Developing Twitter Bot with Language Modeling

ET61002: Language Processing for e-Learning

January 23, 2018

1 Task Description

The objective of this assignment is to test out the language modeling techniques that we have discussed in the class. The primary task is to design a bot that can generate interesting tweets given an entity or topic. The bot will consult a language model for generating the tweets.

2 Steps

You should follow the steps mentioned below:

1. *Corpus Compilation:* The twitter data is in JSON format. So, a corpora has to be created out of the JSON files by concatenating the individual tweets. Remove the URLs if they are present in tweets (you might need to use regular expressions to identify URLs).

2. *Build Language Models:* Use the code snippets used during the tutorial to build n-gram language models.

3. *Experiment:* Use different smoothing techniques, Laplace, Good-Turing and Kneser-Ney and report perplexity values using held-out test. (**Assignment 1**)

4. *Generate tweets:* Use generated n-gram language model to generate tweets for a given entity (e.g. virat Kohli) or a topic (e.g., education)[1]. The length of the generated tweets are restricted to a number (say 10 words). Use the following methods to write your algorithm:

   (a) *Random Generator:* Randomly pick-up the n-grams that contains the topic or entity. Use that as pivot and add prefix and suffix to that (**Assignment 2**)

   (b) *POS-Informed Generator:* In this scheme, the tweets will be generated guided by POS template. Define a set of POS templates and generate random tweets and accept them based on the pre-decided POS templates. (**Assignment 3**).

[1] Searching for n-grams: You might need to work on searching for the n-grams that contain the given entity. For this you may use dictionary data structure in python. A better implementation could be to use text indexing engine like Solr. We shall provide tutorial on that

(c) *Collocation-based Generator:* Here instead of generating tweets for one entity, the generated tweets will involve two entities (e.g., Virat and Anushka) (**Assignment 4**)

3 Evaluation

The evaluation will be done based on two judgments:

- Evaluation of algorithms and implementation (myself and TA)
- Peer Judgment: Top N twits will be voted by all the students in the class using Mentemeter (`https://www.mentimeter.com/`).

4 Dataset

Download twitter dataset from `http://www.cet.iitkgp.ac.in/json/json.gold.tar.gz` . The link will be valid for 1 week. So, download the data as soon as possible.

5 Deadlines and Submission Details

The deadlines for submissions are:

1. Assignment 1: 02.02.2018 [Code and report of the experiments, frequent non-functional unigrams]
2. Assignment 2: 09.02.2018 [Algorithm, code and top 5 tweets]
3. Assignment 3: 16.02.2018 [POS templates code and top 5 tweets]
4. Assignment 4: 09.03.2018 [Algorithm, code and top 5 tweets]