Encountered Problems for TxMM Project

As this is my first ever text mining project I tried to make the best out of it, as it is also one of the subjects that have really intrigued me in my first semester in this master's course. My main difficulty which will be evident in the project but was also recorded in the 'Research Project Timeline' was collecting the Tweets. I could say this difficulty can be divide in two phases.

The first phase I tried collecting COVID-19 related Tweets, obviously after applying for a Twitter Developers account, using Python. After following numerous tutorials and setting up all the required code correctly, I couldn't collect the required amount of Tweets I deemed necessary. That amount being at least over 5,000 Tweets. The maximum number of Tweets I could collect using Python was round about 2,000 Tweets including Retweets. I wasted a good amount of time searching on ways to fix this and thus I started using R to collect my Tweets.

Obviously, the second phase is about trying to collect COVID-19 related Tweets using R. Again after following a good amount of tutorials, I was able to collect about 18.000 Tweets in English. The time period of this part of the project was just before/during 'Black Friday'. After using this collected dataset I faced two problems. The first problem was an encoding problem, in more detail a good amount of text was just unknown characters and the second was that the majority of the Tweets were actually spam bots/auto bots posting about different advertisements and Black Friday deals using the #covid #coronavirus tags.

I could have used this dataset which also had some related Tweets to Donald Trump, but other than Black Friday advertisements and Donald Trump Tweets I wouldn't have any interesting findings and I also didn't want to adapt to a new idea or project title based upon my collected dataset.

Thus I preferred to find a 'ready' made and unseen dataset .That lead me to using a dataset from Kaggle. This dataset is a good fit for my project and also contains a large amount of collected Tweets. The only problem is that it is rather older and the Tweets were collected about 5 months ago.

Another problem occurs when trying to run the topic modelling part of the script. When running the topic modelling part of the script and creating the vector using CounteVectorizer for 100k Tweets and outputted vector of 113GB is created. I cannot analyze a vector that big as I get an error thus for the topic modelling part it will have a different number of analyzed Tweets compared to the sentiment analysis part of the project.

As this is my first Text Mining project I followed a number of tutorials which was time consuming and it took my rather long compared to what I expected to actually build up to the current script I have now.

I also didn't want to add a machine learning aspect to the project. During the course the teachers and TAs always underline the slight difference between machine learning projects and machine learning projects with a Text Mining aspect. Thus I decided not to add machine learning so I

don’t fall into this ‘trap’ and get off track. However an extension of this project in the future could add machine learning and then perform some kind of inference on the test set to determine the sentiment, emotion and topic.