NAME: Xiang Xu

605.741.31
Module 11 Quiz

Suppose our Accumulo system stores entire documents in the value column using the following data structure to represent the Gettysburg Address and the Declaration of Independence:

rowID	family	qual	time	value
gettysburg	speech	script		Four score and seven years ago …
declaration	document	script		When in the course of human events …

Write Map and Reduce pseudocode that determines document similarity according to combinations of three important words. Important words are the words that are not in **stop word** lists. Stop words are words that are common to all documents that do not provide much indication of the topic of the document. Example stop words include *the, and, to, but, because, an, a, ….* Assume that you are given the following array of stop words to use:

```
private String[] stopWords = {"the", "and", "to", "but", "because", "an", "a", …};
```

For all combinations of three important (non-stop word) words, create an Accumulo output table that clusters documents according to the important word triples. Two sample output rows from the MapReduce algorithm applied to the table above will look like:

rowID	family	qual	time	value
government:liberty:people	speech	script		gettysburg
government:liberty:people	document	script		declaration

Make sure that the important words used in the rowID are sorted in alphabetical order so you have only one "government:liberty:people" rowID value, not "government:people:liberty", "liberty:government:people", "liberty:people:government", "people:liberty:government", "people:government:liberty". Also, do not worry about multiple occurrences of any word. Even though the word government appears many times in the Declaration of Independence, just create one output for the "government:liberty:people" triple.

a) Map algorithm pseudocode

```
Mapper(key = family+qual+time+rowid,value = value)
String[] words = value.toString().split("\\s+");
 for i in range(0, len(words[])-2):
   if i not in stopWords[]:
      Keyout = concat (CurrentWord+ : + NextWord + : + NextNextWord)
```

write(keyout + “,” + family+qual+time+rowid , 1)

b) Reduce algorithm pseudocode

```
Reduce(Key: keyword, Value: <list of value>)
        For each value in <list of values>:
        Sum += value;
        Write(BeforeFisrtCommnaOfKeyword, sum);
```

//if sum equals 2 that means Gettysburg Address and the Declaration of Independence
//have one pair of combination non-stop words the same, it will give us an idea how
//similar these two paragraphs are
//if sum is 1 that means it didn’t find similarity.