CS661: Big Data Visual Analytics
Project Proposal Report

AgriViz: Analysing Farmer Queries with Data

Siddharth Kalra 211032
Aniket Suhas Borkar 210135
Chitwan Goel 210295
Kartik Anant Kulkarni 210493
Geetika 210392
Vikas Yadav 211166
Divyansh 210355
Shrey Bansal 210997

Group 3

1 Motivation and Introduction

Agriculture in India employs 45% of India's workforce and contributes to ∼18% of GDP, yet farmers continue to face challenges related to weather, market prices, and government policies. A large section of Indian farmers still rely on informal sources of information for crucial agricultural advice, often leading to misinformation and financial losses. Kisan Call Centers offer a toll-free helpline for farmers to obtain agricultural information. However, this data exists in raw, unstructured formats. By structuring and visualizing this large-scale Kisan query data and combining with the available price data, we can provide valuable insights into the issues faced by farmers of India and empower policymakers and agribusinesses to make data-driven decisions.

2 Data Source and Description

We propose to use 2 different data sources as follows:

Dataset 1: Kisan Call Center Query Data *(Link)*

This dataset contains queries made to Kisan Call Centre. The historical data for the past 15+ years is available on the Indian Government Website with attributes such as Location, Season, Sector, Category and Crops with the query and the provided response. The raw dataset contains more than 4 million data points.

Dataset 2: Crop Price Data *(Link)*

This data set contains crop price and crop production volume data for past 10 years for different crops in different markets of the country. It is provided with attributes such as locations, variety, group, arrivals and price with the time variation. The raw data set has more than 10 million data points.

3 Specific Tasks

1. **Data Collection**: The datasets need to be scraped from the websites mentioned above.
2. **Data Wrangling**: The data has missing values and needs to be cleaned appropriately.
3. **Machine Learning Analytics:** Apply NLP and clustering algorithms for segregating query data into themes and clusters.
4. **Visualization:** Create an interactive dashboard with various plots as detailed in the following section.
5. **Analysis and Insights:** Perform a thorough analysis, deriving insights from the visualizations into the various farmer concerns and suggest possible measures that could be implemented to make the system more effective.

4 Overall Solution

In this project, we will create a **web-based dashboard** for query and price analysis in the **domain of agriculture**. The interactive visualizations will be prepared using the **d3.js** library highlighting farmer concerns, for e.g., geospatial heatmaps of query distributions, time-series analysis of seasonal issues, etc. The **Python**-based backend will provide support for NLP-based topic modeling and help identify trends in queries. The dashboard will provide options for grouping concerns regionally and temporally, while other dynamic charts will analyze the market price fluctuations and farmer queries. These insights can help government officials detect early signs of agricultural distress and enable market regulators to stabilize crop pricing policies.

5 Visualization Description

Our project will include visualizations broadly under the following categories:

1. **Choropleth Map:** It is a thematic map that uses colors and symbols to visualize the data values across a geographic area. Since both datasets involve geographic variation, we can map the data values to the corresponding geolocations, with appropriate granularity. Further, these charts would also be used for overviewing and dynamically filtering into corresponding time series (eg: price, query frequency, etc.).
2. **Clustered query Categorization and Visualization:** Categorize queries into themes (e.g. government schemes, pricing, plant protection, etc.) and visualize them using tree maps and pie-charts corresponding to filtered datasets (interactively).
3. **Word Cloud and Topic Modeling:** Display frequently occurring words from queries using NLP in order to gain insight into the farmer concerns.
4. **Time Series Visualization:** Price analytics and Query frequency distribution evolution with time will be plotted with a filter-and-zoom methodology, to pinpoint relevant parts of the data and derive insights.
5. **Correlation Plots:** Plot correlations between attributes for both datasets. For eg: Price of a particular crop across different districts, price of different crops in the same regions, etc.

6 Work Distribution

Following is the tentative assignment of roles and responsibilities within our team; however, as deemed necessary, we may alter the distribution, while ensuring fair and balanced workload.

1. **Chitwan Goel -** Backend Development and Machine Learning
2. **Kartik Anant Kulkarni -** Machine Learning and Visualization Module
3. **Siddharth Kalra -** Frontend Development and Machine Learning
4. **Aniket Suhas Borkar -** Database Management and Application Hosting
5. **Shrey Bansal -** Database Management and Application Hosting
6. **Geetika -** Backend Development and Application Hosting
7. **Divyansh -** Frontend Development and Visualization Module
8. **Vikas Yadav -** Frontend Development and Visualization Module