Karan Gadiya, Alex Gromadzki, Sean Mullane

Kaggle Competition 3

# Who might care about this problem and why?

As we discussed in class, the use case of Twitter data in crime prediction could be logically expanded and adapted to predict IED detonation in the Middle East. Thus, the initial value of strictly one form of response might not be what makes a solution interesting to a client. While predicting users’ ages off of their blog activity might not be the most pressing issue, we might consider the other expansions to text analysis. For example, advertisers could extract useful information beyond age from similar NLP analyses in order to target what ads are displayed when the user is on the blog site. If we were to expand the scope of this problem, entities such as Facebook or Google might be interested in analyzing what users are posting in order to respectively suggest pages to like or to query.

# What made this problem challenging?

Upon initial data exploration, we noticed that all of the data was not in English. Within the scope of our problem, we did not spend time implementing sophisticated translation techniques that could have otherwise improved the predictability of a user’s age. In addition, we noticed that the age (response variable) in the training set seems to be missing a band between 18-21 and 27-32. This could be an artifact of the competition set up or simply of poor data collection. Given that there were 400,000+ posts, it is highly unlikely to get the clean cutoffs we observed. The actual implementation of the TF-IDF and SVD models caused us relatively less difficulty than determining the correct parameters to feed into our Elastic Net regression. The most frustrating thing about this problem was the time that it took to compute each model.

# What other problems resemble this problem?

On a very general level, we could compare this problem to most applications of NLP. On a more topic-specific basis, it could be useful to adapt this model to predict several other responses. For example, we could predict future blog usage based off of previous blog usage in terms of number of posts and topics. We also could predict future blog topics based off of previous posts’ text content. Beyond the scope of the blogosphere, we could attempt to scrape other forms of social media in order to compare user consistency across different platforms.