Contents

1 Text Analytics - Lesson 2

1.1 Lesson Outline

- The INTs (HUMINT, ELINT, GEOINT, OSINT, etc.)

- Text Analytics - the process of analyzing unstructured text, extracting relevant information, and transforming it into useful intelligence.

- Linux: wget, mv, alias, rm, ~, echo, $HOME, pwd, PS1

1.2 While waiting for class to begin, login, launch a terminal window, and try these commands:

1.2.1 https://github.com/matt-jacobs/modules/blob/master/text-analytics/lesson-02/lesson-plan.md

```
clear
pwd
echo $HOME
cd ~
pwd
cd $HOME
pwd
pwd;pwd;pwd
ls
ls --all --author -1 --group
alias x='ls --all -author -1 --group'
x
echo $HOME > myHome.txt
x myHome.txt
cat myHome.txt
mv myHome.txt anotherPlace.txt
x *.txt
mv anotherPlace.txt myHome.txt
x *.txt
gedit myHome.txt &
rm myHome.txt
export PS1='$ '
x ~/nlp
x ~/nlp/programs/cor
x ~/nlp/programs/core
x ~/nlp/programs/ner
unalias x
history
```

1.3 Stanford Natural Language Processing Tools

1. Visit the NLP Software page. Read about the core and NER.

2. Get some text to analyze:
3. Open up a text editor and have it run in the background

4. Go to a news website (e.g., cnn.com, bbc.com, theonion.com), browse to a news story, copy the text into your text editor, and save the file as ~/nlp/data/news1.text
5. Use wget to get a copy of a different news web page

`wget http://msnbc.com` `mv index.html ~/nlp/data/news2.txt` 2. See what version of Java you have installed - version 1.8 is needed.
`java -version` 3. Make an alias for running the coreNLP program (reference)
`alias nlpCore='java -cp "$HOME/nlp/programs/core/*" -Xmx2g edu.stanford.nlp.pipeline.StanfordCoreNLP -annotators tokenize,ssplit,pos,lemma,ner,parse,dcoref -file'` 4. Execute the java program to analyze the two files you collected

`nlpCore ~/nlp/data/news1.txt` `nlpCore ~/nlp/data/news2.txt`
5. Use gedit to look at the two files that were generated: ~/data.news1.txt.output and ~/data/news2.txt.output
6. Checkout this list of POS (parts-of-speech) tags
7. See if you can figure out how to run the Named Entity Recognizer program using the instructions on this page