Section 1:
In this code, I use three functions to find the term frequency, inverse document frequency, and tf-idf measure. I assume that log in inverse document frequency is the natural log. I plot the tf and tf-idf values. The plots are relatively similar to one another.

Section 2:
In this section, I build a function called loss_SVM that calls on two helper functions to calculate the loss in the model. There are two versions of this code, one where the weights are used and another that also rounds the predicted value to either 1 or -1. The rounded values perform slightly better than the proper model.

Additional Note: The Dockerfile probably will not work. I tried asking ChatGPT to help me write the Dockerfile that will let me run Pyspark. This session was unproductive. I have been informed by the TAs that getting Pyspark to work is “not worth it” after I’ve spent a grand total of 2 hours trying to get Pyspark to work on my local computer. This means I spent 4 hours trying to get Pyspark to work. I’ve gotten my java -version and echos for JAVA_HOME and SPARK_HOME to output something.

Gen AI Disclosure:
I’ve used Google Collab which recommends code after commenting which finishes some of the code that I am writing. It is frequently wrong and requires editing. Other times, it’ll recommend things that are completely wrong.

Furthermore, I’ve used ChatGPT to help me debug and try and get Pyspark to work. This process was basically me saying:
“JAVA GATEAWAY error for pyspark.” Then saying “doesn’t fix the problem” over and over again while recompiling the Dockerfile.

I’ve used ChatGPT to help me debug code that Google Collab’s Gemini failed (I think it’s called Gemini). ChatGPT helped recommend some functions that would work such as withColumn and join. This included prompts like “What function can I use in Pyspark that joins two different dataframes” and “how to select specific columns in a dataframe” where ChatGPT recommended me using (*) to select multiple columns.