Report

Zeljko Kraljevic

June 28, 2016

1 General

$$C = \sum_{t} \sum_{a} \sum_{b \in C_a^{(t)}} \left(\underbrace{log\sigma \left(x_b^{(t)} \cdot y_a^{(t)} \right)}_{u_p} + \sum_{c} \underbrace{log\sigma \left(-x_b^{(t)} \cdot y_c^{(t)} \right)}_{u_n} \right) \tag{1}$$

Setup when training:

- I normalize the time of the whole dataset to be between 0-10 for all datasets
- We don't have alterations currently, everything is trained all the time

2 Subsampling

I still subsample frequent words using $P(w_i) = 1 - \sqrt{\frac{1}{f(w_i)}}$. I also subsample documents in the following way:

- From a training set I always take a fixed number N of (a, b) pairs
- From every document I take a fixed number M of (a, b) pairs limiting the number of paris having the same target with K.
- When choosing a pair from document the closer the words in it are the higher the chance of it being choosen.
- Depending on N, M I calculate the probability of taking a document so that the whole dataset is always equaly present in the subsampled training set.

3 Clustering

The basic formula used for clustering is:

$$p(c \mid d) = \frac{p(d \mid c)p(c)}{p(d)} \tag{2}$$

Given that $p(c)$ and $p(d)$ are the same for every cluster and document, we get:

$$p(c \mid d) \propto p(d \mid c) \tag{3}$$

Now we have the probability of a word given a cluster and we define the probability of a document given a cluster to be the product of all words in the document:

$$p(d \mid c) = \prod_{w_i \in d} p(c \mid w_i) f_c(t_i) \tag{4}$$

Which means:

$$p(c \mid d) = \prod_{w_i \in d} p(c \mid w_i) f_c(t_i) \tag{5}$$

We take a *log* of the probability for convenience:

$$log(p(c \mid d)) = \sum_{w_i \in d} log(p(c \mid w_i) f_c(t_i)) \tag{6}$$

Once this is calculated the document is clustered with:

$$doc\_cluster = \arg\max_{c \in C} log(p(c \mid d)) \tag{7}$$

4 Time Prediction

Time prediction is similar to clustering except we don't use the time limiting function:

$$p(t \mid d) = \prod_{w_i \in d} p(c \mid w_i) \tag{8}$$

We define the log of this probability as: $l_c = log(p(t \mid d))$ Now we can predict the time as:

$$predicted\_time = \frac{\sum_{c \in C} e^{l_c} t_c}{\sum_c e^{l_c}} \tag{9}$$

We multiply this by $\frac{e^{-z}}{e^{-z}}$ where $z = max(l_c)$ and get WHY (I know why, but how to explain it normally):

$$predicted\_time = \frac{\sum_{c \in C} e^{l_c - z} t_c}{\sum_c e^{l_c - z}} \tag{10}$$

5 Finished Tests

Notes:

- Cap
- tau=0
- regularization

Results are in /develop/results/

Notes	Dataset	Iterations	clusters	Tau	Name
Without reg or cap	NIPS	500	300	1	normal
Without reg or cap	NIPS	500	300	0	normal_tau
Without reg, normalization	NIPS	500	300	1	normalization

6 Running Tests

Tests that are currently running, approximately it takes one day for a test to finish.

Notes	Dataset	Iterations	clusters	Tau	Folder
tau=0.01	NIPS	500	300	0	normal_tau_small
tau=0.01	Tweets	500	500	0	tweets_tau_small

7 TODO

- Try using aleterations, not so easy to implement