IST736 Text Mining - HW1 – **Deeper Explanation of Requirements**

An Evaluation of Sentiment Classification Tools

NOTE: I will add notes in red to clarify and better explain the requirements of this assignment. All assignments must adhere to the required format. I am lenient with this first one, but will become strict starting in Unit 2.

Artificial Intelligence (AI) has become a popular topic recently.

As part of this assignment, you will:

1) Assume that you are a consultant at a public relations firm, and a client of your firm would like you to **evaluate the current public sentiment toward AI** in social media like Facebook and Twitter.
   a. **Think about what this means.**
   b. **Interestingly, in real life, I was recently asked to do exactly this. A new program was trying to determine whether to use "AI" in the title of a new class. "AI" has a long history and public sentiment has ebbed and flowed over the past 30 years.**
   c. **So, the goal here is to create (in part) a report that discussed past, present, and perhaps future (predicted) views on "AI" and its applications. Even the word "AI" has gone through changes (such as ML, DM, etc.)**

2) Since there are too many comments on social media, you can't manually collect and analyze them all.
   a. **In other words, you as a human cannot review millions of comments. However, a computer can!**
   b. Fortunately, you have discovered some free **sentiment analysis tools**, and now need to **evaluate** whether they are good enough to do sentiment analysis for your assigned task.
   c. **So – what is meant by "evaluate"? Let's read on…..**

3) You need to **collect a sample data set**
   a. **I suggest that you first build a small one by hand. This way, you can run tests faster and will know if you are doing it correctly.**
   b. Next, ….**choose two tools to compare their effectiveness in sentiment analysis.**
      i. **Here, this does not imply that you need to code this in Python. You certainly can, but that is not required by this assignment. Many students (especially those who have already used Python) do choose to use NLTK and sklearn and to code this. The other option is to use application-type tools that allow you to "paste" in your data and see and evaluate the results.**
      ii. **You can do either or both.**
      iii. **The key here is the "compare their effectiveness".**

1. **Think about how to compare them.**
2. **What measures can you use?**

c. Write a report to describe **(The report should adhere to the required format and should contain at least the following elements in the appropriate locations in the document.)**

i. (1) **your sampling strategy**

1. **What does this mean??**
2. **This will depend on whether you use an application or code this in Python.**
3. **Sampling is "how and where" you got your data and how you formatted it.**

ii. and whether it would result in a representative sample of public sentiment toward AI,

1. **This is about the idea of samples vs. populations and the representative nature of sampling.**

iii. (2) data preparation and system evaluation process, and

1. **This is connected to how and where you got your data, how you formatted it, and which applications or code tools you used.**

iv. (3) your conclusion on whether these tools are suitable for your task.

A few examples of tools:

If you would like to do some programming for this assignment, here are a few popular tools you could consider:

(1) NLTK's built-in sentiment analysis tools. See sample code in http://www.nltk.org/howto/sentiment.html
(2) VADER https://github.com/cjhutto/vaderSentiment

Notice the "if" part of the above. This is where it is clearly noted that coding is not required, but is optional and certainly a great idea.

If you prefer GUI-based tools at this time, consider:
http://sentistrength.wlv.ac.uk/
http://text-processing.com/demo/sentiment/

Suggestions:

1) **Explore NLTK, sklearn, and the two GUI tools noted, as well as Vader. You do not need to "use" them all. However, if is great to explore, discuss, read about, and compare them in your paper.**
2) **Use a comparison table to assist the reader with clarity.**
3) **Start thinking like an expert (not a student).**

Your report should **adhere to the required class format AND should also include….**

- Be formatted as a research paper that includes introduction, method, result, and conclusion sections. **– please adhere to the assignments format**
- ~~Up to 4 pages~~. **– I do not have page counts or requirements in this class**
- One inch margin on all sides.
- ~~12pt Times New Roman or 11pt Arial~~. **– I do not care about font and usually use 11 and Calibri myself.**
- Attach your data file in csv format. **– For all assignments, upload your .docx which is your assignment paper. Then, as a zip, zip up and include any data, Python code as .py, etc)**

Grading criteria:

Report grading is similar to reviewing academic papers. Grading will focus on the following aspects:

(1) The analytical methods are all used correctly, and the interpretation is convincing. (50%)
(2) The report has clarity of presentation. It is well organized, clearly written, ~~within page limit~~ (25%)
(3) The report provides sufficient information for others to replicate the analyses. (Necessary information includes but is not limited to problem definition, data description, description of analytical methods, processes, result interpretation, and conclusion). (25%)
(4) Adhere to format