Explanation_Train_Test

Sumit Sidana

Wednesday 12th July, 2017

Friday 26th February, 2016

1 Explanation: Train and Test

Notations:

$$z = ailment \tag{1}$$

$$t = time(month, week, ...etc.) \tag{2}$$

Perplexity of topic model depends on its ability to predict the *probability of future words.*
Probability of words depend on *probability of topic* with the following formula.

$$P(w) = \sum_z P(w|z)P(z) = \sum_z \underbrace{\frac{n(z,w)}{n(z)}}_{\text{Constant, w} \in \text{second month, n(z,w)} \in \text{train}} \times \underbrace{P(z)}_{\text{Varies with topic model}} \tag{3}$$

So focussing on the only thing in equation 3 which varies: $P(z)$

$$P(z|t) = \sum_{tweet} \sum_{p \in t} P(z|p) \tag{4}$$

$$P(z|p) = \sum_{word} \sum_{w \in p} P(z|w)P(w|p) = \sum_w \frac{n(z,w)}{n(w)} P(w|p) \tag{5}$$

P(z) needs to be calculated on the 1st month and:

- **atam**: Underlying assumption of atam is that topics stay static with respect to time. That is why $P(z)$ of 1st month needs to be used for $P(z)$ of the second month. Because P(z) stays static with time. This is what tm-lda did for lda
- **tmatam**: $P(z)$ of second month needs to be predicted using the $P(z)$ of first month using the transition matrix learnt during training period

- **tatam**: $P(z)$ of second month needs to be computed directly on second month because model itself learnt ailments using the knowledge of time in-built in the model. Can ailments inferred using a time-aware model actual representative of words tweeted about in the time of interest or it just learns noise?