Information Retrieval – Project Presentation
Group 20 - Project 3
Topic - Tag Recommendation with Topical Attention based LSTM

Team Mentor : Suman Kalyan Maity

Team Members:

Rishabh Miglani	13MA20054
Siddhartha Tekriwal	13MA20041
Harsh Khetan	13MA20049
Abhinav Agarwalla	13MA20003
Prithvi Chandak	13MI31020

OVERVIEW

- Study of attention-based LSTM model using topic modelling for tag recommendation.

- Flow structure adopted from the research paper “Hashtag Recommendation with Topical Attention-Based LSTM” by Y Li et.al. in COLING 2016.

- Dataset obtained from Stanford Network Analysis Project (SNAP).

OBJECTIVES

- Investigate a novel topical attention-based LSTM model for the task of hashtag recommendation.
- Goal is to achieve a better accuracy than normal LSTM model using attention.
- Incorporate topic modeling into the LSTM architecture through an attention mechanism and take advantages of both.
- End-to-End LSTM model avoiding hand-crafted features.

DATASET DESCRIPTION

Number of tweets	**4,30,000**
Number of tweets for training	**3,00,000**
Number of tweets for validation	**50,000**
Number of tweets for test	**80,000**

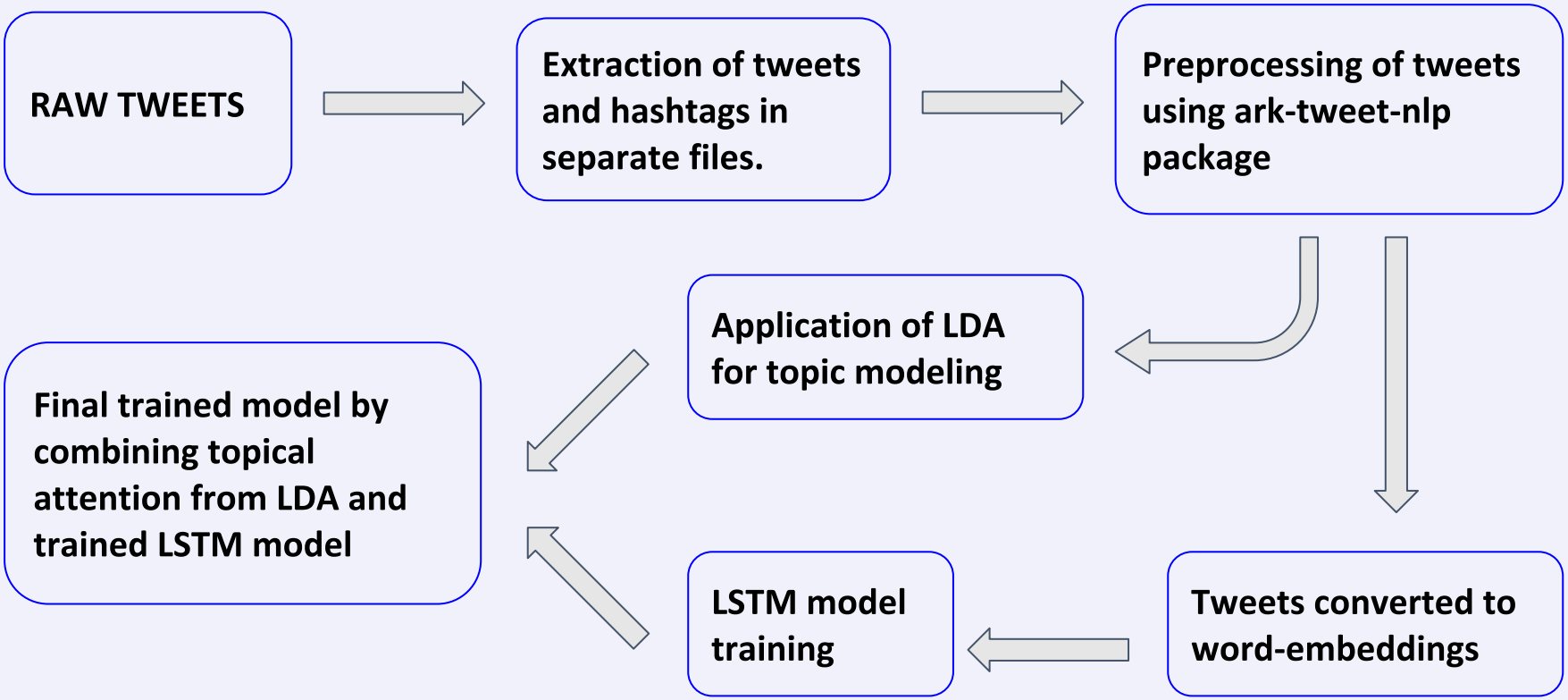
For each public tweet the following information is available:

- **Time (T)** - date and time of tweet
- **Author (U)** - twitter profile link of user
- **Content (W)** - tweet content

Sample tweet :

T 2009-06-30 23:59:48
U http://twitter.com/navarrojuan25
W I am watching Canal 44 live online TV @ http://wwitv.com/tv_channels/b32

APPROACH


RAW TWEETS
Extraction of tweets and hashtags in separate files.
Preprocessing of tweets using ark-tweet-nlp package
Application of LDA for topic modeling
Final trained model by combining topical attention from LDA and trained LSTM model
LSTM model training
Tweets converted to word-embeddings

SELECTION OF TWEETS

- The tweets containing hashtags were only considered.
- The retweets were ignored.
- English tweets were only chosen.
- The hashtags with length less than 3 were ignored.
- Only those hashtags whose frequency is more than 100 are considered.

PRE-PROCESSING

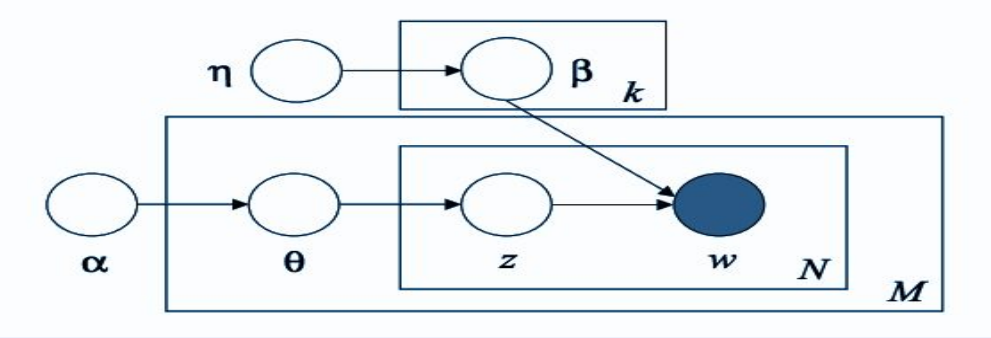
- Parts of Speech tagging was done for the tweets using the ark-tweet-nlp package.
- The words tagged as common noun(N), proper noun(^), nominal + possessive(S), proper noun + possessive(Z), verb including copula, auxiliaries(V), adjective(A) and adverb(R) were kept. Rest were discarded.
- The data was then divided into training, validation and test sets.

TOPIC MODELLING

- LDA (Latent Dirichlet Allocation) algorithm is used for topic modelling.
- It is a generative probabilistic model for collection of discrete dataset such as text corpora.
- The text corpora here is the train dataset and the number of topics are taken to be “100”.
- The sklearn library of python is used for obtaining a matrix of token counts and LDA.

LDA

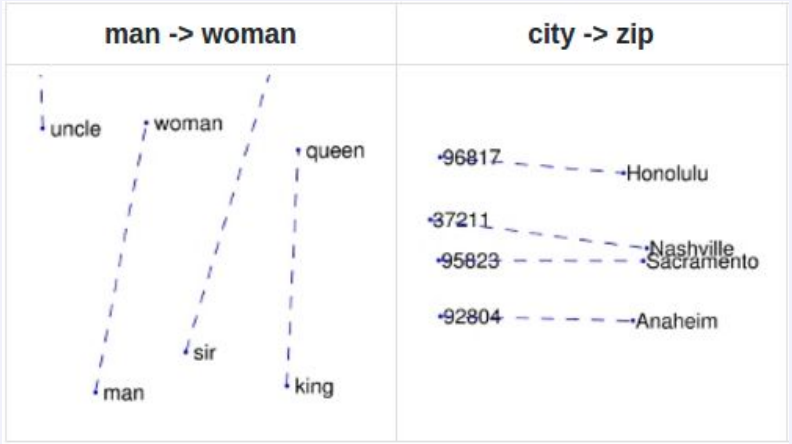
- Where, for each topic k , β_k ~ Dirichlet(η) , k = 1...K
- For each document d, draw θ_k ~ Dirichlet(α) ,d = 1….D
- For each word i in document d:
  - Draw a topic index z_{di} ~ Multinomial (θ_d)
  - Draw the observed word w_{ij} ~ Multinomial (β_{zdi})
  - For parameter estimation , the posterior distribution is:
  - p (z,θ,β|w,α,η) = p (z,θ,β|α,η) / p(w|α, η)



WORD EMBEDDINGS

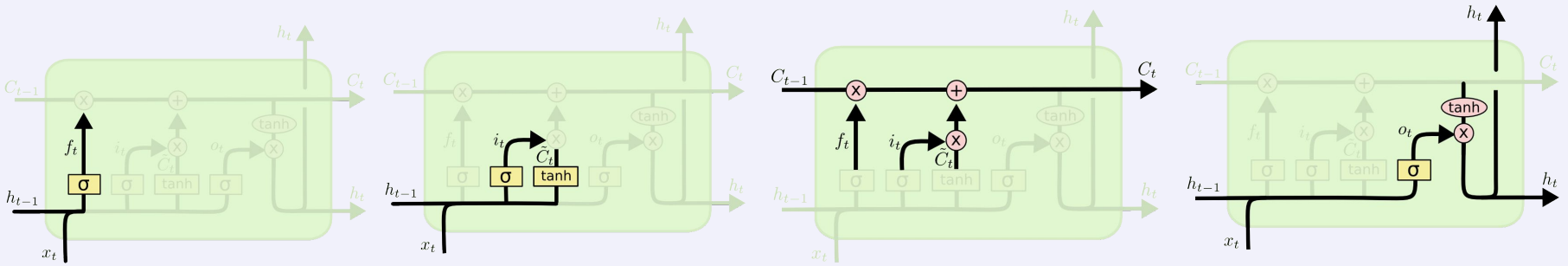
- Each word is represented as a continuous and real-valued vector, for which a representation is learned.
- Quantify similarities between terms.

- Words → GLoVE word embeddings
- Used as input to the LSTM model.



LSTM

- Long short term Memory networks are RNNs capable of learning long term dependencies.
- Overcomes vanishing gradient problem in RNNs
- Input , output and forget gates are used to control the passing of information (cell state) along the sequence.



Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

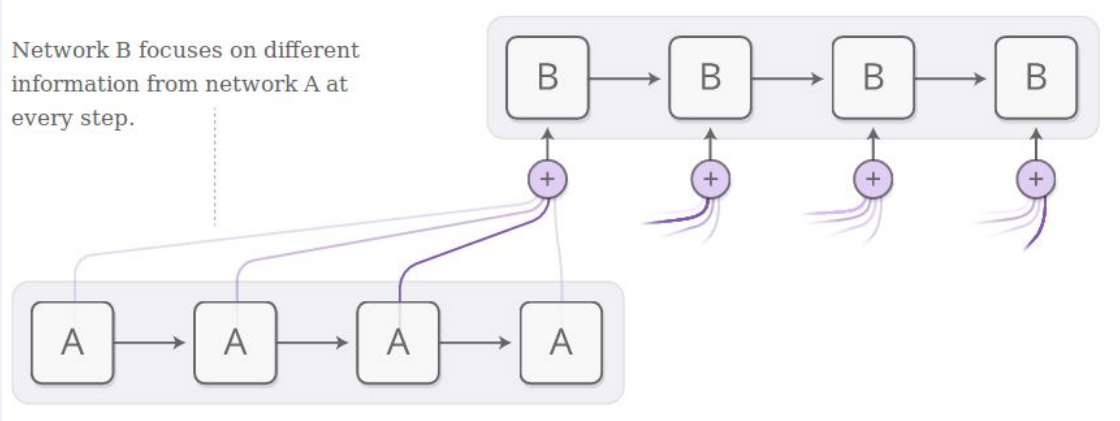
LSTM - Variants

- **Naive LSTM:**
  - Use hidden state of the last LSTM unit
  - Feed it into final output layer

- **Attention LSTM:**
  - Use all the hidden states of the unrolled LSTM
  - Identify weights for each hidden state i.e. **attend** to specific words in the tweet
  - Feed it into final output layer

Attention LSTM

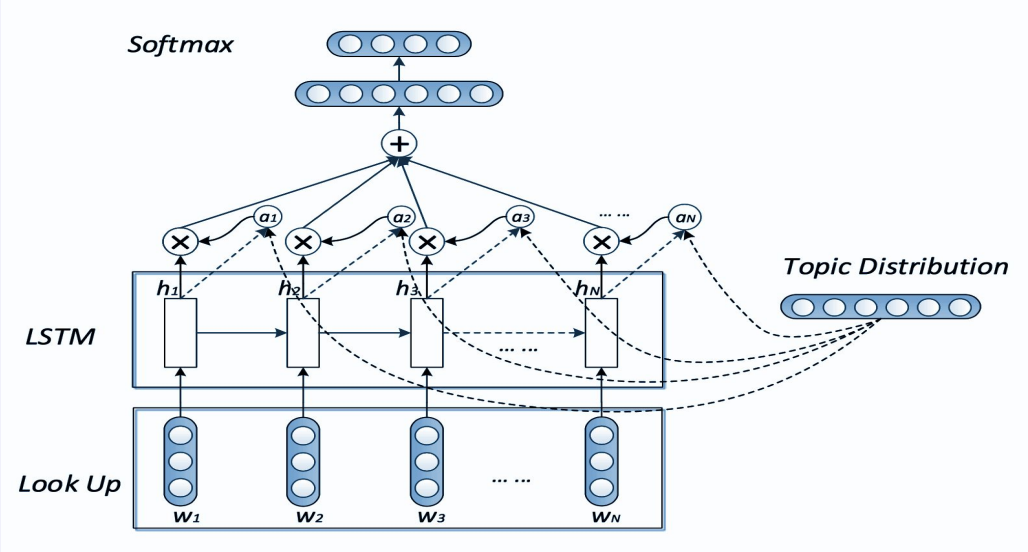
- For producing output at B, a weighted average from all hidden states [h_1, h_2, ..., h_N] from A.
- Weights(a_j) determined using SoftMax operation.
- a_j = softmax(tanh(W * h_j))
- Output: vec = $\sum_{j+1}^{N} a_j h_j$



Source: https://distill.pub/2016/augmented-rnns/

Topical Attention LSTM

- Hidden states from LSTM $[h_1, h_2, ..., h_N]$.
- The external topic layer from LDA is $\theta_s \in R^{K \times 1}$
- $a_j = \text{softmax}(\tanh(U * \theta_s + W * h_j))$
- Output: vec = $\sum_{j+1}^{N} a_j h_j$



RESULTS

Methods	Naive LSTM	Attention LSTM	Topical Attention LSTM
Log-Loss	3.41	3.23	**3.088**
Accuracy	32.5	41.87	**42.91**
Precision	0.3806	0.4751	**0.4864**
Recall	0.33	0.4123	**0.4221**
F-score	0.3536	0.4414	**0.4519**

CONCLUSION

- The increase in performance due to only attention is very substantial. (~ 8% increase accuracy)
- Topical Attention LSTM:
  - performed the best on all metrics
  - was easier to train;
  - yielded results in less number of epochs, even with increased parameters

Number of Parameters

Naive LSTM	Attention LSTM	Topical Attention LSTM
2,13,526	2,13,656	2,43,726

FAILED EXPERIMENTS

- **Spell-Checking and Lemmatization for preprocessing tweets:** Due to this, most of the Proper Nouns were distorted , which were observed later as important features for our model.
- **Initial sampling of data with all processed tweets:** Only 1% of the total hashtags from initial processed data were frequent(>=100 citations) and contributed to around 4.5 lacs tweets from our original 5 lacs records.
- **Hardware Limitations:** Total initial data collected was around 13GB. Processing limitations restricted us to work with only 2GB of data.

FUTURE SCOPES

- **User characteristics**: Taking into account the user profile and other information.
- **Time Variation:**
  - Data not representative of the general tweets appearing over time.
  - Continuous updation necessary.
- **Extension:** to other tag recommendation systems such as stackoverflow , instagram(multi-modal).
- **Hierarchical Approach:** Use of Stacked LSTMs for modelling both tweets and topics using LSTM
- **Twitter-LDA:** for topic modelling in tweets

REFERENCES

- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global Vectors for Word Representation." *EMNLP*. Vol. 14. 2014.
- Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3.Jan (2003): 993-1022.
- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).
- Yoshua Bengio, R´ejean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. journal of machine learning research, 3(Feb):1137–1155.