Information Retrieval in R - STA 521

Abbas Zaidi

1 Agenda

1. The tm package in R
2. The Corpus() Function
3. Data pre-processing for text-mining using tm_map(): case conversion, removal of punctuation, removal of numbers and stemming

2 Lab Tasks

1. Read the text data (provided) on Shakespeare into R with the load() function. The variable shakespeare contains the complete works of William Shakespeare. Apply the function length() to determine how many documents are present.
2. Create a corpus of this text using the following command Corpus(VectorSource(shakespeare)) and store this in the variable corp.
3. Use the tm_map() command to pre-process the data: remove punctuation, convert to lower case, and remove any numbers present in the data. After the processing, create a document term matrix (DTM) of this corpus and store it in the variable dtm. Apply as.matrix() to the final matrix.
4. Set the variable myQuery to the following c("something","rotten", "state","denmark")
5. Write a function called myTextMiner() that accepts as its inputs, a string vector containing keywords (akin to myQuery), and a corpus. The function should then process the corpus to first convert all entries to lower case, then remove all punctuation, then remove any numbers present and finally construct a document term matrix (DTM) that is normalized by the length of each document. Finally based on the query above, compute the Euclidean distance to each document in the shakespeare corpus. Note that the easiest way to do so involves including the query in the DTM. The function should return a subset of the normalized DTM with those columns that are shared with the query, with one additional column that

contains the Euclidean distance for each document that has been normalized by document length. Name this column distanceMetric. **Hint**: Remember to use as.matrix() on the DTM before any calculations.

3 Directions

In general for Labs, at the top of any file you are asked to submit, please list the following:

1. First Name Last Name
2. Lab Date
3. Team Member(s)

With respect to any item for which you are asked to generate any output, please provide the actual R output as a part of your solution and any explanation needed as well. For any functions/ computations that you will write, please list the following as comments before the step in R:

1. Task number and descriptions.
2. Input(s) with descriptions.
3. Outputs(s) with descriptions.
4. Function/ output summary (along with intermediate step comments).

For Lab 3, please provide the following deliverable items:

1. Please provide your solutions using Markdown as a .pdf with the following naming convention: LastName_FirstName_Solutions_Lab3.pdf.
2. Provide your .Rmd file (this **MUST** compile) for the lab using the following naming convention: LastName_FirstName_Solutions_Lab3.Rmd
3. A .csv file containing the final document term matrix from your function with the following naming convention LastName_FirstName_DTM.csv. To directly output a .csv from R use the write.csv() function