Chris Won, Aiden Jung, Alex Yiu, Tiening Chen

Professor Samuel Rebelsky

CSC151.03

Nov 28, 2017

Project Short Report: Trump-Bot

Primary Goal

The team “Aiden and his friends” is writing a program named Trump-Bot, a text generator program that simulates President Donald Trump’s tweets. Our objective is to mimic Donald Trump’s style of speech based on the statistics gathered from his Twitter posts. Initially we planned import Donald Trump’s live tweets from Twitter to Dr. Racket with Scheme built-in procedure get-pure-port from net/url package (https://docs.racket-lang.org/net/url.html), instead we implemented the procedure get-tweets that imports tweets from the newest downloaded file.

New “tweets” were originally designed to based on: the analyses of both semantics; the syntax of the Twitter posts; basic categorization of proper nouns into more specific topics. Quantitative aspects of the data set will be explored to support the persuasiveness of the simulation. However, after team discussions and reevaluations, new tweets will be based on: the frequency of the words and syntax of the Twitter posts.

Data

Originally the team planned to gathered the data directly from Donald Trump’s Twitter account: (https://twitter.com/realDonaldTrump). The raw data will be composed of HTML code of the Twitter page. However, data extracted with this approach from the tweeter page are limited to the current page. With our final version of data set, raw data are imported from (https://github.com/bpb27/trump_tweet_data_archive). Raw data include all previous Trump’s tweets up to date. As the Twitter posts are also embedded within the HTML code, we can remove any unnecessary HTML tags to only extract the information we need with the procedure split.

Algorithm

```
;;; Procedure:
;;;   split
```

;;; Parameters:

;;;   str, a string

;;; Purpose:

;;;   Divide a string by word and create a list of strings each of which contains an individual word.

;;; Produces:

;;;; result, a list of strings

;;; Preconditions:

;;;   [No additional]

;;; Postconditions:

;;;   There are still some unclear words in result because some of the words have punctuations inside.

Description

> The procedure split is designed to extract individual words from string of words. It wipes out punctuations and extracts each word from a sentence as a string so that we can analyze the data and pick random words to generate sentences that mimic Trump's style of speech.

;;; Procedure:

;;;   get-tweets

;;; Parameters:

;;;   path, a string that names a file

;;; Purpose:

;;;   Import a file of Donald Trump tweets and extract only the texts and date of the tweets

;;; Produces:

;;;; result, a list of lists

;;; Preconditions:

;;;   [path has to be valid]

```
;;; Postconditions:
;;;   if path is valid, each list in result should be formatted as ‘("text" "date")
```

Description

The get-tweets procedure is used to convert hash tables in the downloaded files into lists of strings which consist of Donald Trump’s tweets and dates.

```
;;; Procedure:
;;;   avg-word-ct
;;; Parameters:
;;;   lst, a list of lists whose first elements are strings
;;; Purpose:
;;;   Calculate the average number of words in all of the first elements of the lists in lst
;;; Produces:
;;;;  result, a real non-negative number
;;; Preconditions:
;;;   Each list in lst should be formatted as ‘("text" "date")
;;; Postconditions:
;;;   result could be different from the actual average because split cannot wipe out all the
unncessary texts.
```

Description (arith-mean)

The procedure is used to calculate the average length of a single tweet by dividing the sum of all the number of words by the number of tweets.

Analysis

With the implementation of the new pathway for raw data, trump-Bot gain access to all previous tweets. The larger data sets allow us to randomly generates tweets that are closer to real Trump Tweets. By using the procedure split, our algorithm can correctly extract individual words at 90 percent of the time estimated. Although the final version of Trump-Bot is not able to mimic Donald Trump's usage of words, topics and subjects, but team reaches satisfying results. The first procedure creates a tweet based on the frequency of words and the average length of tweets. The second procedure, categorize-sentence, generates a tweet based on the frequency of words usage and the syntax of the Twitter posts.