Lab1: Entropies of English Letters

(1) Simulate Shannon's study on English letters using the WSJ data. Compute the conditional entropy given one previous letter and two previous letters;

(2) List the top five and the bottom five 3-grams from Item (1);

(3) Compute the point-wise mutual information for the word pairs, (wall, street) and (great, wall), using the WSJ data. What can we say about the results? Can you come up two more example pairs to illustrate the important of mutual information in learning something from text data?

(4) Repeat (1)-(3) for 1K and 10K sentences, any difference?

(5) <u>Extra Credit</u>: Can you do Item (1) given three previous letters using 10K sentences? What are the difficulties?