CS21120 Assignment

## Introduction

In this assignment we were tasked to output a list of words that rhymes within a given word.  
To do this we were given a CMU dictionary text file that contains ~135,000 lines of words and pronunciation.

For this coding assignment I didn’t use any coding AI-assistance, only resources online such as java documentation, Stack Overflow, W3Schools and other websites. As for the documentation I will be using a grammar checker website QuillBot.

## Task 1: Implementing IPhoneme

Implementing IPhoneme was relatively simple because I had to create a Phoneme class that complied with the IPhoneme interface, which then generated methods from the interface and added extra codes, such as the constructor arguments are stored as attributes and checks if the vowel or non-vowel phoneme stress value is valid. For example, a vowel phoneme cannot have a stress value of -1, whereas a non-vowel phoneme must have a stress value of -1.

## Task 2: Implementing IPronunciation

Implementing IPronunciation was a bit tricky for me at first because I couldn’t quite understand how the rhyming process worked therefore coding findFinalStressedVowelIndex() took some time.

To find the final stressed vowel, I had two choices as to either go through the list of Phonemes front to back or vice versa. When looking through the CMU dictionary text file, I realized not all primary stressed vowel words are towards the end of the pronunciation of the word, but rather it is randomised. Since the idea is to find the final stressed vowel, I decided to search back to front where upon finding a primary stressed vowel it will instantly return the index otherwise it will go through the whole list of Phonemes and return the index with priority to former highest stress vowel.

Initially I used a LinkedList as the data structure to store the list of Phonemes however, upon reevaluating my completed code, I realized an ArrayList is a more suitable data structure because of two specific operations I used for pronunciation class, a addPronunciation() which adds phoneme to listOfPhonemes and findFinalStressedVowelIndex().

One thing I would like to point out is that when .add() an Phoneme to an ArrayList, it will have a default capacity of 10 and when reading from the CMU dictionary text file, majority of the words will have a list of Phonemes of 10 or less therefore, making the average case time complexity of θ(1) and in the rare cases O(n) will happen.  
As for ArrayList indexing, the average-case time complexity is θ(1) which is constant time.

LinkedList on the other hand, holds references to the first and last element, so adding will have a average-case time complexity of θ(1) however, as for LinkedList indexing, it has to traverse the list until it finds the element, therefore it has a average-case time complexity of θ(n) since it still needs to search for at least half of the list which is still linear time.

As for findFinalStressedVowelIndex() with an ArrayList average time complexity, it would be θ (n) when it comes to both ArrayList and LinkedList, because final stressed vowel can be found anywhere within the list of Phonemes therefore on average it will take n/2 indexing time to find it which is linear time.

## Task 3: Implementing IWord

IWord class was relatively simple to implement because it only needed a constructor and 3 methods so there weren’t many moving components to think about, the only notable method to mention is the addPronunciation() where a passed pronunciation object would be added to a setOfPronunciation. As for the setOfPronunciation implementation of data structure there was HashSet, LinkedHashSet and TreeSet. I decided to use a HashSet because it’s the only data structure that doesn’t consider the order of the set therefore it will be faster compared to other set frameworks, for example LinkedHashSet implements pointers or TreeSet which rearranges the whole data structure.

addPronunciation() average time complexity is θ(1) because storing pronunciation object will be based on the hash code generated by the hashing algorithm.

## Task 4: Implementing IDictionary and parsing the CMU file

Initially I used a HashSet to store all the Word objects because each of their own had their own .getWord() which outputs a English word so every line from the CMU dictionary text file would be split into word and pronunciations seconds. Then the pronunciation section will be stored into a pronunciation object, of which will be stored within a setOfPronunciation within the new Word object.

However upon coding and testing loadDictionary() it would take 4 minute 28 seconds to load the whole CMU dictionary and I deemed that unoptimized and not a suitable solution. The reason why the loadDictionary() was so slow is because, before every word is added to the HashSet it would be checked if it exist within the HashSet and since there is 135,000 words it would take 522,497,908 check operations if newly created Word object English word is the same as the HashSet Word object English word.

Upon revaluating I concluded that a HashMap needed to be used, I did think about using a HashMap at the very start however I didn’t know why I would need a key and value when a Word object already has all the functionality. But if I have a HashMap with a English word as a key and Word object as a value it would make the check exist within the HashMap quicker because the HashMap look up, average-case time complexity is θ(1) and worst-case time complexity is O(n) if there is multiple collisions. Once the HashMap implementation was done and some minor changes happen then the dictionary can load in 600ms which is considerably a lot faster than previously version with a HashSet.

With the implementation of HashMap it allows getWord() average-case time complexity of θ(1) and worst time complexity of O(n) if there is multiple collisions. The addWord() has a average time complexity of θ(1) whereas worst time complexity is O(n^2) because if the dictionary HashMap contains multiple collisions then when checking if the word exist within the dictionary it would take O(n) and then it would .put() the pass Word object into the HashMap which will also be O(n) making it O(n^2) assuming there is multiple collisions within the HashMap.

As for getWordCount() it returns the amount of keys which would take a average-case time complexity of θ(1) and getPronunciationCount() returns the amount of values by iterating through the values and using .getPronunciation.size() to get the amount of pronunciation making average-case time complexity of O(n).

## Task 5: Rhyming!

For the Dictionary class, I decided to use a HashSet to store all English words that rhymes because HashSet will only keep unique elements therefore there won’t be duplicate words and has a .put() with average cast time complexity θ(1) but worst-case time complexity of O(log n) because Java 8 or above will store list of collisions as a tree.

Upon testing rhymesWith() Junit tests I realised testRhymesWith\_OneSyllableCVC() test kept failing and I found out that upon finding the last stressed vowel of both words, past it must have matching Phonemes. So I created isPronunciationRhymesWithPronunciation() which checks if two given pronunciation and finalStressIndex past Phonemes matches.

Another difficulty I came across was when testing testGetRhymes() JUnit specifically testGetRhymes\_Cat() where upon getting all the rhymes with the word “cat” it should have a set of 77 English words that rhymes with “cat” but my code could only find 67 words. I used temporary regex code to find words within the CMU dictionary matches with regex “AE1 T$” and compare it to the 67 words. I found out that the missing words contained two primary stressed vowels within the pronunciation and it would only return the left-most primary stress index, to fix this I altered the code to return the right-most primary stress index.

As for getRhymes() average-case time complexity is θ(n) because majority of the English words within the CMU dictionary only have one so majority of word’s pronunciation will only be check once.

In order to find the English words with the most rhymes, I would use a HashMap containing the English word and number of rhymes respectively as key and value. Then I would create a loop which goes through the dictionary HashMap keys and pass it into the getRhymes(), of which will return a set of English words that rhymes. By using .size() on the HashSet it would return the amount of rhymes then the current English word in the loop and size of rhymes found will be put into a HashMap. Iterate through the whole HashMap entry set and return the key with the highest value.

## Self-evaluation

Task 4 was the most difficult for me because parseDictionaryLine() had a lot of moving components, such as splitting the line into word and pronunciation objects, removing comments, and using HashMap. But overall, I completed everything to the best of my ability in terms of readability and efficiency (loadDictionary() and getRhymes() worst-case scenario ~600ms). With extra JUnit tests, I believe I should get ~75%.