Notebook

March 1, 2020

0.1 Question 0

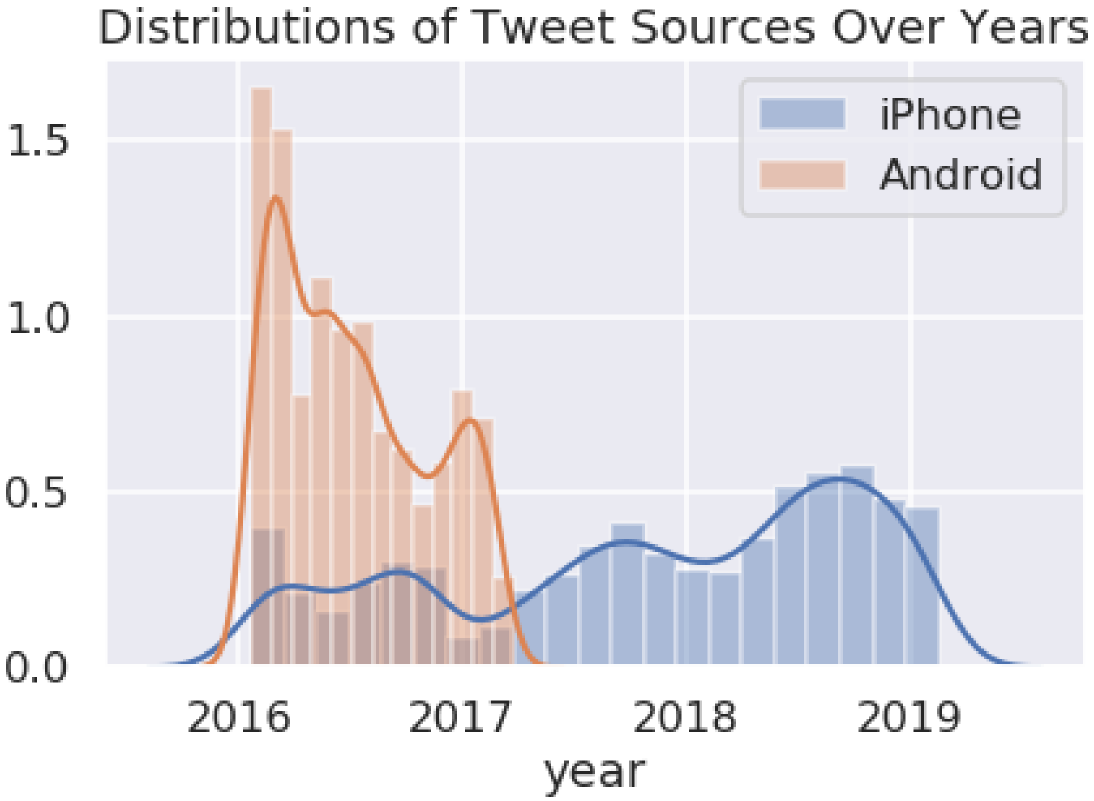
Why might someone be interested in doing data analysis on the President's tweets? Name one person or entity which might be interested in this kind of analysis. Then, give two reasons why a data analysis of the President's tweets might be interesting or useful for them. Answer in 2-3 sentences.

One of the President's rivals might be interested in doing so. With a data analysis of the President's tweets they might

1. have a better understading of President's policy focus in order to find some reasons to criticize him
2. make sure they themselves don't make the same mistake when their own rivals doing a data analysis to them

Now, use `sns.distplot` to overlay the distributions of Trump's 2 most frequently used web technologies over the years. Your final plot should look similar to the plot below:

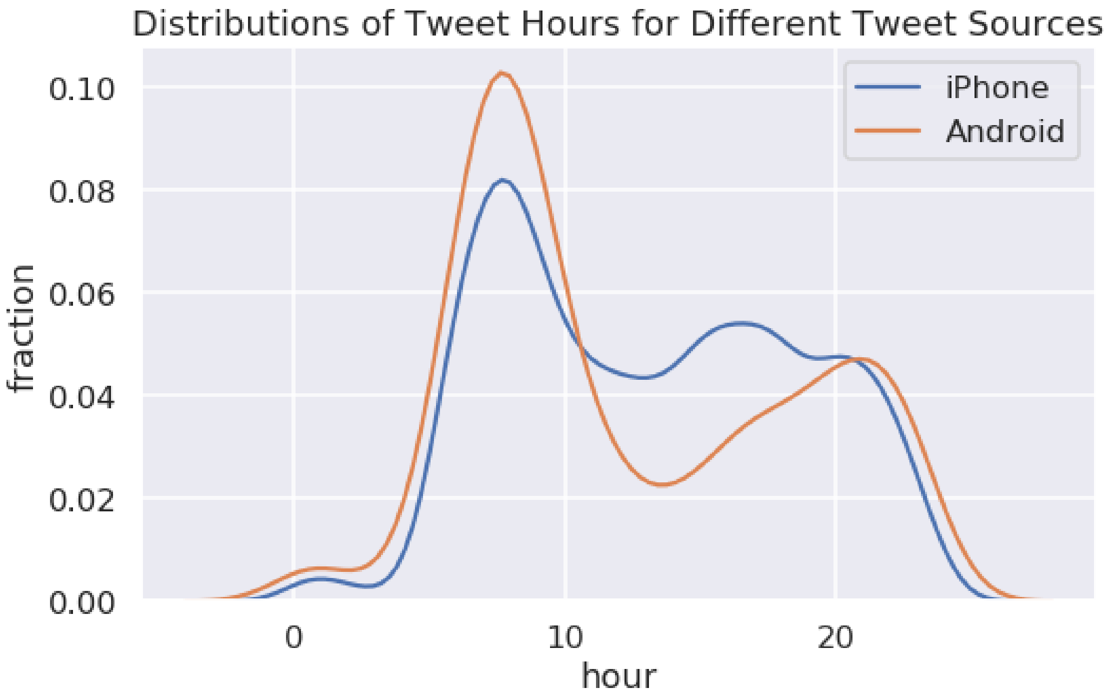
```
In [120]: sns.distplot(trump[trump["source"]=="Twitter for iPhone"]["year"])
          sns.distplot(trump[trump["source"]=="Twitter for Android"]["year"])
          plt.xlabel("year")
          plt.title('Distributions of Tweet Sources Over Years')
          plt.legend(['iPhone','Android'])
          plt.yticks(ticks=np.arange(0.0, 1.7, 0.5))
          plt.xticks(ticks=np.arange(2016, 2020, 1));
```



0.1.1 Question 4b

Use this data along with the seaborn `distplot` function to examine the distribution over hours of the day in eastern time that Trump tweets on each device for the 2 most commonly used devices. Your final plot should look similar to the following:

```
In [237]: ### make your plot here
          plt.figure(figsize=(8, 5))
          sns.distplot(trump[trump["source"] == "Twitter for iPhone"]["hour"], hist = False, label = 'iI
          sns.distplot(trump[trump["source"] == "Twitter for Android"]["hour"], hist = False, label = '
          plt.xlabel("hour")
          plt.ylabel("fraction")
          plt.title('Distributions of Tweet Hours for Different Tweet Sources')
          plt.xticks(ticks=np.arange(0, 24, 10));
```

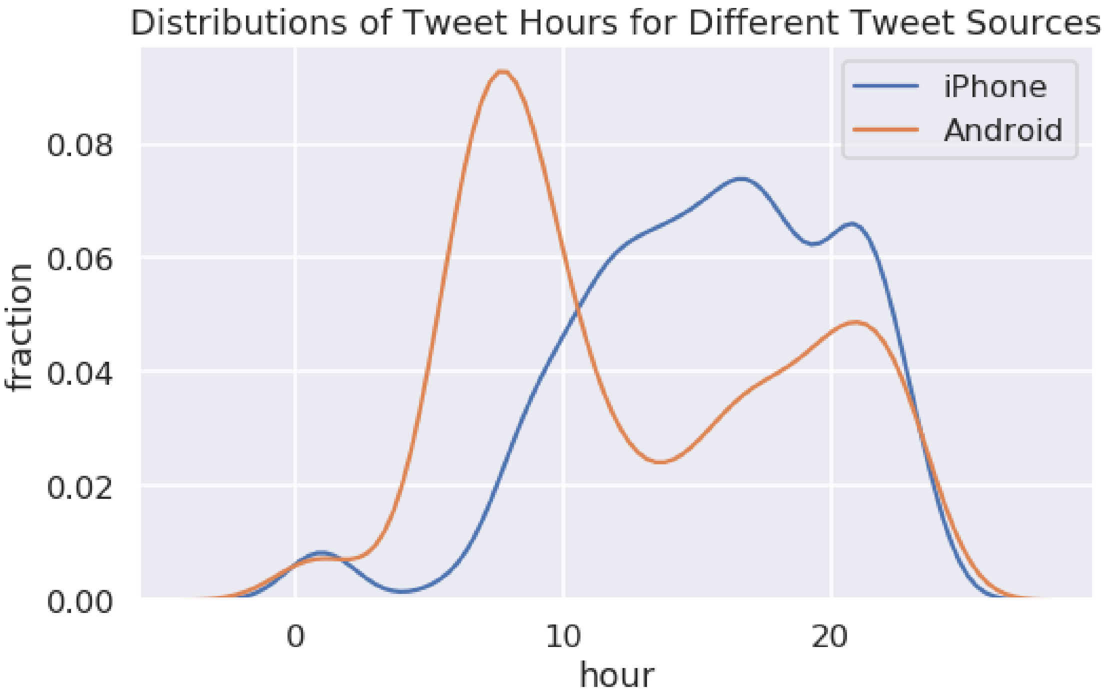


0.1.2 Question 4c

According to this Verge article, Donald Trump switched from an Android to an iPhone sometime in March 2017.

Let's see if this information significantly changes our plot. Create a figure similar to your figure from question 4b, but this time, only use tweets that were tweeted before 2017. Your plot should look similar to the following:

```
In [126]: ### make your plot here
          plt.figure(figsize=(8, 5))
          sns.distplot(trump[(trump["source"] == "Twitter for iPhone") & (trump["year"] < 2017)]["hour"]
          sns.distplot(trump[(trump["source"] == "Twitter for Android") & (trump["year"] < 2017)]["hour
          plt.xlabel("hour")
          plt.ylabel("fraction")
          plt.title('Distributions of Tweet Hours for Different Tweet Sources')
          plt.xticks(ticks=np.arange(0, 24, 10));
```



0.1.3 Question 4d

During the campaign, it was theorized that Donald Trump's tweets from Android devices were written by him personally, and the tweets from iPhones were from his staff. Does your figure give support to this theory? What kinds of additional analysis could help support or reject this claim?

Does your figure give support to this theory?

Yes, it does. As we can see, tweets time from android devices are consistent before and after 2017, as contrast to iPhone.

What kinds of additional analysis could help support or reject this claim?

To help support, find enough cases that multiple tweets are published from different devices at roughly the same time. Show the otherwise to help reject.

0.2 Question 5

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.

0.2.1 Question 5a

Given the above information about how VADER works, name one advantage and one disadvantage of using VADER in our analysis.

Advantage: fast and easy to use.

Disadvantage: might not be accurate under certain circumstances (see below).

0.2.2 Question 5b

Are there circumstances (e.g. certain kinds of language or data) when you might not want to use VADER? Please answer ”Yes,” or ”No,” and provide 1 reason for your answer.

Yes. Certain sentences including logic computation like ’not’ can inverse/change the sentiment of the whole sentence which cannot be measured via summing words together.

0.3 Question 5h

Read the 5 most positive and 5 most negative tweets. Do you think these tweets are accurately represented by their polarity scores?

Yes, I do.

0.4 Question 6

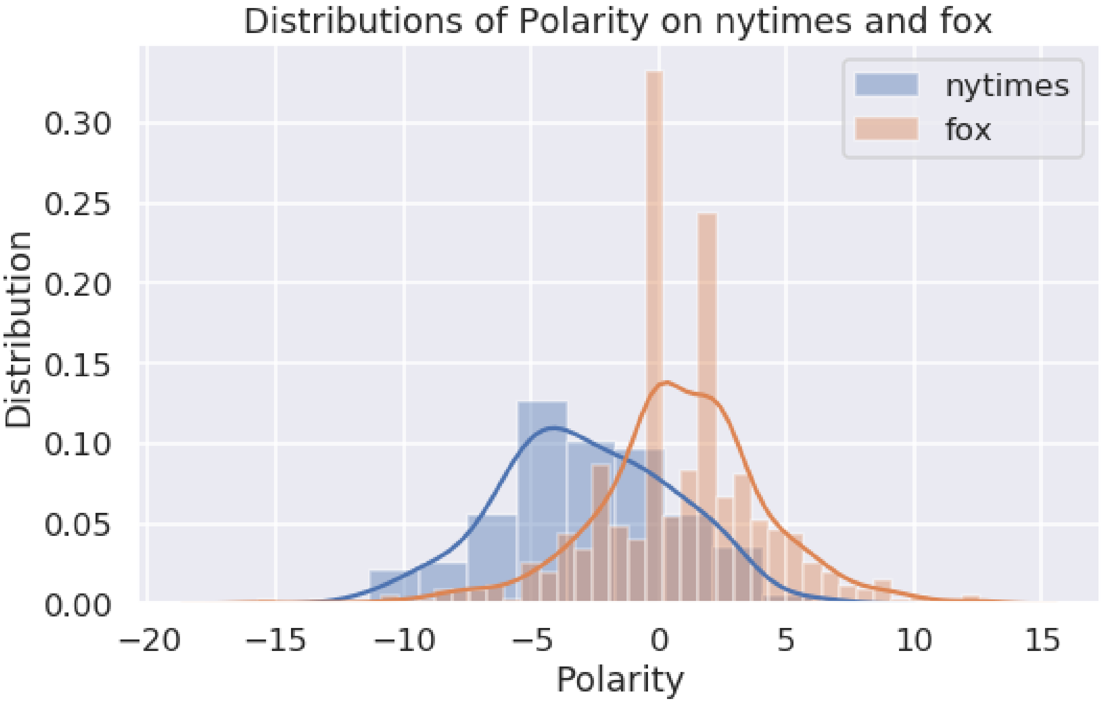
Now, let's try looking at the distributions of sentiments for tweets containing certain keywords.

0.4.1 Question 6a

In the cell below, create a single plot showing both the distribution of tweet sentiments for tweets containing `nytimes`, as well as the distribution of tweet sentiments for tweets containing `fox`.

Be sure to label your axes and provide a title and legend. Be sure to use different colors for `fox` and `nytimes`.

```
In [240]: plt.figure(figsize=(8, 5))
          sns.distplot(trump[trump["text"].str.contains("nytimes")]["polarity"], label = 'nytimes')
          sns.distplot(trump[trump["text"].str.contains("fox")]["polarity"], label = 'fox')
          plt.xlabel("Polarity")
          plt.ylabel("Distribution")
          plt.title('Distributions of Polarity on nytimes and fox')
          plt.legend(['nytimes','fox']);
```



0.4.2 Question 6b

Comment on what you observe in the plot above. Can you find another pair of keywords that lead to interesting plots? Describe what makes the plots interesting. (If you modify your code in 6a, remember to change the words back to `nytimes` and `fox` before submitting for grading).

Comment on what you observe in the plot above.

D. Trump's twitters generally have a negative attitude against NYTimes, and a medorate one against Fox.

Can you find another pair of keywords that lead to interesting plots? Describe what makes the plots interesting.

keys: "love - like - hate" shows that the sentiment method is generally precise

keys: "japan - europe - china - russia" shows that Trump's twitters like Japan over Europe over China over Russia

What do you notice about the distributions? Answer in 1-2 sentences.

D. Trump's hashtaged/linked twitters generally have positive sentiments, while unhashtaged/unlinked generally have moderate sentiments.