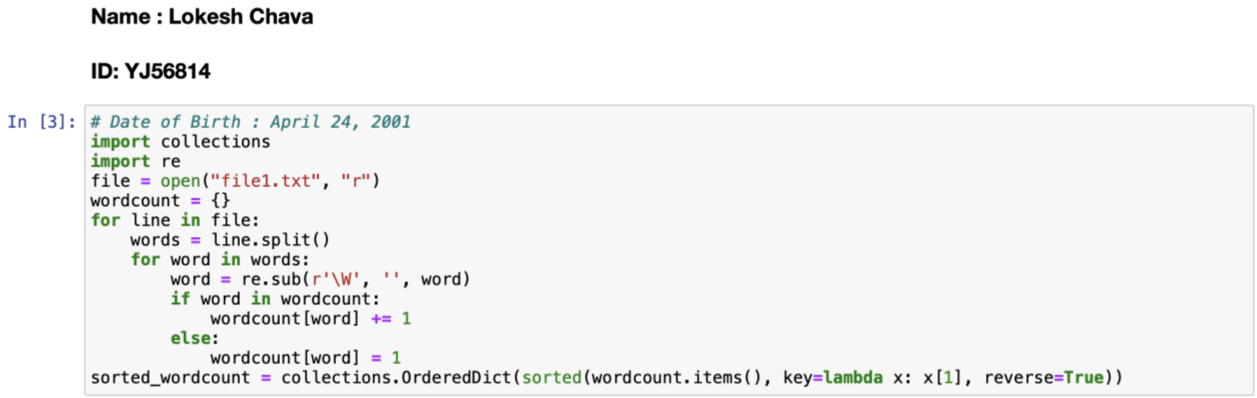
Name: Lokesh Chava
Campus ID: YJ56814
Course: DATA 603: Platform for Big Data Processing

Q1) Write Python code and use MapReduct to count occurrences of each word in the first text file (file.txt). How many times each word is repeated?

Code:

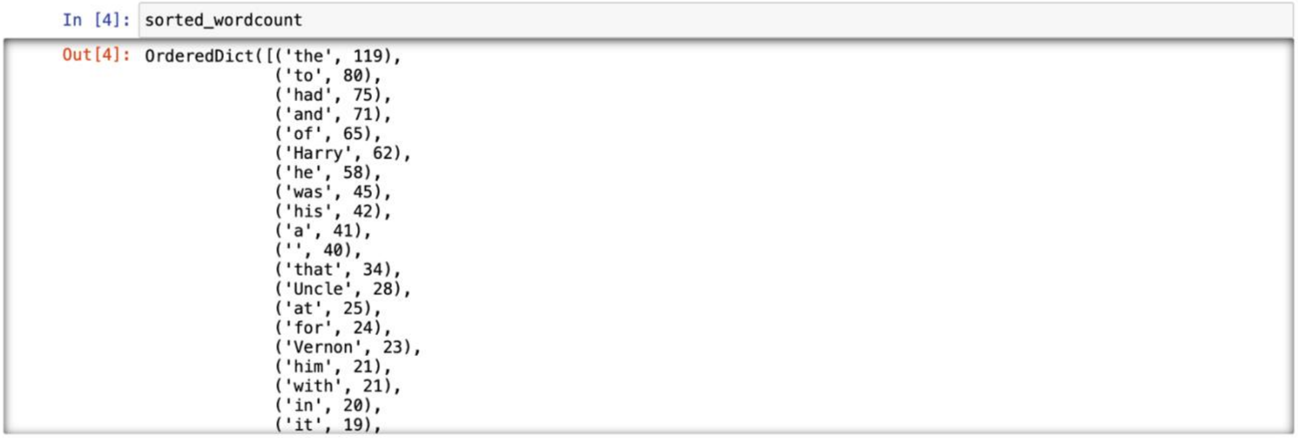


Name : Lokesh Chava

ID: YJ56814

```
In [3]: # Date of Birth : April 24, 2001
        import collections
        import re
        file = open("file1.txt", "r")
        wordcount = {}
        for line in file:
            words = line.split()
            for word in words:
                word = re.sub(r'\W', '', word)
                if word in wordcount:
                    wordcount[word] += 1
                else:
                    wordcount[word] = 1
        sorted_wordcount = collections.OrderedDict(sorted(wordcount.items(), key=lambda x: x[1], reverse=True))
```

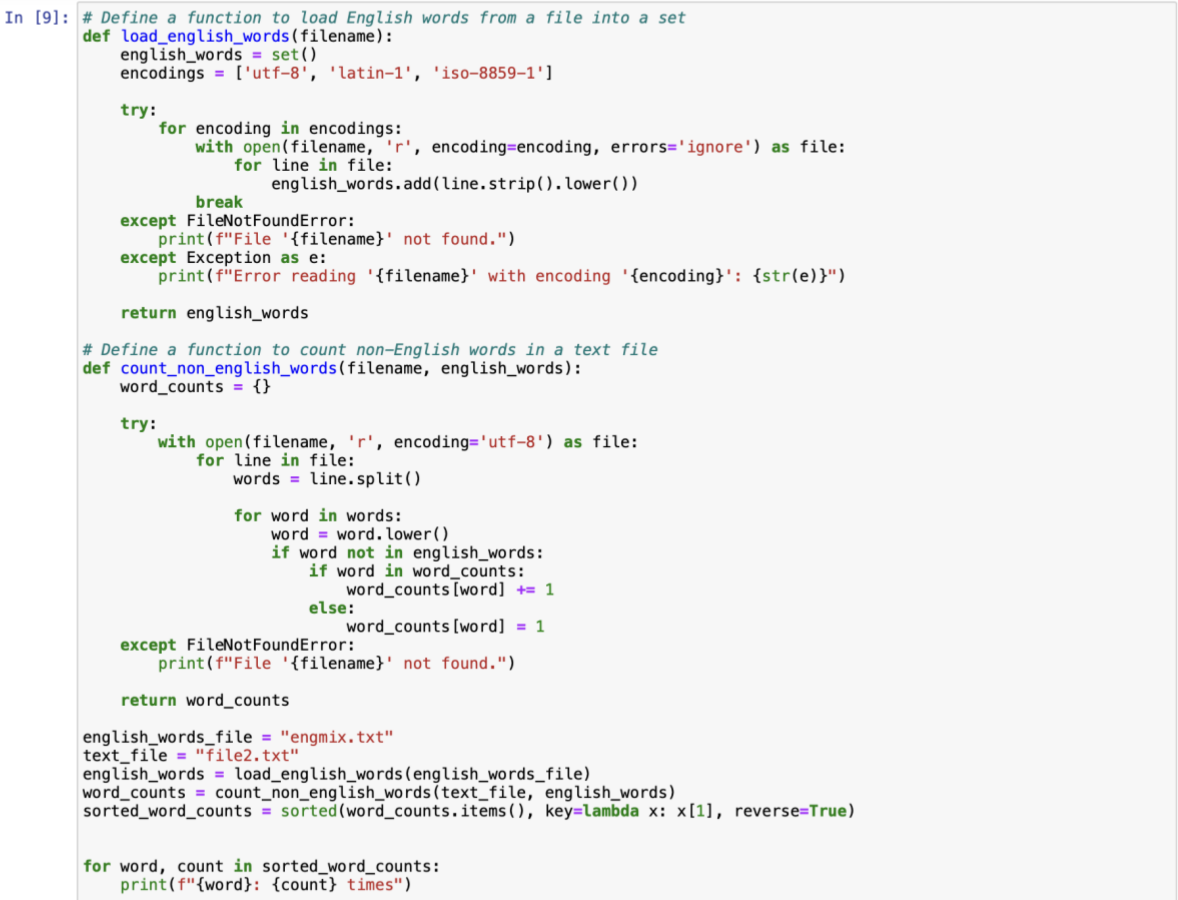
Output:



```
In [4]: sorted_wordcount

Out[4]: OrderedDict([('the', 119),
                     ('to', 80),
                     ('had', 75),
                     ('and', 71),
                     ('of', 65),
                     ('Harry', 62),
                     ('he', 58),
                     ('was', 45),
                     ('his', 42),
                     ('a', 41),
                     ('', 40),
                     ('that', 34),
                     ('Uncle', 28),
                     ('at', 25),
                     ('for', 24),
                     ('Vernon', 23),
                     ('him', 21),
                     ('with', 21),
                     ('in', 20),
                     ('it', 19),
```

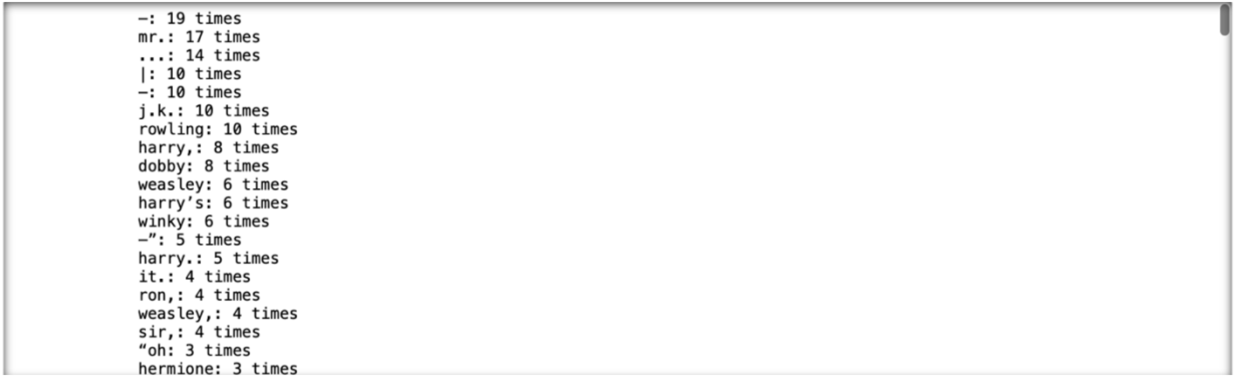
Q2) From the second text file (file2.txt), write Python code and use MapReduct to count how many times non-English words (names, places, spells etc.) were used. List those words and how many times each was repeated.

Code:



```
In [9]: # Define a function to load English words from a file into a set
        def load_english_words(filename):
            english_words = set()
            encodings = ['utf-8', 'latin-1', 'iso-8859-1']

            try:
                for encoding in encodings:
                    with open(filename, 'r', encoding=encoding, errors='ignore') as file:
                        for line in file:
                            english_words.add(line.strip().lower())
                    break
            except FileNotFoundError:
                print(f"File '{filename}' not found.")
            except Exception as e:
                print(f"Error reading '{filename}' with encoding '{encoding}': {str(e)}")

            return english_words

        # Define a function to count non-English words in a text file
        def count_non_english_words(filename, english_words):
            word_counts = {}

            try:
                with open(filename, 'r', encoding='utf-8') as file:
                    for line in file:
                        words = line.split()

                        for word in words:
                            word = word.lower()
                            if word not in english_words:
                                if word in word_counts:
                                    word_counts[word] += 1
                                else:
                                    word_counts[word] = 1
            except FileNotFoundError:
                print(f"File '{filename}' not found.")

            return word_counts

        english_words_file = "engmix.txt"
        text_file = "file2.txt"
        english_words = load_english_words(english_words_file)
        word_counts = count_non_english_words(text_file, english_words)
        sorted_word_counts = sorted(word_counts.items(), key=lambda x: x[1], reverse=True)


        for word, count in sorted_word_counts:
            print(f"{word}: {count} times")
```

Output:



```
–: 19 times
mr.: 17 times
...: 14 times
|: 10 times
–: 10 times
j.k.: 10 times
rowling: 10 times
harry,: 8 times
dobby: 8 times
weasley: 6 times
harry’s: 6 times
winky: 6 times
–”: 5 times
harry.: 5 times
it.: 4 times
ron,: 4 times
weasley,: 4 times
sir,: 4 times
“oh: 3 times
hermione: 3 times
```