

Gender Classification from Tweets

Matthew Stanley Class of 2015.5

Middlebury

Abstract

Twitter is a popular social media platform that produces an enormous amount of user created information and associated metadata. However useful pieces of demographic information such as gender, age, ethnicity, political orientation, are not present as part of the definition of a user. Thus a particularly interesting challenge presents itself, is it possible to automatically determine these properties? In terms of this project only the attribute gender was studied. The goal of this research is three-fold: to create a simple, efficient and accurate framework for collecting and labeling training data, to create an accurate classifier for predicting the gender of Twitter users, and finally to produce a simple interactive web application for showcasing the result of the latter two goals.

Tools



Overview

In order to classify the gender of a twitter user based on the text of their tweets I decided to utilize machine learning techniques. In order to accomplish this my project is broken into a 2-step process: collect a set of labeled training data and implement a machine learning algorithm to learn from that data set and then predict the labels of new tweets.

Tweet Collection

For this project I wanted to implement an efficient but accurate method for producing a labeled training set of tweets. In order to accomplish this I utilized another dataset of baby names recorded as part of the Social Security application process. The dataset provides **1,825,433** names, **93,889** are unique, along with their associated gender and occurrence. This allowed me to create a mapping from first name to gender.



Frequency - Inverse Document

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

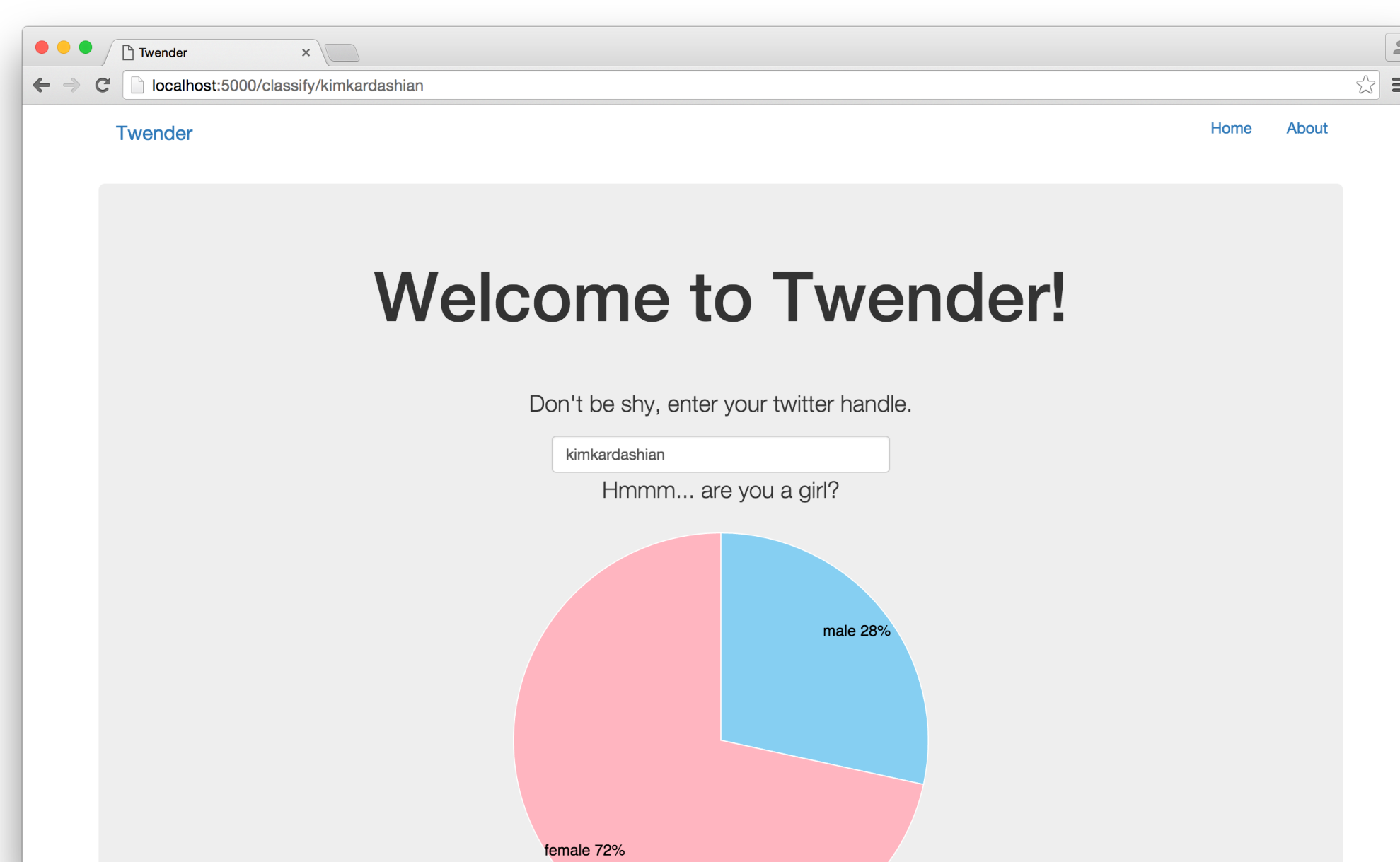
df_x = number of documents containing x

N = total number of documents

Frequency (**TF-IDF**) scheme for feature extraction. In the context of this project terms are equivalent to a single word.

Tweet Application

As an end result I implemented a simple web-based application to be an interactive display of what I had accomplished. The web application uses the flask micro-framework and is written in python.



The application allows the user to enter a twitter handle, it then uses twitter's API to access the specified user's most recent tweets. Finally it utilizes the classifier built for this project to determine the user's

gender, at which point it displays the tweets used in the classification and the percentage classified as male or female.

Results

To allow testing of the resulting classifier I used the same method for collecting the training dataset to create a testing set. I then ran the classifier on that set and calculated the percentage correct. Using this method the classifier achieved

63% Accuracy

Conclusion

In conclusion we can see that the end result shows promise toward achieving the initial goal of the project. From the accuracy determined for the classifier we can also see that there is also room for improvement. For future work I would like to spend time researching an intelligent way of selecting terms and tokens for features. For example an area of potentially rich feature information is the set of socio-linguistic features found in tweets, e.g. user mentions, hashtags and URLs.

Tweet Analysis

The machine learning algorithm I chose was a simple naïve Bayes classifier. I also chose to use Term