Automatic tagging of articles

Maxim Partin   Chubarova Daria

May 2025

Abstract

Modern platforms for introducing text blogs offer their authors a tag system. The author can put a set of specific tags for his article so that it can be found in a search by keywords. Articles on platforms for introducing text blogs can number in the thousands. Examples of popular such platforms are medium, livejournal. The task of setting tags for an article is of no interest to the author of the article. This task is important for the search engine. To relieve article authors from this work, you can automate the setting of tags for articles. This is what the project is dedicated to, the project code can be found here `https://github.com/infamax/hse_nlp_project`.

1 Introduction

First of all, you will need to write the whole report in English, with a few exceptions mentioned below. This section is devoted to a problem motivation. You should answer the question of why the problem you were working on is important. Also, you should describe what is unique in your approach to this problem, what are the differences to other approaches.

1.1 Team

Please list your team members and describe their responsibilities. Example:

Valentin Malykh prepared this document.

2 Related Work

In this section, you will describe in details the existing approaches to the problem you work on. For each approach, you need to provide a reference.

[Levenshtein, 1966] is a sample reference to the previous art. [Левенштейн, 1966] is a sample reference in Russian.

3 Model Description

Here you need to write a detailed description of your approach. It is important to mention that this description should give more details than the descriptions from section 2[1].

You will likely be providing a figure to better present your approach. A sample circle is presented on Fig. 1.

The other possible contents of this section are formulae. They could be on a new line:

$$S = \pi r^2,$$

or they could be inline, e.g. if you want to describe the used variables, like S is an area of a circle, while r is its radius.

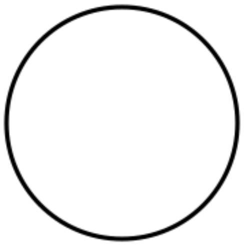

Figure 1: A sample circle.

4 Dataset

In this section, you need to describe the dataset(s) you are working with. An example dataset we will use is WikiText-2. Please mention a paper where it was presented, e.g. WikiText-2 was presented in [Merity et al., 2017]. Please provide guidance on how to obtain the dataset[2]. It is important to mention that the dataset you use must be available for the research purposes. So please make sure about that.

Your description will likely be including a table. On the Tab. 1 you can see the statistics for the mentioned dataset. It is important to notice that there is a split of the dataset and this split is covered by the description.

If you were collecting the dataset on your own please describe the collection procedure, like criteria were used to filter the documents, the pre-processing steps, etc. It is preferable that you release your dataset for the public, but you are not obliged to do this. Please make sure that you have legal rights to collect and distribute the data you were working with. We recommend you to look at

[1] This is an example of internal references and footnotes at the same time.

[2] The one way to do it is to include a link to the website, where it could be downloaded from. Like this.

	Train	Valid	Test
Articles	600	60	60
Tokens	2,088,628	217,646	245,569
Vocabulary size		33,278	
Out of Vocab rate		2.6%	

Table 1: Statistics of the WikiText-2. The out of vocabulary (OoV) rate notes what percentage of tokens have been replaced by an $\langle unk \rangle$ token. The token count includes newlines which add to the structure of the dataset.

C4Corpus and how it is licensed to make your corpora. C4Corpus is described in [Habernal et al., 2016].

5 Experiments

This section should include several subsections.

5.1 Metrics

First of all, you should describe the metric(s) you were using to evaluate your approach. Most likely a metric description will include a formula.

5.2 Experiment Setup

Secondly, you need to describe the design of your experiment, e.g. how many runs there were, how the data split was done. The important details of your model, like hyper-parameters used in the experiments, and so on.

5.3 Baselines

Another important feature is that you could provide here the description of some simple approaches for your problem, like logistic regression over TF-IDF embedding for text classification. The baselines are needed is there is no previous art on the problem you are presenting.

6 Results

In this section, you need to list and describe the achieved results. It is crucial to have the results of the experiments for the other approaches. This is needed to be able to compare your results with some competitors. Most preferably, you should provide some references with results on the same problem.

Almost inevitably the results are presented as a table, but it is also possible to have a graph, i.e. a figure.

You need also to provide an interpretation of the presented results, to describe some features. E.g. your approach shows higher results on the short texts or by one metric instead of another.

Also in this section, you could provide some results for your model inference. The samples could be found in Tab. 2.

Это пример вывода вашей модели на русском.
This is a sample output of your model in English.

Table 2: Output samples.

7 Conclusion

In this section, you need to describe all the work in short: what you have done and what has been achieved. E.g. you have collected a dataset, made a markup for it and developed a model showing the best results compared to other models.

References

[Habernal et al., 2016] Habernal, I., Zayed, O., and Gurevych, I. (2016). C4corpus: Multilingual web-size corpus with free license. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 914–922.

[Levenshtein, 1966] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:709.

[Merity et al., 2017] Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2017). Pointer sentinel mixture models. In *Proceedings of International Conference of Learning Representation.*

[Левенштейн, 1966] Левенштейн, В. И. (1966). Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады Академий Наук СССР*, 163(4):845–848.