NLP course Template

G.G. Varenikov

E-mail: varenikov@phystech.edu

Repository: https://github.com/VGeorgeo/NLP_huawei_final_project.git

Abstract

A properly functioning search engine is an important part of creating websites. Correct recommendations can increase your traffic and profits. The results are based on the quantity and quality of the collected data and the selected recommendation model. In this work, I created a search engine for an aggregator of educational resources. The problem was solved by several approaches that were compared with each other.

1. Introduction

The problem with the implementation of search engine recommendations is quite old. There are many different approaches based on statistics or machine learning (Fig.1). The reviews of methods can be analyzed in these articles [1, 2]. In my opinion, two main approaches should be highlighted:

- **Using models that have been pre-trained on large text data**. Today there are several main architectures. Search engines of global corporations such as Google, Facebook, etc., are using neural networks (transformers) - BERT, GPT, ELMO, and so on [3, 4].
- **Using statistic methods to make suggestions, for example TF-IDF [5].**
- **Using the history of user requests and clicks on a website to make recommendations.**

In given case, we are dealing with a limited set of names and descriptions of websites with an emphasis on training and education. The main problem of this task is the lack of user requests data and website transitions, so we

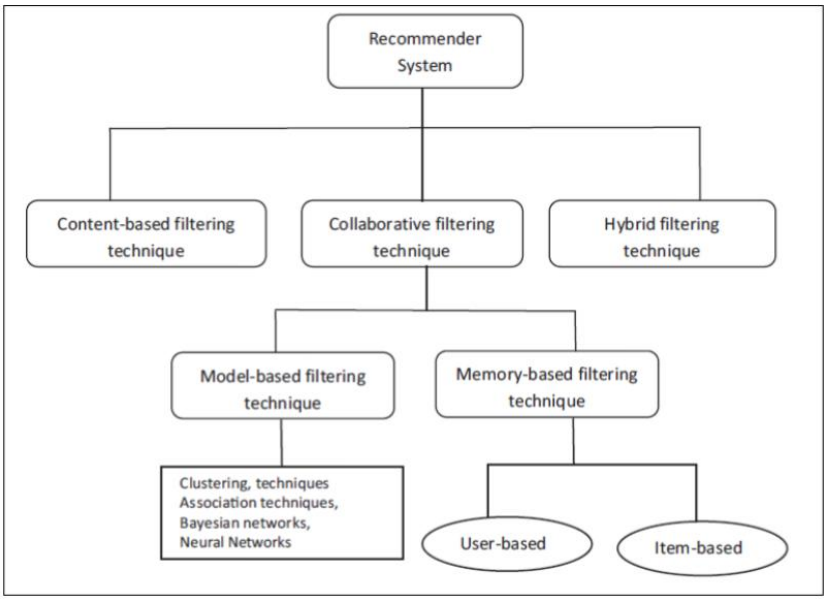

Рис. 1: Types of approaches to recommendations [1].

can use only titles and descriptions. Also, the data is not labelled, so we can't measure the accuracy of suggestions. However, by trial and error when using the current TF-IDF model, cases of unsatisfactory recommendations were identified. They can be used as a approbation of the proposed models. Moreover, the adequacy of the recommendations can be checked 'manually'.

2. Methods and data

2.1. Methods

2.1.1. TF-IDF

Term Frequency — Inverse Document Frequency (TF-IDF) [5]: a method based on statistics and calculating the number of words in a set of documents. The equation of TF-IDF can be written as Term Frequency (TF) * Inverse Document Frequency (IDF), where term frequency is frequency of a word in a document and inverse document frequency is inversion of the frequency with which a certain word occurs in the collection documents. The frequency can be calculated both for one word (unigram) and for several (bigram, for example). In this work we used both variants.

2.1.2. Transformers

Transformers – type of neural network architecture which include attention mechanisms. Usually used in issues of natural language processing, however, it becomes more popular in other areas as well. One of the most popular type of models is based on bidirectional encoder representations from transformers (BERT [3]) which was developed by Google. In this work different pre-trained models are based on BERT architecture: different types of representation-based sentence encoders for Russian language [6] (ruBERT tiny, ruBERT base, ruBERT threeway, ruBERT tiny bilingual) and model of language-agnostic BERT sentence embedding (LaBSE) [7].

2.2. Data

Unfortunately, I am unable to provide a full description of the data frame because it is a corporate secret. The main features that will be used are 'title' and 'description' (if it is available) of courses and other educational resources. The full size of the data frame is more than 50 thousand rows. The main language of the text and requests is Russian. However, the database is very muddled: some rows are uninformative, others have incorrect and incomplete translations.

2.3. Data cleaning and pre-processing

The first step was to remove the unhelpful information from the important data. The text analysis showed the existence of more than 2 languages, so the rows without English and Russian words were deleted. Also, the analysis showed some unique descriptors of uninformative strings which were assembled manually; the rows with these descriptors were deleted too. Similarly, strings of less than 20 symbols were found to be misleading and were cleaned. The problem of incorrect and incomplete translation was solved by translating texts with 2 languages types of letters into English and then into Russian again. The final clean database consists of three columns: 'title', 'description', and 'title + description' (full text). It can be used with pre-trained nlp models like BERT.

However, other approaches require more complicated pre-processing. For example, the TF-IDF method is not able to distinguish between the same basic root words with different prefixes or suffixes. Therefore, another database with text pre-processing was developed. These are the stages of pre-processing:

- Clean data from html and other symbols
- Delete English and Russian stop words
- Make a lemmatization of each word
- Make a stemming of each word

Finally, there are two data frames: without preprocessing and with preprocessing, which will be used to make recommendations after the search query.

3. Results

All models described in the section "Data and methods"have been tested manually for various requests. In this text the phrase "Инженерное налогообложение"(**sequence**) which translates from English as 'Engineering Taxation' was chosen to demonstrate and compare the ability of the selected methods. This phrase was chosen because the previous model had problems with adequate recommendations: it often proposed texts about 'engineering', even though the main point of **sequence** is 'taxation'. Moreover, there is no actual information about the engineering taxation in the database, but it is still important to offer similar overlapping topics.

First of all, the TF-IDF method was applied. Model were fitted on the full text or only on the title. Since there are no any bigram equals or anything similar to **sequence** in the database, the model that collected bigram-only information did not find any similar texts or made bad recommendations. The TF-IDF models that were fitted on titles and only on unigrams proposed quite good suggestions, but some of the recommendations were about the engineering. The best results were demonstrated by models that fitted on the pre-processed full text (collected unigrams or both unigrams and bigrams). Despite the inaccurate suggestions TF-IDF approach can be used as one of the parts of the recommendation system, the similarity values can be multiplied by a coefficient $k < 1$.

Pre-trained nlp models showed more meaningful recommendations. Basically, the text did not contain the words from **sequence** but it was about overlapping topics such as financial accounting, tax law, etc. It is worth noting that, on the one hand, tiny models gave more misleading recommendations, but, on the other hand, the best recommendations are more specific and accurate. The basic models (large) give more general recommendations, but they are

most often correct. RuBERT base, LaBSE, and ruBERT threeway models were chosen as more preferable for given task.

Results of models recommendations
LaBSE fitted on preprocessed data
Основы бухгалтерского учета
Формальный Финансовый учет
Теория и практика эффективной коммуникации
Финансовые технологии
Финансовые технологии
LaBSE fitted on original data
Искусство налогового планирования
Формальный Финансовый учет
Институциональная экономика (Institutional economics)
Математическое моделирование в инженерных науках
Машинное обучение в финансах
Rubert base model fitted on preprocessed data
Финансовый менеджмент
Основы менеджмента
Основы менеджмента
Управление маркетингом
Язык и инструменты финансового анализа
Rubert base model fitted on original data
Налоговое право (Tax law)
Машинное обучение в финансах
Бюджетирование и ценообразование в ДПО
Налоговое право
Правовое обеспечение бизнеса в России

4. Conclusion and discussion

In this work, I tried to solve the problem of a search engine and recommendations based on the given database. The database was cleaned from misleading information and pre-processed. TF-IDF and pre-trained nlp model were applied

to make recommendations. TF-IDF obviously showed muddled recommendations because it bases them only on words frequency. However, it can still be used to search for certain unigrams, bigrams, and so on. Pre-trained models showed more reliable results, but they can be improved too. First of all, I will finalize the model for this database which should increase the adequacy. Also, it has been noticed that pre-trained models often suggest short texts, for example, with only a title. This may mean that the models of long texts are noisy and show less similarity. Thus, reducing the length of the text can improve accuracy. Another option, long text keywords can be extracted and used for searching similar texts.

Results of models recommendations
Rubert tiny bilingual model fitted on preprocessed data
Финансы для всех: Debt
Байесовская метода машинного обучения
Карьера, команда, инвестиции.
fintech - финансовые инновации
Финансовые технологии
Rubert threeway model fitted on preprocessed data
Финансовый анализ
Финансовый менеджмент
Финансовые технологии
Финансовые технологии
Финансовые рынки

Results of models recommendations
TF-IDF fitted on preprocessed full text data with 1,2-grams
Основы налогообложения бизнеса в России
Налогообложение I хозяйствующих субъектов: Корпорации
Налоги и налогообложение: специальные налоговые режимы
Налогообложение хозяйствующих субъектов
Налогообложение в Российской Федерации
TF-IDF fitted on preprocessed title text data with 1,2-grams
Двойное налогообложение в России
Основы налогообложения бизнеса в России
Инженерная механика
Инженерная механика
Инженерная механика

References

[1] M. Zahrawi, A. Mohammad, Implementing recommender systems using machine learning and knowledge discovery tools, Knowledge-Based Engineering and Sciences 2 (2) (2021) 44–53 (Aug. 2021). doi:10.51526/kbes.2021.2.2.44-53.

[2] H. Alharthi, Natural language processing for book recommender systems (2019). doi:10.20381/RUOR-23382.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding (2019) 4171–4186 (Jun. 2019). doi:10.18653/v1/N19-1423.

[4] Y. Kuratov, M. Arkhipov, Adaptation of deep bidirectional multilingual transformers for russian language (2019). doi:arXiv:1905.07213.

[5] S. Qaiser, R. Ali, Text mining: Use of TF-IDF to examine the relevance of words to documents, International Journal of Computer Applications 181 (1) (2018) 25–29 (Jul. 2018). doi:10.5120/ijca2018917395.

[6] A. Golubev, N. Loukachevitch, Improving results on russian sentiment datasets (2020) 109–121 (2020). doi:10.1007/978-3-030-59082-68.

[7] T. Shavrina, A. Fenogenova, A. Emelyanov, D. Shevelev, E. Artemova, V. Malykh, V. Mikhailov, M. Tikhonova, A. Chertok, A. Evlampiev, Russiansuperglue: A russian language understanding evaluation benchmark (2020). doi:arXiv:2010.15925.