Sentiment Analysis Tutorial Answer Sheet

# Part 1 - Graded exercises

## **Graded Exercise 1**

**Take a screenshot of this page and paste it in the answer sheet. *Make sure that the screenshot clearly shows***

* the source file name and the target table schema and name at the top
* the number of rows read from the source file, rows loaded, and rows rejected at the bottom left of the page
* the start and end time at the bottom left
* the number of errors and warning in the right panel. In this case, there are no errors.

Graphical user interface, application, Teams

Description automatically generated

Figure 1: Load Status

## **Graded Exercise 2**

Graphical user interface, application, Teams

Description automatically generated

Figure 2: Query and Result for the number of rows in the Airline Sentiment file

Copy and run the following query from Airline sentiment SQL code file. Paste the output to the answer sheet and ***explain what the query returns****.*

This query returns the number of rows in the file (N = 14640, Figure 2).

## **Graded exercise 3**

Run the **ls()** command in the integrated RStudio environment. Take a screenshot of the returned output*. Paste the output screenshot to the answer sheet and explain what the command returns.*

The “is()” command on the Sentiment data (shown above) returns the objects in the data frame (Figure 3).

Text, application

Description automatically generated

Figure 3: Shows returned results from the is command

Run the **dim(SENTIMENT)** command in the integrated RStudio environment. Take a screenshot of the returned output. *Paste the output screenshot to the answer sheet and explain what the command returns.*

The dim function is used to get the dimensions of the matrix (table, data frame). In the above example, the dim returned the number of rows (n = 14640) and number of columns (n = 15, Figure 4).



Figure 4: Shows returned results from the dim command

## **Graded exercise 4**

Table

Description automatically generated

Figure 5: Document-Term Matrix for the Airline Sentiment data

*What is the document-term matrix, and why is it useful? What is sparsity, and why do we remove sparse terms?*

The document term matrix is a mathematical representation of the frequency in which terms within the documents occur (Figure 5). The rows represent the individual documents. The column represents the terms (or words) and their frequency of occurrence within the document (Zhao, 2015). For instance, in document 10 there was no frequency of the word “ear”, however, in document 2, the word “ear” was used one time.

The document-term matrix is useful because it quantifies the frequency of the words across all the documents. This enables the analyst to:

1. visually inspect the table to see:
   1. if certain words “jump out” as occurring more frequently
   2. if null values or missing words occur across documents
2. move to the next step of analysis and ultimately use quantifiable methods such as Naive Bayesian statical analysis on text documents

Sparsity refers to a relative high number of cells with no data. Empty, sparse, or null data cells can be more profuse in tweets because of the short contextual information in each document. High sparsity can be problematic as it increases the result of improper text classification therefore, sparse terms are removed (Cambria, Das, Bandyopadhyay, Feraco, 2017).

**References**

Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (2017).A Practical Guide to Sentiment Analysis, 5th Ed. Springer, Singapore, Singapore, pp. 142.

Zhao, Y. (2015). *R and Data Mining: Examples and Case Studies.* Elsevier, Cambridge, MA, pp. 106.

## **Graded exercise 5**

Text

Description automatically generated

Figure 6: Word Cloud generated from the Airline Sentiment data

Text

Description automatically generated with medium confidence

Figure 7: Code parameters for development of the word cloud (e.g. colors, frequency etc.) for the Airline Sentiment data

*What are the visual advantages of the word clouds over the word frequency tables? What are the limitations?*

The visual advantages of word clouds are that they show the frequency of terms in a document by showing more frequent words as larger font. For example, in the word cloud in figure 6, “southwestair” and “jetblu” are more frequently occurring words than “plane” and “delay”. The visual word cloud can therefore provide insights of the commonly occurring terms within the various documents which may guide future analytics decisions. For instance, should negative and positive words be frequent then a polarity algorithm may be chosen as a good next step in understanding the data. Furthermore, word clouds can be tweeked to exclude words that do not occur frequently giving a clearer picture to common terms. Colors can be used to make words “pop” and provide another element of visual learning for the analyst (Figure 7).

Another advantage of word clouds is that they are easy to understand and fast to develop (Alida, 2012). The visual representation of the data is engaging to the audience and provides an impact in such a way that does not require an understanding of technical or algorithmic complexities.

The disadvantages (cons) of word clouds can be in data sets where is there is not much difference in the frequency of words. In such cases, word clouds simply appear with many words, all of the same size without many differentiating characteristics.

Size and color should be used to explain and not confuse the audience. The size of the word (i.e. the literal number of letters in a word) can affect the visual within the word cloud making a longer word appear larger than a shorter word with the same frequency of occurrence within the matrix. This can lead to misinterpretation of data.

In a similar vein, color should communicate and not confuse. Often words clouds generate random colors. However, colors can be improved to help explain frequency in three prominent ways according to Burns (2021):

1. Use similar colors to words that are positioned close to one another
2. Highlight words that appear in one sample of text and not in another
3. Show two sample texts in the same word cloud, each with its own color.

A further improvement could be to show words in the same color with the same frequency. For instance, all words with a frequency of 4 are in blue, and a frequency of 10 are in green.

A final important consideration in using word clouds is that counting the frequency of words does NOT necessarily indicate their importance. This can be illustrated in three primary ways:

1. Common words such as “and” “the” are frequently occurring but often meaningless
2. Sometimes less frequently occurring words are more meaningful simply because they are sparsely used. Extreme displeasure usually incurs words that are used only rarely but should be weighted with higher importance in a polarity measure because they are, by their nature, “extreme”.
3. Counting individual words may be misleading. For instance, “not happy” and “happy” if counted separately may lead to an over representation of “happy” and an underrepresentation of “not” (Feinberg, 2010).

**References**

Alida. (2012). *The Pros and Cons of Word Clouds as Visualizations.* Retrieved on July 4th, 2022, from: <https://www.alida.com/the-alida-journal/pros-and-cons-word-clouds-visualizations>

Burns, M. (2021). *5 Ways to Use Word Clouds in the Classroom.* Retrieved on July 4th, 2022, from: <https://www.edutopia.org/article/5-ways-use-word-clouds-classroom>

Feinberg, J. (2010). Worlde. In J. Steele, & N. Llinsky, *Beautiful Visualization: Looking at Data Through the Eyes of Experts* (pp. 37-58). Sebastopol: O'Reilly.

## **Graded exercise 6**

*What information does the dendrogram above reveal about airline tweets? What are the limitations?*

Diagram

Description automatically generated

Simplicifolious clade

Clades

Leaves

Figure 8: Dendrogram of the Airline Sentiment data

A dendrogram is a binary tree which has two branches in each node (Bramer, 2013). A dendrogram places each term (or leaf in Figure 8) in the airline tweets documents in a separate cluster, then merges the terms that are similar at each iteration in the dendrogram. This method continues until each term is joined.

A clade is a branch in the tree (Glen, 2021, see Figure 8), which is lined to one of more successor groups. A single leafed clade is called a simplicifolious clade (see Figure 6). For instance, “jetblu” is a simplicifolious clade which is associated with “flight” and “unit” in one branch and other airlines such as “southwestair” and “Americanair” (among other terms) in another branch. The dendrogram reveals that “southwestair” and “Americanair” are more similar to one another than “jetblu” as they are placed closer together than to “jetblu”.

The dendrogram (Figure 8) should be read from the bottom-up. Each term (leaf) is labeled at the bottom of the figure (e.g., southwestair, guy, make, etc.) and clades are arranged according to how similar (or dissimilar) they are. Differences in height of the clades indicate more dissimilarity. For instance, the “gate” and “agent” clade are dissimilar to “fli” and “airlin”. However, the “gate” and “agent” clade is more similar to “custom” and “service”.

Limitations of this dendrogram are that it is difficult to read and interpret and therefore, to obtain value. This limitation is expounded upon when there are many terms in the document matrix (Bock, 2022). Personally, I found the word cloud much easier to interpret and guide further decision making than the dendrogram in the airline tweets data set.

Glen (2021) reveals that dendrograms can be “just plain wrong” and the analyst must know their data inside and out to detect the incorrect presentation in a dendrogram. This seems impractical as the purpose of a dendrogram is to help the analyst better understand the data.

Bock (2022) further states that dendrograms involve many arbitrary decisions, do not work well with missing data and mixed data types.

**References**

Bramer, M. (2013). *Principles of Data Mining. Springer.* New York, NY. Pp: 323.

Bock, T. (2022). What are the Strengths and Weaknesses of Hierachical Clustering? Retrieved on July 4th, 2022 from: [https://www.displayr.com](https://www.displayr.com/strengths-weaknesses-hierarchical-clustering/#:~:text=The%20weaknesses%20are%20that%20it,the%20dendrogram%2C%20is%20commonly%20misinterpreted).

Glen, S. (2021). Hierarchical Clustering/Dendrogram: Simple Definition, Examples. Retrieved on July 4th, 2022, from: <https://www.statisticshowto.com/hierarchical-clustering/>

# Part 2 – Answers to the Sentiment Analysis Quiz

Put your answers to Sentiment Analysis Quiz questions on this answer sheet.

* For multiple choice questions, enter the letter of your answer into the second column.
* For multiselect questions, enter the letters of your answer choices into the second column.
* For short answer questions, enter your answer into the second column.

|  |  |
| --- | --- |
| Question Number | Your Answer |
| 1 | d |
| 2 | ~~A~~ c |
| 3 | b |
| 4 | a |
| 5 | a |
| 6 | c |
| 7 | d |
| 8 | a, e, f |
| 9 | c, d |
| 10 | ~~A~~ b |
| 11 | e |
| 12 | b |
| 13 | ~~B~~  d |
| 14 | d |
| 15 | ~~C~~ b |
| 16 | c |
| 17 | c |
| 18 | b |
| 19 | d |
| 20 | a |
| 21 | ~~E~~ a |
| 22 | d |
| 23 | b |
| 24 | c |
| 25 | b |
| 26 | d |
| 27 | ~~e, f~~ c, d |
| 28 | d |
| 29 | a |
| 30 | b |
| 31 | ~~G~~ b |
| 32 | c |
| 33 | I thought it was all equally challenging. I have not used the IBM Cloud in the past, so it was all new to me. The most challenging thing is to read the instructions and follow them precisely and VERY carefully. I overcame this by following the instructions, but I had two accounts and so still encountered some confusion. Also, I am not sure how to go outside these boundaries should I need to do something slightly different. The project was challenging simply because it also takes a lot of time to complete, and I was concerned about the hour usage as I am not a fast coder. I can overcome this by practicing my code. |