A2 Report

1.1

a. Look at the contents of the folder "output" - what are the files placed in there? What do they mean?

```
_SUCCESS  part-r-00000
_SUCCESS means our command successed
part-r-00000 is the results of mapreduce
```

b. How many times did the word 'Discovery' (case-sensitive) appear in the text you analyzed?

5

c. In this example we used Hadoop in "Local (Standalone) Mode". What is the difference between this mode and the Pseudo-distributed mode?

Standalone mode is the default mode in which Hadoop run. You can use input and output both as a local file system in standalone mode. I don't need to do any custom configuration in the files- such as mapred-site.xml, core-site.xml, hdfs-site.xml, yarn-site.xml.

Pseudo-distributed Mode is also known as a single-node cluster where NameNode, DataNode, SecondaryNameNode and Jps will reside on the same machine. In pseudo-distributed mode, all the Hadoop daemons will be running on a single node. Such configuration is mainly used while testing when we don't need to think about the resources and other users sharing the resource. Every Hadoop components can communicate across network sockets on a separated Java Virtual Machin which producing a fully functioning and optimized mini-cluster on a single host. In this mode, we have to change in configuration files for all the files- mapred-site.xml, core-site.xml, hdfs-site.xml and yarn-site.xml.

1.2

a. What are the roles of the files core-site.xml and hdfs-site.xml ?

The core-site.xml is for core configuration. It informs Hadoop daemon where NameNode runs in the cluster. It contains the configuration settings for Hadoop Core such as I/O settings that are common to HDFS and MapReduce.

The hdfs-site.xml file is for HDFS configuration. It contains the configuration settings for HDFS daemons; the NameNode, the Secondary NameNode, and the DataNodes. It can change the dfs.replication and value. We can configure hdfs-site.xml to specify default block replication and permission checking on HDFS. The actual number of replications can also be specified when the file is created. The default is used if replication is not specified in create time.

b. Describe briefly the roles of the different services listed when executing 'jps'.

Namenode: is the main server in Hadoop, managing the file system namespace and access to files stored in the cluster.
SecondaryNameNode: Without Namenode, HDFS cannot work. Therefore, the fault tolerance mechanism of the namenode is very important, and running a secondary Namenode (Secondary Namenode) is one of them. It is not a redundant daemon of namenode, but provides periodic checkpoints and cleanup tasks.
DataNode: It is responsible for managing the storage connected to the nodes (there can be multiple nodes in a cluster). Each node that stores data runs a datanode daemon.

jps (Java Virtual Machine Process Status Tool): is a command provided by JDK 1.5 to display the pids of all current java processes.

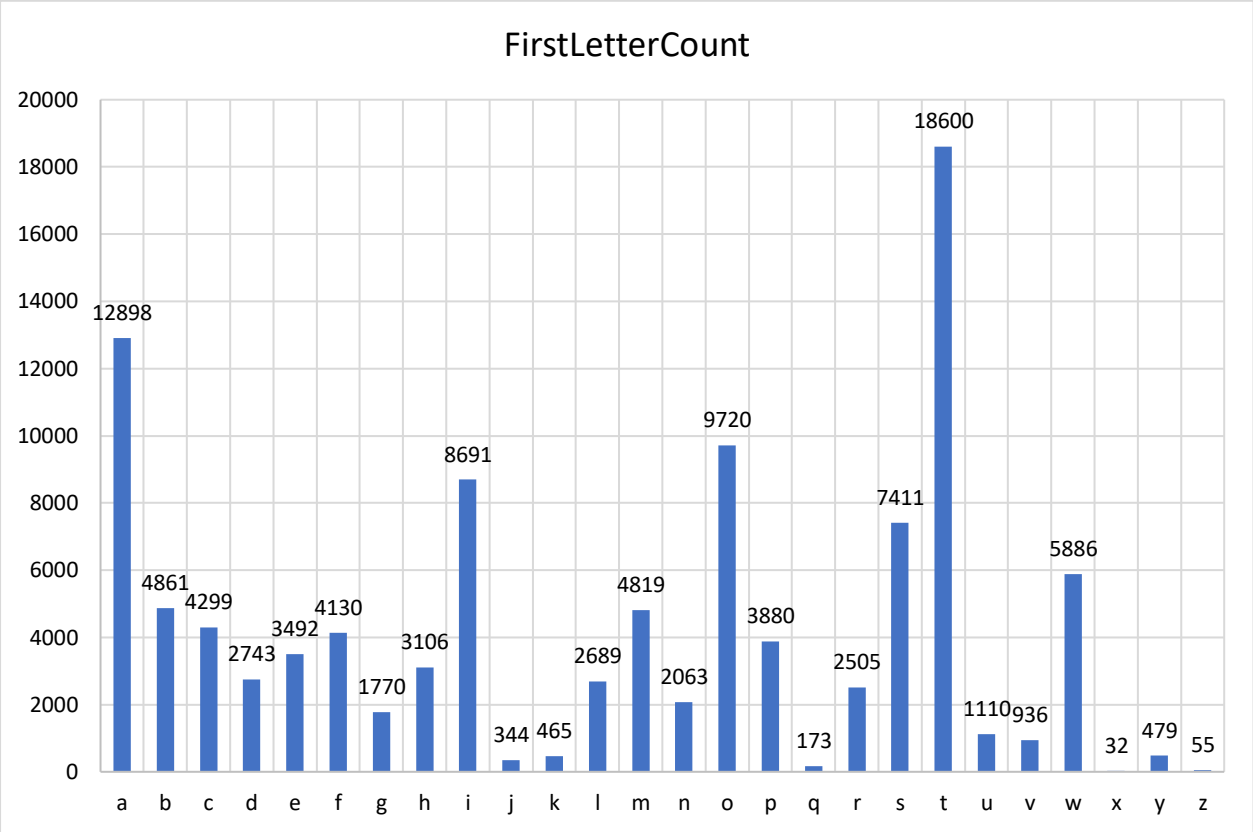
1.3

a. Explain the roles of the different classes in the file WordCount.java.

Driver class (Public, void, static, or main; this is the entry point).
The Map class needs to inherit the Mapper class in the org.apache.hadoop.mapreduce package, which extends the public class Mapper<KEYIN,VALUEIN,KEYOUT,VALUEOUT> and implements the Map function.
The Reduce class needs to inherit the Reducer class in the org.apache.hadoop.mapreduce package, which extends the public class Reducer<KEYIN,VALUEIN,KEYOUT,VALUEOUT> and implements the Reduce function.

b. What is HDFS, and how is it different from the local filesystem on your virtual machine?

The Hadoop Distributed File System (HDFS) is a distributed file system. It is mainly responsible for the data storage of each node, and realized high throughput data read and write. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.



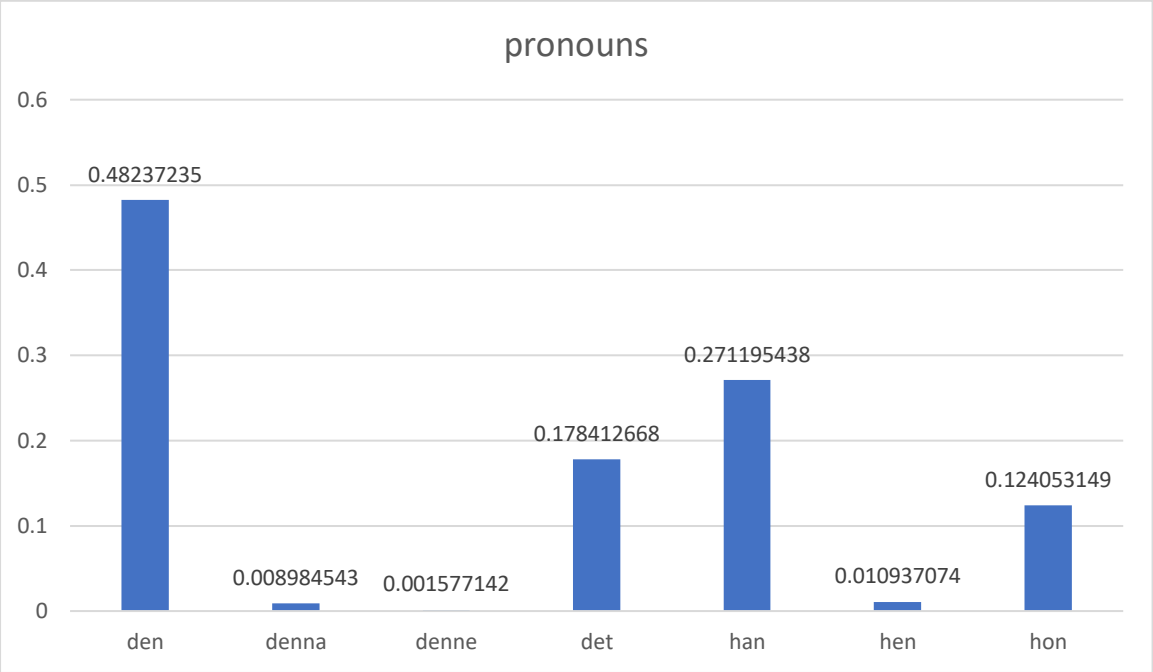
2.1. Answer the following questions:

a. Based on the twitter documentation in the above link, how would you classify the JSON-formatted tweets - structured, semi-structured or unstructured data?

It's semi-structured data.

b. What could be the challenges of using traditional row-based RDBMs to store and analyze this dataset (apart from the possibility of very large datasets)?
RDBMs work perfectly with structured database but not for semi-structured or unstructured data. RDBMs store average size data like Gbs not for Tbs or Pbs. When a size of data is too big for complex processing and storing or not easy to define the relationships between the data, then it becomes difficult to save the extracted information in an RDBMS with a coherent relationship. RDBMs use SQL not for multiple languages like python, java, Ruby. RDBMs can read fast but not for writing. RDBMs is expensive. The throughput of RDBMs are low. RDBMs have to use high-end hardware.

2.1.3 Plot a bar chart visualizing the counts of each pronoun.




```
ubuntu@wangyue-a2:~/pywordcount$ hdfs dfs -cat output4/part-00000
den     1129512
denna   21038
denne   3693
det     417767
han     635025
hen     25610
hon     290480
```

```
ubuntu@wangyue-a2:~/pywordcount$ hdfs dfs -cat output_t/part-00000
t       2341577
ubuntu@wangyue-a2:~/pywordcount$
```