## ***Refined Team Project Proposal***

### *Part 1. General Information*

Name: Vou Quien (Ken) Luy, Daniel W. Zhang, Yunlin Zhang
Project Title: The Political Race and the Wit and Wisdom of Donald Trump
Project Description: This analytic will take tweets from Donald Trump's official Twitter page and data from national opinion polls (e.g. Gallup) in an attempt to find any positive and negative correlations between certain words in the tweets of Mr. Trump and his standing among those public polls. From there, the analytic will attempt to predict his standing in future
Describe who will benefit from your analytic: Public media pundits, political analysts, Donald Trump's social media team

### *Part 2. General Data Source Information*

| **Data Sources**<br>- E.g. tweets | **Data Source Description**<br>*(Provide a short description of the data source.)* | **Data Size**<br>Estimate size, e.g. MB? GB? TB? … |
|---|---|---|
| Data Source 1: Twitter | Primary source tweets from Donald Trump himself, secondary source tweets from other people that talk about his hair, etc as well as satire (e.g. @TrumpsHair) | 10s - 100s MB |
| Data Source 2: CNN/ORC | Republican primary candidate poll data. Separated by state. Interviews with likely republican caucus goers. | ~2.7 MB per file |
| Data Source 3: Public Policy Polling | Poll data regarding how potential voters feel about Trump as a candidate. | ~2 MB per file |

### *Part 3. Detailed Data Source Information*

| **Data Sources**<br>- E.g. Twitter tweets | **Data Characteristics**<br>- Is the data source a realtime source – i.e. are you collecting the data in realtime?<br>- Is the data stored some place and you will collect it periodically? (e.g. a log file)<br>- Is the data static? (e.g. historic data that you will load once) | **Data Frequency**<br>- If realtime data, what is the frequency and volume of data? |
|---|---|---|

| | | |
|---|---|---|
| Data Source 1: Twitter | - We will collect the data weekly. We will store the tweets that match our search in log files. The data will be static. | Weekly frequency, each batch having ~1GB. |
| Data Source 2: CNN/ORC | The data is in a 75 page pdf file. The data is also duplicated in text files. | This is historical data, though we may want to check each week in case there are additional updates |
| Data Source 3: Public Policy Polling | Data will either be in pdf files or raw data if we can get access. | This is historical data, though we may want to check each week in case there are additional updates |