AgroSearch helps farmers

Dmitry Stepakov, Ivan Tsvetcov

May 2023

Abstract

Currently, there are interruptions in the supply of various extracts and medicinal substances. According to statistics for various positions, imports range from 20 to 80% of all raw materials used. The most important task is to help Russian farmers grow raw materials on their land. We are pleased to present a solution that structures information about medicinal plants. Project code link right here: `https://github.com/Brampaap/argo_smart_search`.

1 Introduction

The agricultural sector, especially in the domain of medicinal plant cultivation, is facing significant challenges due to inconsistent supplies of various extracts and medicinal substances. It's worth noting that statistics reveal a considerable reliance on imports, accounting for 20 to 80% of all raw materials employed. In the wake of these disruptions, an imminent need arises to aid Russian farmers in their pursuit to grow these essential raw materials domestically.

Our endeavor is to introduce a solution that structures information about medicinal plants and tailors it to the specific needs of farmers. This addresses three primary issues: matching farmers with suitable crops as per their requirements, efficiently handling voluminous and regularly changing data, and enhancing business profitability via user-specific recommendations.

To solve the first problem, our solution incorporates an advanced attribute-based search mechanism. This intelligent system will assist the farmer in identifying suitable plants based on their unique farming conditions and requirements.

When it comes to data, it's not just about volume, but also the rate at which it changes. To manage this, we propose a flexible and scalable method for replenishing the search knowledge base. This approach ensures the database is always up-to-date and capable of delivering the most relevant results to the farmer.

Finally, our solution aims to enhance business profitability through a sophisticated ranking algorithm that leverages additional features. This algorithm is designed to provide user-specific recommendations, thereby promoting more informed decision-making and consequently boosting profit margins.

In contrast to other approaches, our solution combines a sophisticated attribute-based search, dynamic database management, and a customized recommendation algorithm. This unique blend ensures a more personalized, efficient, and profitable farming experience, addressing the pressing issue of import reliance head-on.

1.1 Team

Dmitry Stepakov involved in data parsing, making use of tools like BeautifulSoup and regular expressions to gather crucial features. This member also worked on formulating the search algorithm and developed a post-ranking model, focusing on the popularity of the crops. His role culminated in creating the user interface, ensuring the solution was accessible and user-friendly.

Ivan Tsvetcov was tasked with feature extraction, leveraging fuzzy matching to maintain precision while dealing with a unstructured data. They also introduced the ArgoBERT classifier, improving attributes classifier problem. His role also included parsing certain features, aiding in enriching our database.

Both members, with their contributions, have strived to develop a solution aimed at supporting farmers in their medicinal plant cultivation efforts.

2 Related Work

At the moment, there is no project in the world that has implemented our task. But there are several similar and related projects.

[learning for LC–MS medicinal plants identification, 2016] Herbal medicines are vigorously marketed, but poorly regulated. Analysis methodology for this field is still forming. One particular analytical task is confirmation of plant species identity for medicinal plants used as ingredients. In this work, machine learning approach has been implemented for LC–MS plant species identification. Samples for 36 plant species have been analyzed. Peak data (m/z, abundance) from respective samples have been used for development of classification algorithms. Namely, logistic regression (LR), support vector machine (SVM) and random forest (RF) techniques were used. For most of used machine learning algorithms, classification accuracy of 95% higher were obtained on cross-validation dataset. Now, massive training datasets are needed for full-scale application of this approach. This work only analyzes a part of the plants according to the available collected data.

[learning to extract specific information from unstructured texts, 2018] Modern linguistic models (ULMFiT, Elmo) use unsupervised learning methods, such as creating RNN attachments in large text corpora, in order to gain some primary "knowledge" of language structures before a more specific controlled stage of learning. In some cases, on the contrary, you need a model trained on a very specific and small data set. These models have almost zero knowledge of general language structures and work only with special text functions. A classic example is a naive sentiment analysis tool for movie reviews or news datasets -

the simplest working model can only work with synonyms of "good" or "bad" adjectives, and some emphasize the presence of words. In this study, the authors used the advantages of both approaches. This approach solves the problem in general, we were inspired by some ideas, but adopted them for ourselves.

3 Model Description

Our approach to this challenge is rooted in designing an interactive and personalized user experience, integrating data science techniques for effective search and ranking, and maintaining an agile database system. The target architecture is shown on Fig. 1.

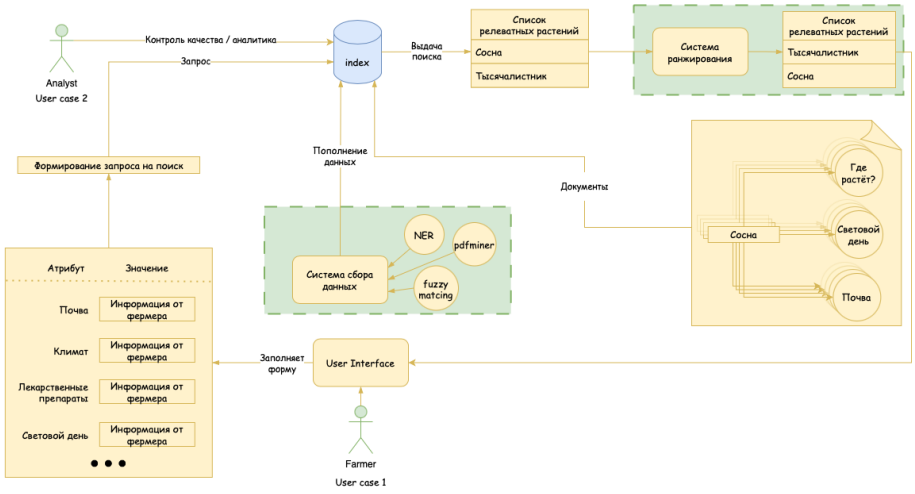

Figure 1: Solution architecture.

Firstly, the user is presented with a set of attributes to fill out, encompassing soil type, climate conditions, and specific plant characteristics. These features provide the baseline inputs to our system, enabling us to generate a personalized request. Once the user provides the necessary attributes, we formulate a query to our database to fetch the relevant documents.

The resultant documents then undergo a ranking process using a model that orders them based on certain business features. For instance, in our solution, the model ranks the documents according to the number of medicines in which the particular plant is used. This ranking not only ensures the relevance of the returned documents but also aligns with the user's potential business interests.

As for the knowledge base or the database, two important considerations govern its design. First, we've developed scalable and flexible data collection and update methods to swiftly replenish the database, accommodating the dynamic nature of the information landscape. Secondly, we've made the database accessible not just to the users via search queries but also to the company's

analytics team for analysis and updates. This dual-accessibility feature allows for a holistic utilization of the data resource.

The technical stack that powers our solution includes Python and its open-source libraries. To convert PDFs into text, we use pdfminer. To extract structured information from Russian text, we employ natasha. Lastly, to enable fuzzy matching for non-precise search terms, we utilize a library aptly named fuzzy. These tools collectively facilitate the creation of a user-friendly, personalized, and business-oriented solution to assist Russian farmers in their medicinal plant cultivation endeavors.

4 Dataset

Our dataset is an amalgamation of data drawn from various sources, processed and structured to serve our search and ranking tasks. Here's a detailed breakdown of our data extraction process and the attributes we managed to fill:

First, we sourced data from the comprehensive "Atlas of Medicinal Plants of Russia" [atl, 2021]. This detailed guide on flora and nature conservation served as a primary source of information about different plant attributes.

Secondly, we supplemented our dataset with data from the State Register of Medicinal Products. This provided valuable insights into the various medicinal substances and their chemical compositions, enabling us to map specific plants to the medicines they contribute to.

Once we had our sources, we performed a thorough extraction process. The reference book was divided into various sections, each dedicated to a different plant and its attributes. We separated the text based on the headings and subheadings, ensuring each attribute of the plant was correctly extracted. Following this, we lemmatized the text within the attributes, preparing it for further parsing.

The parsing tools used include Natasha, which helped extract soil information, and DeepPavlov, which assisted in deriving location data. This rigorous process led to a comprehensive list of attributes, each with a varying degree of completeness across the dataset (Table. 1).

5 Experiments

5.1 Search engine

5.1.1 Metrics

We used two primary metrics to evaluate the efficiency and accuracy of our search algorithm: Average Hit Rate at 5 (Avg Hit@5) and Average Relevance at 5 (Avg Relevance@5).

The table. 2 illustrates the metrics and their corresponding values.

1. Average Hit Rate at 5 (Avg Hit@5): This metric evaluates the accuracy of our search algorithm in terms of how often the correct item (plant entity)

Attribute Name	Description	Percentage of Documents with Attribute
location_feature	Location of the plant	97%
climate_feature	Climatic conditions suitable for the plant	99%
red_book_feature	Indicator of inclusion in the Red Book	99%
chemicals_feature	Chemical substances present in the plant	74%
source_type	Type of raw material the plant serves as	17%
calendar_month	Ideal months for harvesting the plant	17%
max_type	Maximum storage period for the plant	17%

Table 1: Dataset Attributes and Completeness

Metric	Value
Avg Hit@5	0.601
Avg Relevance@5	0.890

Table 2: Performance metrics

appears within the top 5 search results. In our experiments, the average hit rate was approximately 0.601, implying that our model correctly identified the desired plant in the top 5 results in about 60.1% of cases.

2. Average Relevance at 5 (Avg Relevance@5): The Avg Relevance@5 metric assesses the positional relevance of our search results. A score of 0.890 indicates that the correct item is not only likely to appear in the top 5 results, but it's also likely to be ranked highly within those results.

Both these metrics together paint a comprehensive picture of our system's effectiveness at accurately identifying and appropriately ranking the correct plant entities based on the user's input attributes.

5.1.2 Experiment Setup

At the core of our experimental design, we have a procedure that randomly selects a plant entity from our database and subsequently extracts a subset of its associated features. This process is crucial as it mimics the user's input, creating various combinations of plant attributes for our search mechanism to work with.

The search algorithm operates on the basis of a user input which includes a combination of plant attributes. This algorithm operates by comparing the input attributes against the attributes of all plants in our database.

For each plant, the algorithm determines the degree of matching between the input attributes and the plant's attributes. This is achieved by calculating the intersection of the attribute sets. To provide a fair comparison across different attributes, each intersection score is normalized by the total number of input values for the corresponding attribute.

This forms the first part of the plant's score. The second part of the score comes from the popularity of the plant, reflecting user preferences and trends in the database.

The final score for each plant is computed by combining the attribute matching score with the popularity score. This ensures that the results are relevant (high attribute matching score) and reflect user preferences (high popularity score).

Finally, the algorithm sorts all plants by their final scores and retrieves the top N matches, providing a concise list of highly relevant and popular plant suggestions based on the user's input.

Throughout the experiments, the top N results were kept to evaluate the system's performance, with N defaulting to 5. This mirrors a realistic scenario where a user is presented with a handful of top matches for their query. The experiments were designed to be reproducible, allowing for the model's continuous evaluation and improvement.

To guarantee that the experiment was representative and the findings valid, we used a comprehensive and diversified dataset gathered from trusted sources, as described in the previous section.

5.2 AgroBERT

5.2.1 Metrics

The performance of the model was evaluated using Accuracy as the primary metric (Table 3). Accuracy is a widely accepted measure for classification problems and is calculated as the number of correct predictions made by the model divided by the total number of predictions. Confusion matrix presented on Fig. 2.

Metric	**Value**
Accuracy	0.77

Table 3: Performance metrics for BERT-based model

5.2.2 Experiment setup

The model was fine-tuned using a RuBert-Tiny pre-trained model on the texts extracted from the Atlas of Medicinal Plants of Russia. This process resulted in a model that accepts text input (up to 512 characters) and predicts the attribute class to which the text relates.

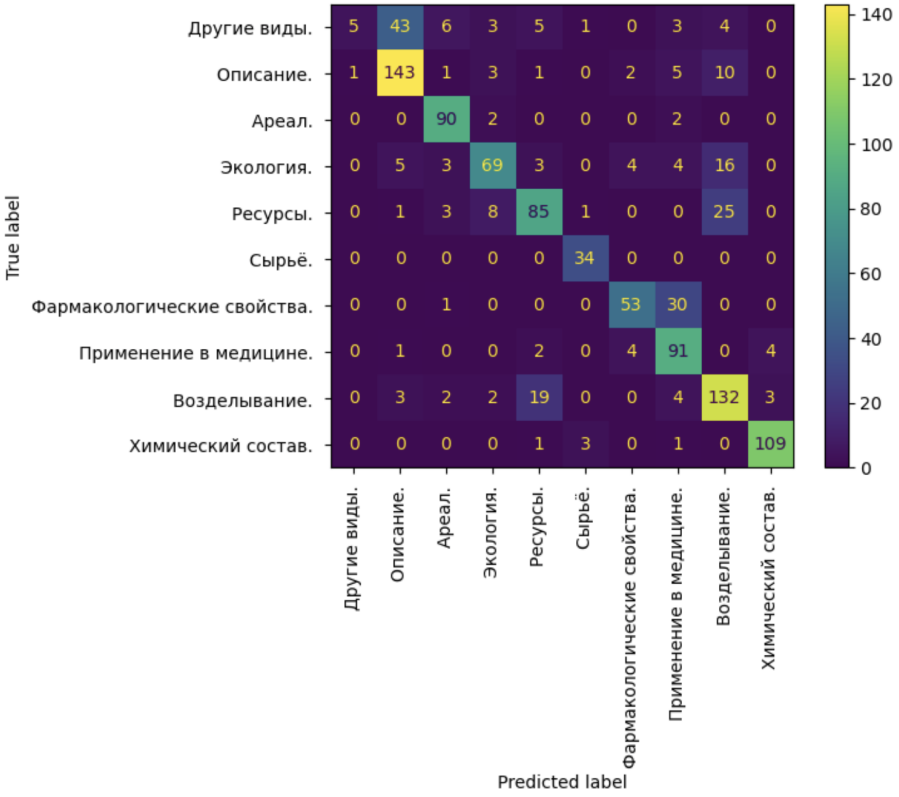

Figure 2: AgroBERT: Confusion matrix.

Moreover, the trained model was used for predicting the corresponding sections of the Atlas on the text data obtained from the internet. This approach enables us to quickly identify the data from which relevant features need to be extracted, increasing the efficiency of the process and allowing for rapid updating of our knowledge base.

The high accuracy achieved by the model showcases the efficiency of fine-tuning pre-trained BERT models using the transformers library in dealing with such classification tasks. This approach significantly enhances the feature extraction capabilities of our system, ultimately enriching our knowledge base and improving the precision of our search results.

6 Results

The comprehensive solution we have developed has proven its utility by producing promising results. Leveraging a meticulously curated dataset, a sophisticated search algorithm, and an advanced BERT-based feature extraction

model, our solution was able to effectively match user requirements with appropriate medicinal plants, facilitate efficient data updates, and optimize recommendations based on business characteristics.

Our work stands as a testament to the potential of integrating advanced NLP and machine learning techniques to address the challenges faced in the agricultural and medicinal sector. The solution excelled in delivering accurate and relevant results while maintaining scalability and flexibility.

One of the most significant milestones for our project was its performance in a competitive setting. Our solution was field-tested in a hackathon, a high-intensity event that places stringent demands on the efficiency, accuracy, and practicality of proposed solutions. Despite stiff competition, our approach distinguished itself, securing the 8th place out of 28 participating teams.

This accomplishment speaks volumes about the effectiveness of our solution and its potential in real-world applications. It also underscores the relevance and urgency of the problem we are tackling - helping farmers grow raw materials locally, which has wide-reaching implications for the economy and the environment.

The achievement in the hackathon also serves as a valuable feedback and motivator for future work. It shows that we are on the right track and encourages us to continue refining our approach, expanding our dataset, and exploring new ways to enhance the accuracy and utility of our system.

The successful deployment and recognition of our solution at the hackathon demonstrates the practicality and efficiency of our approach. We believe that with further refinement and expansion, our system can make significant contributions to the farming industry and ultimately help in the transition towards self-sustenance in raw material production.

7 Conclusion

We offer a new solution in the field of agricultural technologies - an innovative service that will allow farmers to choose suitable medicinal plants for planting on their land. Our unique information retrieval system includes a carefully designed algorithm that uses only a few key parameters, such as location, climatic conditions, as well as the required characteristics of plants, including the harvest period and shelf life, to provide the most suitable options.

The uniqueness of our solution is a full-fledged information search system that not only searches for suitable plants, but also offers to choose those that are most in demand on the market due to the ranking model. We have also automated data collection, so the knowledge base and the search itself can be quickly replenished with new plants and their characteristics.

References

[atl, 2021] (2021). *Atlas of Medicinal Plants of Russia*.

[learning for LC–MS medicinal plants identification, 2016] learning for LC–MS medicinal plants identification, M. (2016). Medicinal plants identification. *Chemometrics and Intelligent Laboratory Systems*, 6:174–180.

[learning to extract specific information from unstructured texts, 2018] learning to extract specific information from unstructured texts, D. (2018). Extracting specific information. *Machine Learning Mastery*, 4.