Data 100 HW3 with HCE components Draft

A note about the dataset -- working with President Trump's Tweets as data in Fall 2020
All data is made from real-world phenomena, be it the movement of the planets, animal behavior, or human bodies and activities. Work with data always has a bearing back on how human beings know and act in the world. The dataset that you're about to work with in this homework consists of a compilation of President Trump's Tweets. It's important to acknowledge that these Tweets are more than just data -- they're the means by which the President expresses his opinions, performs public and foreign policy, and shapes the lives of people in the US and all over the world. More fundamentally, these Tweets are a powerful form of speech that is particularly significant on the eve of the 2020 US Presidential Election. We recognize that working with this data now, even in the context of a technical exercise, is not a neutral activity and one that students may find uncomfortable. We encourage you to observe what you may be experiencing and invite you to consider this as dimensions of data science work alongside your technical lessons and we're glad to discuss these issues together in section.

Question 0

Learning Objective
Students will gain an understanding of how technical data science tools can be applied to real world problems, how they can be used to investigate and expand one's knowledge base - here in a political context.

Q: There are many ways we could choose to read the President's tweets. Why might someone be interested in doing data analysis on the President's tweets? Name a kind of person or institution which might be interested in this kind of analysis. Then, give two reasons why a data analysis of the President's tweets might be interesting or useful for them. Answer in 2-3 sentences.

Solution (+1 Point)
Example responses:

- A news reporter could be interested in looking at Trump's tweets this way. They might be interested in his general word choice, or even the sentiment on his tweets to further an argument.
- An NGO could be interested in reading Trump's tweets this way. They might select tweets by certain topics, and run analysis on those, or want to get a general sense of Trump's opinions on given topics.

- Any response which identifies a person or entity, and then gives two reasons. Answers should be within 2-3 sentences, but no credit taken away for longer responses.

Question 5 (Redo)

Context 9 (in notebook)
What is Sentiment Analysis?
We can use the words in Trump's tweets to calculate a measure of the sentiment of the tweet. For example, the sentence "I love America!" has positive sentiment, whereas the sentence "I hate taxes!" has a negative sentiment. In addition, some words have stronger positive / negative sentiment than others: "I love America." is more positive than "I like America."

We will use the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon to analyze the sentiment of Trump's tweets. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media which is great for our usage.

The VADER lexicon gives the sentiment of individual words. Run the following cell to show the first few rows of the lexicon:

<CODE explaining scores>

The creators of VADER describe the tool's assessment of polarity, or "compound score," in the following way:
"The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). This is the most useful metric if you want a single unidimensional measure of sentiment for a given sentence. Calling it a 'normalized, weighted composite score' is accurate."

As you can see, VADER doesn't "read" sentences, but works by parsing sentences into words assigning a preset generalized score from their testing sets to each word separately.

VADER relies on humans to stabilize its scoring. The creators use Amazon Mechanical Turk, a crowdsourcing survey platform, to train its model. Its training set of data consists of a small corpus of tweets, *New York Times* editorials and news articles, Rotten Tomatoes reviews, and Amazon product reviews, tokenized using the natural language toolkit (NLTK). Each word in each dataset was reviewed and rated by at least 20 trained individuals who had signed up to work on these tasks through Mechanical Turk.

Questions 5a

Learning Objective
Students will understand how the sentiment of a word is context-dependent and consider how this affects the sentiment score that a word is given.

Please score the sentiment of one of the following words:

- police
- order
- Democrat
- Republican
- gun
- dog

- technology
- TikTok
- security
- face-mask
- science
- climate change
- vaccine

What score did you give it and why? Can you think of a situation in which this word would carry the opposite sentiment to the one you've just assigned?

Solution (+1 Point)
Example responses:

- Eg. Gun: -0.8: in most cases the word Gun has negative connotations and hence would receive a score closer to the extreme negative. This word can convey the opposite sentiment if someone is referring to a water gun or toy gun, or if someone is speaking about guns as a supporter of the 2nd Amendment
- Eg. Dog: 0.8: in most cases the word Dog has positive connotations as they are common pets loved by many and therefore receives a score close to the extreme positive. This word can convey the opposite sentiment if someone is being called a dog as an insult.

- Any response which identifies a score, states whether it is negative or positive and gives an example of an opposite scenario, indicating that students have a grasp of the way in which sentiment of a word is context-dependent Answers should be within 1-2 sentences, but no credit taken away for longer responses.

Question 5b

Learning Objective
Students will understand how the sentiment of a word is context-dependent and how the aggregation of individual scores can take away/misrepresent features of the whole tweet.

VADER aggregates the sentiment of words in order to determine the overall sentiment of a sentence, and further aggregates sentences to assign just one aggregated score to a whole tweet or collection of tweets. [This is a complex process and if you'd like to learn more about how VADER aggregates sentiment, here is the info at this link]

Are there circumstances (e.g. certain kinds of language or data) when you might not want to use VADER? What features of human speech might VADER misrepresent or fail to capture?

Solution (+1 Point)
Example responses:

- Yes there are - In cases where there is some sort of word play, for example irony or sarcasm, these statements will be harder to break down word by word and the score may not reflect the true sentiment.

- Any response which identifies a similar concept to sarcasm, irony, figures of speech that couldn't be conveyed in a score. Answers should be within 1-2 sentences, but no credit taken away for longer responses.

Question 5h - we have further expanded on this question after F2020 HW 4

Learning Objective
Students will understand how aggregation processes can take away from the meaning and intended tone of a tweet and how this can be misrepresented numerically by the polarity scores.

Identify and read the 5 tweets scored most positive and 5 tweets scored most negative.

- Do you think these tweets are accurately represented by their polarity scores?
- If you were interested in Trump's attitudes towards a certain subject, what could you learn from these tweets?
- Do you think a “negative” and “positive” score for a tweet is *generally* accurate? Why or why not? (Hint: consider the aggregation process and the role of a specific context when scoring an entire tweet)

Solution (+1 Point)
Example response:

- The positive tweets are accurately represented, they're all happy and positive. Most of the negative ones are too, however in some Trump is tweeting about a negative event in a positive way, therefore the categorization is incorrect. The negative or positive scoring is not always correct as a tweet with mainly negative scored words would result in a total negative score, however the context matters for such tweets.
- Any response which explains the positive and negative tweets, the score accuracy, mentions how context is important to score an entire tweet and that scores can be inaccurate under some specified circumstance. Answers should be within 2-3 sentences, but no credit taken away for longer responses.

END