Topic Modeling for Twitter Accounts using Bayesian Nonnegative Matrix Factorization

Burak Suyunu Şemsi Yiğit Özgümüş

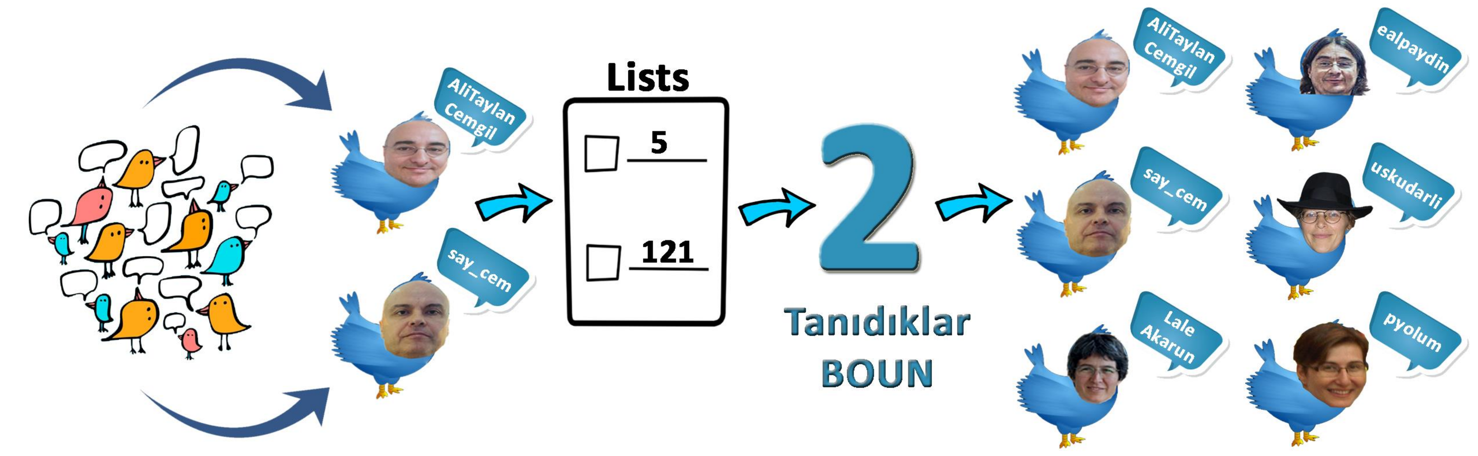
Advisor: Assoc. Prof. Ali Taylan Cemgil

GitHub

OUR MOTIVATION

Makers, scientists, influencers and many other people share their ideas, products and innovations via the most intellectual social network **Twitter**. It is hard to find the information about a **topic** in the giant network of Twitter. Our aim is to find users who are tweeting about the same **topic**. With this aim we want to bring people interested in the same **community** together. There are potential methods like LDA and NMF to tackle this problem, we want to investigate the addition of klb-nmf and to see whether this method is an applicable solution candidate for this problem.

DATASET - SIMILAR-TWITTER

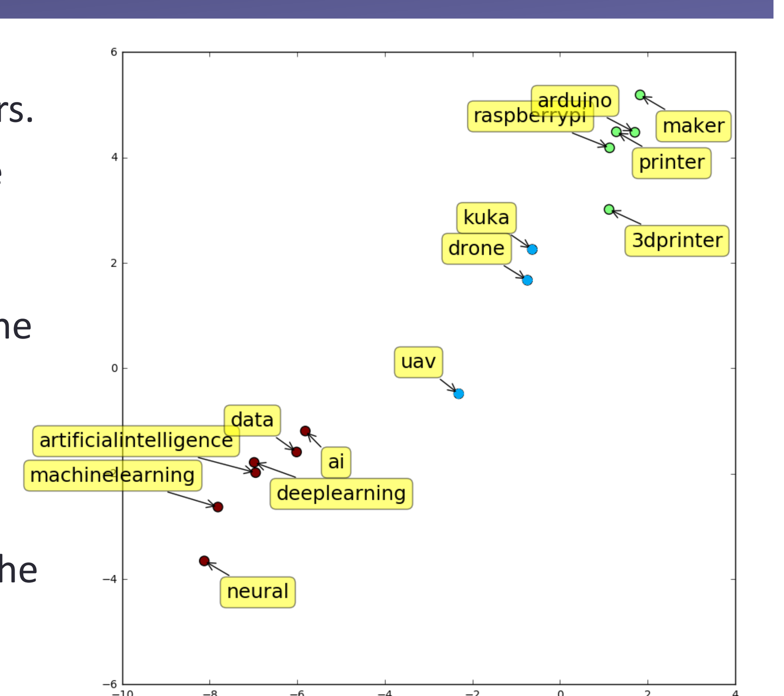


MAINTAINING TWEETS – NLP

- Remove URLs
- Tokenization
- Stop Words
- Remove non-English accounts
- Stemming
- Remove words that appears at most 10 times in the whole corpus

CLUSTERING WORDS - WORD2VEC

- **Word2Vec** uses word embedding to map words to a **vector** of real numbers.
- We applied **k-means clustering** to the vectors to see the relevant words together.
- We chose the word at the **center** of the cluster to represent the other words from the same cluster in the word corpus.
- We **normalized** the number of occurrences in the corpus to handle the problem of less frequent words being more important.



TOPIC MODELING

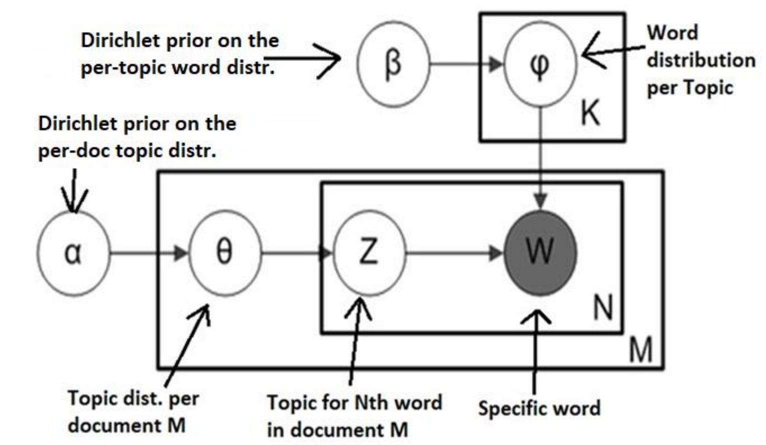
In **machine learning** and **natural language processing**, a topic model is a type of statistical model for discovering the topics that occur in a collection of documents. So we are trying to learn **topic distribution over the vocabulary** or **word distributions of the topics**.

- I like to eat broccoli and bananas.
- Hamsters and kittens are cute.
- Look at this cute hamster munching on a piece of broccoli.

- **Sentences 1** : 100% Topic A
- **Sentences 2**: 100% Topic B
- **Sentence 3**: 60% Topic A, 40% Topic B
- **Topic A**: 30% broccoli, 15% bananas, 10% eat, 10% munching, … **(Food)**
- **Topic B**: 20% kittens, 20% cute, 15% hamster, … **(cute animals)**

LDA (LATENT DIRICHLET ALLOC)

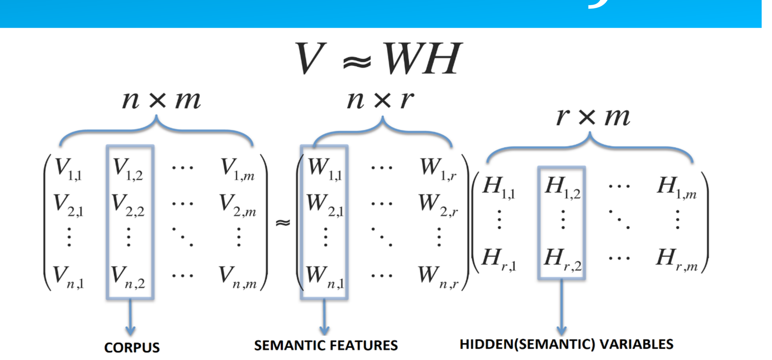
- Assign each word in a document to one of **K topics randomly**
- To obtain a correct distribution, iterate over each document D and for each document iterate over each word W.
- Then, for each topic T reassign the word W to a new topic T':



$$P(Word\ W \mid Topic\ T) * P(Topic\ T \mid Document\ D)$$

NMF (NON-NEGATIVE MATRIX FACT)

- NMF decomposes the data into two **low rank matrices (W, H)** whose product constitutes the data matrix.
- At each iteration, update W and H with additive update rules to minimize the **squared error** to reach a good decomposition.



$$(T, V)^* = \arg\min_{T,V>0} D(X\|TV).$$

$$D(X\|\Lambda) = -\sum_{\nu,\tau}\left(x_{\nu,\tau}\log\frac{\lambda_{\nu,\tau}}{x_{\nu,\tau}} - \lambda_{\nu,\tau} + x_{\nu,\tau}\right).$$

$$T \sim p(T \mid \Theta^t), \qquad V \sim p(V \mid \Theta^v),$$
$$s_{\nu,i,\tau} \sim \mathcal{PO}(s_{\nu,i,\tau}; t_{\nu,i}v_{i,\tau}), \qquad x_{\nu,\tau} = \sum_i s_{\nu,i,\tau}.$$

$$\text{E Step} \quad q(S)^{(n)} = p(S \mid X, T^{(n-1)}, V^{(n-1)}),$$
$$\text{M Step} \quad (T^{(n)}, V^{(n)}) = \arg\max_{T,V} \langle \log p(S, X \mid T, V)\rangle_{q(S)^{(n)}}.$$

- For the general NMF problem, our goal is the following minimisation problem on the left.
- For the definition of this distance our approach is the information (**Kullback-Leibler**) divergence.
- This is the **generative model** that we use in order to describe the NMF process from a statistical perspective.
- You can see the expectation and maximization steps of the EM algorithm to calculate **MLE in T and V** which is equivalent to minimization of the information divergence.

KL-BNMF

- General NMF approaches aim calculating a maximum a posteriori estimate. In contrast, KL-BNMF approach we investigated includes a full bayesian treatment, where the templates and the excitations (T,V in above model) are integrated out.

$$t_{\nu,i} \sim \mathcal{G}\left(t_{\nu,i}; a^t_{\nu,i}, \frac{b^t_{\nu,i}}{a^t_{\nu,i}}\right), \qquad v_{i,\tau} \sim \mathcal{G}\left(v_{i,\tau}; a^v_{i,\tau}, \frac{b^v_{i,\tau}}{a^v_{i,\tau}}\right).$$

This hierarchical model on the left is more powerful than the basic model above.

With given data and hyperparameters, we may wish to calculate the marginal likelihood which can be used to

1. Estimating the hyperparameters given examples of a source class
2. To compare two given models via Bayes factors.

$$\Theta^* = \arg\max_{\Theta} p(X \mid \Theta)$$

$$l(\Theta_1, \Theta_2) = \frac{p(X \mid \Theta_1)}{p(X \mid \Theta_2)}.$$

Variational Bayes method to bound to marginal log-likelihood as

$$\mathcal{L}_X(\Theta) \equiv \log p(X \mid \Theta) \geq \sum_S \int d(T,V) q \log\frac{p(X,S,T,V \mid \Theta)}{q}$$
$$= \langle \log p(X,S,V,T \mid \Theta)\rangle_q + H[q] \equiv \mathcal{B}_{VB}[q],$$

$$q = q(S,T,V) = p(S,T,V \mid X,\Theta)$$

To simplify this distribution, we assume a factorized form:

$$q(S,T,V) = q(S)q(T)q(V)$$
$$= \left(\prod_{\nu,\tau} q(s_{\nu,1:I,\tau})\right)\left(\prod_{\nu,i} q(t_{\nu,i})\right)\left(\prod_{i,\tau} q(v_{i,\tau})\right) \equiv \prod_{\alpha\in\mathcal{C}} q_\alpha,$$

$$q_\alpha^{(n+1)} \propto \exp\left(\langle \log p(X,S,T,V \mid \Theta)\rangle_{q^{(n)}_{\neg\alpha}}\right),$$

$$q(s_{\nu,1:I,\tau}) \propto \mathcal{M}(s_{\nu,1,\tau},\dots,s_{\nu,i,\tau},\dots,s_{\nu,I,\tau}; x_{\nu,\tau}, p_{\nu,1,\tau},\dots,p_{\nu,i,\tau},\dots,p_{\nu,I,\tau}),$$

$$p_{\nu,i,\tau} = \frac{\exp(\langle\log t_{\nu,i}\rangle + \langle\log v_{i,\tau}\rangle)}{\sum_i \exp(\langle\log t_{\nu,i}\rangle + \langle\log v_{i,\tau}\rangle)}, \quad \langle s_{\nu,i,\tau}\rangle = x_{\nu,\tau}p_{\nu,i,\tau}.$$

$$q(t_{\nu,i}) \propto \mathcal{G}(t_{\nu,i}; \alpha^t_{\nu,i}, \beta^t_{\nu,i}),$$

$$\alpha^t_{\nu,i} \equiv a^t_{\nu,i} + \sum_\tau \langle s_{\nu,i,\tau}\rangle, \quad \beta^t_{\nu,i} \equiv \left(\frac{a^t_{\nu,i}}{b^t_{\nu,i}} + \sum_\tau \langle v_{i,\tau}\rangle\right)^{-1},$$

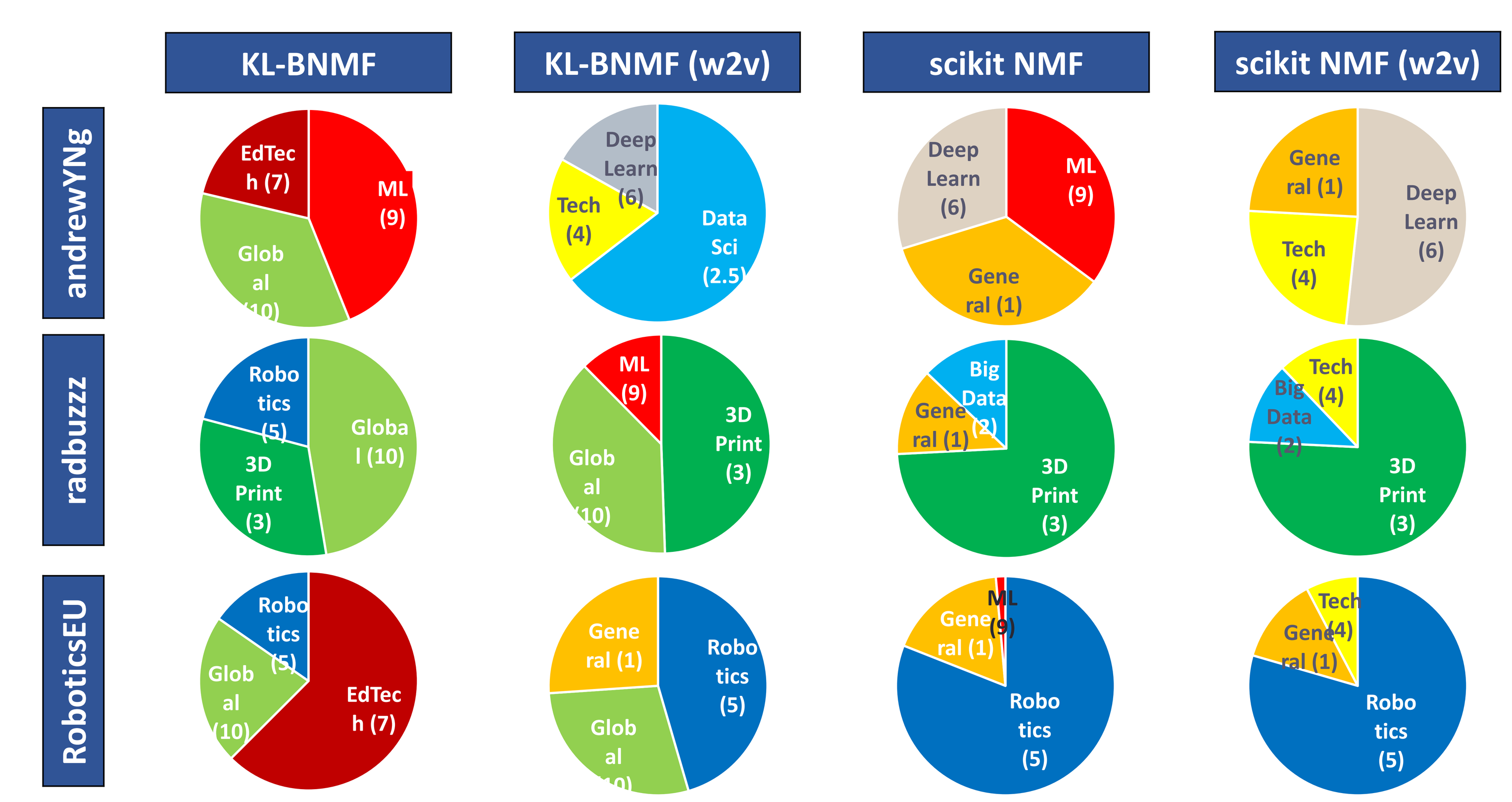
$$\langle t_{\nu,i}\rangle = \alpha^t_{\nu,i}\beta^t_{\nu,i},$$

$$q(v_{i,\tau}) \propto \mathcal{G}(v_{i,\tau}; \alpha^v_{i,\tau}, \beta^v_{i,\tau}),$$

$$\alpha^v_{i,\tau} \equiv a^v_{i,\tau} + \sum_\nu \langle s_{\nu,i,\tau}\rangle, \quad \beta^v_{i,\tau} \equiv \left(\frac{a^v_{i,\tau}}{b^v_{i,\tau}} + \sum_\nu \langle t_{\nu,i}\rangle\right)^{-1}, \quad \langle v_{i,\tau}\rangle = \alpha^v_{i,\tau}\beta^v_{i,\tau}.$$

RESULTS

	Daily (1)	Big Data (2)	3D Print (3)	Tech (4)	Robotics (5)	Deep Learn (6)	EdTech (7)	Politics (8)	ML (9)	Global (10)
scikit LDA	work time think	data bigdata ai	3Dprint 3D print	startup business market	robot manufac us		stem code learn			
scikit NMF	work time look	bigdata analytics data	3Dprint 3D print		robot drone kuka	learn deep neural	edtech stem edchat		datasci ML DeepL	
KL-BNMF	day love time	data bigdata datasci	3Dprint 3D print	business market startup	robot drone ai	learn ai deep	stem edtech code	trump us people	learn ai deep	manufac innov robot
scikit LDA (w2v)	love day us	data bigdata analytics	3Dprint 3D printer	innov join learn	robot ai techn	learn deep machine	code stem learn	trump us science	datasci data ML	
scikit NMF (w2v)	time day today	bigdata analytics data	3Dprint 3D printer	startup bussiness innov	robot kuka automat	learn deep neural	stem science women	trump vote obama	datasci ML bigdata	
KL-BNMF (w2v)	us day join	data analytics bigdata	3Dprint 3D print	startup bussiness tech	robot drone uav	learn deep neural	edtech stem code	trump us people	bigdata ML data	health healthcare innov



CONCLUSION

- We have investigated an hierarchical model with conjugate Gamma priors, and used Variational Bayes algorithm for inference on our twitter data. We reached to similar results with the original NMF approach.
- Considering the topic distribution over vocabulary compared to the NMF, it is fairly successful in finding important topics but, generally it finds more mainstream topics like global issues and daily talks instead of specific topics like python and R.
- Overall, we found that NMF is slightly more consistent than KL-BNMF for topic modelling problem on twitter data.