Assignment 3

L715/B659

Due Tuesday, October 4

1. Obtain a copy of the PAN16 Author Profiling training corpus (`http://pan.webis.de`).

2. Extract the data into a format you can work with for the rest of the assignment. Make sure you are keeping the meta-data (e.g., the category) organized, too.

3. POS tag and syntactically parse at least some of this data with an NLP package of your choice.

4. Select some subportion of the data and qualitatively evaluate what's happening. Do the parses seem accurate? Do they seem helpful to the task at hand? Why or why not? It might help to consult online documentation about what the categories mean.

5. For two distinct files:

   (a) count up the number of instances of *which* and *that*

   (b) count up the number of relative clauses the parser identifies (e.g., number of `rcmod` relations)

   (c) (informally) note whether there is a correlation between the two different kinds of counts. That is: is the identification of syntactic properties doing something that basic word counting isn't?

   Feel free to substitute some other lexical vs. syntactic distinction, if you have a different question you're interested in examining.

Condense all your work into a brilliant, gleaming report!