Modern Data Analysis

Aleksey Ryabykin, Mikhail Osokin

20/12/22

HSE UNIVERSITY

Datasets

We have chosen a dataset consisting of Medium articles data. Our objective is to examine the topics, volume, content, and other factors that can contribute to the popularity of the articles on this platform.

Claps could be viewed as a platform business metric. Because anyone can clap as many times as they want, unlike with likes on other platforms, it's also intriguing to examine the outliers in this feature.

You also can read the variables description here.

Proposed approaches

- Data scrapping (due to the date limits, we can scrap modern articles);
- Exploratory data analysis;
- Hypotheses testing;
- Topic modeling;
- Interpret the relation between topics and claps on the platform (if it exists);
- Build a regression with claps counts as a target variable;
- Build an ensemble model of regressors;
- Explain the results.

Data scrapping

After

9935 instances

All of them exist on the medium

Before

6508 instances

≈5% had been already deleted

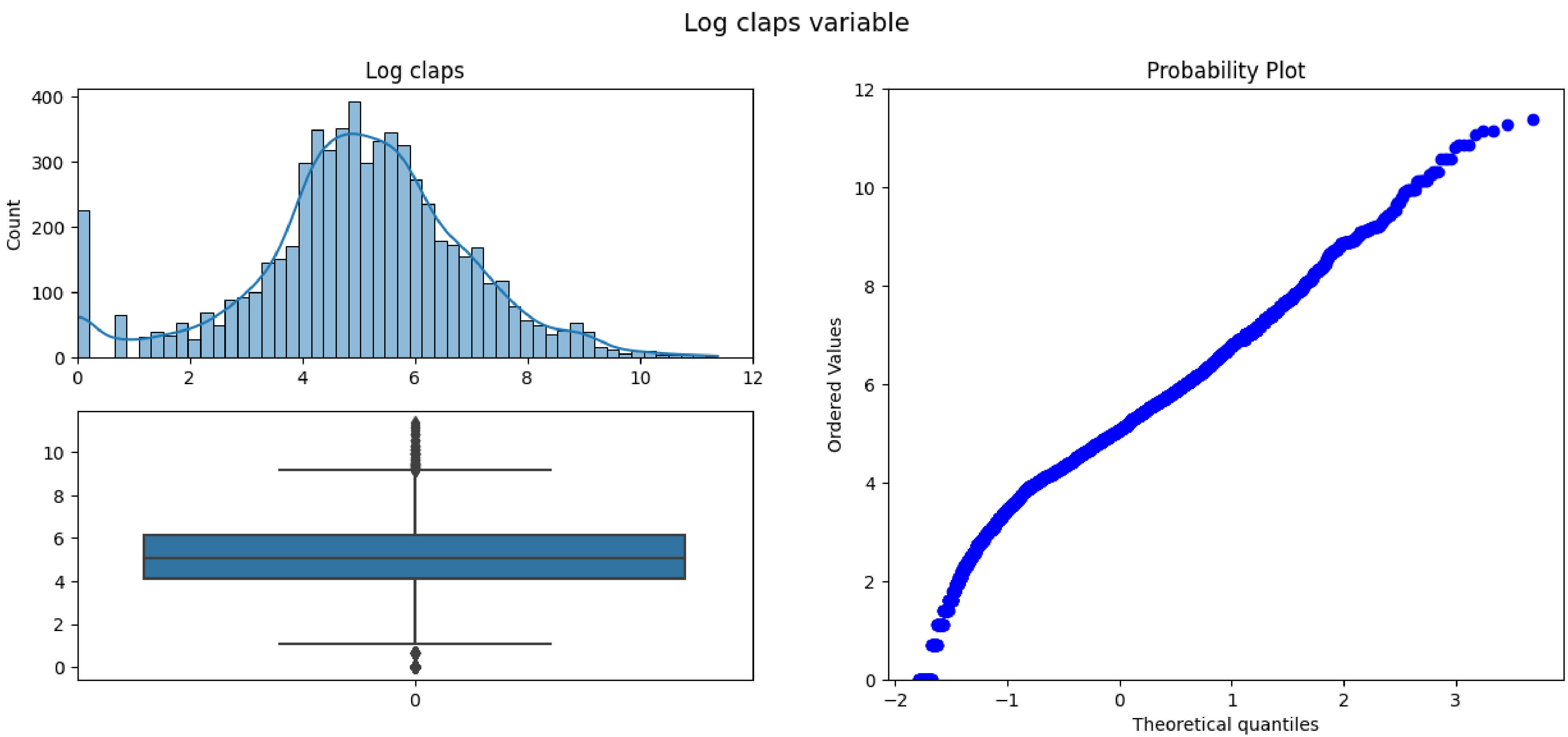
Before

Link to an article	url
Title of an article	title
Subtitle	subtitle
Main image	image
Responses	responses
Reading time	reading_time
Publication	publication
Date of publication	date
Number of claps	claps

After

Title of an article	title	Responses	responses
Link	link	Number of code chunks	n_code_chunks
Publication	publication	Number of bold	bold_text_cou nt
Author	author	Number of italic	Italic_text_cou nt
Followers	followers	Mean image size	mean_image_ width,height
Reading time	reading_time	Number of images	n_images
Number of words	n_words	Number of videos	n_vids
Text of an article	pure_text	Number of references	n_links
Data of publication	date	Claps	claps

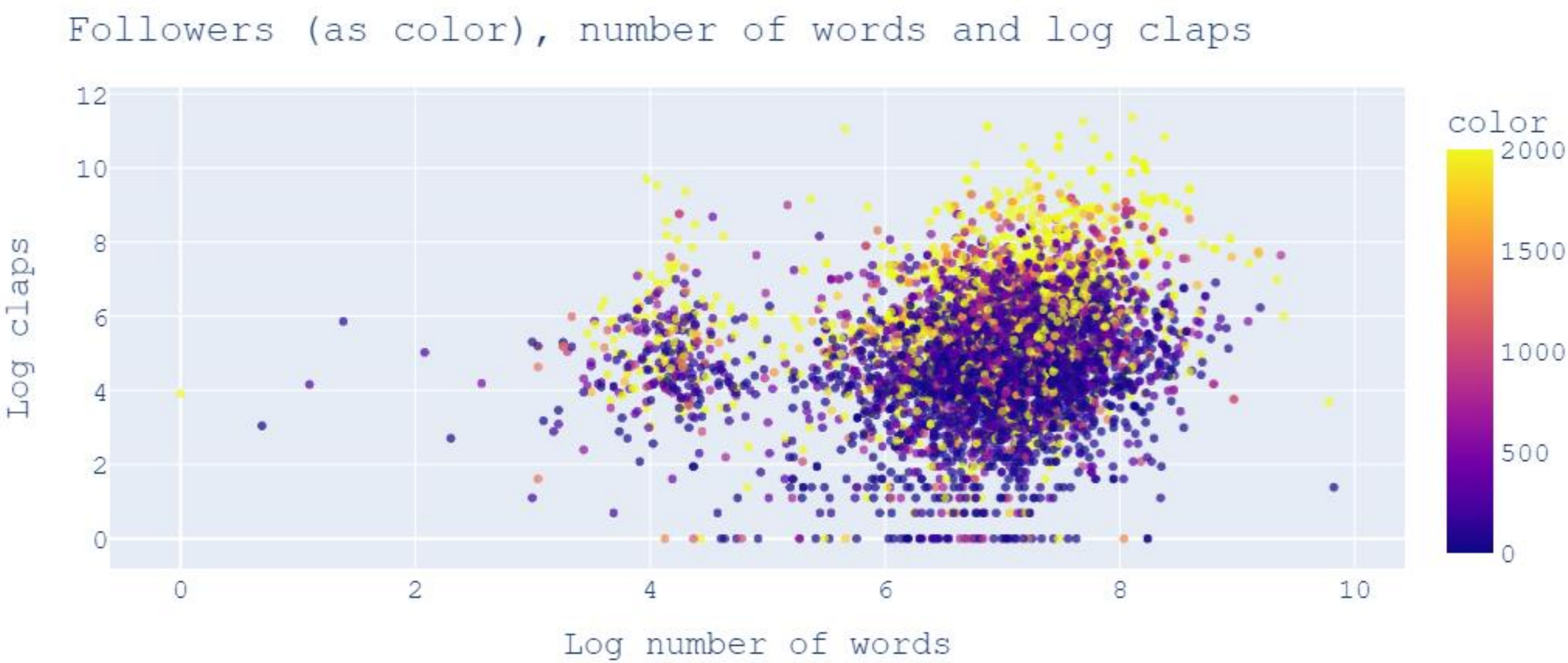
EDA inferences



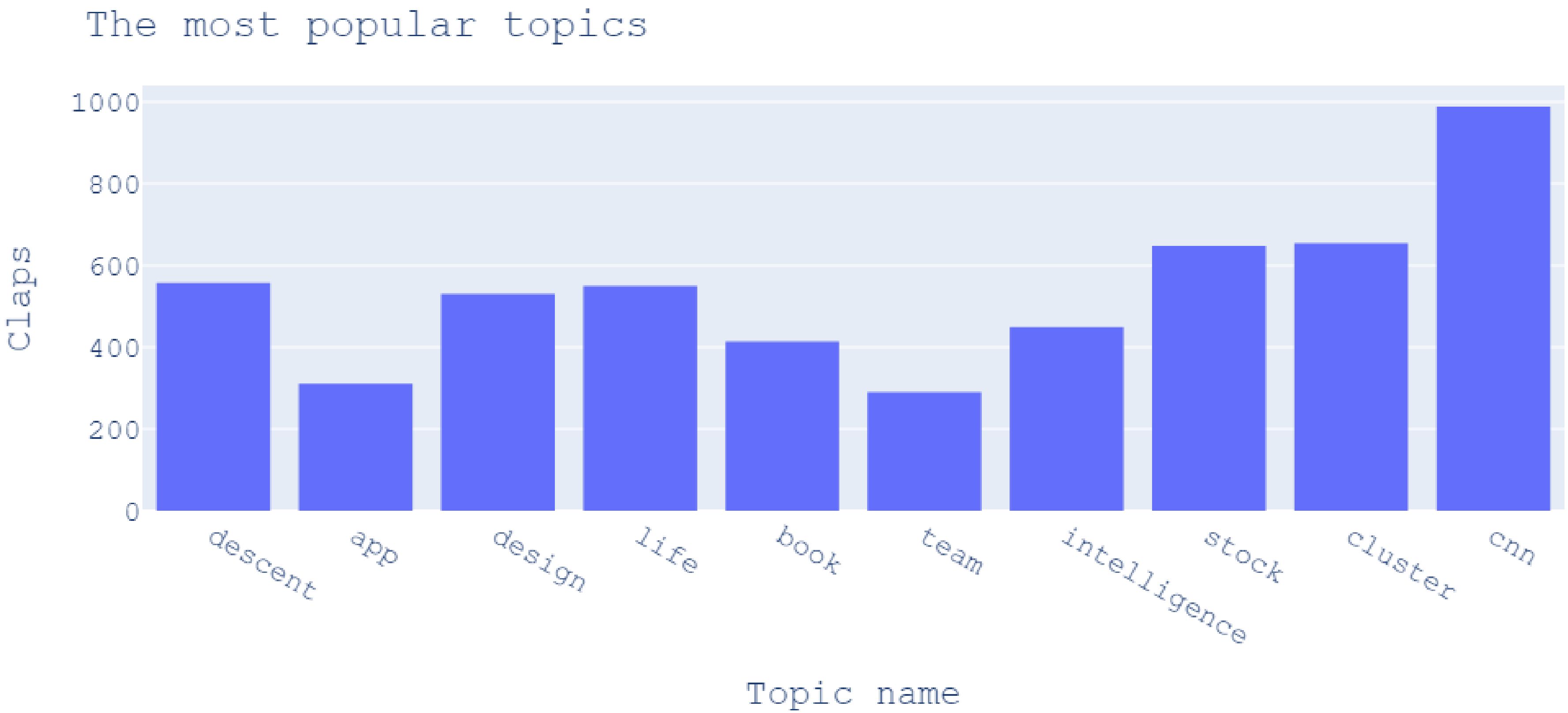
Close to normal distribution

EDA inferences

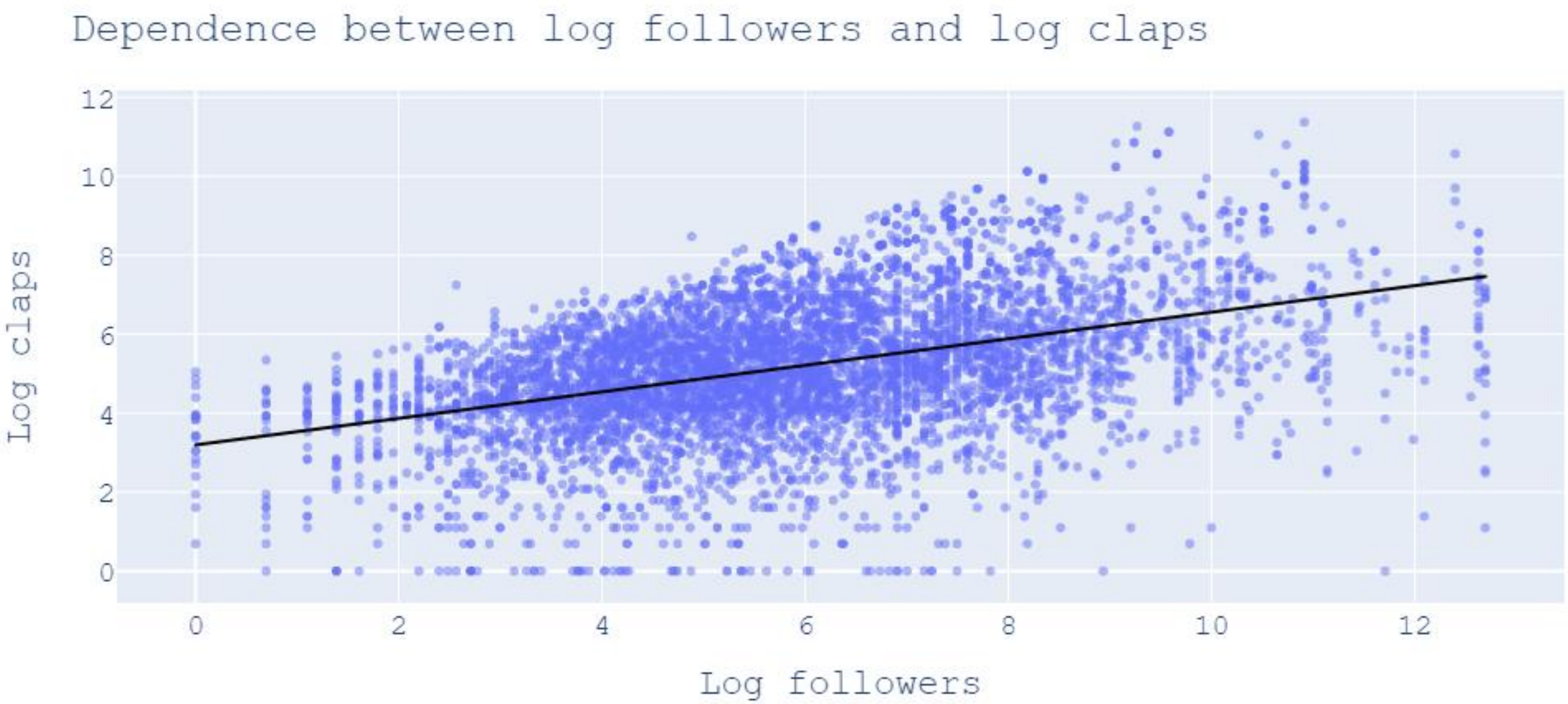
As we see, data splits into two clearly separated groups. Moreover, this plot shows that authors with more followers get more claps. However, there is no visible dependency of number of followers and length of articles.



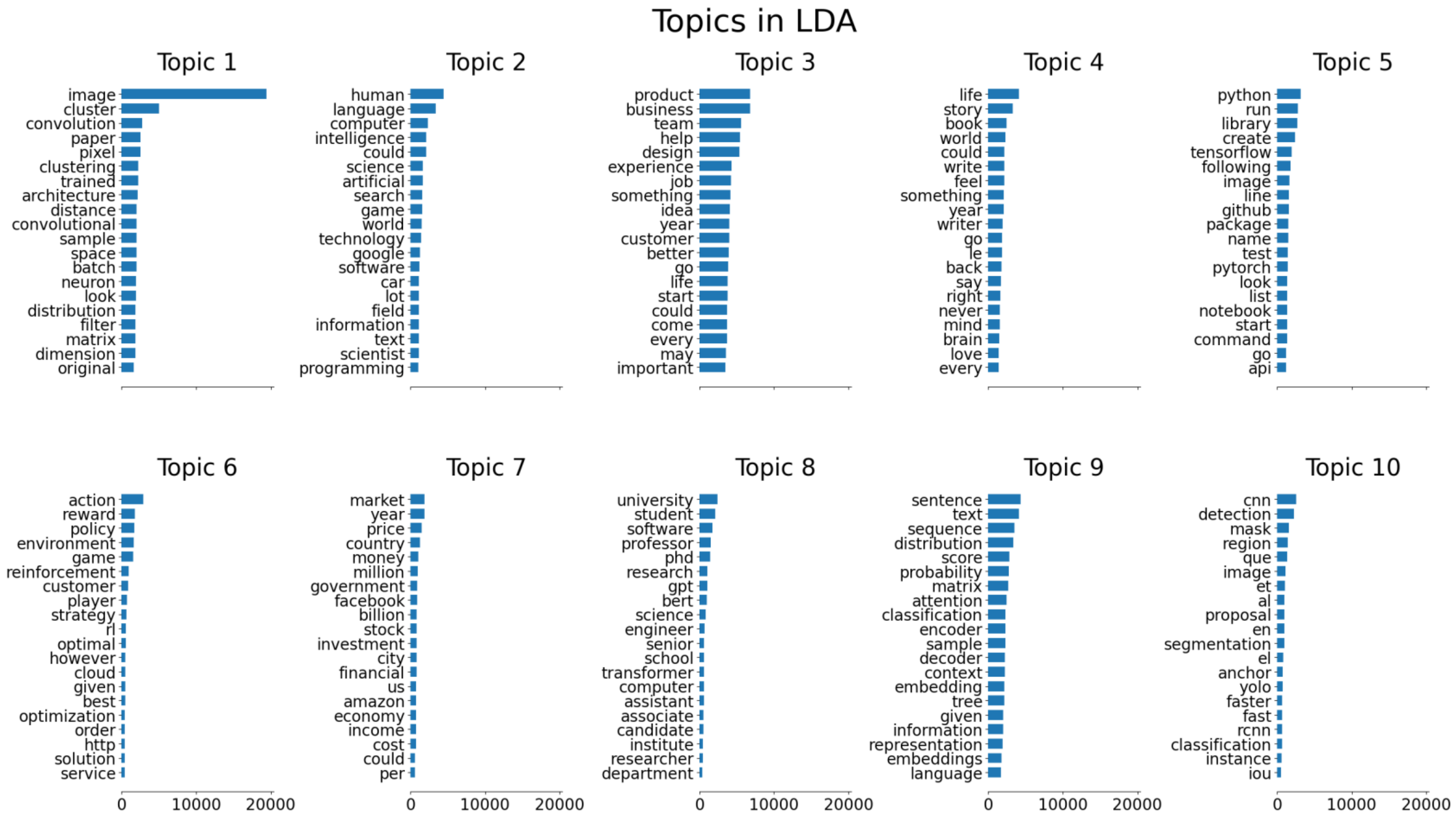
EDA inferences



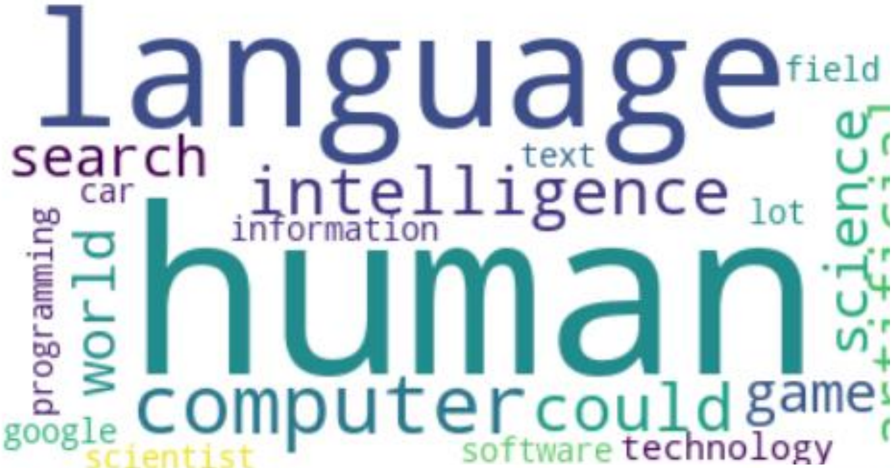
EDA inferences



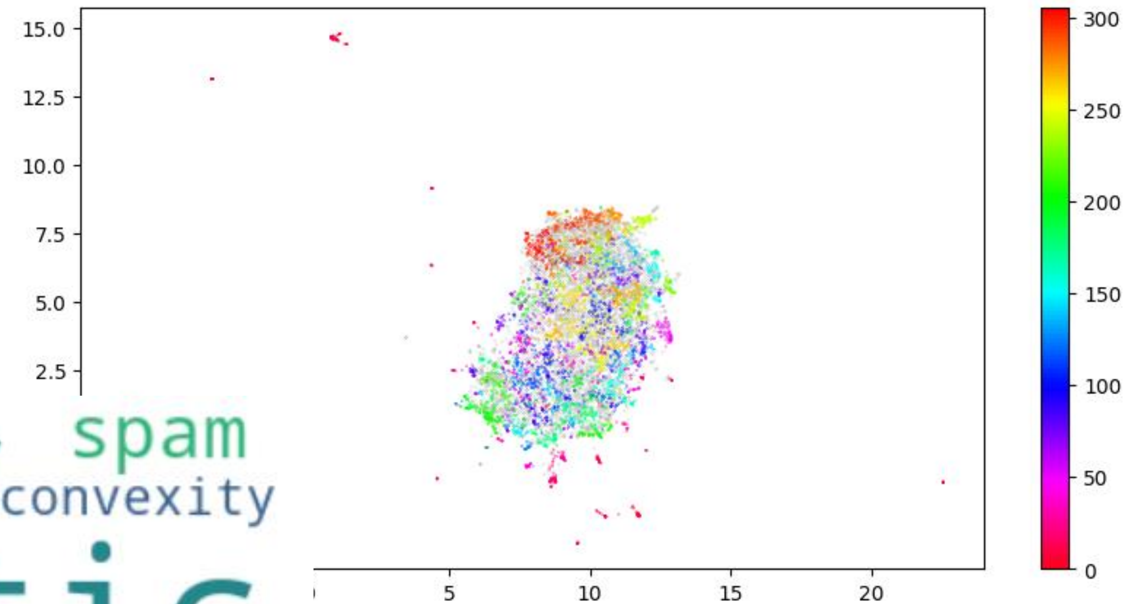
Topic modeling LDA



Topic modeling LDA

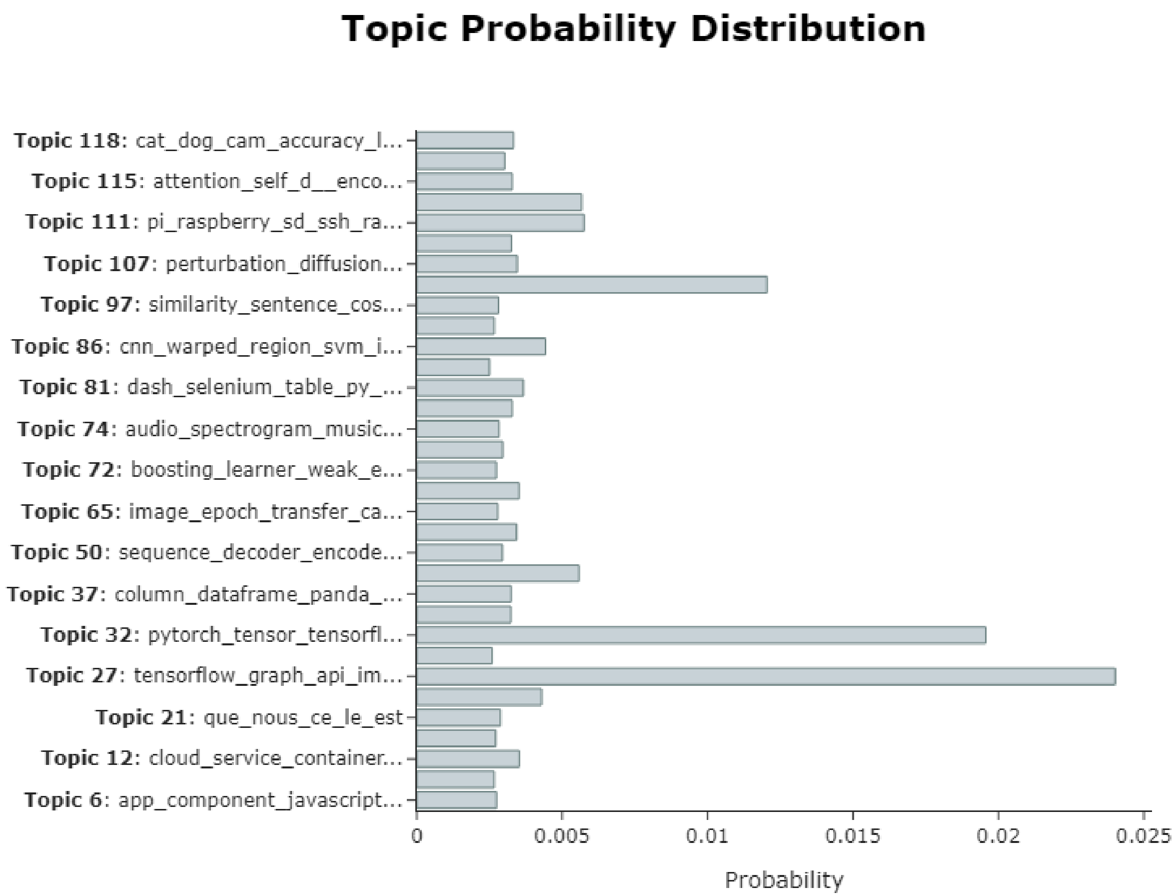


Topic modeling BertTopic (Hand-made)



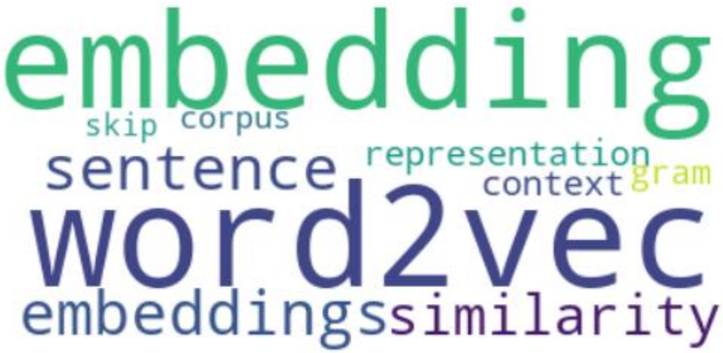
Topic modeling BertTopic

Topic modeling BertTopic



Inferences of topic modeling

All clusters are easy to interpret and useful working with target. For instance, if we do not take the most popular ones, then the variance and expectation are preserved for the same clusters. It indicates a fairly close relationship.



Regression results

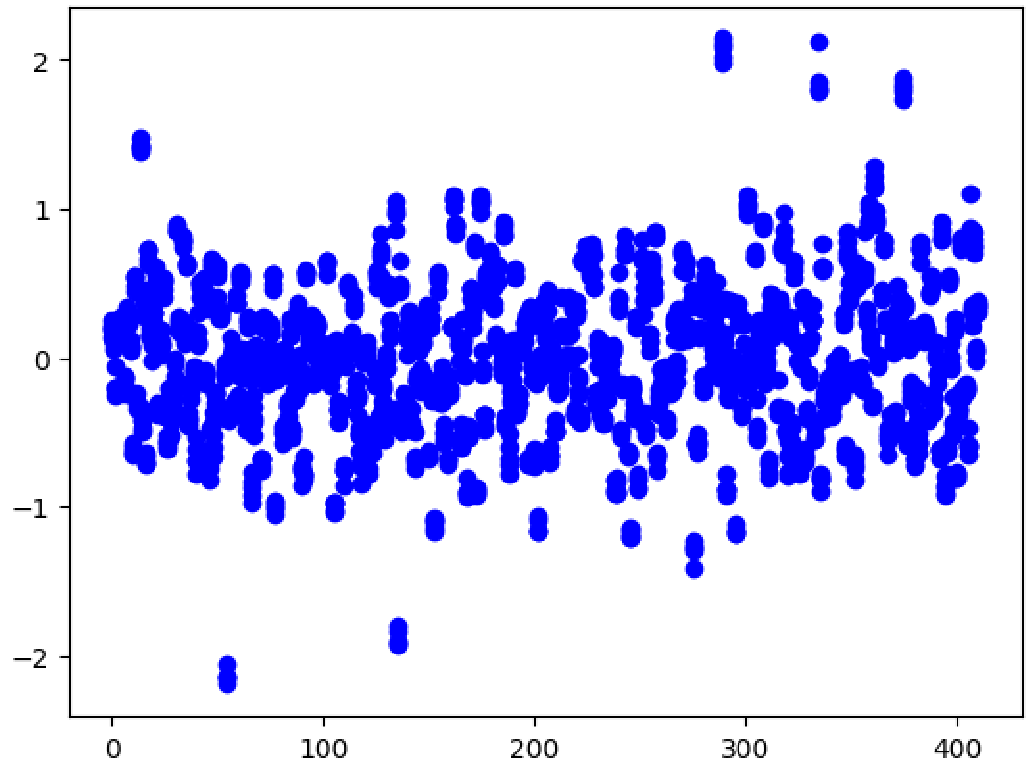
loss	catboost	Xgboost	Lightgbm
mae	307.4	297.3	353
mse	371.67	307.7	408.6
huber	453.2	450.1	476.4

Optuna-tuned ensemble of regressors

optuna	catboost	Xgboost	Lightgbm
282.2	307.4	297.3	353

Error analysis

Regression residuals proved their homoscedasticity following the Levene criteria



	W	Pval	equal_var
Levene	0.008147	0.999869	True

Overview

Time improvement by asynchronization

≈5x

Scrapped articles

4120

Time improvement

≈6.5x

Average quality improvement

≈32%

Number of ensemled models

3

Hypotheses tests

10