Twitter User Gender Classification

Introduction

This problem is to determine whether twitter account belonged to male, female or brand by just looking at twitter profile. So idea is to predict user gender based on tweets & profiles, set of words that strongly predict male or female gender and by stylistic factors like link color and sidebar color.

Competition Goal

This is a classification task and the goal of the competition is to create a predictive model for classifying user's gender.

Dataset : https://www.kaggle.com/crowdflower/twitter-user-gender-classification

Required libraries

We can use the Anaconda Python distribution to install most of the Python packages we need. The primary libraries that we'll be using are:

- **NumPy**: Provides a fast numerical array structure and helper functions.
- **pandas**: Provides a DataFrame structure to store data in memory and work with it easily and efficiently.
- **scikit-learn**: The essential Machine Learning package in Python.
- **matplotlib**: Basic plotting library in Python; most other Python plotting libraries are built on top of it.

Questions

Here are a few questions which comes in mind with this dataset:

- how well do words in tweets and profiles predict user gender?
- what are the words that strongly predict male or female gender?
- how well do stylistic factors (like link color and sidebar color) predict user gender?

The Data

The dataset contains the following fields:

- **_unit_id**: a unique id for user

- **_golden**: whether the user was included in the gold standard for the model; TRUE or FALSE
- **_unit_state**: state of the observation; one of *finalized* (for contributor-judged) or *golden* (for gold standard observations)
- **_trusted_judgments**: number of trusted judgments (int); always 3 for non-golden, and what may be a unique id for gold standard observations
- **_last_judgment_at**: date and time of last contributor judgment; blank for gold standard observations
- **gender**: one of *male*, *female*, or *brand* (for non-human profiles)
- **gender:confidence**: a float representing confidence in the provided gender
- **profile_yn**: "no" here seems to mean that the profile was meant to be part of the dataset but was not available when contributors went to judge it
- **profile_yn:confidence**: confidence in the existence/non-existence of the profile
- **created**: date and time when the profile was created
- **description**: the user's profile description
- **fav_number**: number of tweets the user has favorited
- **gender_gold**: if the profile is golden, what is the gender?
- **link_color**: the link color on the profile, as a hex value
- **name**: the user's name
- **profile_yn_gold**: whether the profile y/n value is golden
- **profileimage**: a link to the profile image
- **retweet_count**: number of times the user has retweeted (or possibly, been retweeted)
- **sidebar_color**: color of the profile sidebar, as a hex value
- **text**: text of a random one of the user's tweets
- **tweet_coord**: if the user has location turned on, the coordinates as a string with the format "[*latitude*, *longitude*]"
- **tweet_count**: number of tweets that the user has posted
- **tweet_created**: when the random tweet (in the **text** column) was created
- **tweet_id**: the tweet id of the random tweet
- **tweet_location**: location of the tweet; seems to not be particularly normalized
- **user_timezone**: the timezone of the user