to the graph, each newsgroup has its own unique set of frequently used bigrams.

This graph helps us understand the topics discussed and the language patterns used in each newsgroup. The length of the bars explains how often each bigram was used.

1. Top terms for each topic (Topic modeling analysis):

A screenshot of a graph

Description automatically generated

This graph helps us understand about Top 10 terms for each topic and the idea of what individuals were talking about in this specific conversation. It shows that they were interested in files, images, and how to make them smaller.

Meanwhile, it is about files and images like more common words are file, image, format, gif, jpeg, and bmp all related to files and images.

The terms size and compression suggest that in the 1990s people were discussing how to make files smaller.

1. Sentiment analysis by newsgroup:

A screenshot of a graph

Description automatically generated

Sentiment analysis by newsgroup graph effectively visualizes the overall sentiment polarity of the discussions within the three analyzed newsgroups.

By examining the sentiment scores, it is easy to understand the emotional tone and attitude expressed in these online communities.

This graph also covered three newsgroups: comp. Graphics, sci. Space and talk.politics.misc.

For comp. Graphics newsgroup has a significantly higher sentiment score which is a more positive tone, the sci. Space newsgroup score is close to zero suggesting a more balanced tone in discussions and for the talk. Politic. Misc newsgroup has a lower sentiment score and a more negative tone in discussions.

1. Average sentiment by newsgroup:

A screenshot of a computer

Description automatically generated

This graph specifically focuses on the fourth facet of talk. Politics. Misc, visualizes the average sentiment score for individual messages within the talk. Politics. Misc newsgroup.

Meanwhile, talk. Politics. The misc newsgroup has a negative average sentiment per message. The length of the bar indicates the magnitude of the average sentiment per message. A longer bar would represent a stronger positive or negative sentiment.

1. N-gram analysis:

A screenshot of a graph

Description automatically generated

In this n-gram analysis, the Top 10 bigrams by newsgroup graph tried facet effectively visualize the linguistic features and subject matter prevalent in the talk.

Politics. Misc newsgroup. By examining the most frequently used two-word combinations.

As well as I can understand deeper about topics discussions and the language patterns employed within this online community.

Conclusion:

This project allowed me to understand 1990s trends and which things have changed over time as well as which is not. Overall, this data set explains certain things like over the period language patterns, slang, human thinking processes, and trends changing all the time. Most technologies are changing nowadays as compared to the 1990s.

I discovered that teaching talk was everywhere, even back then! Computer graphics geeks were interested in image formats and 3D stuff, while space enthusiasts people out over NASA and orbits. I also understand the surprising twist even the politics group had tons of each chatter, it surprised me when I understood this thing.

Meanwhile, data explained to me each group had its selective languages and patterns. For example, space nerds were super positive, graphics geeks were neutral, and the political crowd was not exactly spreading sunshine. In the 1990s individuals were already connecting the ‘tech’ word to big societal issues, especially in the political groups and these people were more forward thinkers in the 1990s.

For this era, I predicted mostly yes word command, but the amount of teach talk in non-tech was about tech’s future impact. I thought this thing because when I scroll data yes is very common on every page. Even they were more used to teach words than I thought.

Overall, I learned it's not easy to predict anything directly, and for this assignment, I thought like any data set we used for analysis, but I found out about a certain amount of data everyone used for analysis and it also depends on their device as well. The internet revolution is fading. Some topics have not changed much but context sure has changed. Even in the 1990s, it was wild to spot internet user information and future impact and discuss stuff we are still grappling with today.