Developing temporal word embeddings with Twitter

Zane Page   
University of Georgia   
Department of Computer ScienceAthens, GA  
zane.page@uga.edu  
  
line 1: 2nd Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID  
  
line 1: 3rd Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORCID  
  
line 1: 4th Given Name Surname  
line 2: *dept. name of organization (of Affiliation)*  
line 3: *name of organization (of Affiliation)*line 4: City, Country  
line 5: email address or ORC

*Abstract*— Human language is evolving, the meanings and associations between words is ever changing. Apple is a good example since once it was mostly known as a fruit, now it is also associated with the tech company. In this paper, we attempt to create dynamic word embeddings using data collected from twitter. By examining what words are closer to others we can infer semantic meaning. Then, creating a positive pointwise mutual information matrix, PPMI, we can attempt to align the time-separated matrices over time and examine how the semantic meaning can change over time. Then in the future, using data collected from twitter we may be able to detect changes in language associated with HIV/AID or the opioid epidemic and be potentially be able to predict outbreaks.

Keywords—natural language processing, big data, word embeddings, twitter

# Introduction

Word Embeddings are a helpful method when exploring the semantic meanings of words. They have many uses in Natural Language Processing like finding the distance between words is used to infer the similarity between words. Therefore, words that are more similar should have embeddings that are closer to each other. *Apple* and *Pear* being closer then *Apple* and *Chair*. Although, a thing that word embeddings alone fail to do is take time into account is time. There is no way to determine whether a word used to be more sim

ilar to one word then it does now.

Creating Dynamic Word Embeddings can allow for a time-aware analysis of semantic meaning between words. With Dynamic Word Embeddings the evolving meanings between words can be discovered. Now, how words like “hip” have shifted from being the body part to also being associated with fashionable. In this paper we developed Dynamic Word Embeddings to try to infer the shifting meaning of words on social media by using data collected from twitter.

By using twitter data, we are able freely collect large amounts of data. We used *Tweepy*, an open source python library that works with the twitter API, to collect tweets. Titter is unique in a many way, of those length, hashtags, and slang are the most prominent. Retweets are unique to twitter, and our data set did initially contain many duplicates from multiple users that retweeted another’s tweet. While data collected from twitter has unique qualities, it allowed for us to easily and quickly collect data nicely formatted(.json) and easy to use.

# Methodology

The data initially collected contained much unnecessary information pertaining to our experiment, so cleaning the data was the first step in our methodology. All 313GB of our data was collected via a twitter query around the name “Trump”. By collecting our data around Trump we were able to collect a large amount of data around a word that is constantly being connected to different topics/words that are changing on an hourly basis. This allowed us to combat one of our worries about collecting data through twitter. The worry being that we would not have enough semantic change over the few months we collected data to create good Dynamic Word Embeddings.

Parsing the data involved removing the user name, tweet\_id, coordinates, retweet count, favorite count, etc. While parsing we decided it would also be the best time to divide tweets into even time segments thus allowing us to only have to save the tweet and freeing us from keeping more than the minimum information needed. We also removed common stop-words and the “RT” from in front of many tweets. We chose to just remove the “RT” and keep the rest of the tweet despite there being multiple tweets that ended up being the same.

Once the data was separated by time, it was then a matter of creating the Dynamic Word Embeddings.

While cleaning, we also organized all the tweets into separate time segments. Each time segment became a separate PPMI matrix and was then aligned for comparison and analysis.

# Analysis

- mention all the software I used here.

# FUTURE

Dynamic Word Embeddings for biosurveillance.

##### Acknowledgment

##### References

1. Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiiong, *“Dynamic Word Embeddings for Evolving Semantic Discovery”, 13 Feb 2018,* [arXiv:1703.00607](https://arxiv.org/abs/1703.00607)