Samuel Valentim-Gervais October 2nd, 2019

Deliverable 1

Dataset Choice

I chose to use this dataset I found on Kaggle: (https://www.kaggle.com/tomlisankie/blog-posts-labeled-with-age-and-gender)

It's a dataset of 600K+ blog posts, labelled with the age and gender of the poster. I chose it because it provides me with two labels for my model to predict, and well… let's just say that lack of data won't be a problem. The dataset weighs in at 751 MB, and for text data, that's absolutely enormous. I'm pretty sure that the sheer amount of data will help me a lot in creating a predictive model. (Although I think I remember hearing something about "never enough data" in one of the lectures, so maybe I shouldn't be so confident, but whatever, it'll have to do ;))

Preprocessing

Yeah, I'll do a couple things to the dataset before I work with it. First off, I'll change the "male" and "female" labels to 1 and 0, since numbers are probably way easier for the algorithm to work with than strings. It saves me having to add some code to convert to/from strings, since it'll already be done. Another thing I *might* do is remove very rare words from the (training) set. I'm not sure if this is ideal (I've never done this before…), but if it is, it should save a boatload of time and space modeling relationships between words that aren't really useful to the model because it'll hardly ever see them.

The model

Maybe two models. I do have two labels, after all. Anyway, I don't know at all what sort of algorithm I'll use. Since I have a *lot* of features, I'm thinking naïve bayes for the gender classifier (it's also the only classifier I learned so far, but I heard it's good on datasets with a lot of features). I'm sure there's something fancier and better adapted to my dataset that I've yet to learn, though. As for the age, well, that's a discrete value, so I'll need some sort of regression algorithm that works with text inputs. But even if it just uses word frequencies as inputs, for example, wouldn't the X matrix just be unworkably huge? One row per word, right? I'll definitely need to learn something fancier for that.

The final project

Posters are boring. I'm building an app. You upload a text file, and it spits out predictions. I'm not sure what I'm going to use as a "medium", but I know that Linux programs that read from STDIN and write to STDOUT aren't the most visually engaging things ever, so I'll have to come up with something for my frontend. A simple webpage with HTML/php will probably get the job done, so for now, that's what I'm planning.