«Сейфуллин оқулары – 16: Жаңа формациядағы жастар ғылыми – Қазақстанның болашағы» атты халықаралық ғылыми-териялық конференциясының материалдары = Материалы Международной научно-теоретической конференции «Сейфуллинские чтения – 16: Молодежная наука новой формации – будущее Казахстана. - 2020. - Т.II. - С.325-327

BASICS OF TEXT CLUSTEING PRINCIPLES USING LATENT DIRICHLET ALLOCATION

Koyshybay Adil

Nowadays it is hard to imagine our world without automation of working process in industry, because robots and, most of the cases, computers are widespread through all the companies. XX century was the beginning of that, when American factories decided to make routine work more automated and to focus all the human resources on creative tasks.

After including computers on work process, company directors met another task to solve: how to automate not routine work process? It is the breakpoint where such definitions as machine learning and its problems were placed.

For a long time, people had to learn how machines work, how to operate them and how to use them as efficiently as possible. Today, it’s the other way round: machines learn to understand processes, interact with their environment and intelligently adapt their behavior. Robotics, sensor technology, big data and artificial intelligence make machines in industrial production smarter than ever before [1].

There are three implementations of creative tasks by machines: machine learning, deep learning and artificial intelligence.

Machine learning has become increasingly popular over the past decade, and recent advances in computational availability have led to exponential growth to people looking for ways how new methods can be incorporated to advance the field of Natural Language Processing.

Often, we treat topic models as black-box algorithms, but hopefully, this post addressed to shed light on the underlying math, and intuitions behind it, and high-level code to get you started with any textual data [2].

My project solves one of the problems of machine learning – text clustering.

Text clustering problem (TCP) is a leading process in many key areas such as information retrieval text mining, and natural language processing. This presents the need for a potent document clustering algorithm that can be used effectively to navigate, summarize, and arrange information to congregate large data sets. The TCP demands a degree of accuracy beyond that which is possible with metaheuristic swarm based algorithms. The main issue to be addressed is how to split text documents on the basis of GWO into homogeneous clusters that are sufficiently precise and functional [2].

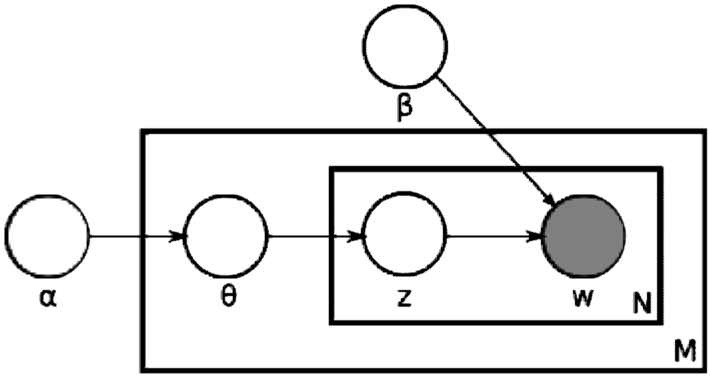
One of the researchers abroad who decided to solve topic model problem is University of Maryland Institute for Advanced Computer Studies. Their work validates the use of topics for corpus exploration and information retrieval. Humans appreciate the semantic coherence of topics and can associate the same documents with a topic that a topic model does. An intriguing possibility is the development of models that explicitly seek to optimize the measures we develop here either by incorporating human judgments into the model-learning framework or creating a computational proxy that simulates human judgments.[5]

In these days there are a lot of topic modelling algorithms, such as Latent Dirichlet allocation, probabilistic latent semantic analysis, Pachinko allocation or Hierarchical latent tree analysis and etc.

Latent Dirichlet Allocation (LDA) is a Bayesian technique that is widely used for inferring the topic structure in corpora of documents. It conceives of a document as a mixture of a small number of topics, and topics as a (relatively sparse) distribution over word types. However, our intuitions tell us that while documents may indeed be conceived of as a mixture of topics, we should further expect topics to be semantically coherent.[3]

With plate notation, which is often used to represent probabilistic graphical models (PGMs), the dependencies among the many variables can be captured concisely. The boxes are "plates" representing replicates, which are repeated entities. The outer plate represents documents, while the inner plate represents the repeated word positions in a given document; each position is associated with a choice of topic and word. The variable names are defined as follows:

M denotes the number of documents, N is number of words in a given document, α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distribution, theta is the topic distribution for document I, k is the word distribution for topic k, ij is the topic for the j-th word in document i (picture 1).



Picture 1 – LDA algorithm for document clustering

Therefore, LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

We can describe the generative process of LDA as, given the *M* number of documents, *N* number of words, and prior *K* number of topics, the model trains to output:

psi, the distribution of words for each topic *K*

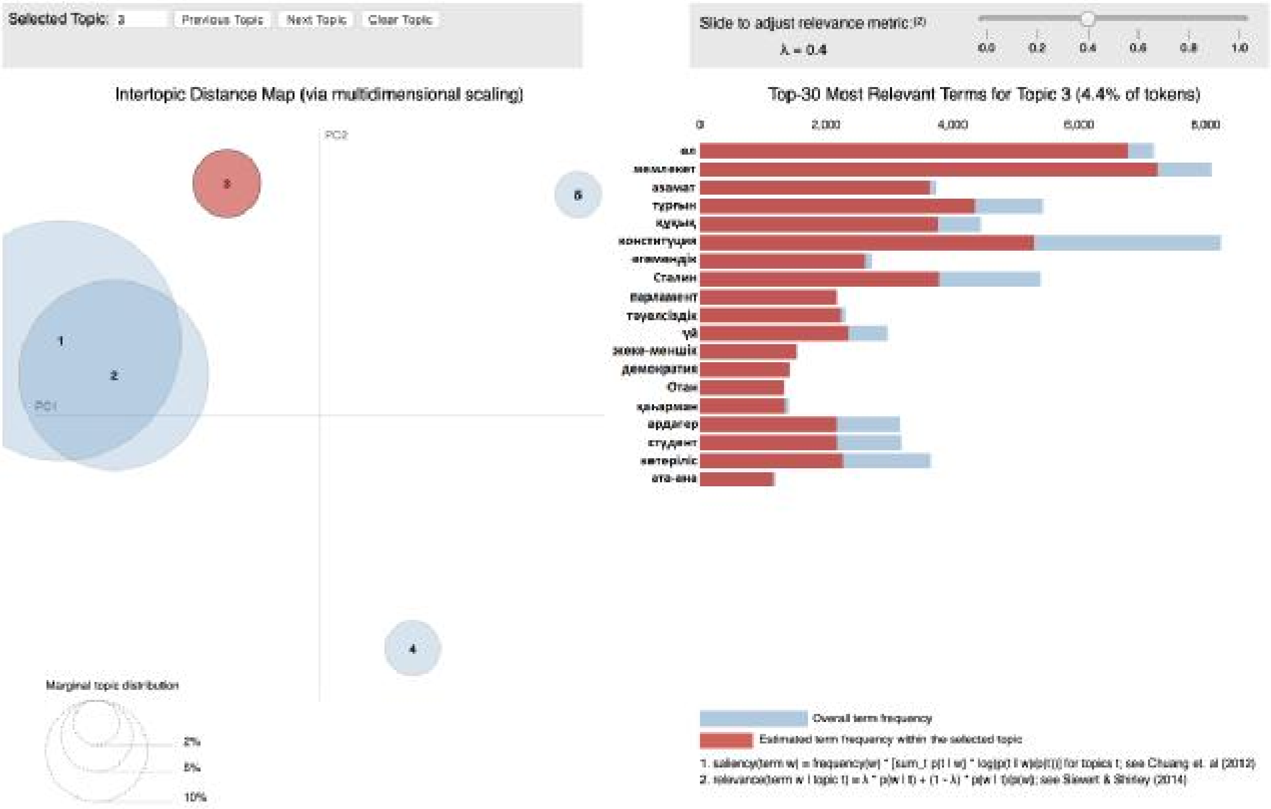
phi, the distribution of topics for each document *i.*

Alpha parameter *is Dirichlet prior concentration parameter that represents document-topic density — with* a higher alpha, documents are assumed to be made up of more topics and result in more specific topic distribution per document.

Beta parameter *is the same prior concentration parameter that represents topic-word density —* with high beta, topics are assumed to make of up most of the words and result in a more specific word distribution per topic [4].

In Python we wrote a system that takes a corpus of documents, tokenizes, lemmatizes and deletes stop words in a corpus, uses Latent Dirichlet Allocation algorithm to convert corpus into vectors and selecting most similar topics using Gensim library's LDA method and visualize the data through pyLDAvis framework which is shown below (picture 2).

In picture 2 it is shown that a user can select several numbers of topics, depends on what was given in the code. After that, it is seen, that there are a frequency slider to show most frequent words in a topic and a coordinate system where small spheres mean a low relevance of topic according to corpus, and big spheres mean a high percentage of relevance. A good formed topic is when a sphere is big and do not cross other spheres.


Selected Topic: 3
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.4
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (4.4% of tokens)
0
2,000
4,000
6,000
8,000
ел
мемлекет
азамат
тұрғын
құқық
конституция
егемендік
Сталин
парламент
тәуелсіздік
үй
жеке-меншік
демократия
Отан
қаһарман
ардагер
студент
көтеріліс
ата-ана
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Picture 2 – Frequency of words after topic modelling using LDA

The absence of text clustering problem made an action to create such project as using LDA for text clustering in Kazakh language. Whole process was made in Python with several NLP frameworks for document processing and to visualize topic modelling.

Our next steps to improve this system will be increasing the accuracy of prediction and predicting themes of unlabeled documents.

Literature

1. Machine Learning in Industry 4.0: Five Use Cases [Электронныйресурс]. URL: www.datafestival.de/en/machine-learning-in-industry-4-0-five-use-cases

2. Rashaideh, H., Sawaie, A., Al-Betar, M.A., (...), Al-Khatib, R.M., Braik, M.A Grey Wolf Optimizer for Text Document Clustering Journal of Intelligent Systems 29(1), c. 814-830

3. Rajarshi Das, Manzil Zaheer, Chris Dyer, Gaussian LDA for Topic Models with Word Embeddings. [Электронныйресурс]. URL: rajarshd.github.io/papers/acl2015.pdf

4. Latent Dirichlet allocation. [Электронныйресурс]. URL: en.wikipedia.org/wiki/Latent_Dirichlet_allocation

5. Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei, Reading Tea Leaves: How Humans Interpret Topic Models, [Электронныйресурс]. URL: users.umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf

Scientific director, PhD Ismailova A.A.