Homework 2 for CMSC 498U/644

Due March 13

1 Problem 1

Read Section 4.5.2 of the textbook *Mining of Massive Datasets* for Alon-Matias-Szegedy algorithm (unfortunately, we were not able to cover it in class). This algorithm estimates the second moment of a stream. You do not need to submit anything for Problem 1.

2 Problem 2

Please show step by step how the algorithm you learned in Problem 1 runs on the following stream (suppose the underlined letters are the ones you sampled):

$$a, a, \underline{a}, a, \underline{b}, a, a, c, a, b, \underline{b}, b, b, a, \underline{c}$$

3 Problem 3

3.1 Part A

Please give the 2-shingles for the following two sentences (treat each word as a token, and a sentence would be a list of words). What is their Jaccard similarity?

- S_1 = I would drink black tea
- S_2 = I would drink green tea

3.2 Part B

What about this pair? Please do the same task as Part A on the following pair of sentences.

- S_1' = I would drink green tea but I would not drink black tea
- S_2' = I would not drink green tea but I would drink black tea

3.3 Part C

What did you learn? What shingle length do we need to distinguish S_1' and S_2'? Show your calculations.