***Use this word document to describe all your steps after each problem. Show all commands & screen pages and describe your steps. Points are deducted for missing steps.***

Describe all steps for each problem set and capture screen shots for each step with a JPG image within this document. Do not forget to put your name on the top of this file for your submitted homework. Also, please replace \_ALL\_ text in red with your information!

Problem 1:

1. If you do not already have a Microsoft account, go to https://signup.live.com and create one;

I’ve already had one： zhao.binb@northeastern.neu

A screenshot of a cell phone

Description automatically generated

1. In a browser, and if you do not already have a free Azure Notebooks account, go to https://notebooks.azure.com and sign up for one using your Microsoft account;

A screenshot of a social media post

Description automatically generated

1. Go to https://notebooks.azure.com/Microsoft/libraries/samples to view the sample notebook library.

There are 9 documents in this sample notebooks:

A screenshot of a social media post

Description automatically generated

1. Click Clone to clone the library to your Azure Notebook account. You can name the cloned library anything you like. Use the default name and remove the option to make your cloned library public.

I cloned the Azure Notebook and used default ID and I removed the option (public).

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. Run “Discover Sentiments in Tweets” and go over all steps. Explain each step.

A screenshot of a cell phone

Description automatically generated

Click the RUN and Discover Sentiments in Tweets.ipynb

A screenshot of a cell phone

Description automatically generated

（The tweet sample csv file）

1. The first step is to download the data into the “sentiment140-subset.csv”

A screenshot of a social media post

Description automatically generated

As the output shown above, the total time spent is 4s

2. then use the ls command to locate the file

A screenshot of a cell phone

Description automatically generated

It’s here!

3. Next we should prepare the data:

A screenshot of a cell phone

Description automatically generated

Import the packages (os & pandas) and prepare the dataset named “mydata” including 160k rows (tweet texts), 2 columns

4. Data Cleaning. Removing useless information on twitter, such as numbers and under scores and transform all text into lower case.

A screenshot of a cell phone

Description automatically generated

5. Using the Natural Language Processing toolkit: NLTK package to tokenize data (split data or sentences into words). Import Regular-Expression Tokenizers (RegexpTokenizer) to split string into substrings.

A screenshot of a cell phone

Description automatically generated

6. Then we have to calculate the value of term frequency-inverse document frequency matrix to use non-negative factorization. This value represents the frequency of the appearance of each word after adjusting stop words with higher frequency but no important meanings, such as “being”, “for”, etc.

A screenshot of a social media post

Description automatically generated

Matrix size: 10000 x 13379

7. We choose NMF for short text analysis:

A screenshot of a social media post

Description automatically generated

Based on the tf-idf matrix we calculated in the previous step, we can implement the decomposition.NMF function from sklearn to see the most frequent word in different topics as shown below:

A screenshot of a social media post

Description automatically generated

8. Data Visualization -- WordcloudA screenshot of a social media post

Description automatically generated

(install wordcloud)

**A screenshot of a cell phone

Description automatically generated**

(upgrade pip version)

A screenshot of a cell phone

Description automatically generated

(reinstall)

Before plot the wordcloud, we should separate the topics into different lists :

A screenshot of a social media post

Description automatically generated

3 topics identified from NMF↓

A screenshot of a social media post

Description automatically generated

A picture containing food

Description automatically generated A picture containing wooden

Description automatically generated A close up of text on a black background

Description automatically generated

The above wordclouds shows the most frequent words are “good, day, work, just…” in topic 1, “quot, love, know, just, http…” in topic 2, and “thanks, haha, following, got…” in topic 3.

A screenshot of a social media post

Description automatically generated

Total points: 10

Put all your steps and screen captures here.

------------------------------------------------------------------------------------------------------------