

Neuron7.ai – Senior Data Scientist – Home Exercise

Background

During the 2016 election, Donald Trump used Twitter very actively. There is a theory out there that, since he was known to use an *Android* phone, that all *Android* tweets were from Trump, and all *iPhone* tweets were sent by someone on his staff.

However, since mid-2017, the *Android* phone data has ceased, but we still want to know which tweets belong to Trump and which don't! Your mission is to build a text-based classifier to help us do this.

Step 1: Retrieve the data

Go here (<http://www.trumptwitterarchive.com/archive>) and download all the tweets from 2016 (Jan-Dec).

Step 2: Develop a classifier for an "original" Trump tweet

- For our purposes we can assume that Android tweets are Trump tweets and iPhone tweets are from his staff. Tweets from other devices (web, ipad, etc.) are not known to be one or the other.
- Using the data from step 1, develop a classifier that distinguishes between the two types of tweets. The classifier can be based on the text/content of the tweet and its metadata.
 - Note that there is no requirement to do anything "fancy" – if it works, that's good enough.
- Quantify: how good is your classifier?

Step 3: Use the classifier to determine

- Who writes the tweets that come from the other devices during 2016. How confident can we be in these results?
- Run the classifier on tweets from Jan-Dec **2018**. Do you think this classifier is still valid, or has something changed [e.g., perhaps Trump is no longer posting tweets]? How reliable is it for this year?

For your submission:

1. Please attach your work products:
 - a. Jupyter notebook with all the code that you ran
 - b. Dataset you used (from the website)
 - c. Visualization of results (if possible, in the Jupyter notebook, but it's ok in a separate Word doc etc.)
2. Please describe your process (either in Jupyter notebook, or in separate file):
 - a. What approaches did you try? Which worked and which failed?
 - b. What algorithms did you use? How did you evaluate performance?

Good luck!