

Social Network Outreach Analysis
Capstone 2 Project Step-By-Step
Springboard Data Science Career Track
Eric Cruz

- 1) Roster download from Web Scraping
 - a) **Octoparse**
 - i) Filenames
 - (1) CornellW.csv
 - (2) CornellM.csv
- 2) Roster Data Wrangling to produce input file for LinkedIn Search
 - a) Groupby() to consolidate same name on multiple rosters
 - b) Supplement Roster downloads with All-time Letter Winner lists
 - c) Parse Out Names for future use
 - i) **Nameparser utility HumanName()**: First, Last, M.I., Suffix
 - (1) Future Development: Nickname Database
 - d) Output file is "FirstLast" list of names
 - i) Future Development: Utilize Married vs. Maiden Name info but not needed in base case due to previous research
 - ii) Analyze gathered LinkedIn Profiles not found in Search Result
 - e) Program File - Capstone2_DataWrangling_Roster_YaleM2.ipynb
 - i) Input
 - (1) YaleM2.xlsx
 - (2) YaleMpdf.csv (letterwinner list if applicable)
 - ii) Output
 - (1) YaleM2_group.csv
 - (2) NamelistYaleM2.csv
 - (3) CriteriaListYaleM2.csv
- 3) LinkedIn Search
 - a) **Selenium with Google Chrome Extension** used in **Jupyter Notebook** script
 - b) Uses personal login but not counted in LinkedIn limits (only restrict outreach attempts), but applied for whitelist crawler exemption 3/15/20. Applied for PeopleSearch API under Marketing Developer application 2/8/20, received denial response 3/30/20.
 - c) Program File - LinkedIn_Selenium_YaleM.ipynb
 - i) Input
 - (1) NamelistYaleM2.csv
 - (2) CriteriaListYaleM2.csv
 - ii) Output
 - (1) YaleMenLinkedIn2b.csv (interrupted due to excess search warning)
 - (2) YaleMenLinkedIn2_old.csv - to re-rerun and use new file
- 4) LinkedIn/Roster Data Wrangling to produce Model Features
 - a) Gender label
 - b) FuzzyWuzzy ratios for partial name matching
 - c) Boolean classification for Tennis, School, SameName, and combination Same/Tennis
 - d) Decade as function of Year
 - e) Program File - Capstone2_DataWrangling_LinkedIn_YaleM2.ipynb

- i) Input
 - (1) YaleM2_group.csv
 - (2) YaleMenLinkedIn_old.csv
 - ii) Output
 - (1) RosterLinkedIn_YaleM2.csv
- 5) Model and Feature Selection for Binary Classification (*Iteratively built into LinkedIn/Roster wrangling*)
 - a) **Random Forest Classifier (RFC)**; KNN, Logistic Regression evaluated
 - b) **Natural Language Processing with Fuzzy Wuzzy python library**
 - i) Use for approximate name matching as proxy for nicknames
 - c) Year, Decade, Fuzzy Scores are numerical features
 - d) Binary classification features are NameMatch, Tennis Activity, School, and Hybrid NameMatch&Tennis
- 6) Model Application
 - a) Train Models on Men's Data and Predict on Women's Data and vice versa
 - b) Combine Men and Women with Gender Classification feature and Re-train
 - c) Apply model to unseen data from sample of other schools
 - d) Program File - Capstone Project 2-rfc-feature-build-CornellMMWW.ipynb
 - i) Input
 - (1) RosterLinkedInLabel_CornellMW.csv
 - ii) Output
 - (1) modelmmww.joblib
 - (2) Xdata12_CornellMMWW.csv (Contains predictions)
- 7) Get Predictions
 - a) Program File - GetPrediction-YaleM.ipynb
 - i) Input
 - (1) Modelmmww.joblib
 - (2) RosterLinkedIn_YaleM2.csv
 - ii) Output
 - (1) Predict_YaleM2_mmww.csv
- 8) Analysis of Model Results
 - a) For each selected feature, use bar charts to get frequency distributions and pie charts to get percentage scores
 - b) Compare true labelled results with model predictions in feedback loop to identify important features
 - c) Supplement Model Predictions with simple feature sorting to identify likely false predictions
 - d) Program File - Capstone2_PredictResults-YaleM2.ipynb
 - i) Input
 - (1) Predict_YaleM2_mmww.csv
 - (2) RosterLinkedIn_YaleM2.csv
 - ii) Output
 - (1) Note that prediction column is appended to Labelled dataframe
 - (2) Only the Cornell Data is used to build the model, and predictions are taken while building - other schools will not build, only run
 - (3) Results in Notebook - need to work on output report

- 9) Deliverable is Roster List with LinkedIn search results, scores, feature data, and supplemental personal information
 - a) Location, Employment, Schools, Activities, Roster Info
 - b) Supplemental Letter Winner list downloads