INLS 560 – Project: SMS Spam Detection

Date Assigned: November 7, 2019

Completion Date: November 21, 2019

Software issues:

If you feel there are mistakes in this assignment, check Piazza for corrections, and report them to us if they have not been made.

Assignment

In this assignment, you will create a python program that automatically predicts whether an email is spam or not spam.

This application consists of several parts.

Parts

---

1. Dataset
2. Python program to compute features
3. Weka
4. Python program to predict whether a SMS message is spam or not spam
5. Your Presentation

Your presentation will be presented at regular class time on April 25.

Dataset

The dataset is on Piazza under a note named: Project Dataset.
Please download the dataset from Piazza.

The rows with ham are not spam.
The rows with spam are spam.

Python Program to compute features

Create a file named predict_sms_spam.py.

This file should contain:

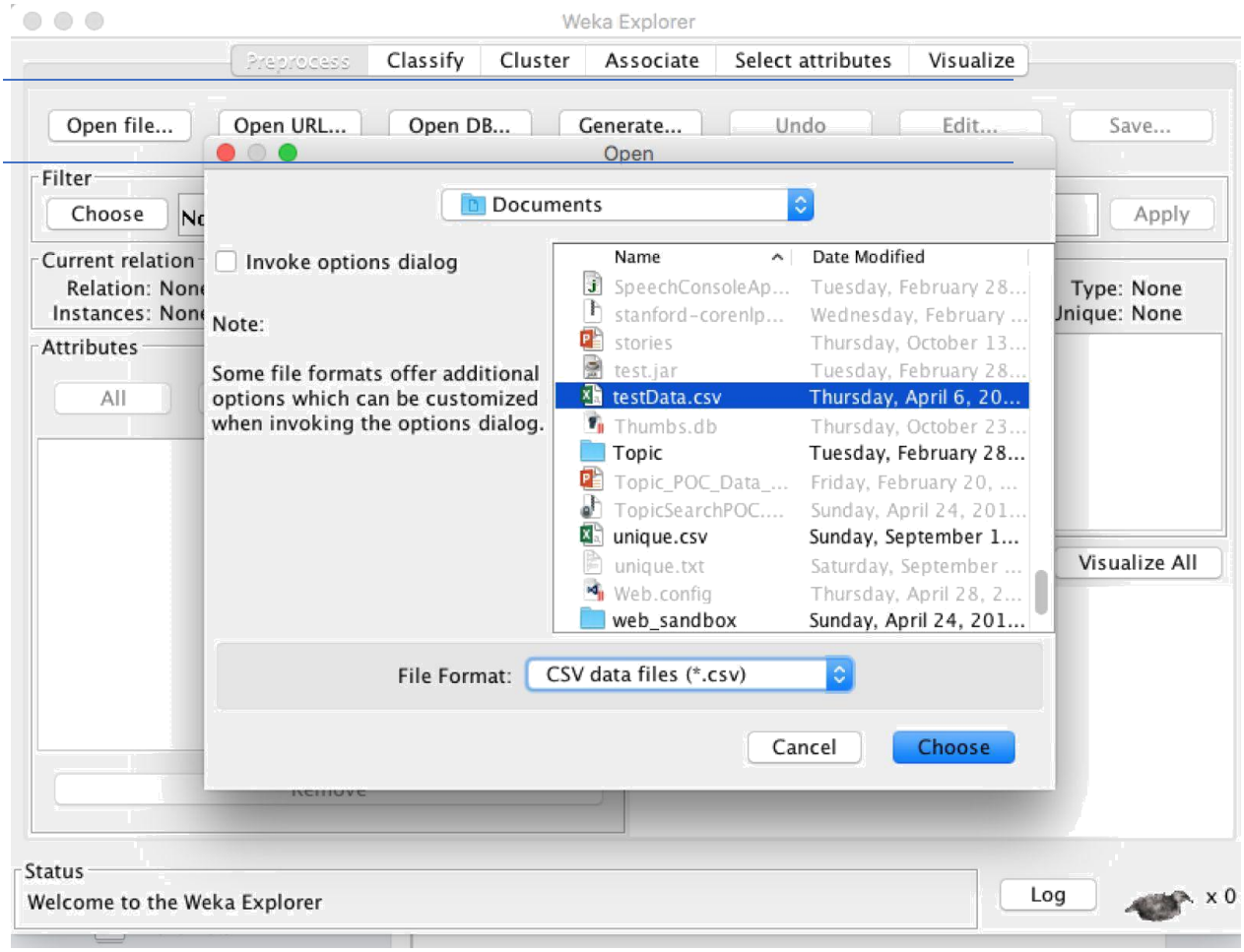
- A main function
- A function for each feature you will compute
- A function or reading in the dataset
- A function for writing the features to a csv file
    - The format for this file should be:
        - doesHaveLinks, doesHaveSpammyWords, LengthOfText, NumberOfSymbols, class (spam or ham)
    - Make sure that your file the headings above in the first row and the values of each column in subsequent rows

You should compute the following features:

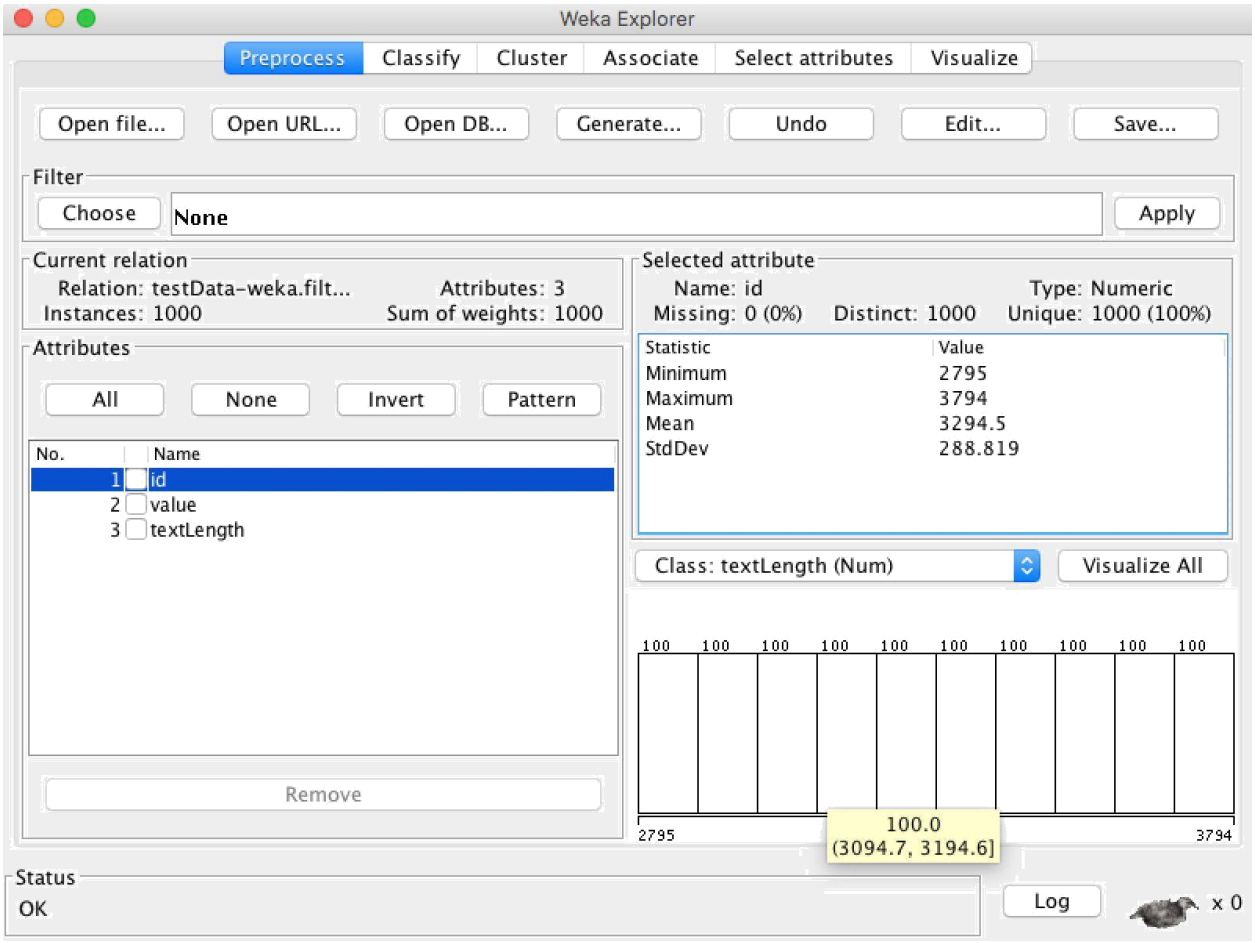
- **doesHaveLinks**: This feature is True if a sms message has links and False if a sms message doesn't have links
- **doesHaveSpammyWords**: This feature is True if the sms contains spammy words and False if a sms message does not have spammy words
    - To determine what words are spammy, look through the dataset and pay close attention to curse words and any words that are in the spam category but not in the ham category.
- **LengthOfText**
    - The number of characters including spaces in the text message
- **NumberOfSymbols**
    - The number of symbols in the text message

Create two additional features of your choice.

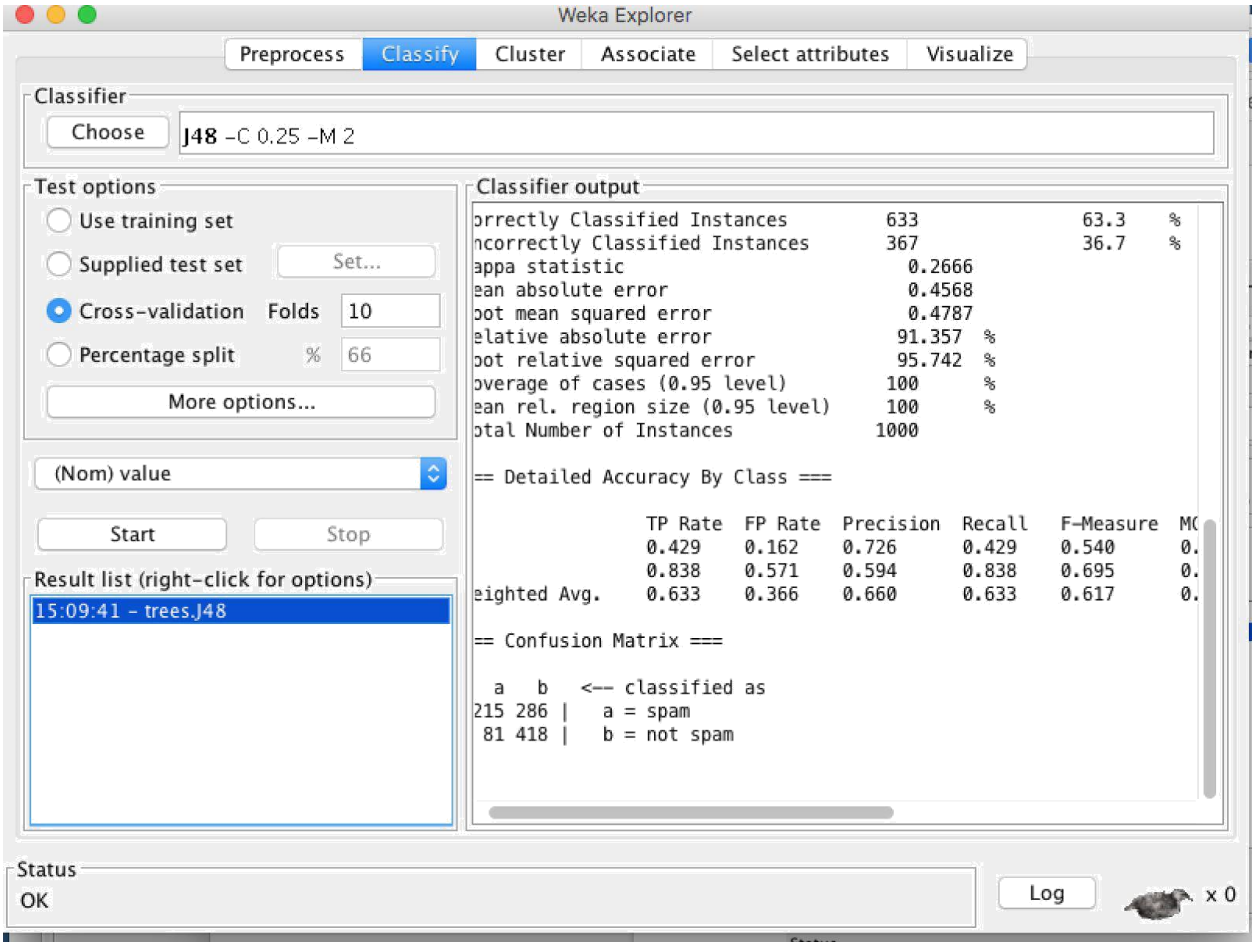
Select the csv file you created in the first python program. Make sure to select CSV data files under File Format.



After you press select, your screen should look similar to the one below (**your attributes will be different**):

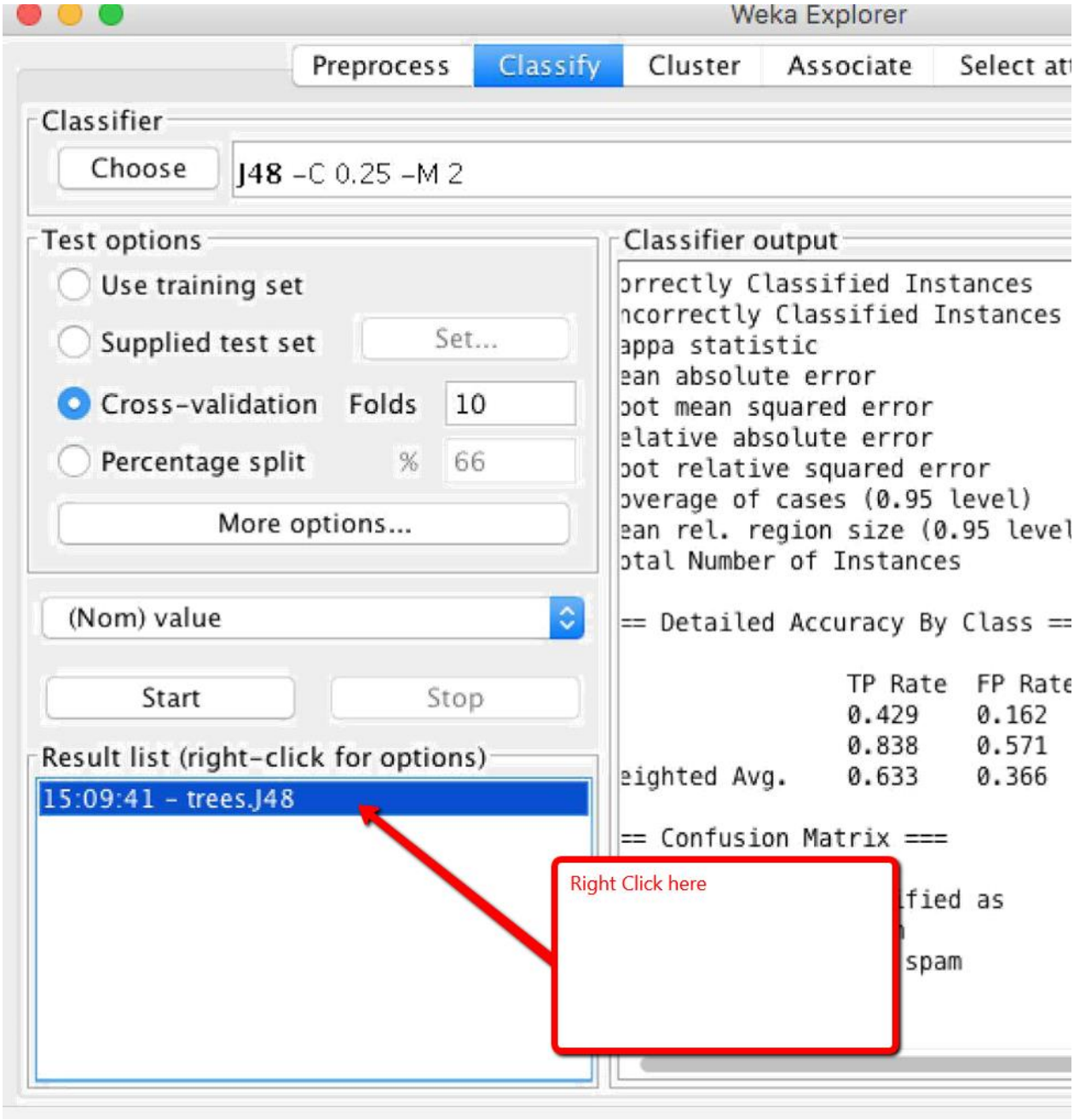


Next select, the Classify tab and **make sure the field (Nom) value is selected.** Then press the choose button, select trees and choose J48. J48 is the decision tree classifier. Finally press Start.

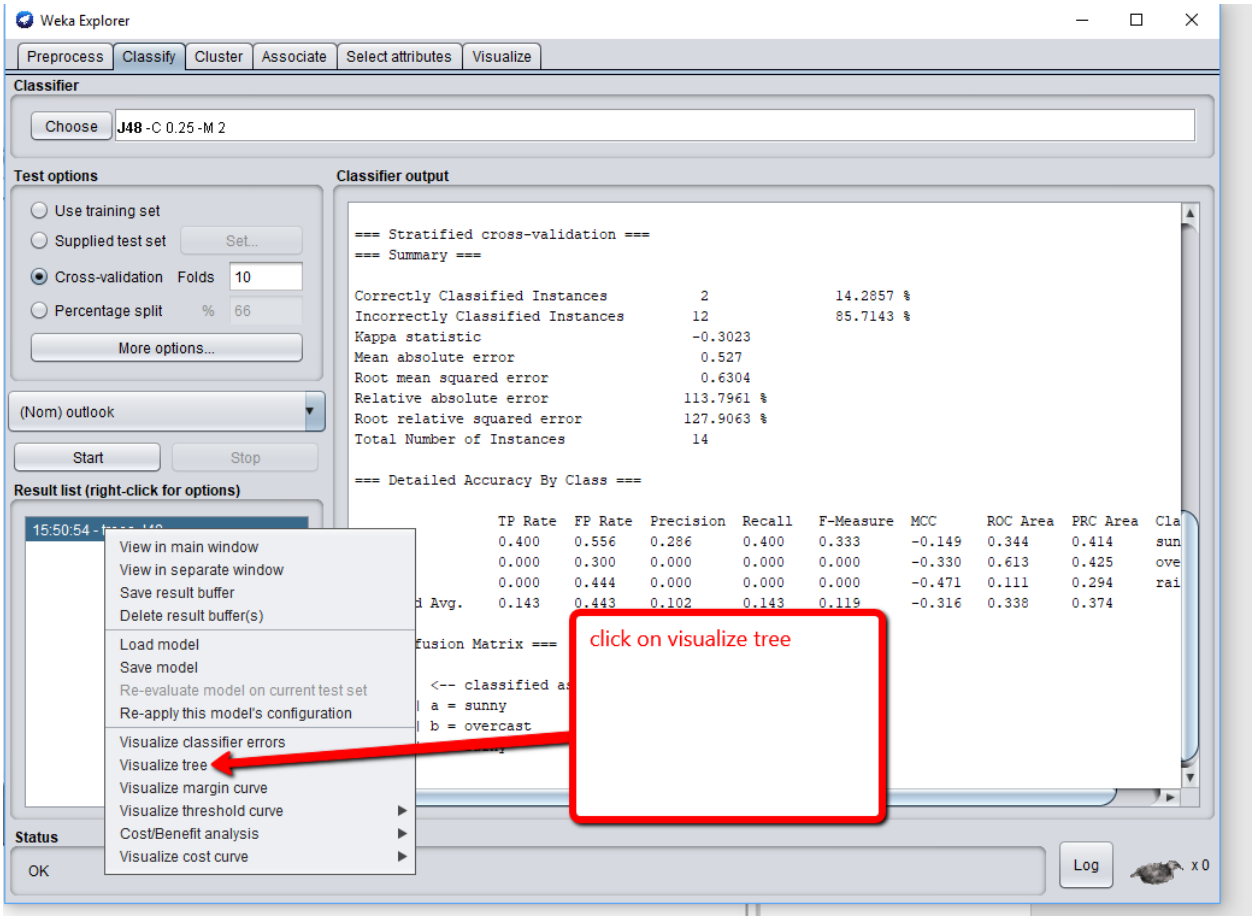


Python program to predict whether a SMS message is spam or not spam (add to your previous program)

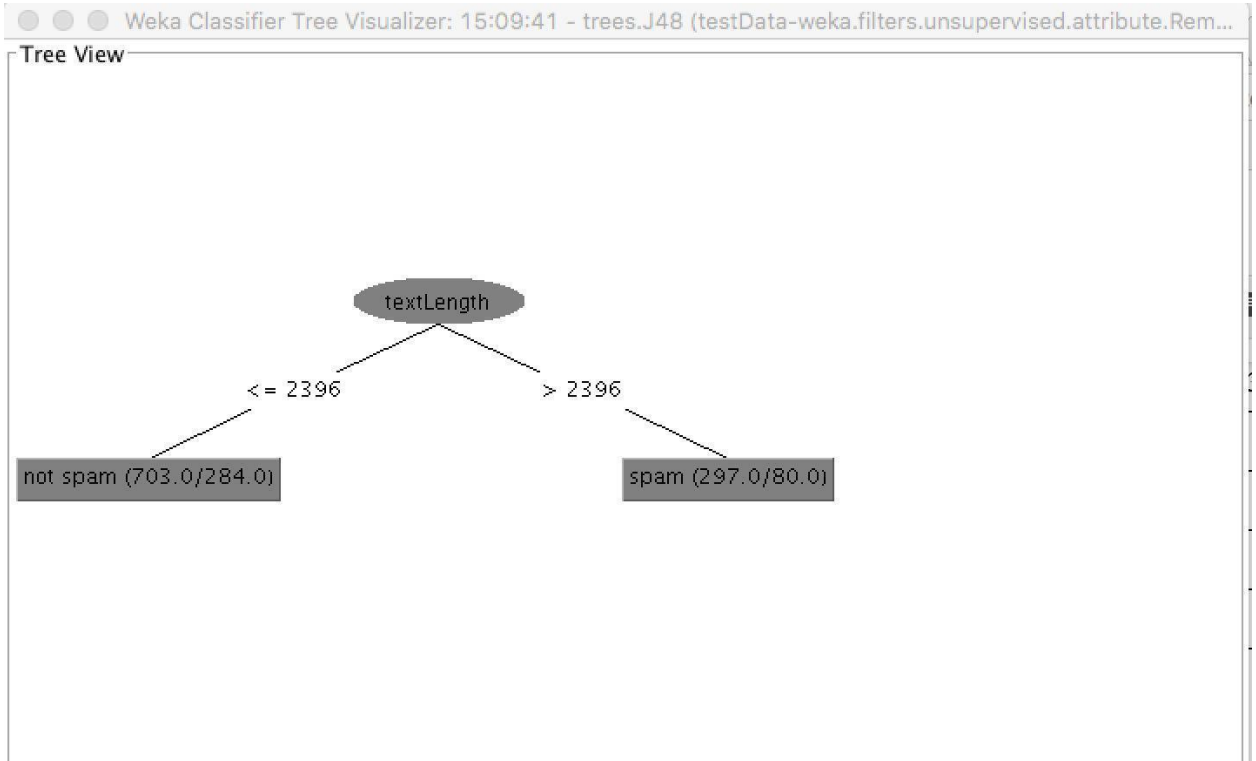
Follow the instructions in red in the image below.



Follow the instructions in red in the image below.



You should see a visualization similar to the one below.



Write down the if/else rules.

For example,

- If the text length is greater than 2396, then the message is spam
- Else if the text length is less than or equal to 2396 then the message is ham (not spam)

You will convert these rules into a function in Python.

Next, create a function named: predict_spam in your file predict_sms_spam.py **(DO NOT CREATE A NEW FILE)**

- This function should take as arguments, each feature
- This function should return “spam” or “ham”

Your Presentation

You will have a maximum of 10 minutes to present your program.

Discuss how you designed your program.

a. Walk us through your program
b. What functions you used and why
c. What data structures did you use to compute your features and why did you choose those data structures?
    i. Did you use lists, dictionaries, arrays?
    ii. What made you choose these data structures?
d. What features did you choose and why?
e. Is there anything you wish you could have done better?
f. What did you learn from this project?