**Final Project**

**DS 710**

The final project for this class is your opportunity to apply what you have learned in this course to answer a question that interests you, by collecting and analyzing real-world data from Twitter.

For your final project, you will submit

a) A 1-page executive summary which reports your question, analysis, and results in a non-technical manner.

* Figures may be embedded with the text or included on a second page.
* In .doc, .docx, or .pdf format.

b) A Python notebook containing the Python code you used to gather data from Twitter and parse it for analysis.

* Do not include your consumer key, consumer secret, access token, or access secret.
* This should be a clean, commented, final version of the code.

c) A .csv or .txt file containing your parsed data for analysis in R.

d) An R script containing the R code you used to analyze the data from Python.

* This should be a clean, commented, final version of the code.

Submit your project to GitHub.

A rubric for the project is available on the next page.

**Executive Summary (40 points)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **A** | **B** | **C** |
| **Introduction** | Clearly explains the question of interest, and why/to whom it is interesting | Explains the question but does not explain why/to whom it is interesting | Question is vague |
| **Data Collection and Analysis** | * Clearly explains what keywords/features used to collect data, and why these keywords/features are appropriate to address the question of interest. * Clearly explains when data were collected, and whether the REST or Streaming APIs (or both) were used. * Method(s) of analysis are appropriate to the question of interest and clearly explained in a non-technical way. | * Explains what keywords/features used to collect data, but does not discuss why these keywords/features are appropriate to address the question of interest. * Explains whether the REST or Streaming APIs were used. * Method(s) of analysis are reasonable for the question of interest, but perhaps not the optimal choice. | * Explains what keywords/features used to collect data, but does not discuss why these keywords/features are appropriate to address the question of interest. * Method(s) of analysis chosen demonstrate weak understanding. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Figures** | * Figures are appropriate to the data and question of interest. * Well-integrated with discussion of analysis and/or results. (For example, “As shown in Figure 1, …”) * Legends or captions used appropriately. * Color used appropriately. * Font size and line widths chosen so that figures are legible when page is viewed at 100% Zoom. | * Figures are appropriate to the data and question of interest. * Legends or captions used appropriately. * Reader has to work to connect the figures with statements made in the text, OR figures introduced in the text with a colon (For example: “The histogram is shown here:”) | * Extraneous figures, or figures which do not highlight the key features of the data that are discussed in the text. * Missing or difficult-to-interpret legends and/or captions. |
| **Results/Conclusion** | * Explains results clearly and accurately in a non-technical way. * Conclusion relates results to larger question or implications. * Avoids using the words, “In conclusion…” | * Results are clearly and accurately explained, but the importance of the results is not necessarily clear. | * Discussion of results is confusing or demonstrates weak understanding. |

|  |  |  |  |
| --- | --- | --- | --- |
| **Writing Style** | * Readable and interesting for a reader who does not know computer programming or statistics.   + You can refer to technical topics (for example, “Using a t-test, I found strong evidence that…”), but don’t get into the nitty-gritty here. * Professional spelling and grammar. | * Overly technical details * Minor spelling or grammatical mistakes | * Unprofessional formatting, spelling, and/or grammar. |

**Parsed Data file (10 points)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **A** | **C** | **F** |
| **Parsed data file** | * Data file is in a .csv or .txt format * Format of data file is consistent with Python code (no editing by hand was necessary) * Format is consistent with R code (no editing by hand is necessary to run R code for this data file) | * Python code, as written, could not have produced the data file   or   * R code, as written, cannot be used to analyze the data file | No data file submitted |

**Python and R Code (50 points)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **A** | **B** | **C** |
| **Python and R Code** | * Code is consistent with analyses described in the executive summary * Comments used appropriately to make code readable * Functions created for effective task management AND/OR evidence of effort put into writing efficient code * Clean, final version of code: When run by the reader, code produces no error messages, and all output is relevant to the analysis in the executive summary * DOES NOT include consumer key, consumer secret, access token, or access secret. | * Code is consistent with analyses described in the executive summary * Comments used appropriately to make code readable * Clean, final version of code: When run by the reader, code produces no error messages * DOES NOT include consumer key, consumer secret, access token, or access secret. | * Reader has to work to see the connection between code and analyses described in the executive summary * Few comments, or comments are unclear * Evidence that programmer overlooked ways to improve task management and/or efficiency |