Background

Coca-Cola launched four new Diet Coke flavors in January 2018: Ginger Lime, Feisty Cherry, Zesty Blood Orange, and Twisted Mango. The new flavors were accompanied by modernized packaging and a new ad campaign. Due to stagnation in sales with Millennials, Diet Coke sought to use this campaign as a way to attract Millennial attention.

The account team has asked the analytics team to perform a series of analysis to determine the effectiveness of the campaign. This includes but is not limited to: identifying areas of improvement, determining whether the target audience was reached, and evaluating how well their set of product influencers performed.

As a data analyst, you are asked to use python(2or3) to perform the following data analysis to support business team. Your final deliverables will include: spreadsheet(csv) files reflecting all the work you would have done and notebooks with your code to generate the csv files.

Analysis

Use Diet Coke Raw Data.csv for the following analysis.

Section 1 Metrics

1. We define “social engagement” as the sum number of “likes”, “comments” and “reposts”. Identify the volume of mentions and total engagement for each month across all platforms and output the result in csv format.
2. Only keep the mentions from Twitter, Instagram, Facebook, News, Blogs and Forums, and group all the related mentions into 3 categories: 1) Social, 2) News, 3) Blog & Forum. Then try to identify which category has the highest sentiment and what’s the total number of positive mentions of that group? Output the result in csv format.
3. Find the top 10 authors with the highest followers for each social platform. Business analyst also want a Url link to authors' profile or mention page, help them on that as well. Output the result in csv format.
4. List all the news websites. (i.e. NYTimes)

Section 2 Topic Analysis

1. Write functions to find the official posts for each brand. Add a new column called “Official” and tag the brand name for each official mention. All the official handles can be found in Brand Official Handle Keywords.csv.
2. Following step 1, tag all the mentions left as “UGC”(User Generated Content). Write functions to identify topic for each UGC mention in only “Social” category by using Emotions Keywords.csv. (The first row of the file is the topic name, from the second row and below are the keywords for each topic) Tag topic name for each mention being matched.
    - (Hint) It is possible that some mentions can be matched by multiple topics. So you need to keep all matched topics.
    - (Hint) Try to use regular expression. Explain the advantages and disadvantages of regular expression.

    - (Hint) Make sure to take care of case sensitivity.
3. Find the number of mentions, total engagement and positive sentiment percentage for each topic.

Section 3 Longform Analysis

In this section, you will be asked to complete the topic analysis and sentiment analysis on longform posts. We define all the mentions from News, Blogs, and Forums to be longform posts. The difficulty of dealing with longform texts is one text can talk about different topics in different segments. To illustrate, take a look at the example below. This piece of text talks topic “Surprise” twice. The first appearance is captured by “thrill” and second by “amazing”. Each highlighted part correspond to one segment.

Hello all! I’ve noticed many of his P50 posts are a bit older now and with so many new products on the market. I’d love and thrill to hear any updated advice on this product. In particular, I already have a routine heavy on acids and actives but would like to incorporate P50. The consultant at the P50 counter at Liberty advised me to use P50W and not to use other Actives. However, I’ve noticed that many on here, and many bloggers, still maintain a heavy arsenal of Actives. Personally, I’m considering adding it in every other night with retinol on other nights, and still using a mild acid in the mornings (Pixi Glow Tonic or Glossier Solution). Also, I am trying to incorporate/use up my existing acid products here and there (Pixi 20% glycolic pads and Goldfaden MD lactic acid treatment which I currently just use here and there). Eventually, I’ll transition out my other acids and use P50 more regularly. it’s so amazing right? What’s your experience with P50? How often do you use it? Do you use other chemical exfoliants? Do you use other Actives like retinol or vitamin c? Any advice would be appreciated, thanks!

In order to calculate the sentiment as accurate as possible, we are going to do the following:

1. Segment the post into individual sentences.
2. Match each sentences against keyword.
3. The matched sentence with 2 sentences before and 2 after combined will be marked as “key content”.

Look back at the example above, this is how the example is generated.
Similar as section 2, in this section, we want you to write functions to complete the following steps.

1. Classifying topics for all mentions from News and Blog & Forum categories by using Emotions Keywords.csv and find the sentiment score for every emotion topic in each longform articles.
2. Please attach all the key content for each topic in all each longform article in a data frame. (For instance, create a new column called “key content” in the data frame and attach the highlighted area in the cell that corresponding to the article showed above)
3. Please included the number of topics matched for each longform article.
4. Perform sentiment analysis on key content.
    - Hint: you may use this https://github.com/youhealthy/vaderSentiment

Hint: It is possible that some longform articles can be matched by multiple emotion topics.

Section 4 Dashboard

In this section, you are required to complete the data visualization for the following charts in a Dashboard.

1. Make a line graph to indicate the weekly volume of mentions change in the data set.
2. Use Bar chart to list top 10 websites by volume of mentions for News category and Blog & Forum categories respectively.
3. Create a bubble chart to show the engagement weekly change in the Social category. The x-axis would be the week, the y-axis would be the percentage of positive mentions of the week, and the bubble size would be engagement.