Requirements Specification

David Lambertson

Twitter Gender Determiner

Below are given the requirements for the project titled Twitter Gender Determiner. Contained are both items the system must/will do/contain while the last two are optional requirements.

1. The system will use the Naïve Bayes Algorithm for data mining.  
   **Description:** Using the Naïve Bayes Algorithm, the system will determine the gender of the user through the given data.
2. The system will be 80% correct.  
   **Description:** Every 8 out of 10 will be determined correctly by the system. I chose this number because through all the research I did, it seemed like a logical number. I read of people reporting 90%, 75% and so on. I thought putting my number in the middle would be a good range for me.
3. The system will be able to determine the gender within 5 minutes per group of data.  
   **Description:**  Each piece is one tweet. The system will not need any longer than 5 minutes (or less) to determine the gender of the user based of the content of the tweet.
4. The system will have access to already labeled tweets of which the training set and testing set will be created.  
   **Description:** I have previously downloaded 20,000 tweets from twitter. This is the batch of data I will be using. I only want 1000 labeled tweets to use, so 667 for training and 333 for testing. I have been going through and labeling these for use.
5. The system will use a collection of tweets.  
   **Description:** The system will determine a user’s gender based off what is contained in a collection of tweets that could range from 2 to 200.
6. Any tweet that is a re-tweet or looks like it is from a business account will not be used within the testing or training set.  
   **Description:** This is to maintain any bias. A re-tweet is not the user’s real words so it could be out of context in trying to determine gender. Business accounts usually are posting to promote their products or business. This could also skew any data while training or testing.
7. The system will only use tweets in the English Language.  
   **Description:** Trying to involve other languages, while it would be cool and an interesting feature, would require some previous knowledge in said language. I do not know any other languages well enough to aid the system, therefore only English will be used.
8. The system may have a user GUI to make it easy to use.  
   **Description:** If I have the time, I may create a GUI to make it easier to use the program and for the results to display more easily. This will be based off how much time I have left in the semester and how many hours more I need.
9. The system may be able to take individual accounts to determine their gender.  
   **Description:** The user of the program could pass a given account name with which the program will then access twitter and download tweets from the given user. Using these tweets, the program will determine the gender of the given twitter user.