Capstone 2 Project Proposal
Social Network Outreach Analysis
Springboard Data Science Career Track

Eric Cruz

The proposal should address the following questions:

- What is the problem you want to solve?

Develop and test a Data Science strategy for Social Network Outreach. The primary issue is how to identify the target members on a popular selection of social media outlets such as LinkedIn, Facebook, Instagram, and Twitter. Due to increasing efforts to limit the ability to compile the data and share the respective API's , a toolkit of scraping and searching tools will be selected for the purpose of digitizing the information and automating the collection of search results. The end goal is to develop a strategy for engaging the target audience on various platforms by analyzing participation rates or other trends. For example, we may expect to observe generational trends, e.g. younger people have left Facebook or reduced usage in favor of Instagram, older people are likely to be on LinkedIn but not Facebook. The actual outreach will be conducted by individuals reaching out to their peers, using curated contact lists generated from the search phase.

- Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

The client at the lowest level cluster is the Men's or Women's college tennis team, with next level groupings to combine the genders for a school's program, and group schools based on Conference and Regional affiliation. The high level client is the College Tennis Alumni Association, which encompasses all the clusters and already functions as a social media outlet with members.

The client will decide on social media outreach strategies to combine with traditional email campaigns on contact addresses, which may or may not have been verified.

- What data are you using? How will you acquire the data?

The data regarding eligibility as a target member will be obtained from the team websites in conjunction with websites which report match results and recruiting information. The eligibility data will be scraped. Other data sources include membership lists from groups on LinkedIn, Facebook, and the College Tennis Alumni Network.

- Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

The compilation of the eligibility list is a straightforward exercise, but the development of tools for identifying the target individual amongst several possible profiles on each social media outlet will be the dynamic challenge, given the lack of access to the API's. By using commercially available scrape and search tools, the product limitations will lead to a roadmap for a script to be developed should access be obtained at a later date.

- What are your deliverables? Typically, this includes code, a paper, or a slide deck.

The deliverables are a set of tools for conducting the research on a list of names, and a report describing the challenges and solutions. The tools will gather information for matching on keywords and photos. A prototype campaign will be conducted on a sample team cluster, and a sample regional target list will be built. The actual outreach campaign completion is out of scope, as the problem to resolve is to develop a scalable strategy, rather than complete the execution.

- ❏ The submission demonstrates that the student has articulated a **problem statement** that is within scope of this course.
- ❏ The submission demonstrates that the student has selected a problem with a **dataset** that is relatively clean, easy to use, and within scope.
- ❏ The submission demonstrates the student is able to justify a significant **problem** to solve for a realistic client (even if imaginary).
- ❏ The submission includes a strong **justification** of how the client can utilize the **outcomes** of the project.

Problem Statement: Using data for College Tennis Alumni Network candidates, collect and label data on targeted individuals, and engage them to expand the network.

Dataset: The data will be collected from websites using scraping tools and other utilities, and may include a comparison of different methods for obtaining and storing data. There will be outreach based on lists for a local alumni organization included in the data collection exercise.

Clients: The clients are the CTAN, or College Tennis Alumni Network, and local level Cornell Men’s and Women’s Tennis Alumni Associations.

Approach: At the local alumni level, there is a list compiled over 4 years, recording attempts to contact via email for a summer event. A fair number have never received a response. The list needs to be reviewed and updated with the current coaches, as the Women’s coach has been replaced. The 50th Anniversary of Women’s Tennis at Cornell will honor Hall of Fame inductees on campus in June 2020. At the CTAN level, the growth strategy will rely on member volunteer committees at the local level, and I will participate in NYC chapter of CTAN, creating opportunities to create outreach campaigns and record the data. Names and College affiliation, including dates played, from the parent company ITA scoreboard online, will be the baseline for membership eligibility and matching will be attempted on social media networks given scraping capabilities from each source. Natural Language Processing techniques will be used to differentiate between many people with the same or similar name, using school names and tennis among other cross reference searches. Sample target collection data includes whether person is found on Facebook, LinkedIn, etc., and whether school website identifies as "Captain". Outreach will be targeted toward captains, asking for their participation expanding the network amongst their peers and beyond if accessible. Data will be collected on network influence statistics.