- Who might care about this problem and why?

  The application of predicting personal identifiable information was widely used across multiple areas. We identified several groups of people that could potentially be interested in the problem, including linguistic experts, social scientists, marketing and advertising, forensic studies. For linguistic experts and social scientists, they try to identify a pre-social media generation and the post-social media generation to identify different language patterns and associate that with social behaviour and media studies. For marketing and advertising, they could potentially identify their product reviews from different aged user groups, in order to improve their products based on feedbacks from different groups. In some digital communities it is easy to hide one's true identity to provide a false age and gender, the analysis of a blogger's writing style and age prediction could be preserved and potentially be used as evidence in federal court to help with fighting against crime and forensic studies.

- What made this problem challenging?

  This problem is particularly challenging due to the size of the dataset itself, being 265 Megabytes. Also, during the data-processing stage, there are several issues such as foreign language, the different spelling of the same word (especially for teenagers), extensive use of punctuation (which we took two different approaches in our coding), the extensive amount of bag of words and the irrelevancy of these bags of words to response variable etc.. We exploited various techniques such as reducing dimensions with pc, elastic nets, getting rid of the vowels consonant variables, tested on smaller samples to tackle these challenges.

- What other problems resemble this problem?
  1. One of this year's capstone projects is the Collective Biographies of Women Project (CBW). This project details cultural representations of women through a large corpus of British and American texts from the nineteenth and twentieth centuries. A central goal of the project is to develop a complete annotated corpus drawn from 1,270 known books, comprising around 13,000 chapters of about 8,000 women. Once complete, this corpus will support the analysis of textual form and content, including narrative structure and biographical detail. Using the text, metadata, and annotation data, CBW seeks to discover information about social networks, cultural patterns, and narrative structures implied by the texts.

  2. There is an analysis on the America's Public Bible (Lincoln Mullen). This explores trends and frequency in use of biblical quotations in newspapers mined from Chronicling America: Historic American Newspapers (LoC).

  3. Utilizing natural language processing to monitor the media reaction to Facebook's disastrous earnings call – News API Monthly Media Review. News API monitors tens of thousands of news sources, analyzes new stories from them as they are published, and indexes them in a vast, live database that people can easily query.

Jing Sun, Murugesan Ramakrishnan, Shaoran Li (C2-3)