HW-10 - Sequential Analysis
(do this or Expectation Maximization on Food)
See the associated Dropbox for due date.
Thomas Kinsman

Homework is to be programmed only in one of the following languages. No other languages will be accepted. Please limit yourself exclusively to: Java, Python, Matlab, or R. The last three have good native graphics and plotting support.

Assume that the grader has no knowledge of the language or API calls, but can read comments.
Use prolific block comments before each section of code, or complicated function call to explain what the code does, and why you are using it. Put your name and date in the comments at the heading of the program.

Hand in IN A ZIP FILE:

1. Your write-up including your results, **HW10_LastName_FirstName_SeqAna.pdf**
2. Your well commented code, **HW10_LastName_FirstName_ SeqAna …**

Worked with Manan Shah.
You are provided with a file of words from the Linux spelling checker, named "words".

Do sequential analysis on the given words, include the special character '^' to indicate the start of the word, and '$' to indicate the end of the word. Ignore all hyphens and non-alphabetical characters. Ignore the case of the letters.

Perform both bigram (two letter) and trigram (three letter) analysis.

1. **Write-UP Should Include the answers to the following questions, with the supported statistics:**
   a. Which is more common, a double n, or a double s? (Report your counts.)
      Count nn : 1187 Count ss : 5547
   b. What is the most common letter for a word in this corpus to start with?
      (Again, report the counts.)
      Starting with s : 11327
   c. What is the most common letter for a word in this corpus to end with?
      Ending with s: 47444
   d. Which vowel is most likely to come after the letter "t"?

Ignoring case starting with t/T	Count
Ta	4466
Te	9582
Ti	10635
To	3608

Tu	1795

e. What letter is most likely to come after the letter "u"?
N -> 3692

f. Which is a more common ending for the words in this corpus: "-er" or "-ed"?
Strings Ending in ed is more common
Ending in er -> 3435
Ending in ed -> 6705

g. Which is a more common combination, "ant" or "ent"?
Strings with ent are more common
Containing ant -> 1410
Containing ent -> 2816

h. What is the most common character to follow the two characters "qu"?
I -> 528

i. Other than the letter u, what is the most common letter to follow the letter "q"?
i-> 10

j. Which is more common, "tio", or "ion"?
tio -> 3533
ion ->4283

k. Which is more common, "ene", or "ine"?
ene ->905
ine ->1826

2. Find three observations that surprised you.
Write three questions for next semester's class, with the answers.
a.
How many clusters does K-means return ? -> K
b.
Which distance metrics should be used when you have data corresponding to a mixture model ? -> Mahalanobis distance.
c.
Given a scenario which clustering technique should be best one to use. Why ?
Step 1-> Plot scatterplot.
Step 2-> If the data is evenly distributed , try K-Means.
Step 3-> If the data is not evenly distributed, try DBScan.
Step 4-> If the results from K-Means are not satisfactory , use EM.

3. Write a conclusion, about what you did, and any challenges that occurred in the implementation of this project. How did you handle the special characters?

In this programming assignment , I learnt more about the regular expressions to be handled in r. We use the package stringr for the same . The str_replace_all function helps us to replace any pattern of characters with any other pattern of characters if required. This assignment gave me

interesting and obvious conclusions like there is no word which start with letters q followed by c.

Code to handle special characters

words_modified<-str_replace_all(words$V1,"[^[:alnum:]]","")

This inbuilt function from R replaces any character which is other than alphanumeric character with a null character.

For handling the case of the characters , I used ignore case=TRUE while calculating the count of occurences

count_s<-gregexpr("ene",str,ignore.case = TRUE).