Assignment 2

L715/B659

Due Tuesday, September 20

You are free to work in groups or free to work individually. You are also free to consult with each other, even if you're working individually.

1. Build an author identification (Y/N) system:
   (a) Read the overview paper of the PAN 2015 Author Identification Task.[1]
   (b) Download the training & test corpora from the PAN 2015 Author Identification Task.[2] You can choose any or all of the four languages to work with for the following.
   (c) Extract a (small?) set of "shallow" features from these documents.
      - You may use `CountVectorizer` as a starting point, but you are required to go beyond that and to use aggregate features (e.g., percentage of tokens which are punctuation).
      - If there are features that you would like to extract, but cannot currently do so, take note of those.
   (d) Using a classifier of your choice & training on a handful of documents, test the classifier on the held-out documents.
   (e) Evaluate the performance of the classifier.
   (f) Condense all your work into a nice & shiny report!
2. Download & install the Stanford CoreNLP tools: `http://stanfordnlp.github.io/CoreNLP/`
   - Feel free to look into a Python wrapper (or other programming language of choice): `http://stanfordnlp.github.io/CoreNLP/other-languages.html`

[1] `http://www.uni-weimar.de/medien/webis/events/pan-15/pan15-papers-final/pan15-authorship-verification/stamatatos15-overview.pdf`

[2] `http://pan.webis.de/clef15/pan15-web/author-identification.html`