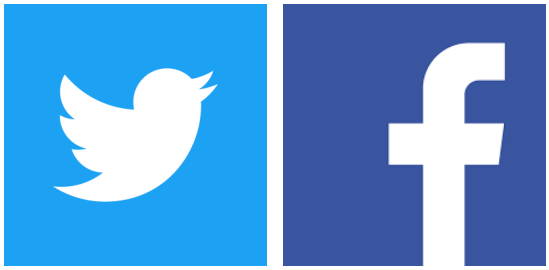
15-110 Hw6 - Social Media Analytics

Hw6 and its checks are organized differently from the other assignments. If you haven't already done so, you should read the **Hw6 General Guide** to understand how this assignment works.

Project Description



In this project, you will analyze a dataset that maps social media messages (tweets and facebook posts) made by politicians over a two-day period to information about the text's bias, partisanship, and message intent. You will use **sentiment analysis** to add additional data on message sentiment and compare politicians across several vectors.

Note: this project uses the **pandas** library to perform data analysis. You will need to install several libraries and learn about pandas DataFrames to complete this project.

In the first week, you'll organize the data by adding several additional columns to the DataFrame, to include information on name, location, position, sentiment, and hashtags. In the second week, you'll analyze political trends by comparing the data across different combinations of features, by state, region, and hashtag. In the third week, you'll visualize these comparisons using matplotlib.

Data source: https://www.kaggle.com/crowdflower/political-social-media-posts

Click on the following links to read the instructions for each week's assignment:

Check6-1 - due Monday 11/18 at noon EST
Check6-2 - due Tuesday 12/03 at noon EST
Hw6 - due Friday 12/06 at noon EST

Check6-1 - due Monday 11/18 at noon EST

In the first stage, you will organize the data for the project by installing necessary libraries, reformatting the CSV data into a pandas DataFrame, and adding additional columns. This reformatting will make analysis easier in the following stage.

Step 0: Written Assignment **[45pts]**

In addition to completing the steps described below, there is a short written assignment on the week's material. You can find the written assignment on Gradescope, and linked on the Assignments page of the course website.

Step 1: Installations **[0pts]**

The first goal of the first stage is to make sure you are able to install the external libraries: pandas, nltk, matplotlib, and numpy. If you have installed them successfully, you should be able to run the starter code without any errors. After you have successfully installed all the libraries, you should proceed to Step 2.

This project will require a couple of installations for your personal machine. **Unfortunately, some of the required installations are not installed on the CMU cluster machines, so you need to own a laptop or desktop computer to complete this project.**

To install these libraries we recommend that you use the 'Manage Packages' feature in Thonny. This tool will manage the installation process for you, which is much easier than trying to install a module manually.

Once you're in the 'Manage Packages' prompt, search for **numpy**, click on the link titled **numpy**, then click on 'Install'. Repeat the process with the search terms **matplotlib**, **pandas**, and **nltk**, installing the identical library each time.

If you're using a non-Thonny IDE, you'll need to use the pip command instead. Open your computer's terminal and run the following commands:

```
pip install numpy
pip install pandas
pip install matplotlib
pip install nltk
```

If an error message occurs, try googling it to find a solution. TAs can also help debug installation errors via Piazza or in office hours.

You can test whether the modules are correctly installed by running the following commands in your interpreter. If they do not give you an error, you're good to go!

```
import numpy
import pandas
import matplotlib
import nltk
```

Note: you may run into a certificate error when running `nltk` if you're on a Mac. If you do, follow these steps to fix the error.

1. Find the Python folder on your computer. It should be the same as the Thonny folder. Make sure it matches the version you run on Thonny (3.7.9 by default).
2. Inside of this folder, find a file called **Install Certificates.command**
3. Double-click on this file. It should open up a terminal window that runs some processes which should download the appropriate certificates to your computer.

If this doesn't work, try running the following in Thonny:

```
import nltk
nltk.download()
```

If that *still* doesn't work, go to office hours to get help from a TA.

Step 2: Learn `pandas` **[0pts]**

This project will use the pandas library to help do statistical analysis of the social media dataset. We have not taught the pandas library in class, so you'll need to do a bit of additional instruction to learn the new syntax.

Watch the linked video and review the linked code and slides to learn the basics of pandas syntax. If you have questions, feel free to post them on Piazza.

Video: https://www.cs.cmu.edu/~110/hw/hw6_pandas_tutorial.mp4
Code: https://www.cs.cmu.edu/~110/hw/hw6_pandas_examples.py
Slides: https://www.cs.cmu.edu/~110/hw/hw6_pandas.pdf

If you'd like to learn more, check out the pandas documentation: pandas.pydata.org

Step 3: Parse Label **[10pts]**

One of the columns in the dataset contains the message's **label**, which describes the person who posted the message. We'll use this column to gather information for four columns in the database: Name, State, Position, and Region.

Write a function `parseLabel(label)` which takes a string (the label) and returns a dictionary mapping three keys to three pieces of information. The key `"name"` maps to the name of the poster; the key `"position"` maps to the position of the poster; and the key `"state"` maps to the state of the poster.

How can we derive this information? We're guaranteed that the label takes the format:

```
"From: <Name> (<Position> from <State>)"
```

So we can use the text that is guaranteed to come before/after each piece of information to isolate the text (using string methods). However, we have to be careful- any of these three pieces of information could be any length, and some may contain more than one word!

For example:
```
parseLame("From: Steny Hoyer (Representative from Maryland)") ->
{ "name" : "Steny Hoyer", "position" : "Representative", "state" :
"Maryland" }
```

To test these functions, run `testParseLabel()`.

Note: to keep the starter file from becoming too crowded, all the tests have been moved to the file `hw6_social_tests.py`, which you can open and read while debugging. The functions in this file are called at the bottom of `hw6_social.py`. **If you change the filename of `hw6_social.py`, you will need to change the filename imported by `hw6_social_tests.py` in the first line of the file too**.

Step 4: Find Region **[5pts]**

We can get the poster's name, position, and state from the label, but to get their region we'll need to take an extra step. Write a function `getRegionFromState(stateDf, state)` that takes `stateDf`, a DataFrame, and a string of a state name to search for, and returns the corresponding region. `stateDf` has rows where each state is in the column `"state"` and its corresponding region of the US, like Northeast or South, is in the `"region"` column.

The pandas way to do this is to create a subset of the DataFrame containing only rows where the `"state"` column holds the state. Then you can index into the first (and only) row in the resulting DataFrame and return its `"region"` value.

Alternatively, you can iterate over each row in the DataFrame and check whether the state in the `"state"` column matches the provided state. Once you've found the row with the matching state, index into the `"region"` column and return the value at that position immediately.

To test this function, run `testGetRegionFromState()`.

Step 5: Parse Message **[15pts]**

We've finished analyzing the label, but we can still gather data from the posted message itself! Given a social media message, what hashtags are included inside that message? Write the message `findHashtags(message)` which takes a string (a social media message) and returns a list of strings- the hashtags included in that message, in order of appearance.

To do this, you'll need to find all the starting places of hashtags (look for the '#' character), and iterate over the characters that follow that starting index until you reach an 'end' character. We've provided a list of end characters in the global variable `endChars`; just check for whether each character is in that list, and stop when you reach one of them.

For example:
`findHashtags("I am so #excited to watch #TheMandalorian! #starwars")`
`-> [ "#excited", "#TheMandalorian", "#starwars" ]`

To test these functions, run `testFindHashtags()`.

Step 6: Identify Message Sentiment **[10pts]**

We can also analyze the message to determine its **sentiment** - whether the message is generally positive, negative, or neutral. Write the function `findSentiment(classifier, message)` so that it returns a string descriptor of the sentiment.

First, to get the actual sentiment score of the message, run the method `polarity_scores` on the `classifier` variable with the message as the argument. This method returns a dictionary with multiple results, but you only need to use one - index into the key `"compound"` to get the message score.

To get a descriptor, you'll need to map the score to a string. The string should be `"positive"` if the classifier predicts the message is positive (greater than 0.1), `"negative"` if it is negative (less than -0.1), and `"neutral"` if it is neutral (between -0.1 and 0.1). The generated string should then be returned.

For example, given the message `"Had opportunity to participate in panel on importance of disaster mitigation"`, the sentiment classifier returns `0.0516`, so findSentiment should return `"neutral"`.

To test this function, run `testFindSentiment()`.

Note: you might be curious about how the sentiment classifier works. Sentiment Analysis infers whether the statement has a positive, negative, or neutral connotation from the words in the sentence. This is a really hard problem. Think about the statement "That's great!" You could use it to mean it really is great, or you can use it in response to something that is really horrible.

The sentiment classifier is trained using thousands of example sentences in many different contexts in order to help it understand whether particular words or phrases are generally positive, negative, neutral (or could be used both positively and negatively). This classifier generates a number for the text given to it, where negative numbers are associated with negative sentiment, and positive numbers are associated with positive sentiment.

Step 7: Add Columns to DataFrame **[15pts]:**

Finally, we need to take all the new information we can extract and add it to the DataFrame. Write a function `addColumns(data, stateDf)` that takes in a DataFrame `data` and a DataFrame `stateDf` and destructively adds six new columns to `data` based on the functions you wrote above. Return `None` at the end of the function.

1. Create six new empty lists for names, positions, states, regions, hashtags, and sentiments.
2. Create a sentiment classifier using the function `SentimentIntensityAnalyzer()` (which has been imported from `nltk`) and store it in a variable.
3. Iterate over the `"label"` column in `data`. For each row's label:
    a. call `parseLabel` on the label to get the name, position, and state
    b. call `getRegionFromState` with `stateDf` and the parsed state to get its region
    c. append each of the extracted values to their respective lists
4. Iterate over the `"text"` column in `data`. For each row's text:
    a. call `findHashtags` on that text to get its hashtags
    b. call `findSentiment` on the classifier generated in Step 2 and the text to get the sentiment label
    c. append each of the extracted values to their respective lists
5. At the end of the loop, set `data["name"]`, `data["position"]`, `data["state"]`, `data["region"]`, `data["hashtags"]`, and `data["sentiment"]` to the respective lists.

To test this function, run `testAddColumns()`.

You have now finished the first stage of the project. To see the DataFrame you've created, run `runWeek1()`.

Check6-2 - due Tuesday 12/03 at noon EST

In this stage, you will dive deeper into actually analyzing the social media messages in the CSV. You'll group data by state and region to identify trends in types of messages, and you'll analyze different hashtags to determine which are most popular and what common sentiments are for different hashtags.

Before you start this second stage, go to the bottom of the starter file and uncomment the four test lines associated with Week 2.

Step 0: Written Assignment **[45pts]**

In addition to completing the steps described below, there is a short written assignment on the week's material. You can find the written assignment on Gradescope, and linked on the Assignments page of the course website.

Step 1: Organize Data to Compare by State **[10pts]**

Next, we will write functions that analyze the DataFrame and create dictionaries mapping states or regions to data about that state or region. We'll use these in Week 3 to create graphs that show trends in the data.

First, write the function `getDataCountByState(data, colName, dataToCount)` that takes the DataFrame, a column name (like `"sentiment"` or `"message"`), and a column value (like `"negative"` or `"attack"`), and returns a dictionary that maps state names to the number of message that had that value in the listed column. For example, `getDataCountByState(data, "sentiment", "negative")` would return a dictionary mapping states to the number of messages from that state that had negative sentiment. You should only include states that match at least one column of that type.

To solve this problem, start by filtering the DataFrame down to just the rows where the value in the column `colName` matches the value `dataToCount`. You can do this with a Boolean operation index, or by iterating over all rows and checking their values.

Then simply iterate over the relevant values in the `"state"` column and generate a dictionary that maps each state to a count of how often it occurs. We've done this in class before- you can use a familiar algorithmic pattern here.

This function should also handle one special case. If both `colName` and `dataToCount` are empty strings (""), just return a dictionary mapping each state to the number of data points it has (in other words, iterate over all rows instead of a subset of rows).

To test this function, run `testGetDataCountByState()`. We'll generate count dictionaries to count negative sentiments, attack messages, and partisan bias.

Step 2: Organize Data to Compare by Region **[10pts]**

Next, we want to organize data by region to find how many of *every* possible value occurs within a given column. Write a function `getDataForRegion(data, colName)` that takes a DataFrame and a column name (a string) and returns a nested dictionary. The keys of the outer dictionary should be regions, which each map to an inner dictionary. The keys of the inner dictionary should be the different values that occur in the column for that region.

For example, if we call `getDataForRegion(data, "message")`, the inner dictionaries may have the keys `"attack"`, `"constituency"`, `"information"`, and more! Each message type should map to the number of times it occurred in that region in the dataset (which you can count by iterating over all rows in the DataFrame).

Setting up a nested dictionary isn't too different from setting up a normal dictionary. The only big change is, if `d` is your outer dictionary, you'll need to set `d[key] = { }`. Then you can add new key-value pairs directly to `d[key]`.

To test this function, run `testGetDataForRegion()`.

Step 3: Get Hashtag Counts **[10pts]**

The next three steps will build towards one central goal: determining what the most popular hashtags in the dataset are, and whether they are used in an overall positive or negative sentiment.

First, write the function `getHashtagRates(data)` which takes the DataFrame and returns a dictionary mapping hashtags (strings) to counts of how many times they're used in the entire dataset (integers).

To do this, you should loop over the `"hashtags"` column, then iterate over the hashtags in the list (if there are any) to update the number of times each hashtag has been seen in the overall dictionary.

To test this function, run `testGetHashtagRates()`.

Step 4: Find Most Common Hashtags **[15pts]**

Next, write the function `mostCommonHashtags(hashtags, count)` which takes the hashtag dictionary generated in the previous function and the number of top hashtags to find (integer) and returns a new dictionary mapping just the top-count hashtags to how often they each occur. For example, `mostCommonHashtags(hashtags, 5)` would return a dictionary of five key-value pairs, the five most popular hashtags in the dataset.

One way to do this is to create an empty dictionary which will hold the known most-common hashtags. Then repeatedly search for the most common hashtag that is not already a key in that dictionary. Once you've gone through all the hashtags and found the one with the highest appearance rate, add it to the dictionary. Continue the process until the length of the dictionary is equal to `count`.

For example, using the rates determined from the main dataset,
`mostCommonHashtags(df, 7)` should return
`{ "#Obamacare" : 61, "#IRS" : 26, "#RenewUI" : 21, "#jobs" : 20, "#Benghazi" : 20, "#ObamaCare" : 20, "#SOTU" : 20 }`

To test this function, run `testMostCommonHashtags()`.

Step 5: Get a Hashtag's Sentiment Score **[10pts]**

Finally, write the function `getHashtagSentiment(data, hashtag)` which takes the DataFrame and a hashtag (string) and returns a 'sentiment score' (float) for that hashtag.

You will need to iterate over the rows of the DataFrame to find every row that contains the given hashtag in its `"hashtags"` column. For each row containing the hashtag, get its sentiment value. If it is `"positive"`, that maps to 1; `"negative"` maps to -1; and `"neutral"` maps to 0.

Note: a message can possibly contain a hashtag multiple times. When this happens, only count the sentiment **once** for that message.

To get the sentiment score, average all the individual hashtag-message scores together and divide them by the total number of messages that contained the hashtag. If this number is positive and close to 1, the messages were generally very positive; if it is negative and close to -1, they were generally very negative. A score close to 0 might indicate a neutral hashtag, or one which has a mix of positive and negative sentiments.

To test this function, run `testGetHashtagSentiment()`.

Now that you've written all the analysis functions, you can run `runWeek2()` to check out your results!

Hw6 - due Friday 12/06 at noon EST

In the final part of this project, you will visualize all of the data you have analyzed in the previous stages. To do this, you will use matplotlib to create graphs and charts.

Before you start this last stage, go to the bottom of the starter file and uncomment the two test lines associated with Week 3.

Step 0-A: Complete Check-in 1 **[20pts]**

If you got a perfect score on Hw6 Check-in 1 (the project part), congratulations; this step is already done! Go to the next step.

Otherwise, go back to your Gradescope feedback on Check-in 1 and use it to update your Check-in 1 code. This is your chance to implement any features you might have missed before, and fix any code that isn't working.

Step 0-B: Complete Check-in 2 **[20pts]**

If you got a perfect score on Hw6 Check-in 2 (the project part), congratulations; this step is already done! Go to the next step.

Otherwise, go back to your Gradescope feedback on Check-in 2 and use it to update your Check-in 2 code. This is your chance to implement any features you might have missed before, and fix any code that isn't working.

Step 1: Review Provided Code **[0pts]**

Drawing graphs with matplotlib requires a lot of setup code, and we want you to draw quite a few graphs, so we've provided a few functions for you to use, to simplify things. You will write some functions that call the functions we've provided.

You should definitely know what each of the following functions do, and we recommend that you look over their code at the bottom of the file quickly, to get a sense of how they work.

- **`sideBySideBarPlots`:** generates two bar charts side-by-side, for easy comparison. The x values come from xLabels, the y values come from valuesList, and labelList is used as a key.
- **scatterPlot:** generates a scatterplot of numerical data. xValues and yValues provide the position of each data point, and labels provide a descriptive term.

Step 2: Graph State Count Dictionaries **[15pts]**

Write a function `graphStateCounts(stateCounts, title)` which takes a dictionary mapping states to some kind of count (integer), and a title (string), and displays a simple bar chart of the count data using Matplotlib, returning nothing.

Make a list of the dictionary's keys and another list of the dictionary's values - the keys (states) will be the labels of the graph, and the values (numbers) will be used to determine how tall the bars should be.

If you use the default settings, the labels on the graph will all overlap. Check out the `plt` method `xticks` - it has a keyword argument that will let you change the label orientation to be vertical instead.

To test this function, check that the first graph produced by `runWeek3()` has 50 bars, that the largest bars are Texas and California, and that the smallest bars belong to Montana, Maine, and Hawaii. The second graph produced (narrowed down to only Facebook data) should have only 49 bars - South Dakota should be missing, as it only has Twitter data.

Step 3: Graph Top N States **[10pts]:**

Write a function `graphTopNStates(stateCounts, stateFeatureCounts, n, title)` which takes a dictionary mapping states to total number of messages, a dictionary mapping states to the number of messages that match a certain feature, an integer `n` for the number of top states to select, and a title, and displays a graph (returning nothing). The function should find the top `n` states that have the highest **rates** of a feature occurring overall to provide to the graph as data.

Note that a feature's rate is different from its count! If we just looked at counts, the biggest states (Texas and California) would always end up at the top. Instead, you should divide each state's feature count by the total number of messages (from `stateCounts`) to get the frequency rate for that feature.

To find the top `n` states, use a similar approach to what you did in `mostCommonHashtags` before. Just make sure to check rates instead of counts!

Once you've created the dictionary of the top states, call `graphStateCounts` with it to graph the top `n` states. Don't forget to include the provided title!

To test this function, check the third, fourth, and fifth graphs produced by `runWeek3()`. The third (graphing attack rates) should have five bars: Louisiana (which should be at around 0.14), Kentucky, Georgia, South Dakota, and Indiana. The fourth (graphing policy rates) should have five bars: Wyoming (which should be around 0.6), Maine, Alaska, Mississippi, and Michigan. The fifth (graphing nationally-focused messages) should have five bars: Vermont (which should be close to 1.0), Maryland, Connecticut, Mississippi, and Utah.

Step 4: Graph a Region Comparison **[20pts]:**

Write the function `graphRegionComparison(regionDicts, title)` which takes a nested region dictionary and a title and displays a graph (returning nothing). The function should take the region nested dictionary and create new lists of the data within that will be used to produce a side-by-side bar chart that can compare all four regions at once. You will need to make three lists: a list of **feature names**, a list of **region names**, and a list of **region-feature lists**.

To generate the list of feature names, iterate over the regions in the region dictionary, then iterate over the features in each region's inner dictionary. Construct a feature list over time, adding any features you find that are not yet in the list.

To generate the list of region names, simply iterate over the keys of the region dictionary and add those keys to a list.

To generate the list of region-feature values, iterate over the regions of the region dictionary and create a new temporary list for each region. This temporary list will hold the counts of how often each feature occurred in the region (with the value at index i corresponding to the feature at index i of the feature list). To produce this list, iterate over the **feature list**. If the feature you're looking at is in the region's inner dictionary, add the count to the list; if it isn't, add 0 (as the feature never showed up). When the temporary list has been fully constructed, add it to the main region-feature list (which is a 2D list).

Once you have all three lists, call the function `sideBySideBarPlots(xLabels, labelList, valueLists, title)` to produce the side-by-side graph. The feature list should be the `xLabels`, the region list provides the `labelList`, and the region-feature list should be the `valueLists`. Don't forget to include the title as well!

To test this function, check the sixth and seventh graphs produced by `runWeek3()`. The sixth chart (message types by region) should show 9 clusters of four bars each, with the South region containing a much larger number of certain types of messages (policy, attack, personal, support, information, media) than the other three regions. The seventh chart (position by region) should show two groups of 4 bars, and should demonstrate that there are many more messages by Representatives than Senators (especially in the South).

Step 5: Graph Hashtag Sentiments to Frequencies **[15pts]:**

Finally, write the function `graphHashtagSentimentByFrequency(data)`. This function will use the functions you've built so far to generate a graph that compares hashtags' frequency (how often they occurred) to their average sentiment score.

Start by generating a dictionary mapping hashtags to their counts (with `getHashtagRates`), then get the top 50 most common hashtags (with `mostCommonHashtags`).

You'll need to generate three lists - a list of hashtags, a list of frequencies, and a list of sentiment scores. You can build up these lists simultaneously by looping over the hashtags in the most-common dictionary (the dictionary gives you the hashtag and count, and you can call `getHashtagSentiment` to get the sentiment).

Once you have the three lists, call the function `scatterPlot(xValues, yValues, labels, title)` to draw the chart. (It may take a while, since the functions are not particularly efficient, but should not take more than a minute). You should use the frequencies for the x values, the sentiment for the y values, and the hashtags for the labels. You should also write your own title to describe the chart.

To test this function, check the eighth (and final) chart produced by `runWeek3()`. The scatter plot should have one hashtag all the way to the right (above 60 on the x axis) and the rest clustered between 0 and 30 on the x axis. The general trend should be that the messages are neutral to positive - few messages should be below -0.25. If you want to investigate the left side more closely, use the zoom feature built into matplotlib by clicking on the magnifying glass icon and drawing a rectangle around the part of the chart you want to zoom.

Optional final step: you'll notice that some data point labels overlap, so that you can't read them. This happens when two or more data points share the same frequency x sentiment pair. If you want to fix this, you can modify the `graphHashtagSentimentByFrequency` function to **merge** data points that share the same frequency x sentiment, by combining hashtags together into a single string. If only one data point/label is drawn, you'll be able to read all the hashtags easily.

Congratulations- now you're done! Feel free to try running some of the graphing functions or analysis functions on different aspects of the dataset, to investigate different political trends.