1/24/2018

CSE 5243
Instructor: Jason Van Hulse
Homework 1

Due Date: 2/7/2018 5:30pm (submitted through Carmen).

Objective: In this lab, you will analyze one or two datasets that have been collected regarding "TED Talks" at Kaggle (www.kaggle.com). TED (**www.ted.com**) is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks. TED talks are well known for being inspirational and though-provoking. The objective of this assignment is to explore the data from TED videos to determine if there are interesting observations or insights that you might uncover. There are not a predetermined set of questions that you need to answer; instead, it is up to you to use your critical thinking and problem solving skills to interactively explore the data, pose and answer interesting questions. This is an iterative, interactive data exploration project.

Data Description: These datasets contain information about all audio-video recordings of TED Talks uploaded to the official *TED.com* website until September 21st, 2017. The *TED_main* dataset contains information about all talks including number of views, number of comments, descriptions, speakers and titles. The *transcripts* dataset contains the transcripts for all talks available on TED.com.

Additional information regarding this problem domain, go to www.kaggle.com and search for "Ted Talks" datasets. If you do not already have an account at Kaggle, you may need to create one to download the data.

Collaboration: For this assignment, you may work either as an individual or as part of a team (maximum of 2 people total). If you work as part of a team, the expected amount of work increases (see below for additional details). You may review notebooks posted at kaggle.com for reference or learning, but please do not simply copy/paste other people's code.

If you are working as an individual: your analysis should focus solely on the **TED_main** dataset. Some of the questions you might want to answer (this is just a sample of questions you might address; you should expand your analysis above and beyond just these questions):

- Which are the most popular talks?
- Is there a relationship between the popularity of talks and the length of the talks?
- Is there a relationship between talk popularity and speaker occupation?
- What conclusions can you derive from the description and/or the ratings?

If you are working as a team: In addition to working with the TED_main data, join the transcripts dataset. Convert the transcripts to a 'bag of words' format and look for additional insights from the occurrences of individual words.

What you need to turn in:

1) **Code** - please submit to Carmen any code that you used to process and analyze this data. You do not need to include the input datasets.
2) **Written Report**
   A. The report should be a maximum of 8 pages (13 pages for teams of 2).
   B. The report should be well-written. Please proof-read and remove spelling and grammar errors and typos.
   C. The report should discuss your analysis and observations. Present charts and graphs to support your observations. If you performed any data processing, cleaning, etc, please discuss it within the report.
   D. The written report can be in the form of a Python or R Notebook or as a Word or PDF Document.

Grading Criteria:

1. **The readability of your report (30%)** - is it well organized and does the presentation flow in a logical manner; are there many grammar and spelling mistakes; do the charts/graphs relate to the text, etc…
2. **Data Methodology (10%)** - Did you outline your data processing steps? Did you clean any data (removing noise, outliers) or handle missing values? Did you eliminate any records, and why? Describe how you processed the text data. If you used 'off the shelf tools' say in Python, please describe them briefly.
3. **The depth of your analysis (55%)** - did you find novel and/or interesting insights, or did you solely focus on simple summarizations of the data? Did you draw and present potential conclusion or observations from your analysis of the data?
4. **Code (5%)** - Since this is exploratory in nature, the code you turn does not need to be efficient, neat or clean. But we will check that it is there and that it contains a flow of data transformations and analysis.

In addition - please hand in the written report in class on the assignment due date.

How to turn in your work on Carmen:

Please choose one of the programming languages from: Python, R or Java. All the related files except for the data will be tarred in a *.zip file or *.tgz file, and submitted via Carmen. Please use this naming convention:
”Project1_Surnames_DotNumber.zip" or “Project1_Surnames_DotNumber.tgz.” The submitted file should be less than 5MB.