COMP 5481

LAB 2 – Predict Patterns

An “n-gram” is a contiguous sub-sequence of *n* items in a given sequence. For example, given the sequence “ALIGATOR”, its only 5-grams are ALIGA, LIGAT, IGATO and GATOR. There are special names for the first few n-grams: unigram for 1-gram, bigram for 2-gram, and trigram for 3-gram.

The Problem:

Write a program that, given a paragraph, will find the most-frequently appearing n-gram in it. We are interested in n-grams consisting of letters only. More specifically, you need to find “the single letter that appears the most (Unigram)”, “two consecutive letters that appear the most (Bigram)”, and “three consecutive letters that appear the most (Trigram)”. If there is a tie for most number of occurrences, (e.g., several bigrams appearing the most), print the one that appears first alphabetically (i.e., the smallest in string comparison). Note that "consecutive" letters means one letter immediately after another letter, i.e., no other characters (spaces or other separators) in between.

The Input:

The first line of the input is an integer p (0<p<51) indicating the number of lines in the paragraph. The following p input lines provide the text for the paragraph. Each of these input lines will contain only lowercase letters, spaces, commas and periods. Assume that these input lines will not exceed column 70 and that each line will contain at least one letter. (Note that the only separators are spaces, commas, periods, and end-of-line.) $(p+2)^{th}$ line or the last line in the input is an integer n (0<n<4). It indicates n-gram to be printed as the output (1->Unigram, 2->Bigram and 3-> Trigram).

The Output:

Print apt “n-gram” text (Unigram/Bigram/Trigram) followed by most-frequently appearing n-gram. Note that for a string such as “aaaaaa”, some interpretations view it as having three copies of “aa” and some view it as having five occurrences of “aa”. Use the latter view for this problem (same concept applies to trigrams as well).

No	Sample Input	Sample Output
1	2 abcd z z. z,z. z z. z,z. 1	Unigram z
2	3 a a. a,a. bc bc abcd abcd abcd 2	Bigram bc
3	1 abababababababababa 3	Trigram aba