**Assignment 2**

Text mining and analysis using tm package in R

I have completed this assignment on 2 short stories

Story 1: Goldilocks and the three bears

Story 2: The three little pigs

Note: The text files are in French language. They have been placed in a ‘texts’ folder.

The assignment has been completed in the following steps:

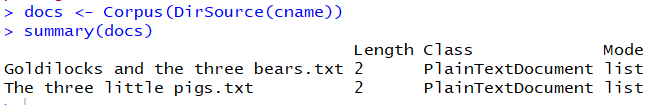
Step 1: Import required libraries



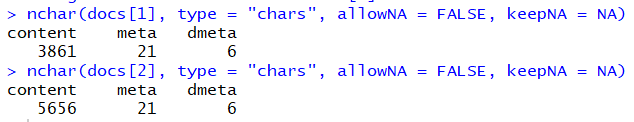
Step 2: Assign the right path to the ‘text’ folder



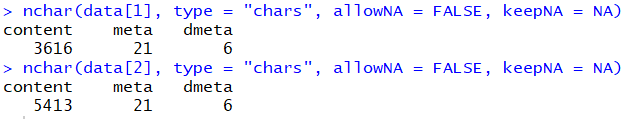
Step 3: Load documents



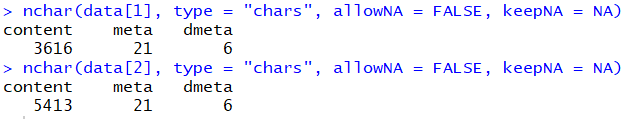
Step 4: Check the document and get the original character count



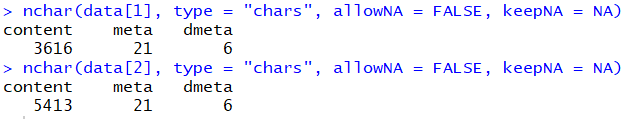
Step 5: Remove punctuations and get the character count



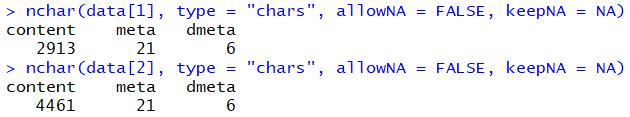
Step 6: Remove special character and get the character count



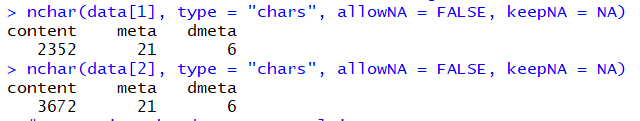
Step 7: Remove numbers and get the character count



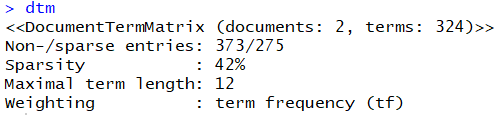
Step 8: Convert to lowercase, remove stopwords and get character count



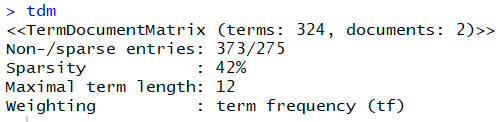
Step 9: Stem the words and Strip unnecessary white spaces and convert the document to plain text



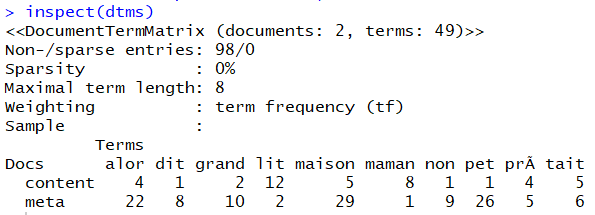
Step 10: Create document term matrix and export it to an excel spreadsheet



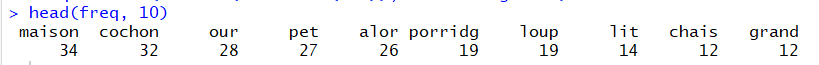
Step 11: Create a transpose of the dtm to get a term document matrix and export it to an excel spreadsheet



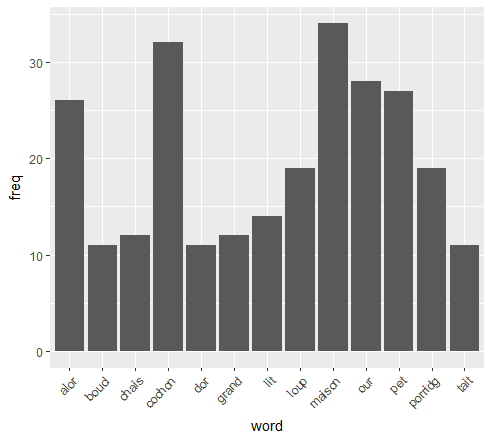
Step 12: Remove sparse words



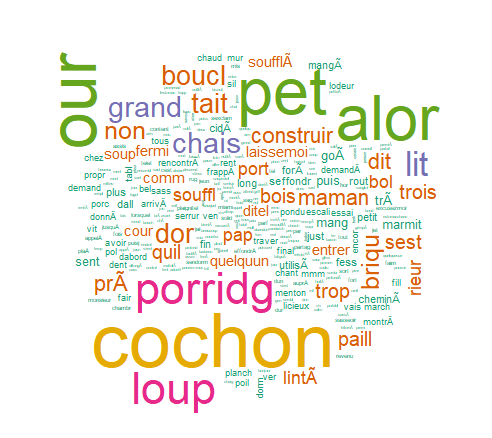
Steps 13: Get the word frequencies and plot a graph to display the top 10 words with highest frequencies



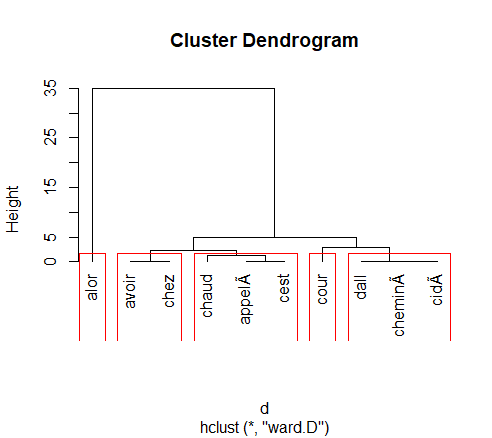
Step 14: Plot a bar graph to display words with frequency > 10



Step 15: Create a wordcloud



Step 16: Create a cluster dendrogram



Step 17: KMeans clustering

