**User Generated Content Analytics (MIS 381) F2017**

**Assignment #2 (Individual work)**

**Date due: 12th October by 11:59 p.m. on Canvas**

In this assignment, you have been hired as an analytics consultant by the two main cryptocurrencies, Bitcoin and Ethereum, to provide insights based on the wisdom of the crowd.

1. Use the Python script tweetstream\_location.py (on Canvas) to obtain a total of about 10k tweets on cryptocurrencies such as Bitcoin and Ethereum. You should use keywords like Bitcoin, Ethereum, Cryptocurrency, and Blockchain. Note: The script is being supplied to you for your convenience. You can use any other script to get the job done.
2. For each cryptocurrency, find **THREE** key issues mentioned by the public in the tweets (provide frequencies of mention).
3. Perform a sentiment analysis on each currency and issue and complete the table:

|  |  |
| --- | --- |
|  | **Sentiment** |
| Issue 1 (Bitcoin) | ? |
| Issue 2 (Bitcoin) | ? |
| Issue 3 (Bitcoin) | ? |
| Issue 1 (Ethereum) | ? |
| Issue 2 (Ethereum) | ? |
| Issue 3 (Ethereum) | ? |

1. Use the location data to find out the OVERALL sentiment about Bitcoin and Ethereum in the US versus the rest of the world and show the results in a table.
2. What insights would you provide to the two cryptocurrencies based on your analyses in B, C and D above?

**Steps in Assignment #2**

1. First obtain your authorization codes from Twitter using the primer posted on Canvas.
2. In collecting 10k total tweets on cryptocurrencies, get them in chunks of 2500 or 3000 tweets. Wait an hour or so, make another request. Make sure the output files names are changed in the script to avoid overwriting. Then merge all data in a single CSV file.
3. Make sure that the actual tweet is in the third column. Don’t create columns 1 and 2 as blank, fill them up with something. At this point, it is good to make location data a part of the tweets themselves. There are multiple ways of doing this. You can Google for a list of cities and states in the US, and replace all these names in the location column with the word USA001 (or something unique that is unlikely to be used by any author of a tweet) using find\_and\_replace.py. Similarly, all other locations outside the US could be replaced by rest\_of\_the\_world.
4. Now merge the tweets column and the location column, or write some simple code in Excel that will append the word USA001 or rest\_of\_the\_world in the tweets column. ***You will have to do this without help from either me or Aditya, it is a part of the assignment.*** Again, make sure that the tweets are in column 3. At this time you should delete columns 4, 5, etc.

1. Do a word frequency analysis with wordfrequency.py. This will give you an initial idea of the important “issues” – some issues may have to be merged as well. E.g., references to “fraud” and “cheat” are talking about the same thing. Run wordfrequency.py again. Make your final selection of issues based on highest frequencies. I would suggest choosing 6 issues at this time.
2. Run a lift analysis of (Bitcoin,Ethereum,issue 1,issue2,issue 3,issue 4,issue 5,issue 6) and find top three issues for each currency. Note that an issue can be associated strongly with both currencies.
3. Once you have identified the issues, now you have to use parserforsentiment.py to extract mentions of a currency and an issue. E.g., Bitcoin and fraud. The script asks you for (i) brand (bitcoin in this case), (ii) attribute (fraud in this case) and the number of words to retain on the right and left hand sides of the issue (the assumption being that people express their emotions within a close proximity of the attribute). I would recommend trying 3 or 4 words. **Note that both inputs (brand and attribute) must be typed in lowercase.** The output file should be labeled carefully, e.g., bitcoin\_fraud.csv.
4. Now bitcoin\_fraud.csv can be used with sentiment.py (on Canvas) for sentiment analysis.
5. For the last question involving US vs. the rest of the world, use parserforsentiment.py again on the entire dataset, but now brand=bitcoin (or ethereum), attribute=USA001 (or rest\_of\_the\_world) and # words = 20 (so that the whole tweet is included). The output of this script should be a file which has tweets on bitcoin from the US. Now use sentiment analysis on this file.
6. Don’t ignore the last question about what advice you can give to the currencies. Remember, this is the most important output of the whole exercise – that is, to perform the analysis, extract deep insights and offer them as actionable advice. Stay dispassionate about the topic, don’t let your personal preferences and opinions bias your analysis in any way.
7. **Deliverables include (i) Your answers to the questions, and (ii) the final master data file (after all replacements).**