DS8004 Project Presentation 1: Twitter topic classification

Anil Trikha, Pengshuai Shi

Ryerson University

February 7, 2017

Overview

Introduction and problem

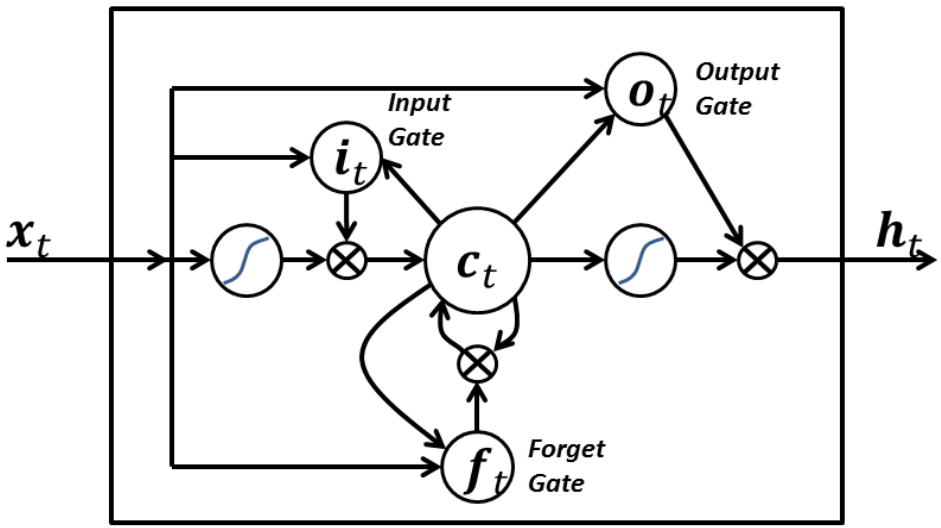
- People use the hashtag symbol (#) before a relevant keyword or phrase in their Tweet to categorize those Tweets and help them show more easily in Twitter search.
- Hashtagged words that become very popular are often Trending Topics.[1]

Example of a Tweet with a hashtag:

Only 14.6% of tweets contain hashtags, a reliable hashtag classification system could help researchers to label the rest of tweets. And existing machine learning models are mostly conventional and reaches a bottleneck because they are using TF-IDF which ignores the semantic information.[2]

Proposed solution

Recurrent neural networks (RNNs) have recently achieved promising results in many ML tasks, and inspired by the recent improvement of document level sentiment classification, a LSTM-RNN model is proposed to learn semantic tweet representations. [2]



[3] *Wiki*

Steps

1. Obtain tweets using Twitter REST/Search API
2. Data preprocessing, e.g. tokenization, word representation
3. Utilize CNN to compose sentence representations
4. Use LSTM to encode the intrinsic relations between words
5. Compare result with conventional methods such as SVM

Schedule

1. Researching on how LSTM is implemented.
2. Choose which tool to use.
3. Try to reproduce some basic result in the paper [2].

Expected outcomes

1. Be able to produce topic label when we feed in a tweet.
2. Obtain similar result as stated in the paper.
3. Produce the most trending topic during the time of the project.

References

- Twitter hashtag
  `https://support.twitter.com/articles/49309`
- Jia Li, Hua Xu, Xingwei He, Junhui Deng and Xiaomin Sun, *Tweet modeling with LSTM recurrent neural networks for hashtag recommendation*, 2016
- Long short-term memory
  https://en.wikipedia.org/wiki/Long_short-term_memory