Text Analytics Part I: Time Series Analysis

1. **Defining the problem and solution**
   We want to see how a certain topic trends over time (In a real world dataset we would take already trending text surrounding the topic and use the most frequently used words). For this small scale dataset we are looking for the use of the words “apple” or “orange”.

2. **Choosing Relevant Data**
   We really only need two columns for this problem set “authored date” and “content”

   Table 1)

Authored Date	Content
2-12-2021	**I like apples a lot, but I only have an orange.**
2-12-2021	**Oranges are my best friend**
2-13-2021	**Do you like apples? I only have an apple.**
2-14-2021	**Oranges are my favorite**
2-14-2021	**Are we talking about fruit?**

3. **Breakdown the Problem:**

   Time series is usually run on quantitative data, so we want to quantify the content. Therefore we need to explore some NLP methodology. We will start out with asking a few questions about what we want to capture, such as:

- **What are some basic NLP Methodology we can use?**
  Typically we will lowercase all characters, remove punctuation, and remove stop words. This works for our current project.

  Result:
  Table 2)

Authored Date	Content
2-12-2021	**like apples lot only orange**
2-12-2021	**oranges bestfriend**
2-13-2021	**like apples only apple**
2-14-2021	**oranges favorite**
2-14-2021	**talking fruit**

- **Do you want to include plural use of the word apple and orange i.e. “apples” and “oranges?**
  The answer to this is yes. So we can either lemmatize the strings or we can manually account for the singular and plural versions of each word. The better practice would be to

apply lemmatization to each row in content. Once we lemmatize the content, the only variation available should be apple or orange.

Result:
Table 3)

Authored Date	Content
2-12-2021	like apple lot only orange
2-12-2021	orange bestfriend
2-13-2021	like apple only apple
2-14-2021	orange favorite
2-14-2021	talk fruit

- **If the word is used multiple times in one sentence, do we want to count duplicates?**
  For this problem set, we really want to see how much unique content related to the topic is trending. Therefore, we should delete duplicates of the same word. We can do this by creating a list of unique words per row.

  Result:
  Table 4)

Authored Date	Content
2-12-2021	[“like”, “apple”, “lot”, “only”, “orange”]
2-12-2021	[“orange”, “bestfriend”]
2-13-2021	[“like”, “apple”, “only”]
2-14-2021	[“orange”, “favorite”]
2-14-2021	[“talk”, “fruit”]

  Now we want to focus on looking at the use of each word by date. Therefore, we can group the data frame by date and collapse the content into a larger list of words. This time we want to allow for duplicates so we can quantify the frequency of the topic by each date. However, this leads to another question before we group rows by date:

- **If two of the trending words are in the same row, do we want to account for both of them?**
  Logically, if we are looking for one topic and two trending words are appear in that same topic, then they both apply to the “one” topic, meaning we would not want to account for both of them. At this point I can use a simple sql query to create a new dataframe (table) that drops any rows that do not contain at least one of the trending words.

  Result:
  Table 5)

Authored Date	Content
2-12-2021	[“like”, “apple”, “lot”, “only”, “orange”]
2-12-2021	[“orange”, “bestfriend”]

2-13-2021	**[“like”, “apple”, “only”]**
2-14-2021	**[“talk”, “fruit”]**

I can now simply group by date and create a new column that sums the count of each row.

Result:
Table 6)

Authored Date	Content	Content Count
2-12-2021	**[“like”, “apple”, “lot”, “only”, “orange”, “orange”, ”bestfriend”]**	**2**
2-13-2021	**[“like”, “apple”, “only”]**	**1**
2-14-2021	**[“talk”, “fruit”]**	**1**

Now that my data is prepped, I can visualize it.

Part 2: Attempting to Implement the Solution

Check my Github for the full code.

Continuation of this Project and Next Writing Topics:
Text Analytics Part I: Web Scraping for Analyzing Trending Topics
Text Analytics Part III: Content Clouds to Visualize Trending Topics
Text Analytics Part IV: Entity Extraction using Spacy