**Introduction**

In our experiment, we used Tweets of selected accounts and author topic model to find topic (word distribution). Topic trends (time series) are identified for twitter text across time. These topic trends are then used as input features for neural network, LSTM to predict Bitcoin price.

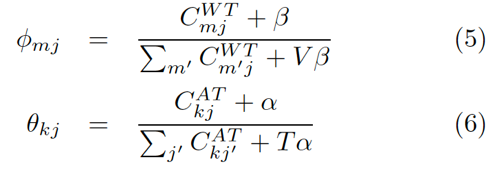
**Author Topic Model**

Topic in this experiment is defined as a distribution over words. At the same time, author’s interest is defined as distribution over topics. The generative mechanism is described below.

First, an author x is randomly chosen from author list . Secondly, x is substituted into topic by author distribution which used Dirichlet Distribution as its prior. Sampling from this distribution, we obtain topic z. After obtaining it, topic z is substituted into word by topic distribution which used Dirichlet Distribution as its prior.

In order to estimate and . This experiment’s reference paper used Gibbs sampling. The first step would be assigning words to random topic and authors. Next, apply equation 4 to every word in the text and make new assignment accordingly. Then, repeat this step 1000 times and save and at the end according to (5) and (6).

A close up of a clock

Description automatically generated

**Implementation**

In this experiment, Number of documents D = 50811, Number of author A = 9. Number of topics is chosen to be 100 and 1000 iterations before and is saved. Size of is 1 as each tweet only has 1 author. Number of unique words is denoted by V=65160. For (4)(5)(6), is set to be 0.01 while is set to be 0.25.

Therefore, size of is V\*T and that of is A\*T; C++ is chosen to implement this model because large scale computation is expected; if V becomes larger would become a sparse matrix and in order to loop through all document each iterations, there is 50811\*1000 operation using equation (4) which involves counting; and then there is resampling for every word.

After 20 hours of running time of MacBook Pro with 2.3Ghz CPU, result of and is obtained. This process is repeated twice once for cleaned Text keeping emoji and once for cleaned text removing emoji. The purpose is to test which contains more information for Bitcoin price prediction.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Topic | Word | Prob | Author | Topic | Prob |
| 0 | bitcoin | 0.0117378 | anondran | 0 | 0.103402 |
| 0 | just | 0.00946777 | anondran | 45 | 0.0278108 |
| 0 | will | 0.00728084 | anondran | 90 | 0.0211368 |
| 0 | lol | 0.00725315 | anondran | 17 | 0.0156039 |