Assignment Report on

"Tweet Generation based on Trending Topics using RNN"

Submitted in complete fulfilment of the requirements for ***Sem VII***

Topics in Deep Learning

Bachelor of Technology
in
Computer Science & Engineering

Submitted by:
Aditya Venkatesh Marada – 01FB16ECS032
Priya S Nayak – 01FB16ECS277

Under the guidance of

Srikanth H R
Professor
PES University, Bengaluru

August - December 2019

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

Problem Statement

Twitter has always been the favorite app for any Machine Learning/Data Science enthusiast. With its easy-to-access API, data gathering is a cake-walk.

A trend on Twitter refers to a hashtag-driven topic that is immediately popular at a particular time. These tweets change day-by-day and, if you are particularly popular, might stay utmost for 2-3 months.

Trends like #womeninstem, #ootd, #tbt, #vegan, #traveltuesday are ever-green and have millions of tweets attached to them. Influencers would love to have tweets written for them in order to be *trending* or featured amongst other big names on the list of *trending*.

This project aims at generating a single tweet with the help of using tweets from an existing trend.

Approach Taken

Dataset

This project relied heavily on the Twitter API in order to get the data. The process of data gathering and cleaning has been shown in Fig. 1

The tweets were selected up until 3 months in order to extract maximum data without losing the integrity of the *trend*.

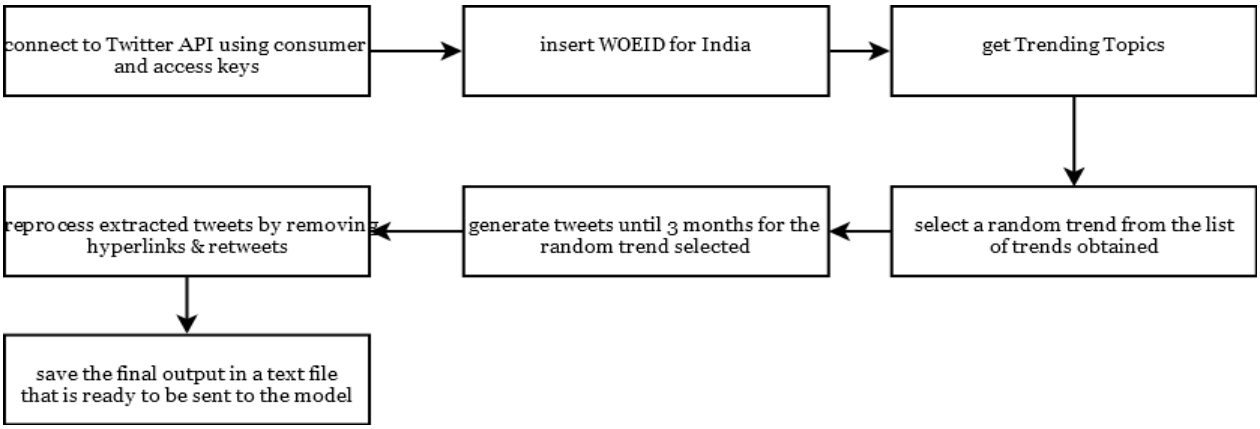

Fig. 1 - Data Generation Process

Approach 1 -

Originally, this project was going to focus on using Torch-RNN to generate the tweets. However, Torch-RNN is an existing library. Usage of this would mean no access to hard code or access to particularly complex code. The only tweaks that could be made would be to the input data and nothing else.

This approach was similar to using *WordPress* or *Drupal* to create your own website. The website is created but you have minimal control over the details.

Due to this major disadvantage, this approach was nipped in the bud.

Approach 2 -

Introduction

Usage of multi-layer recurrent neural networks seemed like a simple yet effective approach to take. This approach focuses on how Tensorflow RNN was used to achieve satisfactory results.

After the tweets are generated and preprocessed, they are ready to be used by the model. However, before that is done, there is a bit of RNN-Preprocessing that has to be done.

RNN Preprocessing & Code

The data is loaded and code is used to split words from the text. Each sentence is split into multiple words and "space" is used as a delimiter. In future cases, to modify the model, the delimiter can be changed.

To build the vocabulary, mapping is done based on previous sentences. This mapping is *from word to index* and requires the mapping of *index to word* and *word to index* of the previous sentence.

Python is helpful here as the mapping can be done using the *Collections* library with the *Counter* module.

A tensor is created with the vocabulary and the text. Since the tweets that are generated differ each time the code is run, the shape of the tensor, too, changes.

Batches are created and the tensor is modified to follow the real number of batches. The words and vocabulary generated are saved because they will be required during the generation phase.

Summaries of the model are saved for understanding in future iterations.

Training & Generation

For each epoch that is run, the learning rate is assigned and variables (like pointer for batches, state of the model, and variable to calculate speed) are reinitialized. The epoch is looped over all batches and, in some cases, a message is displayed on the console.

The model is saved for generation, later.

Generation uses the saved vocabulary file and saved model to restore the parameters.

Results are sampled from the model.

In order to obtain better results -

1. Number of words are restricted to 10. A tweet can only contain upto 280 characters. In order to match these specifications, roughly 10 words are chosen.
2. The starting word for the RNN is the name of the topic itself. This has proven to give better results then when the words were random ("Well", "today", "when", "so", etc.)

This approach was largely inspired from Andrej Karpathy's blog.

Architecture of the Model

Number of Layers	2
Model	LSTM (interchangeable with GRU/Vanilla RNN)
Batch Size	50
Sequence Length	25
Learning Rate	0.002
Decay Rate - for RMSProp	0.97

The batch size is used to calculate the number of batches that are required by using the tensor size.

Softmax is used and the loss function is calculated using the targets, batch size, vocabulary size, and sequence lengths. The cost is calculated by dividing the loss with the batch size and sequence length

Adam is used as an optimizer since it combines the best properties to provide a great optimization algorithm.

GPU memory of 66.6% was used and the code was run on Google Colab.

Results

Result 1 -

Trend Chosen - #IamWithDevendra

Number of Tweets - 1521

Sample Tweets -

- We are Missing You! #iamwithdevendra #PawarSoniaSena #chanakya
- #IamWithDevendra Oh! Gazab!!!! Sanjay ji seems really happy after the government formation! Via WhatsApp!
- My Kind of Supporter. I concur! #iamwithdevendra
- #IamWithDevendra Bad day for Maharashtra today seeing unholy alliance after experiencing the same in my Karnataka.

Generated Tweet -

Devendra Media Body Dynast.Lets considers stands Expect @Ashtalakshmi8, MLA "Lust Anmol

Result 2 -

Trend Chosen - #JINFESTA_D1

Number of Tweets - 1040

Sample Tweets -

- this will forever be iconic #JINFESTA_D1
- Ready for your birthday @bts_twt# #BTS # #JIN # # #JINFESTA_D1
- Jinnie!! The effort you put into showing your love to armys, you really disserve the whole world for it
- The biggest heart of the sweetest soul #JINFESTA_D1 #SEOKJIN #JIN

Generated Tweet -

Well confidence hardworking, even b"that ill youuuu all-time crazy joy. eventtt!

this is for #JINFESTA_21

Result 3 -

Trend Chosen - #republicsummit

Number of Tweets - 389

Sample Tweets -

- Union Home Minister and BJP National President Shri @AmitShah will address #RepublicSummit on 27 November 2019 at 1
- @narendramodi ji heard you at #RepublicSummit You are so balanced Pranam
- @narendramodi Very happy to see you live on #RepublicSummit Sir @PMOIndia and Our Arnab Goswami Sir is too good and
- Delighted to be at the #RepublicSummit. Watch my speech.

Generated Tweet -

Republic such LOL#PMModi you build 99% how he Naxals it? your