Jose Miguel Leyesa

ID no. 12387673

CSC615M

* The raw gz compressed files can be found in the *raw log gz* folder.
* The unzipped log files along with a python script to merge the logs into one txt file can be seen in the *unzipped + merger script* folder.
* The source code of the project can be found in the *project* folder.

**Source Text:**

The source text is a compilation of automatic logs generated by the game *Minecraft* while playing on the server *Hypixel* on the game mode of *Skyblock*. The data cleaning was designed around chat messages that are generated in this context.

**Source text size:**

* raw\_600Kwords – 608,232 words
* raw\_5Mwords – 5,387,455 words
* raw\_10Mwords – 10,007,913 words

**Cleaned text size (using default config):**

* output\_600Kwords – 90,650 words
* output\_5Mwords – 731,251
* output\_10Mwords – 1,518,039

**Dependencies:**

For compiling and making jar:

* JDK 17.0.2
* Java 17.0.2

For running jar only:

* Java 17.0.2

**Instructions:**

*The following can also be viewed in README.md*

**# CSC615M MCO1**

**## Text Cleaning Project**

**#### If the .jar is broken (skip this part if the file is working)**

1. Run the batch file `RemakeJar.bat`

**\*\*or\*\***

1. Go to the file folder and make sure that there are 4 java files and 1 manifest.txt file

2. Open the command line (same directory)

3. Compile the app by typing `javac \*.java`

4. Create the .jar file by typing ``jar cfm CSC615M\_Text\_Cleaning.jar manifest.txt \*.class``

5. You should see the executable jar in the same file directory.

**#### To use the app**

1. After opening the executable .jar file, you will see an interface with two input fields for input and output filenames, customize as preferred.

2. Select your configuration for text cleaning, then click **\*\*Clean text\*\***

3. Wait for the program to clean the text (this might take a while depending on the size of the input)

4. Save the result into a .txt or .csv file (depending selected configuration on step 2) - by clicking click on the **\*\*Save to file\*\*** button. (this might also take a while depending on the size of the input)