Notebook

October 16, 2019

0.1 Question 0

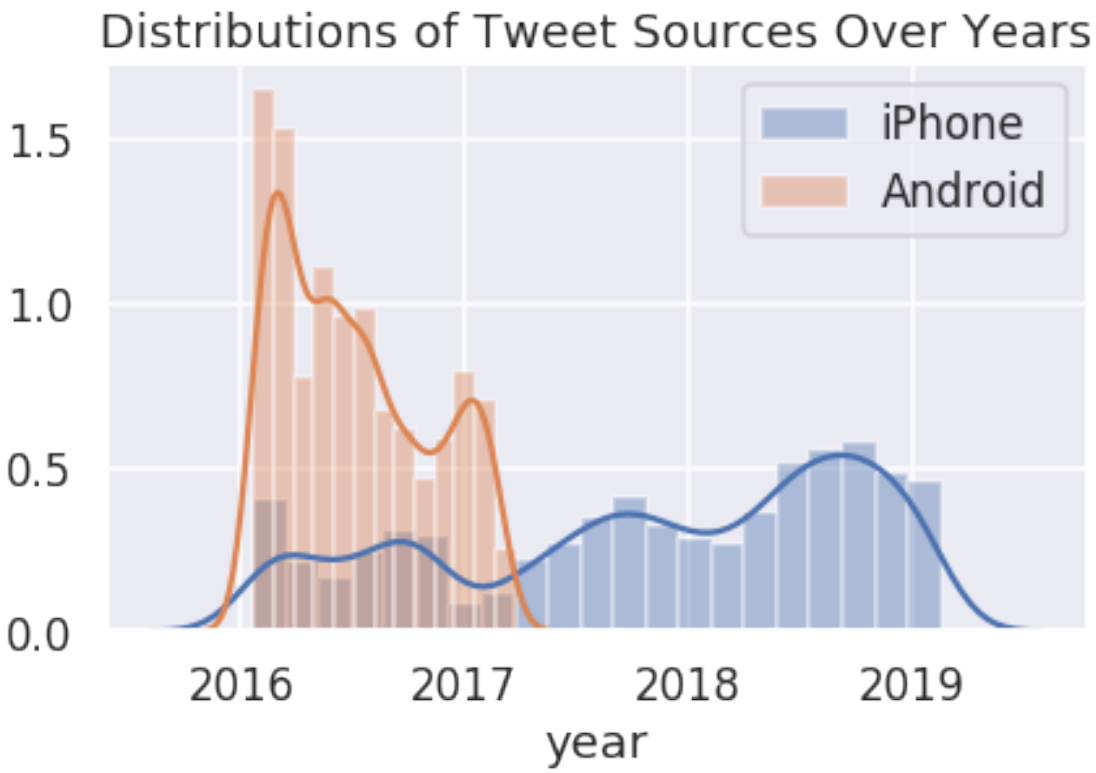
Why might someone be interested in doing data analysis on the President's tweets? Name one person or entity which might be interested in this kind of analysis. Then, give two reasons why a data analysis of the President's tweets might be interesting or useful for them. Answer in 2-3 sentences.

To analyze if his state of mind is ok, to analyze if he is sending important information by some code or something. To compare and contrast how Trump uses Twitter as compared to other presidents/presidential candidates. How many times did he discriminate people or contradict himself are very interesting questions. Opponents would be interested in this.

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 [19]: sns.distplot(trump[trump['source']=='Twitter for iPhone']['year'],label='iPhone')
         sns.distplot(trump[trump['source']=='Twitter for Android']['year'],label='Android')
         plt.legend()
         plt.xlabel('year')
         plt.title('Distributions of Tweet Sources Over Years')
```

```
Out[19]: Text(0.5, 1.0, 'Distributions of Tweet Sources Over Years')
```

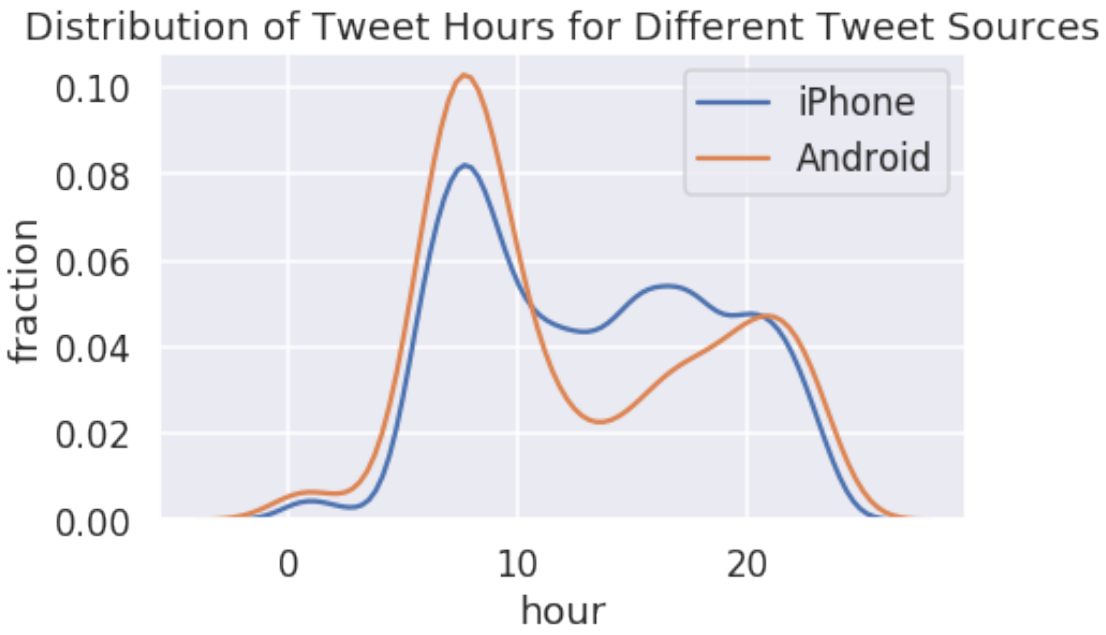


0.1.1 Question 4b

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 [24]:
```
### make your plot here
sns.distplot(trump[trump['source']=='Twitter for iPhone']['hour'],label='iPhone',hist=False)
sns.distplot(trump[trump['source']=='Twitter for Android']['hour'],label='Android',hist=False)
plt.xlabel('hour')
plt.legend()
plt.title('Distribution of Tweet Hours for Different Tweet Sources')
plt.ylabel('fraction')
```

Out[24]:
```
Text(0, 0.5, 'fraction')
```



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 [25]: trump.query('year < 2017 and source== "Twitter for iPhone" ')['hour']

Out[25]: id
         786204978629185536     9.013333
         786201435486781440     8.778611
         … Omitting 6 lines …
         786340623804751872    17.996667
         786285509668696065    14.346667
         Name: hour, Length: 1856, dtype: float64
```

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?

Pre 2017, there were 2 times when the android phone was most actively tweeting, morning and after dinner. Whereas the iPhone was more active during the evening time. We can't exactly be sure that this proves that Trump tweeted from the Android device and we need to see what kind of tweets, their grammar, sentiment, aim, were made during the morning vs evening vs night hours to make sure we can prove/disprove this hypothesis.

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.

VADER uses a very simple scoring range that might result in even findings for 2 completely different tweets and it fails to consider contextin it's training data and we're testing it on tweets of the infamous Donald Trump. VADER does cover emojis and takes data from a large amount of sources which should remove any bias from any one of the training data.

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. Sentiment analysis of Research papers. It will be weird. Some language is not meant to be positive or negative but rather just plain information reported as it was seen.

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?

Yeah the positive tweets are actually pretty positive, discussing wins for different teams, thanking supporters, governers and other people. Wish all his tweets were like that. His negative tweets are sometimes mistaken for negative due to the amount of negative terms in them, but some of them are kind of positive. For example, conquering hate together to defeat anti-semitism and stopping heroin to come into the country from China are positive-ish tweets containing a lot of negative words.

0.4 Question 6

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

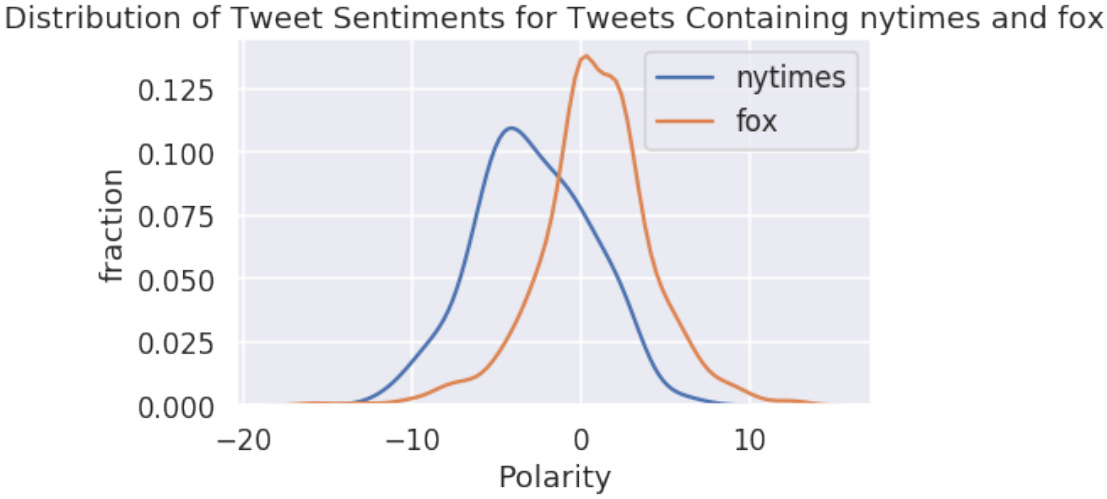
0.4.1 Question 6a

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. Your colors don't have to match ours, but you should use different colors for `fox` and `nytimes`.

```
In [88]: sns.distplot(trump[trump['no_punc'].str.contains('nytimes')]['polarity'], hist=False, label='ny
         sns.distplot(trump[trump['no_punc'].str.contains('fox')]['polarity'], hist =False, label='fox'
         plt.legend()
         plt.xlabel('Polarity')
         plt.ylabel('fraction')
         plt.title('Distribution of Tweet Sentiments for Tweets Containing nytimes and fox')
```

```
Out[88]: Text(0.5, 1.0, 'Distribution of Tweet Sentiments for Tweets Containing nytimes and 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).

The tweets containing fox are much more positive than the ones which contains nytimes. Trump really likes Fox News and hates NY Times as has been established by the graph above and by observing what he's said about the Times in the last 3 years. Russia and Borders have somewhat of a negative polarity attached to them.

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

The ratio of hashtag/link tweets to no hashtag/link tweets about 1:2 by observation although it looks closer to 1:1.8. the hashtagged links are most positive than the ones with no hashtags, which are sort of evenly spread amd symmetric about 0 polarity. The hashtag/link tweets are 0-2 polarity mostly with some with higher positive polarities.