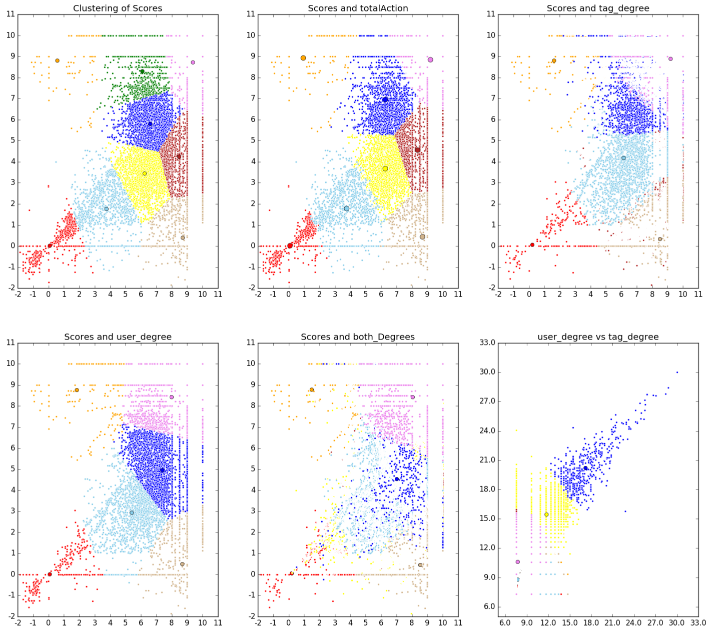
Project Ultra

Phase-1.v4 Phase-2.v1

Fan Yu
University of Michigan
Email: fanyuchn@umich.edu

March 18th, 2016



1 Project Design

Project Ultra has its namesake from WW II designation by the British military intelligence for its secret signals intelligence operations. As such, this project is intended to analyze large amounts of data available from social network and to active prediction model. An "active" prediction model should not only capable of predicting how a system will behave based on observation, but also capable of predicting potential outcomes of inputs one might be able to insert into the system. In other words, compared with "passive" prediction model which is akin to weather forecast, "active" prediction model is more similar to weather forecast PLUS rainmaking, with an additional layer of control over the system.

Project Ultra is intended to have at least three phases. Phase-1 would be data acquisition, data filtering, data cleaning and data storage. Phase-2 would be preliminary analysis, classification of users and information token, and most importantly, sentimental analysis. Phase-3 would build on top of Phase-2, with additional insights borrowed from "relevant" mathematical disciplines, among which would include dynamic game theory and network theory. Of all three phases, phase-3 would be the center piece.

At current stage, this project is designed to use data from Twitter API as an example, tracking both users' behavior and correlations between different information. Section 2 will demonstrate latest results from Phase-2.

With current results, we are capable of recommending twitter users as potential targets, from a purely statistical way, that are more popular among other twitter users yet so-far less biased against either side of the campaign. We are also capable of filtering campaign slogans (hash tags that are directly related each candidate) that are less used, less popular and that are actually distracting energy and efforts of a campaign.

2 Preliminary Results

So far most of the effort has been on Phase-1 and Phase-2 of the project, as Phase-3 depends on high quality results from Phase-2. We managed to analyze 4-days' worth of data collected within the U.S. from March 4th 5:00 AM to March 8th 5:00 AM. In total, we analyzed ∼8 million tweets. With the dynamic filter set to focus on tweets related to Presidential Campaign (with keywords ”trump” and ”hillary”), we are tracking 22554 related hash tags and 7022 users[1].

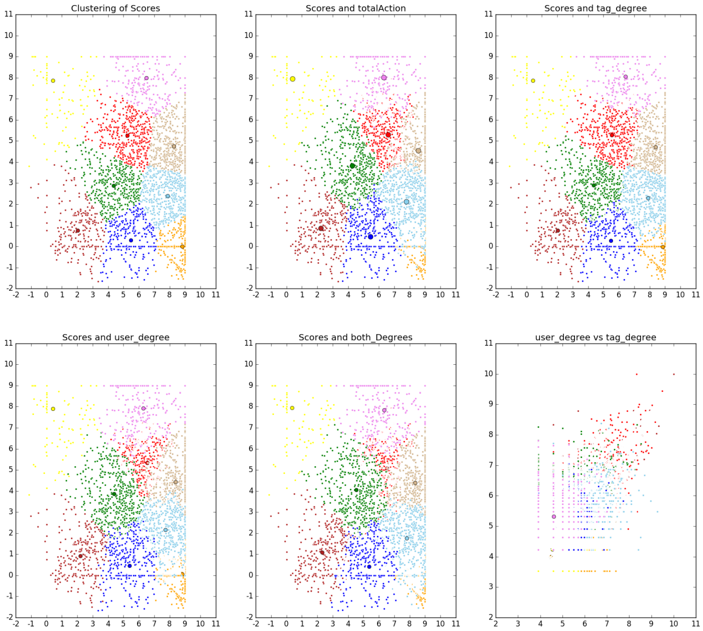


Figure 1: Classification of Users. Using K-means algorithm. All plots except the bottom right has X axis as scores of keyword ”trump” while Y axis as scores of keyword ”hillary”. The bottom right plot has X and Y axis as absolute degree of user and tag respectively. From top left to bottom middle, classification is performed using: scores, scores and totalAction, scores and tag degree, scores and user degree, scores and both degrees. Please note that bottom middle and bottom right are the same classification plotted in different reference frames. All axis are normalized.

[1] Please note that both hash tags and users are tracked under two criteria: 1st, they are related to keywords; 2nd, percentage of their activities related to keywords exceed a certain threshold.

Our objective is to classify hash tags and users according to their relevance with respect to the two candidates (who apparently attracted more attention), judging by hash tags and users behavior statistics. Thus, we assign a pair of scores to each hash tag and user that we are tracking, to represent how closely related to our candidates. The variables that we used to describe and distinguish different hash tags and users involve: scores, totalCall/totalAction (how active each tag/user was during the 4 days' period), tag_degree (how many distinct tags each tag/user related to), user_degree (how many distinct users each tag/user related to). Thus when we are performing classifications of hash tags or users, we are not only considering how closely they are related to keywords, but also how active they are, how widely they are connected to other hash tags and users.

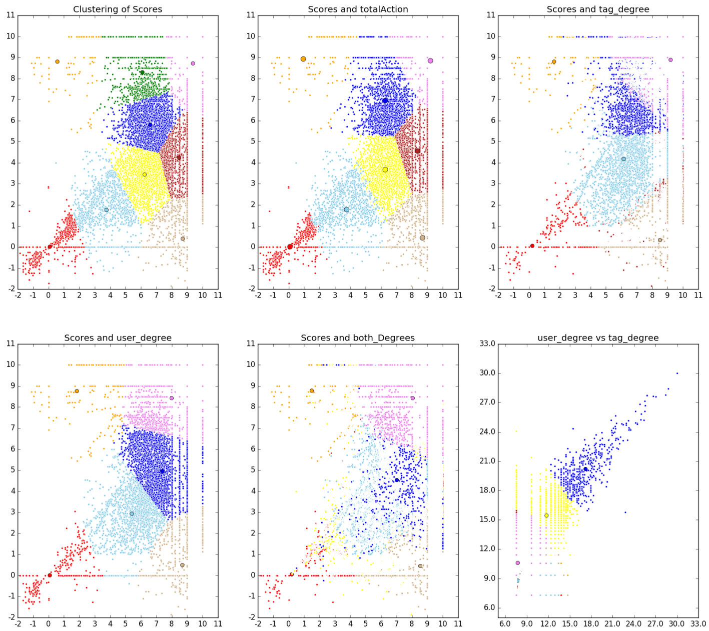


Figure 2: Classification of Hash Tags. Using K-means algorithm. All plots except the bottom right has X axis as scores of keyword "trump" while Y axis as scores of keyword "hillary". The bottom right plot has X and Y axis as absolute degree of user and tag respectively. From top left to bottom middle, classification is performed using: scores, scores and totalAction, scores and tag degree, scores and user degree, scores and both degrees. Please note that bottom middle and bottom right are the same classification plotted in different reference frames.All axis are normalized.

Figure.1 and Figure.2 are our results for users and hash tags respectively. One might be surprised by the similarity between the tow, as there are a lot more users and hash tags closely related to ”trump” than to ”hillary”, demonstrating the extend to which Donald Trump has made himself a celebrity through out the campaign. However, there is a crucial difference between users and hash tags. From the bottom right plots of both Figures, one could observe that, while users share generally comparable connectivity with hash tags and other users, the hash tags are considerably polarized in their connectivity[2]. Thus, when classifying users, we applied equal weight to all variables. But we applied 3 times more weight to degrees of users and tags when classifying hash tags.

The fact that matters revolving around Donald Trump are more hotly discussed could also be seen from Figure.2. Most of the blue group of hash tags are scattered more closely to X axis around the middle (in the bottom middle plot), indicating a higher relevance to keyword ”trump”. In addition, among the dots with their X or Y values strictly equal to 10[3], more blue dots are related to ”trump” also.

Even with preliminary analysis like these, one could achieve amazing result. For example, one could discover twitter users who are more popular among other twitter users yet so-far less biased against either side of the campaign[4]. These twitter users could be prime targets for advertising campaign because: 1st, they have already demonstrated interests in the presidential campaign (their scores are higher than average); 2nd, they are degrees of users and hash tags are also above average, indicating active discussion or debates with other twitter users.

One could also filter out campaign slogans (hash tags that are directly related each candidate) that are less used, less popular and that are actually distracting energy and efforts of a campaign[5]. Those less popular, less used hash tags should no longer be used by people who are waging the campaign, as they reaches out to a narrower range of twitter users.

[2] Please note that, while ALL axis are normalized, the scores are normalized linearly while degrees of users and tags are normalized in Ln(). Thus, the long tail in the bottom right plot of Figure.2 signals that some tags are used significantly more frequently than others.

[3] Those are hash tags that contain keywords ”trump” or ”hillary”.

[4] In this case, the Red and Green dots in bottom middle plot of Figure.1.

[5] In this case, those dots in the bottom middle plot of Figure.2 that has X or Y value equal to 10 and that are NOT blue.