The criteria of grading for STATISTICAL TOOLS AND APPLICATIONS / INTRODUCTION TO DATA ANALYTICS COMP1003_COMP1433_20212_A

Q1:

(a) (10'): correct answer (5') + formula derivation (5') (The formula derivation here means you need to conduct derivation similar to the derivation in the given answer. Besides, illustrating this is a geometric distribution problem to get the answer is also ok.)

(b) (10'): correct answer (5') + formula derivation (5') (The formula derivation here means you need to conduct derivation similar to the derivation in the given answer. Besides, illustrating this is a geometric distribution problem to get the answer is also ok.)

(c) (10'): use two numbers to represent the Agrippa and Ptolemy (2') + correctly sample 500 cards (4') + run the experiments many times (3') + calculate the probability (1')

Q2:

(a) (10'): load the file (5') + name the columns (5')

(b) (10'): use tokenize_words() function to process the tweet (5') + name the columns (5')

(c) (10'): process the situation where a word appears many times in one tweet (3') + the method of constructing the vocab (6') + print the count of tokens in vocab (1')

(d) (20'): get the tokens of positive, negative and neutral tweets (3') + get the number of positive, negative and neutral tweets (3') + calculate the word probability for each sentiment (3') + add-1 smoothing (2') + the process of the word which does not appear in the vocab (directly ignore) (1') + calculate the log-likelihood of predictions (5') + correct prediction results for the given three test cases (3')

Q3:

(20'): load the file (3') + initialize the centroids with the three points (40,40), (100,0), (0,100) (3') + calculate the distance of each data to each centroid (4') + update the centroid after each iteration (4') + run the algorithm after 1,000 iterations (early stopping the algorithm when it converges would be better, but this would not affect your score) (3') + visualize the correct clustering results (3')