University of Science and Technology of Hanoi

Distributed System

PRACTICAL WORK 4: WORD COUNT

Group members:

Nguyen Phuong Thao (BI9-212)

Doan Tuyet Mai (BI9-162)

Trinh Thao Phuong (BI9-191)

Phung Kim Son (BI9-202)

Pham Minh Long (BI9-146)

Hanoi, Mar 2021

1. Why you chose your specific MapReduce implementation?

We choose to implement by Python because it supports dictionary datatype which is extremely convenient in implementing the Map Reduce. By keeping each word and its occurrence by a key:values pair, we can track which word was counted and combined the occurrence list obtained from Mapper in the Reducer easily.

2. How your Mapper and Reducer work?

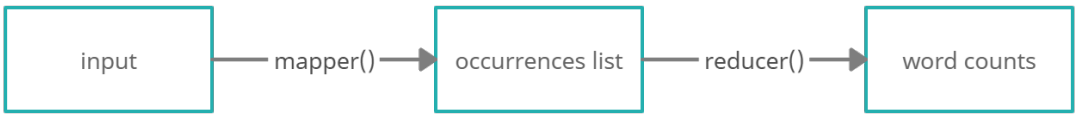

Figure 1: Mapper and Reducer

2.1 Mapper

After received the input texts, we run the `strip()` function to remove the leading and trailing whitespace. We also remove the punctuation by doing some `sub("[^\w\s]", "", stc)` method. After that, we split the sentences into words.

```
def mapper(file_name):
   f = open(file_name, "r")
   occurence = []

   for line in f:
      sentences = line.split(".")
      for stc in sentences:
         # Remove leading and trailing whitespace
         stc = stc.strip()

         # Remove punctuation
         stc = re.sub("[^\w\s]", "", stc)

         # Split the line into words
         words = stc.split()

         for word in words:
            #print("%s\t%s" % (word, 1))
            occurence.append([word, 1])

   return occurence
```

2.2 Reducer

After get the words which are the results of the mapper process, we count the number of each words.

```
def reducer(occurence):
   combined_occ = {}
   for i in range(len(occurence)):
      word = occurence[i][0]
      if word not in combined_occ.keys():
         combined_occ.update({word: 1})
```

```
    else:
        occ = combined_occ[word]
        combined_occ.update({word: occ+1})
return combined_occ
```

3. Contribution

Student	Student ID	Contribution
Pham Minh Long	BI9-146	Write report
Phung Kim Son	BI9-202	Research for different MapReduce implementation
Trinh Thao Phuong	BI9-191	Draw figures, briefly description
Doan Tuyet Mai	BI9-162	Implement code for Word Count (MapReduce)
Nguyen Phuong Thao	BI9-212	Explanation of map reduce