Topic mining for short-text documents on micro-blogging site by combining Doc2Vec and clustering techniques

Thuong-Hai Pham

Faculty of Information and Communication Technology
University of Malta
Msida MSD 2080, Malta
Email: thuong-hai.pham.16@um.edu.mt

ICT5901 - Research Methods, 2017

Outline

Introduction

- Topic model for micro-blogging site
- Latent Dirichlet Allocation (LDA)
  - state-of-the-art in unsupervised topic model
  - not trivial task applying directly on short text
- LDA Variants
  - Author-Topic model
  - Twitter-LDA
- Features learning (doc2vec) & clustering

Outline

Mathematical background

Exchangeability

$(x_1, x_2...)$ is an infinitely exchangeable sequence of random variables if for any permutation π,

$$p(x_1, ..., x_n) = p(x_\pi(1), ..., x_\pi(n))$$

Theorem

De Finetti, 1935 A sequence of random variables $(x_1, x_2, ...)$ *is infinitely exchangeable iff, for all n,*

$$p(x_1, x_2, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta) P(d\theta)$$

Outline

Latent Dirichlet Allocation (LDA)

- Generative probabilistic model
- Documents are represented as random mixtures over latent topics
- Bag-of-words (BoW) assumption

LDA generative process

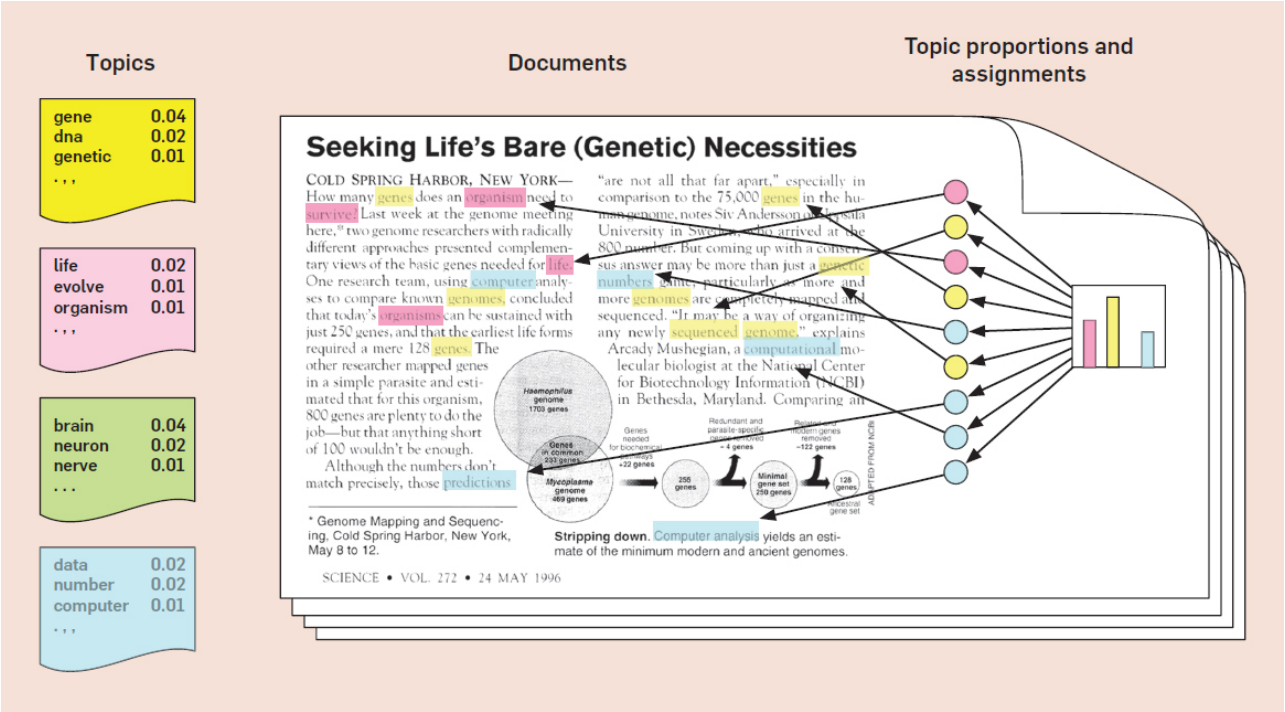

Figure: LDA generative process[1]

[1] http://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-

LDA graphical model

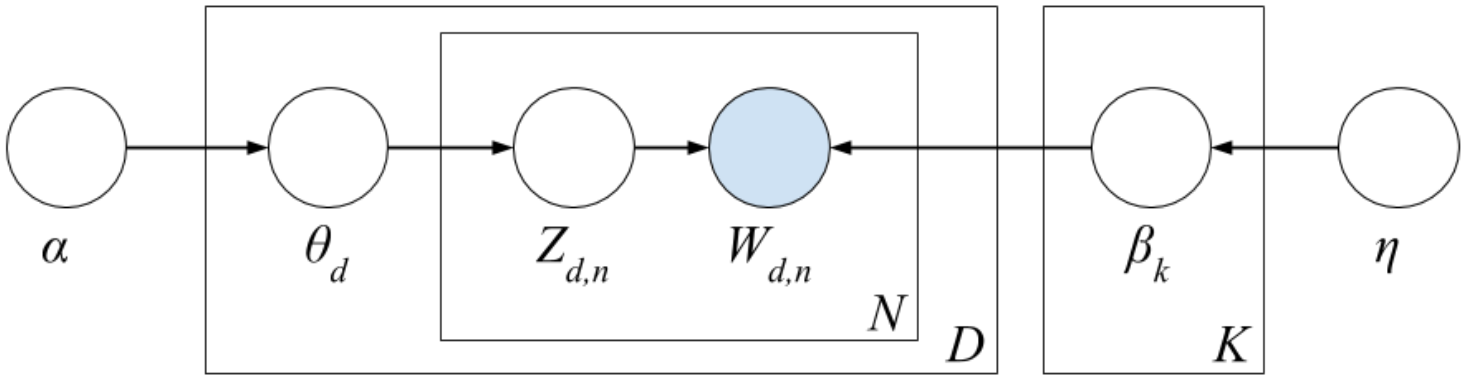

Figure: LDA graphical model

Outline

Author-topic model

- Excluding the topic proportions for each tweets
- Aggregating all tweets of a Twitter's user into a single document (Weng et al. 2010; Hong and Davison 2010)
- Efficient on a specific task (e.g. topic-sensitive influencers mining (Weng et al. 2010))

Twitter-LDA

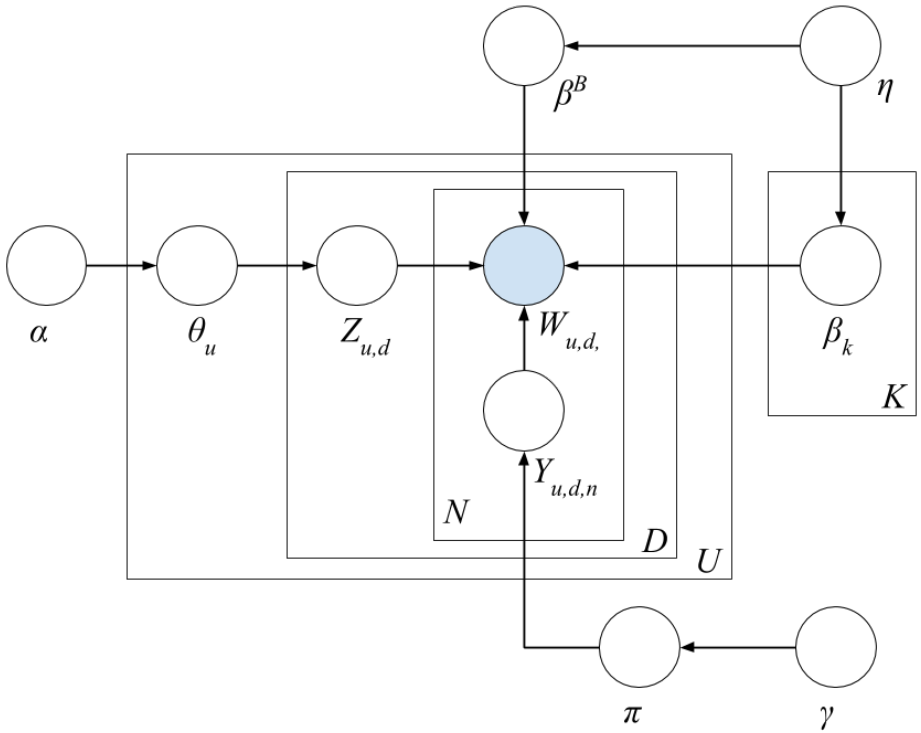

Figure: Twitter-LDA graphical model

Outline

Evaluation

- Intrinsic: view the problem as document modelling (Blei et al. 2003) by measuring perplexity of held-out set C'

$$perp(C') = exp\left\{-\frac{\sum_{d=1}^{D} log(p(W_d))}{\sum_{d=1}^{D} N_d}\right\}$$

- Extrinsic: measure LDA performance on some secondary tasks, such as corpus comparison or topic-sensitive influencers mining (Weng et al. 2010)
- Human evaluation: human judges assign a score on each topic, ranging from 1 (meaningful) to 0 (nonsense)(Zhao et al. 2011)

Outline

Doc2vec as features learning (Le and Mikolov 2014)

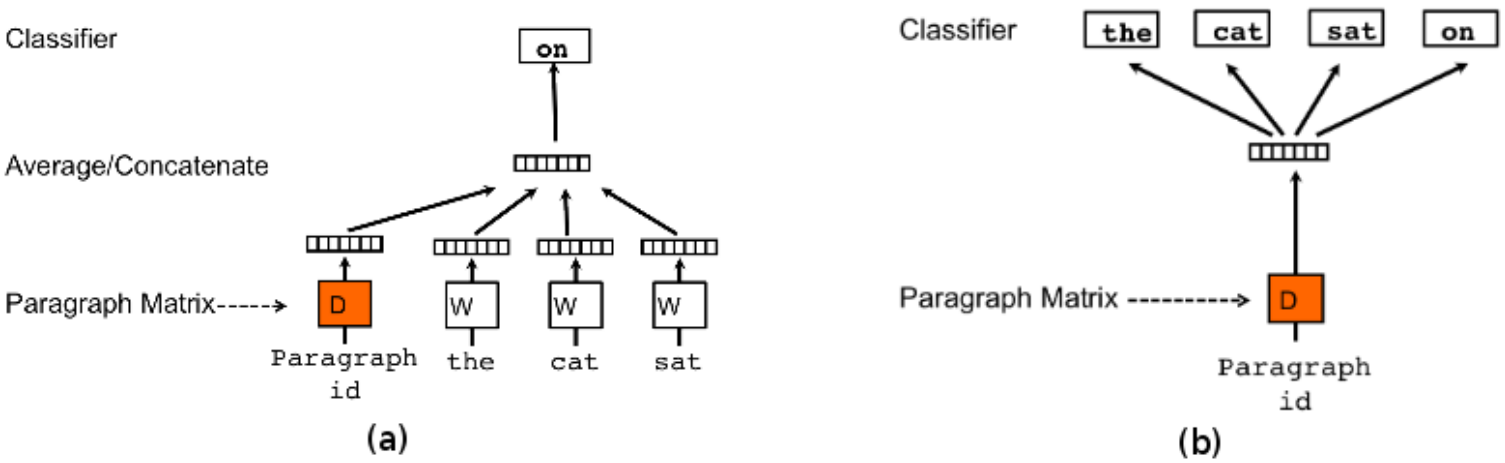

Figure: Framework for learning paragraph vector, (a) Distributed Memory, (b) Distributed Bag of Words

Clustering

- K-means: traditional clustering
- DBSCAN: no K specified
- C-means: fuzzy/soft clustering

Evaluation

- Perplexity does not measure how meaningful the topics discovered are
- Human evaluation strategy (Zhao et al. 2011)
  - Two distinct judges
  - Score from 0 to 10 for meaningfulness
  - Score for mutual agreement on topic assignments

Outline

Timeline

WBS	Task description	Start date	Finish date	Progress
1	**Data acquisition**	**1/30/2017**	**2/1/2017**	**75%**
1.1	Extract data from Twitter archive	1/30/2017	2/1/2017	100%
1.2	Cleaning tweets data	2/1/2017	2/1/2017	0%
2	**Implementation**	**2/2/2017**	**2/9/2017**	**0%**
2.1	Develop code for LDA	2/2/2017	2/4/2017	0%
2.2	Develop code for author-topic model	2/6/2017	2/6/2017	0%
2.3	Develop code for clustering: doc2vec & K-means	2/7/2017	2/9/2017	0%
3	**Experiments**	**2/13/2017**	**2/21/2017**	**0%**
3.1	LDA	2/13/2017	2/21/2017	0%
3.2	LDA with author-topic	2/16/2017	2/21/2017	0%
3.3	Clustering	2/20/2017	2/21/2017	0%
3.4	Clustering with author-topic	2/20/2017	2/21/2017	0%
4	**Evaluation**	**2/22/2017**	**2/23/2017**	**0%**
4.1	1st judge topic assignment	2/22/2017	2/22/2017	0%
4.2	2nd judge topic assignment	2/22/2017	2/22/2017	0%
4.3	Mutual agreement scores	2/23/2017	2/23/2017	0%
5	Final Report	2/24/2017	2/25/2017	0%

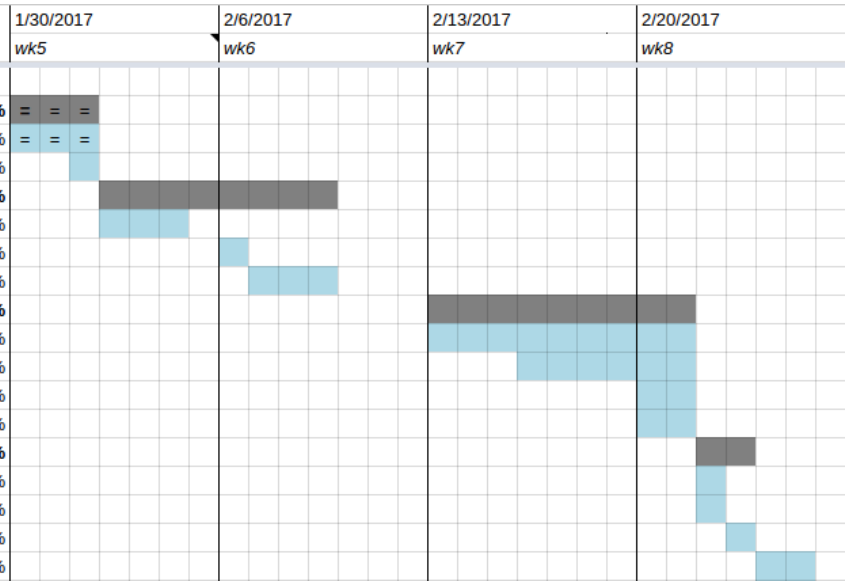

Figure: Project timeline

Resources

- Data: Twitter stream archive[2] in July 2016
- Compute engine: Google cloud n1-standard-4[3]
- Code library: LDA and Doc2Vec model in Gensim[4] library
- Evaluation: 02 human judges

[2] https://archive.org/details/archiveteam-twitter-stream-2016-07
[3] operates with 4 virtual CPUs, 15GB RAM, 200GB hard disk drive
[4] https://radimrehurek.com/gensim/

Expected result

Meaningfulness

Proposed method of combining Doc2Vec and K-means achieves approximate meaningfulness score in compare to traditional LDA method

Advantages

- Eliminate assumptions in LDA variants
- Capability of running on short-text documents (social networks)

Difficulties

- Noise in unconventional language usage on Twitter
- Twitter stream archive is too large to consume at once
- Meaning of topics discovered can be unclear or hardly interpretable

References I

- Blei, David M et al. (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022. ISSN: 15324435. DOI: 10.1162/jmlr.2003.3.4-5.993. arXiv: 1111.6189v1.
- Hong, Liangjie and Brian D. Davison (2010). "Empirical study of topic modeling in twitter". In: *Proceedings of the first workshop on social media analytics*. ACM. New York, New York, USA: ACM Press, pp. 80–88. ISBN: 9781450302173. DOI: 10.1145/1964858.1964870.
- Le, Quoc V and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents." In: *ICML*. Vol. 14, pp. 1188–1196.

References II


Weng, Jianshu et al. (2010). "TwitterRank". In: *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, p. 261. ISBN: 9781605588896. DOI: 10.1145/1718487.1718520.


Zhao, Wayne Xin et al. (2011). "Comparing twitter and traditional media using topic models". In: *European Conference on Information Retrieval*. Springer. Springer Berlin Heidelberg, pp. 338–349. DOI: 10.1007/978-3-642-20161-5_34.