IoT Analytics (EE8225)
Fall 2020
Assignment 2
Deadline, Tuesday NOV 13, 2020

Please download the case study for this assignment, the case can be found on the D2L, in the library E-reserve, you will find a case study called "UNDERSTANDING TEXT MINING AND SENTIMENT ANALYSIS IN HOTEL BOOKING". The dataset reviews.csv required for this assignment can be found in folder "assignments". This exercise introduces sentiment analysis through the vector space model and promotes discussion of its potential applications, specifically for hotel booking.

Goal: After working through the assignment questions, students will be able to do the following:

- Understand the importance of data management.
- Recognize the difference between unstructured and structured data.
- Gain a foundational understanding of the field of text mining, particularly of sentiment analysis.
- Learn the steps needed to complete a vector space model and apply them accordingly.
- Understand term frequency–inverse document frequency (TF-IDF) and its application.
- Evaluate both the findings of their text analytics and their implications for strategic decisions in the recommendation systems industry.

Programming: **If you plan to use R, below are relevant readings.**

- W. N. Venables, D. M. Smith, and the R Core Team, *An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics*, April 26, 2019, accessed July 2, 2019, https://cran.r- project.org/doc/manuals/r-release/R-intro.pdf.
- • Ingo Feinerer, *Introduction to the tm Package: Text Mining in R*, December 21, 2018, accessed July 2, 2019, https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf.

Assignment Questions

1. How many words were typically contained in the reviews? (Hint: Use descriptive analytics)
2. What are the most frequently used terms in the reviews? (Hint: perform text preprocessing first, show your steps, snapshots, then answer the question)
3. Calculate the term-frequency matrix. Why can we not use term frequency alone to rank or weight the terms?
4. Build the TF-IDF matrix. What can you observe? What are the terms with the highest collective weighting (importance), using both term frequency and inverse term frequency? Were these different from the most frequent terms? Why?
5. What is the final decision the professor needs to make after the analysis?
6. If you are asked to label the reviews to GOOD or BAD, what will you do? Write your conceptual model, then implement an algorithm that will help you in this task, finally how many good reviews and how many bad reviews?
7. In all of the above questions, do not forget to add snapshot of your analysis/results (if needed)

Submission Guidelines

Please note that this is individual submission, any copy of code, or answers will be considered as plagiarism

To Submit: Zip all of your word/pdf/code files into a folder lastname.firsname.id.zip and upload to folder "Assignment 2"