**ALY6040 90248 Data Mining Applications SEC 01 Summer 2023 CPS [BOS-D-HY]**

**Module 5 Assignment — Technique Practice**

A black and white logo

Description automatically generated

**NORTHEASTERN UNIVERSITY**

**College of Professional Studies, Boston, MA, 02215.**

**Submitted by**

Hanchan Xu  [xu.hanc@northeastern.edu](mailto:xu.hanc@northeastern.edu)

Zihan Ma [ma.zihan1@northeastern.edu](mailto:ma.zihan1@northeastern.edu)

**Instructor**

 Prof. Kasun Samarasinghe

**Date**

08/09/2023

**Text Mining Analysis Report on "I Have A Dream" Speech Data**

**Introduction**

The importance of extracting valuable insights from textual data cannot be overstated in today's data-driven world. Text mining, a multi-faceted approach to this, can reveal patterns, associations, and themes that are often hidden in vast amounts of textual data. In this report, we perform an in-depth analysis of the iconic "I Have A Dream" speech using text mining techniques for understand how data mining processes work and what insights they can yield.

**Code Walkthrough**

The code provided primarily focuses on the text preprocessing and analysis using the tm, wordcloud, and other associated packages in R.

Loading Required Libraries: The initial lines load necessary libraries, like tm for text mining and wordcloud for visual representation.

Reading the Text: The readLines function imports the text of the speech. The warning indicates there might be an issue with how the file ends, which often doesn’t impact analysis.

Creating a Corpus: A corpus is a collection of documents, in this case, segments of the speech. The inspect(docs) function provides an overview of this corpus.

Text Preprocessing:

Special Character Removal: The tm\_map function with the toSpace content transformer replaces various special characters with spaces.

Text Conversion: The entire corpus is converted to lowercase to maintain consistency.

Removing Numbers & Stop Words: Numbers and common English words that don't add value to frequency analysis are removed.

Punctuation Removal: All punctuations are discarded using the removePunctuation function.

White Space Elimination: Extra white spaces are removed for a cleaned data set.

Term Document Matrix (TDM): The TermDocumentMatrix function converts the corpus into a matrix format where each term is associated with its frequency.

**Analysis**

**1. Word Frequency Analysis**:

* **Process**: The Term Document Matrix (TDM) was computed using the **TermDocumentMatrix** function. By converting the corpus into matrix format, each term's frequency within the speech was identified.
* **Insight**: From the list of words and their respective frequencies, certain terms stand out prominently, including 'will' (17 times), 'freedom' (13 times), and 'ring' (12 times). Their high frequencies underscore the speech's themes of determination, the call for liberty, and the echoing demand for justice.
* **Depth**: Diving deeper, other significant terms like 'dream' (11 times) and 'day' (11 times) also emerge. These words suggest aspirations for a better future and an imminent call to action, respectively. Together, the frequency of these words paints a picture of the speech's impassioned call for change.

**2. Word Cloud Visualization**:

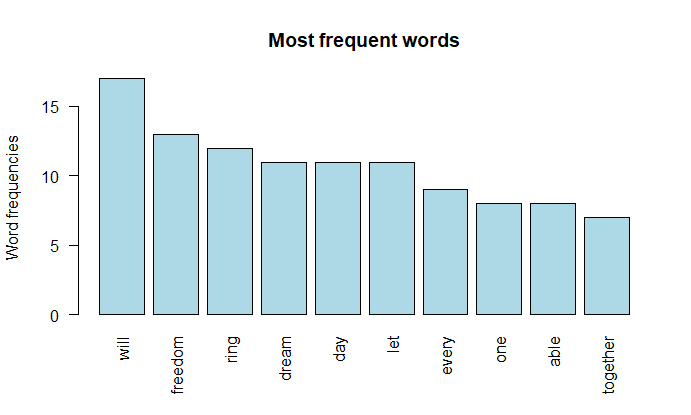
* **Process**: A visual representation was constructed through a word cloud. This offers a direct visualization of word frequencies by adjusting the size of words proportionally to their frequency.
* **Insight**: The word cloud emphasizes terms like 'freedom', 'will', and 'dream' as they appear larger than others. This tool visually reinforces the most important or recurrent themes within the speech without delving into numerical details.
* **Depth**: Colors, rotation, and positioning within the word cloud can influence the viewer's perception. The use of the color palette 'Dark2' and the setting **rot.per=0.35** ensures that the cloud is both aesthetically pleasing and readable. Words are effectively scattered, making it easier to discern even those terms that might appear less frequently.

**3. Term Association**:

* **Process**: The **findAssocs** function computes how often terms co-occur within the speech, suggesting a connection or common theme.
* **Insight**: The term 'freedom' is notably associated with words like 'let' (0.89) and 'ring' (0.86). This hints at a recurrent theme within the speech - the call for freedom to resonate throughout the nation.
* **Depth**: Additional associations with 'freedom', such as 'mississippi', 'mountainside', and 'state', indicate the geographical scope and depth of this call for liberty. These associations also hint at specific challenges or contexts where the call for freedom was especially poignant.

**4. Frequency Bar Plot**:

* **Process**: The bar plot visualization was generated to compare the top 10 most frequent words directly.
* **Insight**: By juxtaposing words like 'will', 'freedom', 'ring', 'dream', and others, the bar plot offers a clear hierarchical understanding of term significance within the speech.
* **Depth**: The choice of 'lightblue' as the bar color and the explicit labeling on the y-axis ensures that the viewer can quickly understand the word frequencies. Beyond just understanding, the juxtaposition allows for comparisons, potentially highlighting the relative importance or focus the speaker placed on certain themes or ideas.



**Interpretation and Recommendations**

**Interpretation:**   
The application of the code on Martin Luther King Jr.'s "I Have A Dream" speech provides not just a simple overview, but a magnified, quantitative lens into its core essence. At the heart of the speech are the recurring terms 'freedom', 'dream', and 'ring'. Their recurrence isn't coincidental; it offers profound insights into King's desire for liberty, his profound sense of hope, and a resounding call for action. Additionally, the strong associations found between words like 'freedom' and 'ring' reveal the interconnectedness of concepts and provide a cohesive narrative thread running throughout the speech. In the broader context, this is a testament to how, through text mining, one can delve into a sea of words and emerge with nuggets of thematic gold.

**For those analyzing such texts in the future:**

1. **Advanced Data Collection Methods**: Using a local file, as done in the current analysis, is efficient for singular texts. However, for more expansive analyses, especially those comparing multiple speeches or texts, leveraging online data scraping techniques would be invaluable. This would enable analysts to tap into a plethora of data sources ranging from historical archives to modern-day transcripts.
2. **Synonym Analysis Integration**: The English language is rich and diverse, often offering multiple words with similar meanings. By integrating a robust synonym checker into the analysis pipeline, words that carry the same or similar meanings can be grouped. This ensures that the thematic weight of a particular concept isn't diluted across synonyms, providing a richer and more accurate term frequency.
3. **Deep Dive with Sentiment Analysis**: Beyond just word frequencies and associations, understanding the emotional undertone of a text can add a layer of depth to the insights. Sentiment analysis tools can dissect the text and offer insights into the predominant emotions conveyed—be it hope, despair, joy, or anger. In the context of historical speeches, this could reveal the speaker's emotional state and the intended emotional impact on the audience.
4. **Comparative Data Incorporation**: By juxtaposing the word frequencies from one text with another, researchers can unveil fascinating insights. For instance, comparing the "I Have A Dream" speech with another significant speech from the same era might highlight common societal themes or unique, speaker-specific patterns. This not only aids in understanding the zeitgeist of the period but also in discerning the distinct voice of individual orators.
5. **Leverage Visualization Tools**: The current analysis employed word clouds and bar graphs. Future analyses can benefit from more advanced visualization tools, such as network graphs for word associations or heatmaps for term frequency distributions across different sections of the text.
6. **Incorporate Contextual Analysis**: While individual word frequencies are insightful, often the true meaning lies in the context. Phrase or n-gram analyses, which consider sequences of words, can offer more contextual insights, revealing commonly used phrases or rhetorical devices that might be pivotal to the message.

Additional Data Incorporation: By comparing word frequencies with other speeches or texts, one can discern common themes or unique patterns specific to each document.

**Conclusion:**

By diving deep into the technicalities of the provided R code, we've uncovered the methodologies employed in text mining. This exercise has demonstrated the power of textual analysis, not just for understanding content but for revealing intricate patterns, associations, and themes. The insights and recommendations provided here are not just applicable to historical speeches but can be extended to any large body of text, from corporate reports to literature, showcasing the versatility of text mining.