**Assignment 5: Text Mining & Word Cloud Fundamental**

Yash S

2025 – 05 – 14

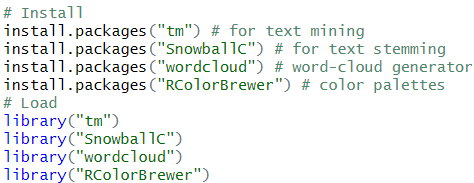
Northeastern University: College of Professional Studies

**Introduction**

**Introduction**  
In the realm of analytics, **text mining** has emerged as a powerful tool for deriving **insights from unstructured textual content**—a crucial capability in today’s data-driven landscape. This report applies **R-based text mining techniques** to dissect one of history’s most iconic speeches: **Martin Luther King Jr.’s "I Have a Dream."** The goal is to methodically **clean, process, and analyze the speech**, uncovering its core linguistic patterns and thematic structures. Using **visual tools such as word clouds and bar charts**, we aim to surface recurring keywords that define the speech’s impact. Guided by the **CRISP-DM framework**, the analysis progresses through structured stages—**data understanding, preparation, modelling, and interpretation**—ensuring that both technical rigor and contextual relevance are maintained throughout. This exercise not only deepens our technical proficiency but also highlights the intersection of **social impact and data science.**

**Dataset Overview**  
The dataset used in this analysis comprises the **full transcript of Dr. Martin Luther King Jr.’s iconic 1963 "I Have a Dream" speech**, obtained from a **publicly available online source**. Recognized globally as a pivotal moment in the **civil rights movement**, the speech is filled with powerful language addressing **justice, equality, freedom, and racial harmony**. Its **rich linguistic structure and emotional weight** make it an excellent case for applying **text mining techniques** to explore **word frequency patterns and thematic associations**, offering both **technical insights and cultural significance**.

**Package Installation and Loading**

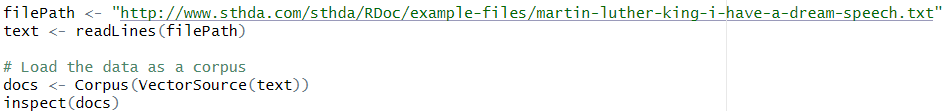
****

* Before diving into text mining, it’s essential to **install and load the necessary R packages** to support preprocessing and visualization tasks.
* ****: Installs the **Text Mining** package, which provides core tools for handling and cleaning textual data.
* ****: Adds stemming capabilities, helping reduce **words to their root forms** for more consistent analysis.
* ****: Enables the creation of **word clouds**, a useful tool for visualizing word prominence.
* ****: Provides **color palettes** that enhance the visual quality and readability of plots.
* Once installed, each package is activated using **library()**, making their functions accessible in the current R session:
  + ****
* These libraries together form the **foundation for text preprocessing, transformation, and visualization** in this project.

**Data Analysis**

**Explanatory Descriptive Analysis (EDA)**

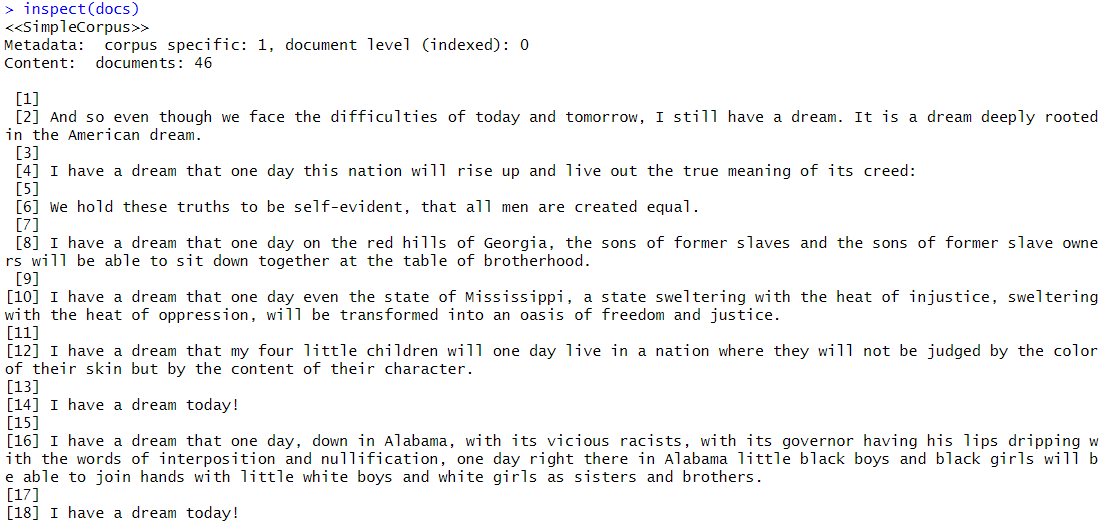
**Data Import and Corpus Creation:**

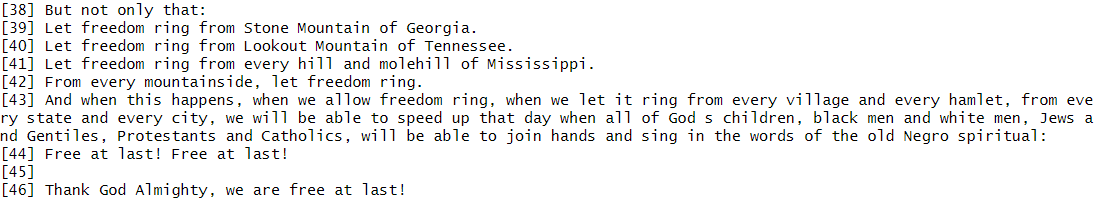
****

Code Explanation-

* ****  
  *Defines a* ***string variable*** *storing the* ***URL*** *of the source text, enabling real-time access to the speech directly from an online repository.*
* ****  
  *Uses the* ***readLines() function*** *to read the file line-by-line and store it as a* ***character vector****, maintaining the original structure.*
* ****  
  *Transforms the* ***character vector*** *into a* ***corpus object*** *using Corpus() and VectorSource(), which is required for advanced* ***text mining operations****.*
* ****  
  *Displays the* ***internal structure*** *of the corpus to validate successful import and ensure that each document line is properly indexed and accessible.*

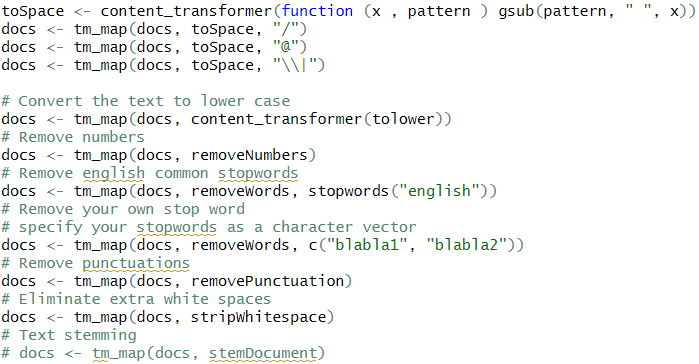
Terminal Output-





The output confirms that the **corpus has been successfully created**, containing **46 individual documents**, each representing a line or paragraph from the speech. Displayed as a <<SimpleCorpus>> object, this structure allows us to perform targeted text mining operations on each segment. What stands out is the consistency and clarity of the formatting—each entry in the corpus retains the **original narrative flow**, such as repeated phrases like *“I have a dream”*, which are **crucial linguistic anchors** for both frequency and association analysis. This step validates that the **data import and corpus initialization were executed correctly**, setting a solid foundation for all downstream analytics.

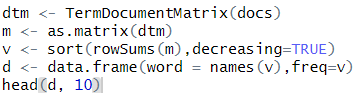
**Text Preprocessing:**

****

Code Explanation-

* ****  
  *Creates a* ***custom transformation function*** *that replaces any matching* ***pattern*** *in the text with a* ***single whitespace****, using gsub() internally.*
* ****  
  *Replaces all* ***forward slashes ("/")*** *in the corpus with a space to eliminate non-informative symbols during analysis.*
* ****  
  *Substitutes the* ***"@" symbol*** *with a blank space, typically useful for cleaning email references or social mentions.*
* ****  
  *Replaces the* ***pipe symbol ("|")****, which is often used as a delimiter, ensuring it doesn't interfere with tokenization.*
* ****  
  *Converts all text to* ***lowercase*** *to standardize words like “Freedom” and “freedom” into a single token for accurate frequency analysis.*
* ****  
  *Removes all* ***numeric values*** *from the text, which are generally irrelevant in speech-based thematic analysis.*
* ****  
  *Filters out* ***common English stopwords*** *(e.g., “the”, “is”) that provide no semantic weight for content-based analysis.*
* ****  
  *Allows for removal of* ***custom stopwords****, letting the analyst eliminate additional non-contextual terms based on domain expertise.*
* ****  
  *Strips away* ***punctuation marks*** *such as commas and periods to support clean word tokenization.*
* ****  
  *Reduces multiple or irregular* ***white spaces*** *to single spaces, ensuring proper text formatting for matrix construction.*
* ****  
  *Optional step to perform* ***stemming****, reducing related terms (e.g., “dreamed”, “dreaming”) to their* ***root form ("dream")****, though currently commented.*

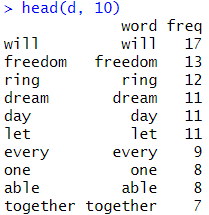
**Term Frequency Analysis (Extraction & Structuring):**

****

Code Explanation-

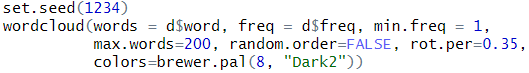
* ****  
  *Creates a* ***term-document matrix*** *where rows represent unique words and columns represent documents, capturing the* ***frequency of each word per document****.*
* **** *Converts the sparse* ***TDM object*** *into a* ***standard R matrix****, enabling easier manipulation, summarization, and aggregation using native matrix functions.*
* ****  
  *Calculates total frequency per word using* ***rowSums()****, then* ***sorts them in descending order*** *to prioritize most frequently occurring terms.*
* ****  
  *Creates a* ***data frame*** *where each row contains a word and its frequency, making it ready for plotting and tabular analysis.*
* ****  
  *Displays the* ***top 10 most frequent words*** *and their counts to provide an at-a-glance summary of dominant speech terms.*

Terminal Output-



The output of reveals the **top 10 most frequent words** from the term-document matrix, arranged in **descending order of frequency**. Words like **"will" (17 times)**, **"freedom" (13 times)**, and **"ring" (12 times)** appear most prominently, reflecting the **core thematic elements** of Dr. King’s speech—emphasizing **agency, liberty, and unity**. This frequency snapshot not only validates the success of the preprocessing pipeline but also highlights the **linguistic anchors** that shaped the emotional and rhetorical power of the text.

**Word Cloud Generation from Frequency Data:**

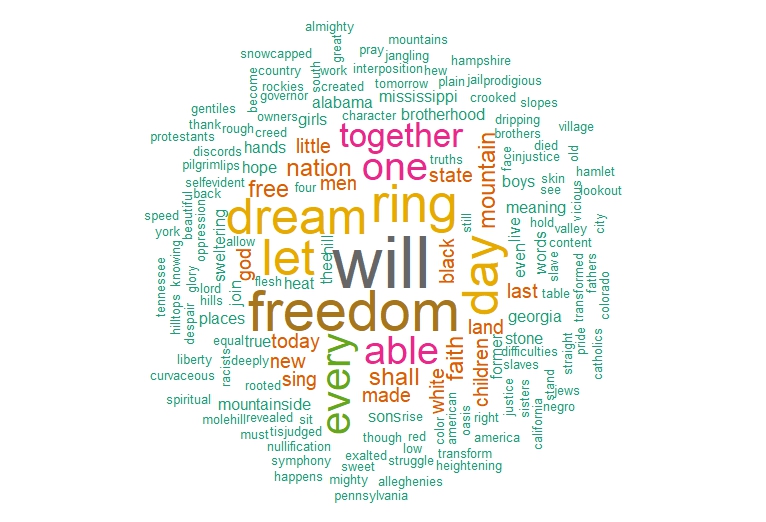
****

Code Explanation-

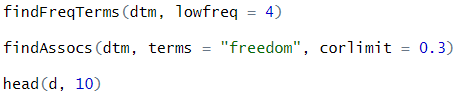
* ****  
  *Ensures* ***reproducibility*** *of the word cloud layout by fixing the random number generation, so visual results stay consistent across executions.*
* ****  
  *Draws words from the d data frame; only includes words with a* ***minimum frequency*** *of* ***1*** *to maximize visual inclusiveness.*
* ****  
  *Limits the total number of displayed words to* ***200****, balancing detail and readability in the cloud.*
* ****  
  *Ensures* ***most frequent words are centrally positioned****, improving visual focus and making key terms more noticeable.*
* ****  
  *Rotates* ***35% of the words vertically****, adding diversity to the layout and reducing visual monotony.*
* ****  
  *Applies a visually appealing* ***8-color palette*** *from the RColorBrewer package using the* ***"Dark2"*** *scheme for contrast and clarity.*

Visualization Output-

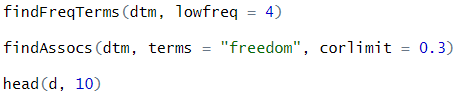
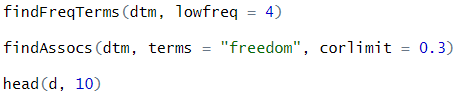
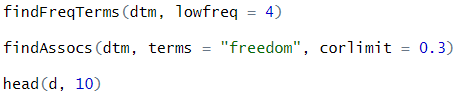
The generated **word cloud** visually represents the **most frequent terms** from the cleaned corpus, with larger font sizes indicating **higher frequency**. Words like **“will”**, **“freedom”**, **“dream”**, and **“ring”** stand out prominently, reflecting the **rhetorical anchors** and **thematic emphasis** of Dr. King’s speech. This image effectively transforms raw frequency data into a compelling visualization that highlights the speech’s emotional and ideological core—**liberty, unity, and hope**. The color diversity, applied using **RColorBrewer’s “Dark2” palette**, enhances contrast and readability, making the cloud both informative and engaging.

****  
A vibrant word cloud highlighting the most frequent and emotionally charged terms in Dr. King’s speech, emphasizing themes of freedom, unity, and hope.

**Exploring Frequent & Associated Terms:**

****

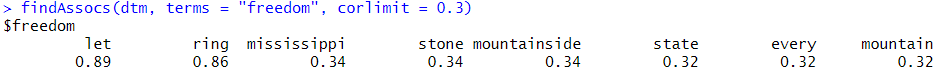
Code Explanation-

* ****  
  *Returns all terms from the* ***term-document matrix*** *that appear* ***at least four times****, helping filter out noise and low-impact words.*
* ****  
  *Identifies terms with a* ***correlation ≥ 0.3*** *to* ***"freedom"****, revealing words that frequently appear in similar* ***semantic or contextual proximity****.*
* ****  
  *Displays the* ***top 10 most frequent words*** *again for reference, reaffirming the dominance of key terms prior to visual or contextual analysis.*

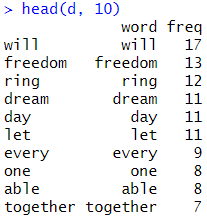
Terminal Output-



The output displays all words from the corpus that appear **four or more times**, indicating their **relevance and recurrence** in the speech. Words such as **“dream”**, **“freedom”**, **“ring”**, and **“together”** reflect the speech’s central themes of **hope, unity, and civil rights**. Filtering by frequency helps eliminate noise while highlighting **semantically important** terms that warrant deeper analysis or visual representation, such as in bar plots or word clouds.



The output identifies words with a **correlation coefficient ≥ 0.3** to the term **“freedom”**, indicating how frequently they co-occur across the speech. Strong associations with words like **“let” (0.89)** and **“ring” (0.86)** reinforce the phrase **“let freedom ring”**, a rhetorical pillar of the address. Other connected terms such as **“mountainside”**, **“every”**, and **“state”** reflect Dr. King’s vision of **universal liberty across all geographic and societal boundaries**. This analysis not only confirms the **semantic focus** of the speech but also maps out its **narrative structure and emotional cadence**.



This output displays the **top 10 most frequent words** in the processed corpus, with **“will”**, **“freedom”**, and **“ring”** leading in frequency. These terms not only reflect the **linguistic rhythm** of Dr. King’s oratory but also emphasize the **core themes of aspiration, liberty, and unity**. Their high occurrence validates the **effectiveness of preprocessing**, and their presence offers a rich foundation for deeper text mining techniques such as sentiment analysis, topic modelling, or rhetorical pattern detection.

**Barplot of Most Frequent Terms:**

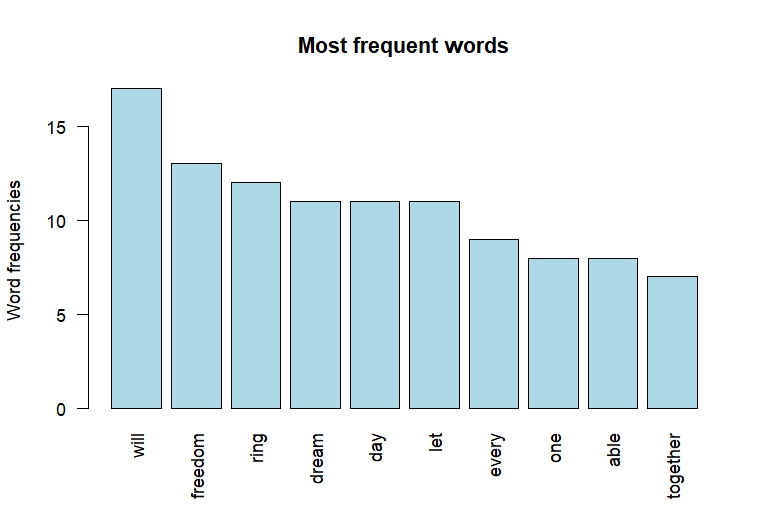
****

Code Explanation-

* ****  
  *Generates a vertical barplot for the* ***top 10 most frequent words****, based on the freq column of the cleaned data frame d.*
* ****  
  *Rotates the* ***axis labels vertically (90 degrees)*** *for better readability, especially when plotting longer or overlapping words.*
* ****  
  *Defines the* ***word labels for each bar****, aligning them directly with their corresponding frequency value.*
* ****  
  *Sets the* ***bar color*** *to light blue for visual appeal and to maintain a non-distracting, clean academic presentation.*
* ****  
  *Adds a* ***main title*** *to the chart, clearly indicating the content and purpose of the visualization.*
* ****  
  *Labels the* ***y-axis*** *to clarify that the bar heights represent the* ***frequency*** *of each word in the speech.*

Visualization Output-

This barplot visualizes the **top 10 most frequent words** from the processed speech corpus, with **“will”**, **“freedom”**, and **“ring”** standing out as dominant linguistic elements. The vertical orientation of the **x-axis labels (via )** enhances readability, and the **light blue bars** offer a clean and professional aesthetic. Each bar’s height directly reflects the **term’s frequency**, reinforcing the thematic emphasis of **hope, liberty, and unity** found in Dr. King’s speech. This plot serves as a clear, quantitative complement to the more abstract representation offered by the word cloud.

****A clear barplot showcasing the **top 10 most frequent words** in the speech, emphasizing core themes like **freedom**, **will**, and **unity** through frequency counts.

**Conclusion**

**Conclusion:**

This project successfully applied **text mining techniques** to uncover rhetorical patterns and core themes within Dr. Martin Luther King Jr.’s historic speech. The analysis demonstrated how **frequent term analysis**, **association mining**, and **visualizations** like **word clouds** and **barplots** can illuminate the linguistic backbone of powerful narratives. Words such as **“will”**, **“freedom”**, and **“dream”** not only appeared often but served as **emotional anchors**, supporting the speech’s structured call to action. These findings underscore the value of **data mining in interpreting historical texts**, enabling analysts to draw meaningful conclusions from **unstructured language data** while opening new avenues for comparative rhetorical studies.

**Key Findings:**

The dominance of **“freedom”**, **“dream”**, and **“ring”** affirms that the speech centers around **civil liberties and collective progress**. Association mining revealed strong correlations between **“freedom”** and terms like **“let”**, **“ring”**, **“state”**, and **“mountain”**, emphasizing the strategic use of **imagery and repetition** to reinforce themes of **unity and justice**.

**Recommendations:**

To build on this analysis, the following steps are recommended:

* **Apply stemming** to consolidate variations of similar words (e.g., *ring*, *rings*) and improve frequency accuracy.
* **Conduct sentiment analysis** to quantify the emotional tone and rhetorical intensity of the speech.
* **Compare with other speeches** to identify thematic or stylistic shifts over time in civil rights discourse.
* **Use topic modeling (e.g., LDA)** to cluster recurring concepts and uncover underlying thematic structures.

**Works Cited**

* Bluman, A. (2018). *Elementary Statistics: A Step-by-Step Approach* (10th ed.). McGraw-Hill.
* Kabacoff, R. (2022). *R in Action: Data Analysis and Graphics with R and tidyverse* (3rd ed.). Manning Publications.
* King Jr., M. L. (1963, August 28). *I Have a Dream [Speech transcript]*. American Rhetoric. <https://www.americanrhetoric.com/speeches/mlkihaveadream.htm>
* Fellows, I. (2014). *Wordcloud: Word clouds*. R package version 2.6. <https://CRAN.R-project.org/package=wordcloud>
* Feinerer, I., Hornik, K., & Meyer, D. (2008). *Text mining infrastructure in R*. Journal of Statistical Software, **25**(5), 1–54. <https://doi.org/10.18637/jss.v025.i05>