Kisan Query Analysis

CS685: Mid-term Project Report (Group 1)

Aniket Sanghi
Roll no: 170110
sanghi@iitk.ac.in

Neil Shirude
Roll no:170429
neilrs@iitk.ac.in

Paramveer Raol
Roll no:170459
paramvir@iitk.ac.in

Sarthak Singhal
Roll no:170635
ssinghal@iitk.ac.in

Aman Thakur
Roll no:170083
amankt@iitk.ac.in

1 Datasets

We are using datasets containing various queries asked by Indian farmers in the Kisan Call Centre (KCC) from various districts (regional call centre).
Some of the important fields in the datasets are

- Category
- Query Text
- KCC Answer
- State Name
- District Name
- Date

We will also be using the neighbor-districts mapping and state-district mapping for analysis at state-level and neighbor-level for the various parameters involved.

2 Data Retrieval

All datasets required for the project are assembled at this government website. We are scraping the data using selenium in python.

```
https://data.gov.in/resources-from-web-service/6622307
```

There are about 80,000+ datasets on this portal where each dataset corresponds to various queries asked by farmers related to various fields in a particular district over a period of one month. The datasets in the portal are updated every day.

3 Broad Aims

1. **Generation of FAQs** (Frequently Asked Questions) based on the relevant queries: The target will be to formulate region specific and overall most relevant FAQs with limited data, covering all major aspects.

2. **Analysis of effectiveness of government schemes**: The target is to find a pattern between schemes effectiveness and the relevant queries asked regarding the scheme. Then we aim to determine which policies need more awareness.
3. **Analysis of issues pertaining to plant protection** based on relevant queries: The objective would be to outline which plants and specifically which diseases were the most bothersome for farmers.
4. **Analysis of weather related queries** based on the queries asked about weather. The goal will be to summarize in which areas and in which months, the farmers are more concerned about the affect of weather on their crops.
5. **Online multi-label classification of queries**: The aim is to classify queries of farmers based on the query's context and simulate a real-time classifier that will be helpful in the assignment as well as in real life applications. The queries are related to Nutrient Management, Plant Protection, Weather, Market Information, Fertilizer Use and Availability, etc. The data can also be classified based on agriculture and horticulture related queries.

4 Methodology

1. **Generation of FAQs**
   - Clustering Queries into various sets based on similarity in query objective
   - Assigning relevant weights and finding out the weighted frequency for each set. Queries more recent and having more relevance will be given more weight
   - At the end relevant cut-offs will be chosen to formulate concise FAQs covering all dimensions from crop to government schemes
2. **Analysis of effectiveness of government schemes**
   - Classification of queries based on different government schemes.
   - Find the number and context of queries for each scheme for every region.
   - Decide relevant cutoffs and health matrices to get effectiveness of the schemes at state-level and neighborhood level.
3. **Analysis of issues pertaining to plant protection**
   - Group queries for each crop with the possible disease queries asked by farmers
   - Analyse region specific crop problems encountered by grouping data based on neighborhood and state.
   - The query frequency would be normalized and the significant (crop, disease) pairs would be determined by deciding cutoff frequencies to get diseases and problems having significant impact on the farmers.
4. **Analysis of weather related queries**
   - The normalized count of the number of queries corresponding to each state and each month related to weather will be calculated.
   - Thresholds will be decided to pick only more concerned states and the corresponding months, and a summarised set of relevant careful advice will be formulated.
5. **Online multi-label classification of queries**
   - Use of key-word searching for labelling queries (creating multi-hot vectors for each query denoting the presence or absence of keywords)
   - Use of large language models such as BERT to capture the context of a given text and extract semantic features
   - Methodology 1: Training 2 separate multi-class classification models on semantic features and keyword vectors of query text-
     (a) Horticulture vs Agriculture
     (b) Nutrient Management vs Plant Protection vs Weather vs Market Information vs Fertilizer Use and Availability vs Others
   - Methodology 2: Training 1 multi-label classification model that performs both tasks.

Find other patterns wherever possible

5 Results Expected

1. **Generation of FAQs**
   - The FAQs will be concise report of most important/relevant queries
   - The obtained data will help plot out the critical areas needing more information spread
2. **Analysis of effectiveness of government schemes**
   - The analysis will render the schemes which have been less queried and which have been more queried. This will be used to quantify the awareness and effectiveness level of the scheme.
   - The inferences obtained can be used to determine which schemes require more resources, which schemes are effective and for other schemes' related logistics.
3. **Analysis of issues pertaining to plant protection**
   - The analysis will provide details of most prevalent and devastating diseases for each crop across regions.
   - The obtained data will help find out the plant varieties needing more attention.
4. **Analysis of weather related queries**
   - The analysis will give the idea of the states and on which months the crops are more prone to being affected due to weather.
   - Using this the authorities can inform the concerned farmers about the weather conditions beforehand which can help them plan the farm production schedule, improve the productivity, prevent losses, protect their health, etc.
5. **Online multi-class classification of queries**
   - The online classification system can be used to assist Kisan Call Centre executives in entering query data.
   - This system can eliminate the errors in data entry, reduce the processing time for each query, resulting in increasing the efficiency of call centres.
   - The system potentially can be further expanded to classify farmers' queries in real-time and redirect them to the relevant Call Centre executives. This will eliminate the need for digital assistants, which are intrinsically difficult-to-use for uneducated people.

6 Evaluation Mechanism

1. **Generation of FAQs**
   - Relevance of the algorithms and statistics used can be evaluated
   - Testings on small sample of data where the FAQ obtained is easier to evaluate
2. **Analysis of effectiveness of government scheme**
   - Check the correctness of the procedures and statistics used.
   - The obtained results and patterns can be subjectively evaluated based on relevance.
3. **Analysis of issues pertaining to plant protection**
   - Relevance of algorithms and statistics used can be evaluated
   - Small sample of data can be used for testing
4. **Analysis of weather related queries**
   - The relevance of the obtained results can be evaluated.
5. **Online multi-class classification of queries**
   - The dataset available on the website contains the class for each sentence. But, we believe that this class information is not collected in real time, but is manually entered later.
   - The class information present in the dataset can be considered as the true values and accuracy, precision and recall of the model can be calculated for evaluation.
   - There are some errors in the classes in the data available on the website. An error-free test dataset will be manually created and classification scores can be evaluated on this dataset.

References

- https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/#3-extracting-embeddings
- https://data.gov.in/resources-from-web-service/6622307