MSDS 7330 File Organization and Database Management Homework Python

Name: Cory Nichols

This is a homework assignment for MSDS 7330, File Organization and Database Management. Your final submission shall be a single pdf document that includes this document plus your answers to each of the questions.

Collaboration is expected and encouraged; however, each student must hand in their own homework assignment. To the greatest extent possible, answers should not be copied but, instead, should be written in your own words. Copying answers from anywhere is plagiarism, this includes copying text directly from the textbook. Do not copy answers. Always use your own words. For each question list all persons with whom you collaborated and list all resources used in arriving at your answer. Resources include but are not limited to the textbook used for this course, papers read on the topic, and Google search results. Note that ‘Google’ is not a resource. Don’t forget to place your name on the document.

This assignment was adapted from the requirements of Dr. Eric Larson.

Python

In this assignment you will be using python to investigate the scrabble dictionary. All the words in the scrabble dictionary are available from

http://www.puzzlers.org/pub/wordlists/ospd.txt

**Using Python with Scrabble Word List**

The text file contains a list of the words separated by newlines/carriage returns. Your first portion of the assignment is to find out how many unique anagrams are in the dictionary using python. Note that this can be done efficiently in about 30 lines of code. However the method you use to store the words can wildly change the complexity of the problem. This can be completed procedurally or using object-oriented coding practices: either are fine implementations.

To find out what an anagram is, check out the Wikipedia page (http://en.wikipedia.org/wiki/Anagram). Essentially, anagrams are words with the same letters like ape and pea. This pair of words forms one anagram.

Turn in your python code used to answer the following questions.

Question 1: How many unique anagrams are in the scrabble word list?

There are 65,783 unique anagrams in the scrabble word list. The anagram was assumed to be the sorted word. For example amry is the anagram for army and mary. This means there can be repeats of the same anagram for different words. To find these uniquely, a python dictionary was used with key:value pairs.

Collaborators: None

Resources: None

Question 2: What is the anagram with the largest number of words in the word list? How many words are in this anagram?

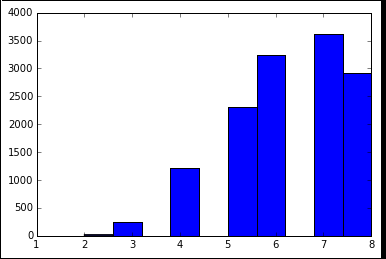
The anagram with the largest number of words is aeprs, with 11 words. The first word to form the anagram is NOT counted in my model, as the anagram must be sourced from an original word.

Collaborators: None

Resources: None

Question 3: Use matplotlib (or Excel) to visualize a histogram of the size of the anagrams. That is, make a bar chart where the X-axi from from X = 1 to X =max anagram size. Each bar will represent the number of anagrams of length X.

Export from matplotlib using scipy library and the hist( method:



Collaborators: None

Resources: None

**Using Python with Databases and Scrabble Word List**

Create a database using python with one table. The table will have three columns: Unique Anagram identifier (i.e. the sorted letters in the anagram), number of words in the anagram, and the actual words as a comma separated value string.

Use python to create and populate the database. You can use MySQL to create a bare bones database, but python should create the table and setup the variable types. Alternatively, you can use sqlite3 that ships with python.

Now create queries to the database that answer the following questions (which are the same as Questions 1 and 2). Turn in your python code used to answer these questions with the database.

**Question 4:** How many unique anagrams are in the scrabble word list?

I created the MySQL database in python. There are also 65,783 unique anagrams. This is not a case sensitive count. E.g., count distinct would count Pears and pears as one distinct value in SQL. Pears anagram of Paers and aeprs are two different anagrams in python – which makes logical sense.

SELECT COUNT(anagram\_id)

FROM ANAGRAMS

Collaborators: None

Resources: None

**Question 5:** What is the anagram with the largest number of words in the word list? How many words are in this anagram?

Aeprs is the anagram with the largest number of words with 11.

SELECT anagram\_id, num\_words

FROM anagrams

WHERE num\_words in (SELECT max(num\_words) num FROM anagrams);

Collaborators:

Resources:

Using Python with Databases and Merriam-Webster English Language Dictionary

Use python to create a new table in the database and again find anagrams in the word list using three columns. However, instead of using the small dataset of scrabble words, use the Merriam-Webster English language dictionary ( approximately 300,000 words).

Now create queries to the database that answer the following questions (which are the same as Questions 1 and 2). Turn in your python code used to answer these questions with the database.

Question 6: How many unique anagrams are in the Merriam-Webster word list?

There are 219,768 unique anagrams in the MW word list.

Collaborators:

Resources:

Question 7: What is the anagram with the largest number of words in the word list? How many words are in this anagram?

Eerst is the anagram with the most amount of words (8). The first word is not considered as it is considered the formative word of the anagram.

Collaborators: None

Resources: None