Tools for (Field) Linguists

Gina

April 30, 2011

Concordia

Mission:

Learn a few key words and a bit of programming so you can be resourceful, and save time (and pull in the big grants).

Advice:

- Reduce, reuse, recycle,
- Avoid toy world projects

Roadmap

- Git
- Wiki
- GATE
- Groovy

Git - 10mins

- Create a GitHub account
- Fork ToolsForFieldLinguistics
- Clone it onto your computer
- Look at the files, what is in where?

Git 30mins

- Now that you know what a repository looks like:
  - Create a new repository in your account (for the mini project you will do today)
  - Follow the instructions to create the repository on your computer and push it to GitHub

Wiki 10mins

- Create a wiki for your mini project
- Write a sequence of steps to get you from your data to your goal

GATE

- Open Source
- Very complex
- Lots of things to do and try
- We will only use GATE to highlight words in context, and to run Groovy scripts on our corpus

GATE 30mins

- Lets see what the built-in ANNIE pipeline can do on English text
- Create a new Language Resource
- Click on the Green Flower
  - It will load all the processing resources for ANNIE and the Annie pipeline application
- Run the application
- Look around at the information GATE tagged for you
- Notice the Tokens, thats where noun/verb information is stored
- Add a Snowball stemmer into the mix

GATE 5mins

- Create a new Language Resource with the
  - URL : http://blogworkorange.blogspot.com/
  - Encoding: utf-8
- Right Click on the Document
  - Create a corpus with this document
- Create a new Processing Resource
  - Document reset
  - GATE Unicode Tokenizer
- Create a new Application
  - Put the Unicode Tokenizer to the right pane
  - Run this application

Lunch

Groovy

- Click on the black P to load plugins
  - Check “Load now” for Goovy
  - Click OK
- Click on the Tools menu
  - Groovy Tools
    - Groovy Console
- Now you're read to write some Groovy!

The End Result

- Open ToolsForFieldLinguistics/src/com/fieldlinguist/groovyInGate/ExtractWordsOrderBySuffix.groovy
- Run it (CTRL+R)
- If it worked, your output will be in your GATE folder
  - Words_function_vs_content.html
    - To see what this looks like, copy it into the ToolsForFieldLinguistics/src/com/fieldlinguist/javascript/tabletobarpielinegraph, then open it in a browser
  - Words_function_vs_content.txt
  - Words_to_look_for_suffixes.txt

Groovy Practice – Maps,loops,output,regex 10mins-1hr

- Go to your github fork so you can see the visual commit history
- Use tutorial.groovy to build your own spanish “Morpheme Finder”
- Start with commit 1, work your way until the final commit

Groovy Practice – Teams of 2 1hr

- As a team of two, grab some data (you can use one of the corpus in cesine/CorporaForFieldLinguistics or build your own from a blog, rss feed, twitter, or opensubtitles.org
- Decide on a Research Question that you want to investigate
  - How can you load that text into GATE
  - What kind of rules would you like to write to highlight the key data?
  - How can you write some Groovy to find what you want?