Running JGAAP Experiments Using the UNIX Command Line

Guide by Luke De Vos for EVL Labs

Why Command Line Experiments?

JGAAP's GUI may be used to conduct experiments very well. However, many experiments can be run automatically using JGAAP through the command line. This convenience quickly becomes necessity as the scale of a research project grows.

Configuration Files

Since we are running experiments without the GUI, we do not have buttons to tell JGAAP what settings to use and what files to experiment on. Instead, we tell JGAAP what to do with "configuration files". We need two types of config files: settings and corpus. They will be .csv files, which we can edit as spreadsheets. We fill in the spreadsheet's fields with the settings and files we normally would specify with the GUI.

First, we'll look at the settings configuration file.

The settings configuration file follows this format:

	A	B	C	D	E	F
1	Experiment Set Name					
2	Experiment Name	Canonicizer(s)	Event Driver(s)	Analysis Method	Distance Function	Path to Corpus Config File

The "Experiment Set Name" will be included in the name of the file JGAAP writes experiment results to.

An example of an experiment settings configuration file:

	A	B	C	D	E	F
1	ExpSet1					
2	Exp	Unify Case&Strip Numbers	Word Lengths	Leave One Out Absolute Centroid	Cosine Distance	/home/luke/EVL/Project1/FileCorpus.csv
3	Exp	Unify Case&Strip Numbers	Rare Words	Leave One Out Absolute Centroid	Cosine Distance	/home/luke/EVL/Project1/FileCorpus.csv

The above example gives JGAAP directions for two experiments. Any number of experiments can be run by filling more lines with settings in this format. If 10 lines of settings are filled, 10 experiments will be run.

More than one setting can be selected in a given slot by separating them with "&". For an example, look at column B. Two settings specified with "&" does not mean two experiments are run on

using one line; it means both canonicizers are used in that single experiment. The same strategy can be used for any settings the JGAAP GUI would allow you to select multiple of.

The final column must provide a path to the "corpus" configuration file. The corpus configuration file is another .csv file. It provides the author, path, and description of each file to be experimented with.

Corpus configuration file format:

	A	B	C
1	Author	Path to Text File	FileName by Author

Example corpus configuration file:

	A	B	C
1	Faulkner	/home/luke/EVL_Labs/Project1/Faulkner/text1.txt	text1.txt by Faulkner
2	Faulkner	/home/luke/EVL_Labs/Project1/Faulkner/text2.txt	text2.txt by Faulkner
3	Hemingway	/home/luke/EVL_Labs/Project1/Hemingway/text1.txt	text1.txt by Hemingway
4	Hemingway	/home/luke/EVL_Labs/Project1/Hemingway/text2.txt	text2.txt by Hemingway

As with the experiment settings configuration file, any number of files can be experimented on by repeating this pattern.

Note that this corpus configuration file includes files from more than one author. This is no problem for JGAAP. To the contrary, this will often be necessary.

Writing the Corpus Configuration File

https://github.com/devosl99/corpusPopulator

Corpus configuration files can be written manually, but the script from the link above can write a corpus config file with all the files in a given directory automatically. This will save a great deal of time.

NOTE: script assumes the author's name is the name of the directory passed to it. Therefore, the name of the directory must match what you substitute for "Author" in column C.

To use, navigate to the directory containing the script and enter on the command line:

```
./[scriptName] [path to directory] >> [corpus.csv]
```

For example,

```
./corpusPopulatorV3.sh Project1/author1 >> project1Corpus.csv
```

Do not forget the "`./`" preceding the line.

After the above example command is entered, all the text files by author1 will be included in the project1Corpus.csv file in the proper format. You can add more files or more authors by repeating this command with a different directory.

Writing the Settings Configuration File

The settings configuration file is easily written manually. Simply refer to the JGAAP GUI for the names of the canonicizers, event drivers, etc., and fill the fields appropriately.

The event driver field is likely to be the only field that changes between experiments. As a general tip, you can write one experiment line, copy-paste the line as needed, then manually change the event drivers.

Once you have your experiment settings and corpus configuration files set up, you can get experimenting!

Running Experiments

Navigate to the directory containing your JGAAP .jar file

The .jar will be something like this: "`JGAAP-7.0.0-alpha-3.jar`"

Enter on the command line:

`java -jar` [.jar file name] `-ee` [path to settings configuration file]

For example,

```
java -jar JGAAP-7.0.0-alpha-3.jar -ee project1/experiment1Settings.csv
```

Once this command is entered, JGAAP will run the experiments in the passed configuration file. As the experiments run, JGAAP writes information about the status of the experiments to standard output.

Once the experiments finish, a folder named "tmp" will appear in the same directory the JGAAP .jar file is in. Within tmp is a series of subdirectories for each setting specified in your config files. Navigate all the way through each setting as desired until you arrive at a text file named after the experiment set name specified in the experiment settings config file. Your results file will look something like this:

```
1 ARose1.txt by RF /home/luke/EVL_Labs/Faulkner_Hemingway/Short_Story_Work/Fit_Texts/RF/ARose1.txt
2 Canonicizers:
3         Normalize Whitespace
4         Punctuation Separator
5         Strip AlphaNumeric
6         Unify Case
7 EventDrivers:
8         Character NGrams n : 2
9 Analysis:
10        Leave One Out Absolute Centroid with metric Cosine Distance
11 1. RF 0.0158138462601477
12 2. RH 0.049166610284078094
```

Line 1: the text file analyzed, the author, and text file's absolute path. (as per corpus config file)

Lines 2-10: the settings used in this specific experiment as determined by the settings config file.

Lines 11-12: The author following the "1." is the author whose writing style most closely matches the writing style in that experiment's text file using the above settings. The number following it is the "distance" from the writing style; a lower number means less distance, or difference, from that writing style. The authors included here are the authors of the files from the corpus configuration file.

Extracting Results

https://github.com/devosl99/getResults

The above script can be used to automatically extract useful information from the results files. The script will total the number of correct and incorrectly identified files, output the names of incorrectly matched files, and display the number of ties. Due to the improbability of ties, a tie usually suggests a problem in the data or analysis methods.

NOTE: script only works for results files concerning experiments with two authors.

To use, enter on the command line:

```
awk -f [scriptName] [path to results file]
```

The results file will likely be a very long path through the tmp folder. You can use the TAB key to auto-fill folder names as appropriate to make the process of writing out the path easier.

For example:

```
awk -f getResultsV4.awk tmp/Normalize Whitespace Punctuation
Separator/POS/Leave One Out Absolute Centroid-Cosine
Distance/FF_FH_Analysis_01FFvsFH2019-07-01.txt
```

Your results will be written to standard output. They should look something like this:

```
luke@luke-VirtualBox:~/EVL_Labs$ awk -f getResultsV4.awk tmp/Normalize\ Whitespa
ce\ Punctuation\ Separator\ Strip\ Alphanumeric\ Unify\ Case/POS/Leave\ One\ Out
\ Absolute\ Centroid-Cosine\ Distance/Fakes\(short_stories\)Exp2019-07-03.txt
FF vs FH
True Author: FF
    Correct: 14
  Incorrect: 1
True Author: FH
    Correct: 13
  Incorrect: 2
Incorrectly Matched Files:
  F11.txt.fitted
  H10.fitted
  H15.fitted
```

As with all standard output, it can instead be written to a file by ending the above command with "`>> [fileName]`".

Conclusion

You're now ready to do research with JGAAP through the command line and build upon the magnificent colossus of human knowledge. If there are any problems with the guide or the provided scripts, or you have any questions, please let me know at devosl@duq.edu.