Dig Data Management Assignment 2

信科四 郑元嘉 1800920541

Task1 - Word Count Task on 新概念英语第二册.txt Based on Hadoop MapReduce

In this task, we count the words in 新概念英语第二册.txt applying Hadoop MapReduce framework.

How I Implement

In this task, we use Python to go through it. Hadoop provides a streaming method to make it possible to run MapReduce tasks for all programming languages.
Here shows the simplistic codes of mapper,

```
#!/usr/bin/python3.5
import sys
from nltk.tokenize import RegexpTokenizer

tokenizer = RegexpTokenizer(r'\w+')

for line in sys.stdin:
    words = tokenizer.tokenize(line.strip())

    for word in words:
        print ('%s\t%s' % (word, 1))
```

It's quite simple and straightforward. We can design any parsing pattern to build up our key-value pairs. We simply use library 'nltk' in Python to tokenize English sentences.

```
#!/usr/bin/python3.5
import sys

current_word, current_count = None, 0

for line in sys.stdin:
    word, count = line.strip().split('\t', 1)
    count = int(count)

    if current_word == word:
        current_count += count
    else:
        if current_count:
            print('%s\t%s' % (current_word, current_count))
        current_count, current_word = count, word
if current_count:
    print('%s\t%s' % (current_word, current_count))
```

We need to treat the key-value pairs as string and do customized parsing here, and then we do a simple counting task. Note that the standard input key-value pairs are sorted by key before passed into reducers, it helps simplify the logic of counting.

Result of Word Count

Top 10

```
 1  the     819
 2  a       418
 3  to      390
 4  and     278
 5  of      274
 6  I       266
 7  was     229
 8  in      219
 9  he      167
10  it      154
```

It makes sense to note that they're almost stopwords.

Last 10 (English only)

```
 1  added       1
 2  actual      1
 3  actresses   1
 4  actress     1
 5  Across      1
 6  acquire     1
 7  accustomed  1
 8  Accurate    1
 9  accurate    1
10  account     1
```