**SOFTWARE IMPLEMENTATION LINK:**
https://colab.research.google.com/drive/1ifKkuqZDRo6b617OlWY2lbnjR7871uX1?usp=sharing

DATA301 REPORT

Abstract / Summary


- The research question is to analyse the tweets of AppleMusic and Spotify. How their tweets differ from each other, to see if they follow what's trending and the techniques they might use in their tweets. The subset of this Twitter data is retrieved through Twitter API. I will be using sentiment analysis to determine how their tweets differ, cosine similarity if they follow trends and TF-IDF to see the techniques they might use in their tweets. Apple Music and Spotify are the two most used music streaming apps. The significance of this project is to see any differences or similarities in their marketing strategy in managing their twitter account.


Introduction

- The Twitter API needs to be authenticated first to access twitter data. The Twitter API pulls twitter data in a json format with specific arguments. I set the arguments to match for AppleMusic through arguments 'screen_name' to match their username and 'count' for how many tweets I want to pull. The data is then changed from json to an rdd format. As for the algorithms I will use, sentiment analysis will take any starting and return an integer score to determine if the tweet is positive, negative or neutral. Cosine similarity will take two lists then return how similar they are. TF-IDF will take in strings and perform which is more frequent in the combined tweets. In addition, I will use spark map-reduce operations as well to parallelize the data, and python to process the data sequentially, useful for defining functions.

- This research is relevant to me as I used to use Spotify last year, yet I am now using AppleMusic because I saw an ad in twitter talking about their spatial audio and rich dolby sound. That's what got me to switch due to their better music. I wonder how many people also got 'switched' through AppleMusic's advertising and this puts importance on how advertising can do to a company. I think in this modern age, almost everyone is subscribed to AppleMusic or Spotify and have a Twitter account. This makes it easier to share this interest with them, especially students around my age.

- The research question is to analyse the tweets of AppleMusic and Spotify. How their tweets differ from each other, to see if they follow what's trending and the techniques they might use in their tweets. This question is relevant to the dataset because I can pull the tweets for both Apple Music and Spotify and do analysis on them. I can then do sentiment analysis on the tweets, put it into a list both for Apple Music and Spotify then do cosine similarity on them to determine if their tweets follow what's trending. I

can also do TF-IDF to determine what words they use the most and see if there's any pattern in their advertising technique.

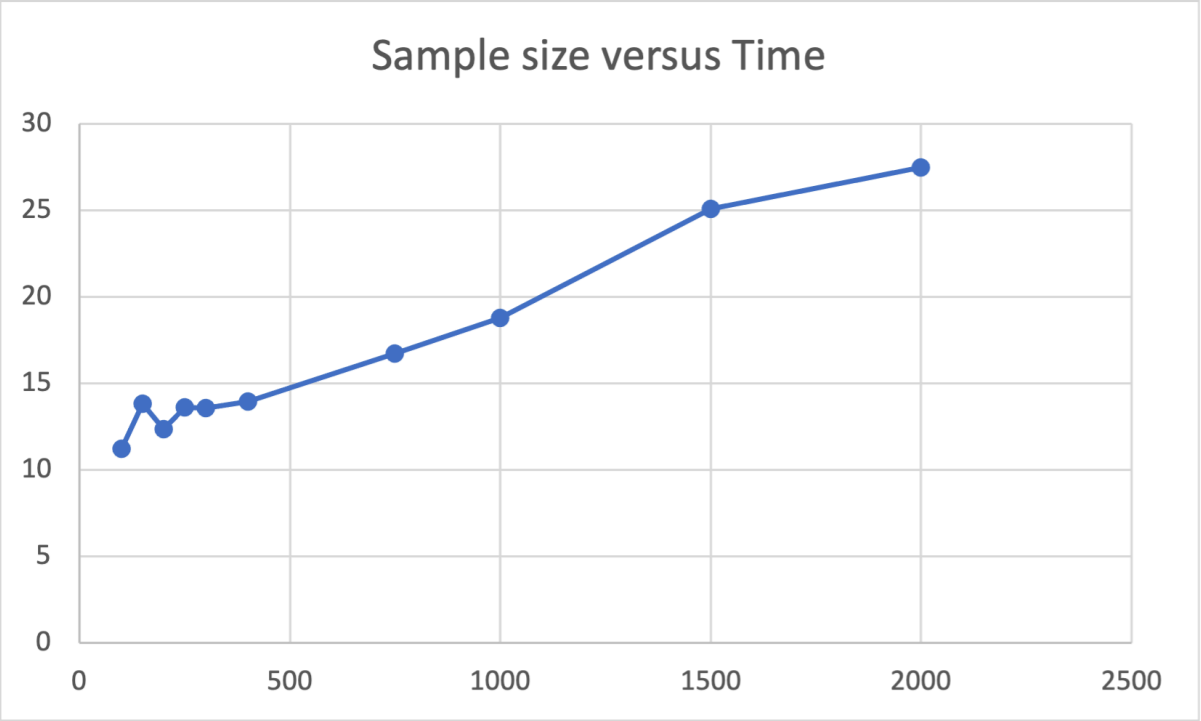
Experimental Design and Methods

- Installing and configuring Spark and Python would be first. Then importing specific libraries like textblob for sentiment analysis. I also needed to configure Twitter API and get access before I can start pulling data. The data pulled should also have a 'max_id' where it will pull tweets that are not greater than the specified 'max_id'. This keeps the data consistent and is not dependent on their recent tweets meaning recent tweets will not be pulled as they are greater than the 'max_id'.
- The twitter data is in a json format so I have to convert it to an rdd format to do spark map/reduce operations. The twitter data would then have to be cleaned, removing any emojis, https and unusual features like extra whitespaces. This is done so textblob can do sentiment analysis on the tweets more efficiently and be able to return a better score. The next step is then to visualise the data through a dataframe, useful for looking at the tweets and their respective sentiments/score. We can also take the tweets with the most positive score and most negative score for further analysis. This is both done for Apple Music and Spotify. In addition, we can also get the overall sentiment for their tweets by getting their percentage. The sentiments of their tweets are then compared through a cosine similarity to see if they follow a trend. This means that if they do follow what's trending, then their tweets would have similar contents therefore would return a higher similarity. The last step is to perform TF-IDF on the tweets. First is to get the TF (term frequency) of each word in their tweets to do a word count then perform IDF for all tweets. Overall, the end result will be a TF-IDF where it tells the importance of a word in a corpus of documents. This is done to determine what word appears most in their combined tweets and see its importance.

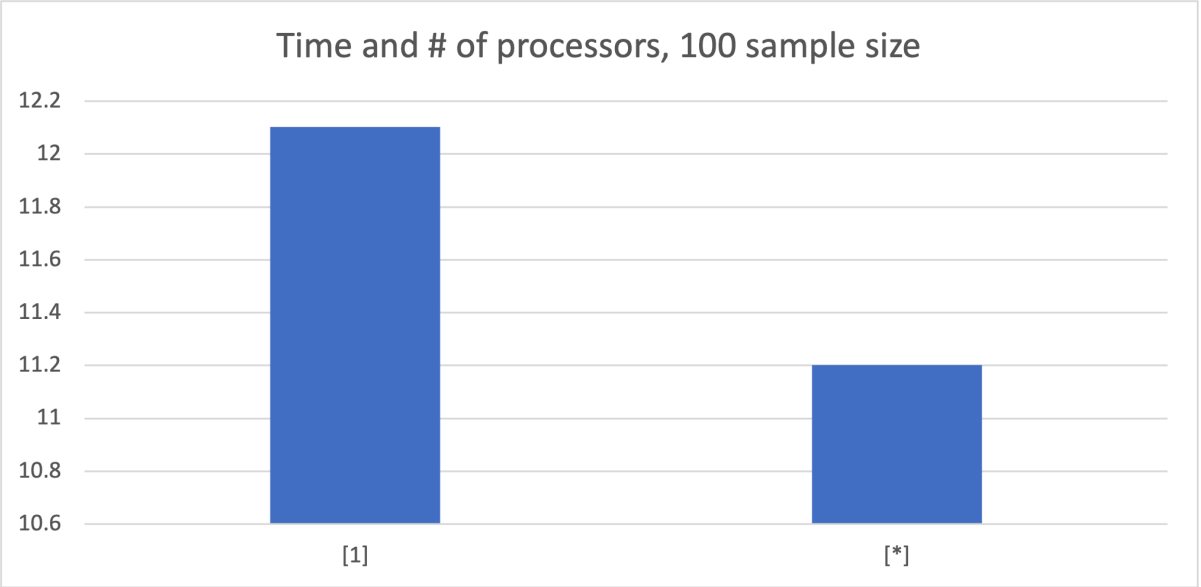
- `import tweepy # accessing twitter API and its data`
- `from textblob import TextBlob # sentiment analysis on the tweets`
- `def remove_emoji(string): # to remove emoji for sentiment analysis`
- `def clean_twts(tweet): # to clean the data for sentiment analysis`
- `def polarity_twts(tweet):`
- `return TextBlob(tweet).sentiment.polarity # sentiment analysis in action taking the tweets and returning a score/sentiment`
- `def pos_scores(sentiment): # getting the total number of positive tweets, useful for getting the percentage of positive tweets`
- `def neg_scores(sentiment): # same as above, but for negative tweets, this and pos_scores will be used to determine the general`

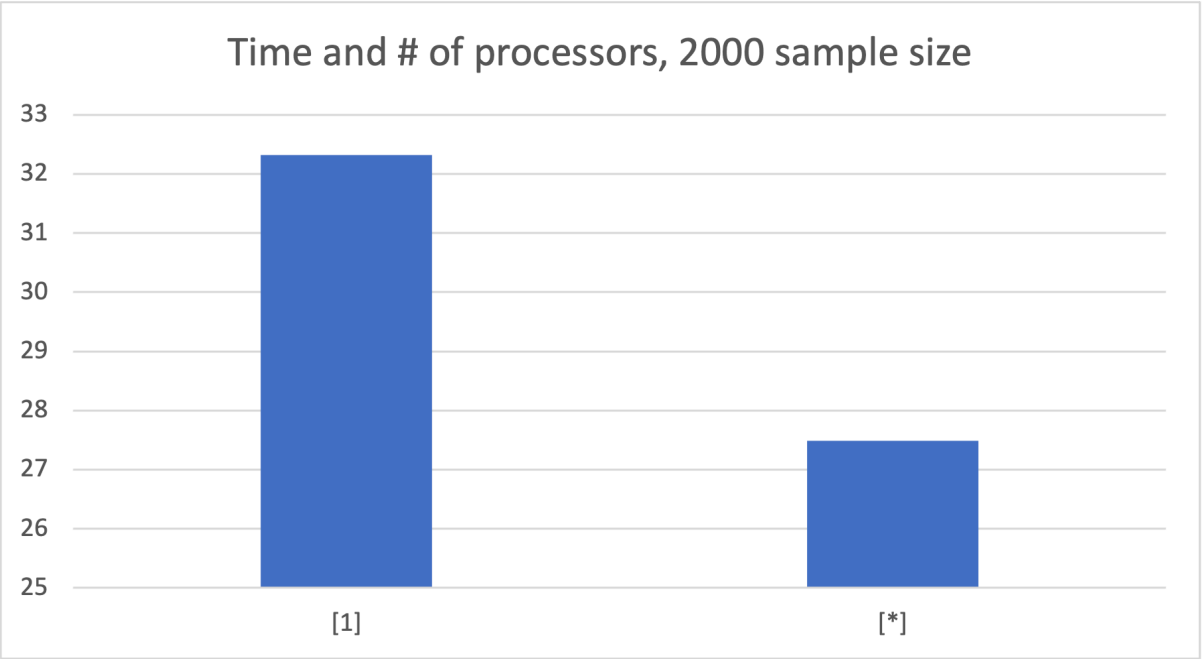
sentiment of tweets of Apple Music and Spotify, if it is positive, neutral or negative, maybe come up with a pattern.

Results




The general trend is that the time it takes to process the data increases as the sample size increases as well implying a negative relationship.






Both sample sizes 100 and 2000 tell us that as there are more processors, the time it takes to process the data decreases implying a negative relationship. This means that the data is parallelised/spread into more worker nodes, so more worker nodes are able to process at the same time decreasing the time elapsed.

- The research question is to analyse the tweets of AppleMusic and Spotify. How their tweets differ from each other, to see if they follow what's trending and the techniques they might use in their tweets. Apple Music and Spotify tweets differ from each other. With Apple music, 56% of their tweets are positive, 8% are negative and 36% are neutral implying that AppleMusic's tweets are pretty positive. This could be because Apple Music is an extension of Apple and positive tweets will help its image not just as a music streaming app, but also as a business. With Spotify, 51% of their tweets are neutral, 41% are positive and 8% are negative. Their tweets are fairly neutral and this could be because Spotify is not an extension of a big company, more of a like a stand-alone company. It does not have to keep a 'positive' image like Apple Music to help boost sales. Although they differ, both of their tweets stay away from being negative as that could push away users.
- Apple Music and Spotify do not follow trends with their tweets. Their cosine similarity is only about 0.25 similar. My reasoning behind this is that if they do follow what's trending, then they would post similar content in their tweets hence would have a similar sentiment. Both of them tweet different contents and the next part of this analysis is to see what content they tweet more often. The TF-IDF for Apple Music shows that the content/words that are of more importance are 'spatial audio' and 'dolby'. They really capitalise and focus on those as that is their selling point and that's what got me to switch to them. As for Spotify, it is '#spotifyisland' which is their recent thing that 'brings new experience for fans and artists to Roblox.' Overall, both music streaming apps have a different delivery of their tweets and have different contents, but one thing that is similar is that they focus more on services they offer and highlighting their importance as a result. Perhaps, this is the reason why the two

are seen as distinct although they offer the same songs and how they are still relevant both to the outside and the inside of the Twitter community.

Conclusion (suggest 3 paragraphs total, one for each prompt)

- I was able to answer my question of analysing the tweets for Apple Music and Spotify. I was able to tell that their tweets are different through sentiment analysis. Apple Music generally tweets positively, whereas neutral for Spotify. I was also able to tell that their tweets are not similar through cosine similarity implying they don't follow trends. Furthermore, I was also able to see a technique they used with their tweets which is highlighting their services/selling point. Overall, I was able to analyse their tweets and somehow determine their advertising strategies.

- Small sample size. This research has only 100 tweets each from Spotify and Apple Music. The implication of a small sample size is that there is a possibility for bias. Yes this was enough to answer my research question, but if the sample size is about 500, it may have different results and will affect how the question is answered.

- A future direction for this project is to pull the number of likes, retweets and comments per tweet. This is to determine which content or tweet generally gets more interactions. This then can help get the average number of likes, comments and retweets for each tweet. Very useful when comparing them to see which has more interactions in general.

Critique of Design and Project

- The cosine similarity could have more in it. Yes I was able to answer my question , but I felt like it was not enough. It was simply taking two lists of sentiment scores and comparing them.
- Perhaps, to improve this is that I can take the positive tweets for Apple Music and do the same for Spotify. Do a cosine similarity on them and they might return a higher result. If they do return a higher result, then that would mean that their positive tweets are fairly positive. I would do the same for negative tweets and neutral tweets. There could be a pattern as well to help explain why Apple is generally positive and Spotify is generally neutral with their tweets.

Reflection

- Mapping an rdd dataset to a function
- Cosine similarity
- TF-IDF
- List comprehension
- Twitter API

- Pandas for visualisation
- ‘Re’ for removing certain characters

- I learned in this project that both Apple Music and Spotify have different ways of getting their content across to their users in twitter, but they still remain relevant and popular. In terms of the skills I have learned, it is definitely how to access Twitter API. I can now confidently pull tweets and do analysis on them. I am also able to use re and pandas. I didn't usually use them before and now I know how to. This project definitely expands my skills as  a data scientist.

References


- https://newsroom.spotify.com/2022-05-03/spotify-island-brings-new-experiences-for-fans-and-artists-to-roblox/  # for the definition of #spotifyisland
- https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-user_timeline # where I got the idea of max_id
- https://www.youtube.com/watch?v=FmbEhKSpR7M&list=WL&index=3 # helpful for getting me started
- https://www.youtube.com/watch?v=ujId4ipkBio # helpful for getting me started
- https://poopcode.com/how-to-remove-emoji-from-text-in-python/ # where I got the code for how to remove emojis
- https://stackoverflow.com/questions/11331982/how-to-remove-any-url-within-a-string-in-python # how I got the code for how to remove specific character in string


- I have not shared any code, or worked directly with a student.