Hyegi Bang

Software Design

2 March 2018

Text Mining

Twitter-Markov

Project Overview

The project uses Twitter as the data source for collecting a given user's tweets from their timeline. Markov chain, a probabilistic technique to forecast the value of a variable that is solely influenced by the previous state, was also proceed throughout. The project was conducted to analyze user's tweet and generate a sentence that the user might say.

Implementation

To establish a Markov chain, data plays a significant role. Tweets were extracted as a string and later stored in a txt file to first be categorized as either a prefix or a suffix. Suffix map is a dictionary with every two consecutive word, prefix, as a key and the predicted word as a value in order to randomly generate the third word based on the prefix. However, since for the first few words, there are no data to compare to, words are simply randomly generated and stored. After enough data is stored to compare, mapping begins. Every word it generates, a new tuple is created; the suffix becomes the prefix, paired with its consecutive word and generates a new word.

Markov Analysis was implemented as a class to organize and define data and procedure with better extensibility. Also, "self" , an instant variable from class, is useful for the project especially with random generator. Another design choice made was to extract hashtags, retweets and url in order to precisely represent user's language. Documenting only the user's word would result a more accurate representation of a user since every individual has their own patterns in their language

Results

"people in chicago he created kits full of socks toil chris long gave his paychecks from the first responders and people helping each other out that is exactly why we need to protect it for kids like bill it is about the road ahead so share your getcovered story "

-Barack Obama

"i knew very well interview a panel of people around the world of why it is a total political and legal bust just confirms all big cpac straw poll results 93 approve of the russian meddling in the election i said was to look at the whi it was my"

-Donald Trump

"from another 25 engine test learn more about one of the most advanced weather satellites ever operating in tandem have you ever seen a rocket roll out to the red planet this year now coverage from the countdown to liftoff of goess latest weather satellite"

-NASA

Barack Obama is likely to tweet a topic about the community and people with the usage of we and you as the pronoun. However, Donald Trump is heavily focused on himself, I, and the political relationship with Russia. NASA's upcoming tweet predicted to be about a weather satellite, which turns out to be very convincing since according to research NASA had launched a new weather satellite at 5: 02 p.m. Eastern, March 1.

Reflection

From a process point of view, I was able to successfully complete the analysis that I have planned at the beginning of the project. I was able to be more motivated with the project since it was more relatable and I was able to choose the direction. I am eager to personally expand on this project to merge this project with sentiment analysis or cosine similarity based on the result generated. Something frustrating throughout was collecting the data from twitter. The instructed github page about twitter was not clear and did not find it useful. Going forward, I would find a less straightforward topic. I personally found Markov had to go "different" in terms of both process and result. Something useful I found were the packages and their documentation as they provide guidance for me.