Quiz #0

Instructions

Question 1

7 / 7 pts

Word Count:

Given a file (/tmp/foxy.txt), which contains a set of records, where each record has a list of words separated by a single space:

```
cat /tmp/foxy.txt
a fox jumped and jumped
red fox jumped high
a red high fox jumped and jumped
red fox is red
```

Write a working python program which reads the given file and outputs the frequency of all unique words.

Your output will look as:

```
fox  4
jumped 5
...
```

Your Answer:

```
file = open("foxy.txt", "rt")
text = file.read()
# print to see what the text contains
text

def Freq_Unique_Words(str):
# split the string into list of words
str_list = str.split()

# find set of unique words
unique_words = set(str_list)

# for each unique word, find the count of occurrences
for words in unique_words :
print(words, str_list.count(words))
```

# call the function over our text
Freq_Unique_Words(text)

#---------- OUTPUT ----------

```
fox 4
a 2
jumped 5
and 2
is 1
red 4
high 2
```

Question 2

3 / 3 pts

What if your file has 100 billion records? how do you do that?

Your Answer:

Hadoop allows distributed processing of large datasets across clusters of computers. A Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. One of the modules of the Hadoop framework is Hadoop MapReduce.

MapReduce is a programming model and it is designed to compute the large volume of data in a parallel fashion. Basically Hadoop MapReduce parallel processes large datasets.

Mapper Phase:

The text from the input text file is tokenized into words to form a key-value pair with all the words present in the input text file. The key is the word from the input file and value is ‘1’. After the map phase execution is completed successfully, the shuffle phase is executed automatically wherein the key-value pairs generated in the map phase are taken as input and then sorted in alphabetical order.

Reducer Phase:

In the reduce phase, all the keys are grouped together and the values for similar keys are added up to find the occurrences for a particular word. It is like an aggregation phase for the keys generated by the map phase. The reducer phase takes the output of the shuffle phase as input and then reduces the key-value pairs to unique keys with values added up.

Quiz Score: **10** out of 10