Semantic search model for entertainment documents

I chose the role of an expert.

1 Planning the industrial research project

Before planning the research, the analyst and (**expert**) discuss the key issues. After the long dash — our remarks.

1. Goal of the project. (**Expected development result.**) — Expected research objective.

   The goal of the project is to create a new semantic search model to improve the search for documents from the entertainment section in the mobile application of a Russian company.

2. Applied problem solved in the project. (**How will the result be used?**) — How to illustrate the result?

   The project solved the following problem. Application clients use the search in the "Entertainment" section, where they can buy tickets for various events: concerts, performances, exhibitions, movies in the cinema, etc. However, the conversion from the request to the purchase is quite small. Preliminary analysis showed that the problem is related to the quality of the search: it is not able to find documents semantically close to the request, since it is built on the basis of a full-text engine. It was necessary to develop a model for semantic search that would complement the basic search, which would allow finding relevant documents for more complex requests and thereby improve the quality of the search and the conversion from the request to the purchase.

3. Description of historical measured data. (**Formats and timing.**) — Algebraic data structure.

   The data that was available within the project are search logs of the mobile application. They contain information about search results for queries: what query was sent, what results were generated for it, what documents from the results were clicked on, and other information that is less useful for searching. The data was collected from approximately October 2023 to February 2024.

4. Quality criteria. (**How is the quality of the obtained result measured, what is in the report?**) — Error function to optimize.

   The recall@3 metric was chosen as a quality criterion - a metric that shows what share of the total number of relevant documents for a query falls into the top 3 elements of the search results. This is due to the fact that documents in the search results are displayed as a block of size 3, and therefore it is important that as many relevant documents as possible fall into the top 3 elements of the search results.

5. Project feasibility. (**How to show that the project is feasible, list of possible risks.**) — Error analysis plan.

   The project is feasible because it pursues a specific business goal, the achievement of which will allow:

   (a) Make the client's life easier, more pleasant and more convenient.

   (b) Earn more money for the company by increasing the conversion to purchase.

   The risks include a deterioration in the quality of the new search relative to the basic full-text solution.

6. Conditions necessary for successful project implementation. (**Organization of work.**) — Requirements for the data set.

   To successfully implement the project, we need a dataset of several thousand labeled request-document pairs, with an assessment of their relevance, as well as a GPU to conduct experiments and then launch the service with a trained model.

7. Solution methods. (**Procedure libraries.**) — Hypotheses, optimal probability models.

   The currently State-Of-The-Art model for semantic search, called e5-multilingual-large, was chosen for the solution. This model is a transformer encoder pre-trained on text similarity, search, and ranking tasks. As part of the project, the model was further trained using fine-tuning with contrastive-learning methods: it learned to bring vectors closer together for relevant query-document pairs and to move vectors apart for irrelevant pairs.

2 Research or development?

In other words, novelty or technological advancement?

Analyst: What impact will the research have on the field of knowledge? How useful will it be?

Expert: (**How long will the model be used? What will replace it in the future?**)

The model is set up for monitoring, which evaluates it once a week on an offline sample of real search logs, on which the model has not yet been trained. If the target metric drops below a certain threshold, then an automatic pipeline is launched to retrain it on this sample as a training one.