5-2023

TREES RECOMMENDATION IN AGROFORESTRY ECOSYSTEM USING NLP

Omkar Sawant

TREES RECOMMENDATION IN AGROFORESTRY ECOYSTEM
USING NATURAL LANGUAGE PROCESSING

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Information Systems and Technology

---

by

Omkar Prakash Sawant

May 2023

TREES RECOMMENDATION IN AGROFORESTRY ECOYSTEM
USING NATURAL LANGUAGE PROCESSING

A Project

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Omkar Prakash Sawant

May 2023

Approved by:

Dr Sepideh Alavi, PhD, Committee Chair

Dr Conrad Shayo, PhD, Committee Member

ABSTRACT

Agroforestry farming is one of the challenging sectors to grow crops or farming of variety of trees from the ancient days due to erosion and desertification. This Culminating Experience Project explored how recommendation systems can be developed and used in agroforestry. The research questions are Q1. What methods can be used to improve the accuracy and reliability of soil-based agroforestry tree species recommendation systems? Q2. How can agroforestry tree species recommendation systems be tailored to the needs of different stakeholders, such as smallholder farmers or agribusinesses? Q3. What will be the top three, tree recommendations using natural language processing based on varying soil content? Data was collected from two datasets the Agroforestry Database and the European Commission's extension of the periodic Land Use/Land Cover Area Frame Survey. The findings are: 1) Various Natural language processing techniques such as cosine similarity, count vectorization, and TF-IDF can significantly enhance the system's ability to analyze and process large amounts of Data collection, validation, and monitoring to improve the accuracy and reliability of soil-based agroforestry tree species recommendation systems. 2) Cosine similarity achieve to recommend tree species based on soil test report data collected by the European Commission's extension of the periodic Land Use/Land Cover Area Frame Survey and tailored based on various soil properties helps the smallholders, stake holders, farmers to best decisions to increase their growth. 3) Natural language processing techniques such as cosine similarity, count vectorization, and TF-IDF can be employed to

analyze soil data and identify the tree species that are most appropriate for different soil types. The conclusions are: 1) The integration of text-based processing and machine learning methods has significant potential to refine agroforestry tree species recommendation systems to generate reliable and acceptable recommendations. 2) The system's recommendations can become more relevant, practical, and acceptable, leading to higher adoption rates and better outcomes.3) Develop The proposed agroforestry tree species recommendation system provides top three trees' recommendations using cosine similarity, TFIDF and Count vectorization techniques. Furthermore, areas for future research that emerged from this study include the need to improve the sustainability and productivity of agroforestry practices, enhance ecosystem services, and promote economic, social benefits and identify additional strategies for improving the accuracy and reliability by getting additional feedback about the tree's recommendation from the stakeholders and farmers directly in design and development.

DEDICATION

For my beloved father

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Agroforestry refers to land-use systems that include trees along with crops, animals, and/or other vegetation on the same plot of land which unites to provide numerous outputs while protecting the resource base (Pantera et al., 2021). Farmers who practice agroforestry cultivate not only crops but also trees and are encouraged to create a small-scale forest (Rosati et al., 2021). The National Resources Conservation Service (Soil health) of the United States Department of Agriculture defines soil health as the soil's continued capacity to serve as an immensely essential ecosystem that supports plants, animals, and people. Due to soil erosion and desertification, the productivity of some fields has decreased by 50% (Eswaran et al., 2019).

The study finds that over the past few decades, a variety of anthropogenic activities have considerably degraded 25% and moderately affected 50% of the world's soil, respectively (Srivastava et al., 2019). Agroforestry is beneficial to the environment as it contributes to soil fertility (Ramos et al., 2015), soil quality, increased carbon sequestration, and improved agroecosystem quality (Ospina et al., 2017). It is also economically beneficial as it provides farmers with multiple sources of income such as firewood, timber and fodder throughout the year.

Agricultural recommendation systems based on artificial intelligence/machine learning are growing in popularity as they can process large amounts of data and make more accurate predictions (Priyadharshini et al., 2021). The development of agriculture recommendation systems is gaining popularity is propelled by the rising need for data-driven decision-making in agriculture ecosystem and the requirement to mitigate the challenges faced by farmers and agribusinesses, such as increasing production costs, declining crop yields, and increasing soil erosion (Patil et al., 2020). Natural language processing (NLP) is a division of artificial intelligence that studies how computer and human language interact (Kumar et al., 2021). NLP algorithms can be used to extract information and insights from large amounts of unstructured text data such as agricultural reports, scientific articles, and farmer surveys (frontier).

Problem Statement

Despite the many benefits of agroforestry, many small-scale farmers lack the resources and knowledge necessary to adopt and effectively implement these practices (Pantera et al. 2021; Jawo et al. 2023). As a result, we miss opportunities to improve soil health, increase yields and reduce the environmental impact of agriculture. Lack of support for agroforestry farmers weakens the potential of agroforestry to chip into sustainable agriculture and food security (Iyabano et al. 2023). Therefore, we propose a natural language processing recommendation system that examines data about soil and outputs tree recommendations tailored to soil type, soil content, and farmers' needs

(Gram et al., 2018).

Research Questions

To gain a better understanding of the overall project needs, the following questions will be answered:

1. What methods can be used to improve the accuracy and reliability of soil-based agroforestry tree species recommendation systems?
2. How can agroforestry tree species recommendation systems be tailored to the needs of different stakeholders, such as smallholder farmers or agribusinesses?
3. What will be the top three, tree recommendations using natural language processing based on varying soil content?

This project is organized as follows: Chapter 2 reviews the literature. Chapter 3 discuss the research methods used to answer each question. Chapter 4 is a discussion of data collection, analysis and findings. Lastly, Chapter 5 includes the discussion, conclusion, and areas for further study.

CHAPTER TWO

LITERATURE REVIEW

Barrios et al., (2018) reviewed six studies of agroforestry and tree cover transitions from three continents to explore the potential for tree integration into agroecosystems and sustainable land use. was investigated. This result indicates that increased tree cover can significantly improve ecosystem function and contribute to the conservation of biodiversity and ecosystem services (Barrios et al., 2018). However, adverse effects have also been identified in certain contexts, highlighting the need to understand which tree species to include and how to manage them in different socio-ecological contexts. The research also highlights the potential of natural language processing techniques to enhance the accuracy and reliability of soil-based agroforestry tree species recommendation systems through effective data collection, validation, and monitoring. To ensure successful agroforestry implementation, local and scientific knowledge must be combined, and tree-based management options adapted to the context to minimize trade-offs and maximize synergies and complementarity knowledge-exchange needs to be done. The author's research applies to this project to some extent. However, their research is less focused on natural language processing techniques using recommendation systems. There is no direct research related to the use of natural language processing methods such as cosine similarity, TF-IDF and count vectorization leveraged for building recommendation systems in the agroforestry domain.

Therefore, we proceed with the research question and the review of the literature that serves the project's goal of building a recommender system for recommending trees in the field of agroforestry.

Gram et al., (2018) developed agroforestry species recommendations tailored to the farmers' need, climate conditions, local context and soil properties. The research used Bradley-Terry approach to rank tree species by ecosystem services along with the local knowledge and farmers preferences customized to recommend tree species in the agroforestry domain for smallholder farmers.

Agriculture is a major industry in India, employing over 58% of the labor force. The failure of farmers to select suitable crops based on soil conditions, sowing time and geographical location has led to serious problems such as suicide, abandonment of agriculture, and relocation to urban areas using traditional and unscientific methods. leading to results (Ahmad et al., 2017). The research focuses on machine learning which considers many environmental parameters and soil properties and can assist farmers in making informed crop selection decisions (Kumar et al., 2015). Precision agriculture, along with current agricultural technologies such as artificial intelligence and machine learning, can expedite utilization, enable data-driven decision-making, and increase crop production, thereby supplying to a country's socioeconomic growth (Priyadharshini et al., 2021).

Brandt (2023) used text mining to evaluate the potential of machine learning to identify restoration policies across categories such as agroforestry, soil erosion, and forest conservation. They achieved this F1 score of 0.83 measured on 14 policy agendas of 31 policy documents using an unsupervised neural search machine learning architecture and cosine similarity and TF-IDF techniques. Restoration Policy (Brandt, 2019).

Muhirwa et al., (2022) investigated the implementation of a water, energy and food (WEF) nexus process in Africa and found that despite growing interest, implementation is limited. A study of 237 WEF-related application projects (demonstrating the empowerment of owners in the agriculture and food industries, the empowerment of agroforests, and the promotion of the agro-silvo-pastoral system) identified three resource sectors that fit the nexus approach: Only 26 initiatives were identified. As per the statistics of term frequency and inverse document frequency (TF-IDF), the word "production" was particularly significant and reliable in adopting the nexus strategy. Using cosine similarity technique, a low similarity score of 0.25 was obtained. The study also demonstrated an imbalance in comprehending and applying nexus notions between academic articles and project implementers reports. Effective policies, creative techniques, and coordinated efforts from different stakeholders are required to achieve successful implementation.

Cosine similarity, Pearson correlation, and Euclidean distance are the most frequently used similarity metrics in recommender systems (Schwarz, et.al, 2017). Pearson correlation is commonly used in rating-based systems, whereas cosine similarity is popular in text-based systems (Islam, et.al, 2022).

Euclidean distance is a simple measurement used in content-based recommendation methods (Schwarz, et.al, 2017). Cosine similarity is a very famous similarity distance measure between the two vectors (Sitikhu, et.al, 2019). Computes the cosine of the angle between two vectors in high-dimensional space and note downs, how similar they are with respect to direction (Peipei, et.al, 2015).

Regarding text classification models, Anjuma et al. (2021), Wendland et al. (2021) compared the impact of feature extraction and text vectorization techniques such as count vectorization and TF-IDF on machine learning algorithms. The results obtained showed better performance and accuracy when TF-IDF was used as the vectorization technique.

Lovedeep et al., (2022) compared the performance of text clustering applied following two vectorization techniques TF-IDF and Doc2Vec. They performed their experiments on a set of tweets and obtained better reliable results with TF-IDF as a vectorization technique (Lovedeep et.al, 2022).

Another research used a mixed methods approach to assess farmers' perceptions of agroforestry, barriers to implementation, and welfare impacts. Results show that agroforestry farmers report significant improvements in 8 of the 18 indicators tested. A major impediment to implementation is the lack of political support and initial investment. They seek coordinated strategies and funding to integrate smallholder farmers and interdisciplinary collaborations and leverage existing expertise to understand the strains of rural communities in forest landscape restoration strategies (Shennan et al., 2022).

A study highlights the importance of agriculture to the Indian economy and how uninformed sowing decisions of farmers can adversely affect local economic conditions and the farmers themselves. To address this issue, the authors propose applying a systematic machine learning model using a dataset combining rainfall, climate, soil properties and fertilizer data from India. Using machine learning models, the authors aim to recommend the top best crops for a given season based on various criterion such as farm location, sowing time, soil properties, and climate which can help Indian farmers in making informed decisions about their crops and avoid the financial and emotional consequences of wrong decisions. The authors emphasize on the importance of using correct machine learning models to solve problems and improve farming practices in India (Patil et al. 2022).

Different algorithms have given rise to various machine learning models used for crop prediction based on environmental determinants like soil characteristics, temperature range, precipitation stats and geographical situation (Priyadharshini et al., 2021).

However, because tree species affect soil properties that can affect crop yields in agroforest ecosystems, farmers and agroforesters interested in maintaining or enhancing site productivity and reducing soil erosion is an important and pressing concern for Planting trees without considering soil properties can lead to reduced crop production in agroforestry ecosystems.

Therefore, building on these previous studies, we continue our research goal in this project to build a machine learning model that uses natural language processing to recommend trees based on soil properties.

CHAPTER THREE

METHODOLOGY

We introduced the following natural language processing methods used for agroforestry tree species recommendations: Finally, we used cosine similarity to advance both count vectorization and word frequency (inverse document frequency).

Q1 What methods can be used to improve the accuracy and reliability of soil-based agroforestry tree species recommendation systems?

To answer this question, the following methods will be used.

Count Vectorization

Count vectorization is a text pre-processing method that converts numerical characteristics from text data. Count vectorization can be applied to plant species names in the setting of agroforestry (Temgoua, et.al, 2018). Developing a list of different plant species names represented as a vector of numbers representing the frequency of each term in the vocabulary (Soons et al., 2016). The resulting vector matrix is known as the document-term matrix.

Term Frequency Inverse Document Frequency (TF-IDF)

The TFIDF methodology is a method for judging the importance of terms in a text. TFIDF is an acronym for Term Frequency-Inverse Document Frequency (Bruno et.al, 2014). The TFIDF number is determined by multiplying the word frequency (TF) by the inverse document frequency (IDF). The frequency of a term in a document is expressed as TF, while IDF is the logarithm of the total number of documents divided by the number of documents containing that term (Donghwa et.al, 2019). The tf-idf method can be used in the context of agroforestry to weight the significance of plant species in document term matrices (Shin et al., 2021).

Q2 How can agroforestry tree species recommendation systems be tailored to the needs of different stakeholders, such as smallholder farmers or agribusinesses? To answer this question, we created new features based on different farmers & stakeholder factors.

Separated the soil into various texture groups based on the values of sand, clay, and silt, denoted as i, j, and k, respectively. For example, powdery soil is defined as having mostly sand and very little sediment and clay ($k + 1.5*j < 15$). Similarly, loamy soil is defined as having approximately equal quantities of sand, silt, and clay. Various plants or trees have different soil texture needs, it is critical to comprehend the texture of the soil to select the right trees or management practices (Charles et.al, 2009; USDA, 2017).

Based on the pH in CaCl2 values, added eight new categories range from very acidic to very alkaline. Here are two features among eight classes. In general, CaCl2 pH greater than 5 and less than 8, is the ideal soil pH range for plant growth. When CaCl2 has a pH above 4 and below 5, this is generally suited for very acid-sensitive crops such as some legumes and barley, and more acid-tolerant plants continue to be adversely affected (Belinda, 2000).

Created two new categories that characterize the soil's nitrogen concentration based on N (nitrogen), P (phosphorus), and K (potassium) numbers. Nitrogen is an important nutrient for plant development that can affect food production, quality, and soil health Four new categories represent soil richness ratios based on N (Nitrogen), P (Phosphorus), and K (Potassium) values. Soil fertility ratio is a measure of the proportional availability of nutrients in soil that can be used to predict agricultural development and yields (Wassen et al., 2003).

Q3 What will be the top three, tree recommendations using natural language processing based on varying soil content?

To answer this question, the following method will be used.

Similarity measures

Similarities like Euclidean distance that measures the distance between two vectors in Euclidean space and Pearson correlation coefficient that measures the linear correlation between two vectors are commonly used in collaborative filtering recommendation systems (Fkih, 2022).

Jaccard similarity measures the similarity between sets of items and is commonly used in recommendation systems that deal with binary data, such as user-item ratings (Scott et.al, 2022).

Cosine Similarity

Cosine similarity is commonly applied in NLP applications such as text classification, information retrieval, and recommendation systems (Alanzi et al., 2019). It can be used to measure the similarity between two text documents, or two words based on their word embeddings, and to identify documents or words (Xin et al., 2016).

The working of cosine similarity involves the following steps:

Vector Representation: Entities to be compared, such as documents and words, are represented as vectors in a high-dimensional space (Jaegul et.al, 2013). Each dimension of the vector corresponds to a particular feature or aspect of the represented entity.Compute cosine similarity: Cosine similarity between two vectors is computed as the dot product of the two vectors divided by the product of their magnitudes (Sunita, et.al. 2004; Alfirna et.al. 2016).

Mathematically, it can be expressed as:

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 1: Cosine similarity formula (Lahitani et al., 2016)

where x and y are the two vectors, "." represents the dot product, and "|| ||" represents the magnitude of the vector.

Interpretation of Cosine Similarity: The resulting cosine similarity values range from -1 to 1, with a value of 1 indicating that the two vectors are identical and a value of 0 indicating that they are orthogonal (that is, they are not similar at all) and a value of -

1 indicates that the vectors are orthogonal, that they are completely different.

Advantages of cosine similarity include computational efficiency, the ability to handle high-dimensional data, and the ability to obtain semantic similarities between entities based on word embeddings. It is also a widely used and well-established measure of similarity in NLP and machine learning (Grigori et.al, 2014; Tao et.al, 2017).

To answer our research question, we choose text processing methods such as cosine similarity, count vectorization and TF-IDF (Lahitani et al., 2016). To create our recommendation engine, we collected a dataset from an extension of the European Commission's topsoil survey. It has a data frame of 19969 datasets with 15 properties related to soil properties and can be easily adapted to the needs of smallholder farmers (LUCAS, 2015). Another dataset with limited access to specific soil properties and records of 303 tree species that fit the purpose perfectly (Orwa, 2009; LUCAS, 2015). In order to enhance the model, we utilized two different embedding methods count vectorization and TF-IDF. In the next chapter, we will start with an introduction to datasets.

CHAPTER FOUR

DATA COLLECTION AND ANALYSIS

The agroforestry dataset used in this study was compiled from two datasets: the Agroforestry Database (Orwa, 2009) and the European Commission's topsoil survey (LUCAS, 2015). The finished collection includes 19,967 geo-referenced samples. The table below contains a summary of the information.

S.no	Attribute name	Description	Range of value
1	course	coarse fragments	0 to 83
2	clay	clay size distribution	0 to 79
3	silt	slit size distribution	0 to 92
4	sand	sand size distribution	0 to 99
5	pH_in_H2O	Ph contains in H2O	3.21 to 10.08
6	pH_in_CaCl2	Ph contains in CaCl2	2.57 to 9.25
7	OC	organic carbon (g/kg)	1 to 586.8

8	CaCO3	carbonate content (g/kg)	0 to 944
9	N	total nitrogen content (g/kg)	0 to 38.6
10	P	phosphorous content (mg/kg)	0 to 1366.4
11	K	extractable potassium content (mg/kg)	0 to 7342
12	CEC	cation exchange capacity (cmol(+)/kg)	1 to 234
13	Sample_ID	Sample Id's	
14	GPS_LAT	Location Latitude information	
15	GPS_LONG	Location Longitude information	

Table 1: Examination of soil dataset (LUCAS, 2015).

Data Information

The raw data in a CSV file consists of sentences, which are large sequences of words and punctuation marks that need to be converted into vectors of numbers without losing content. Ideally the model wants a vector of fixed dimension. Records contain categories such as 'Notes' and 'Point ID' that are not very useful and are removed during processing. The soil dataset contains 19969 rows, and the trees dataset contains 303 values. "Tags" of soil datasets are formed by concatenating different columns in the dataset and performing various actions on the columns (LUCAS, 2015).

The attribute in the dataset represents a comprehensive analysis of various physical and chemical properties of samples that have been collected and measured. The attributes that have been analyzed, include the percentage of coarse fragments from the range 0 to 83, three attributes of particle size distribution such as clay size distribution from the range 0 to 79, slit size distribution from the range 0 to 79 and sand size distribution from the range 0 to 99, pH in both CaCl2 and H2O respectively from the range 2.57 to 9.25 and 3.21 to 10.08, organic carbon content ranging from the 1 to 586.8 , carbonate content ranging from the 0 to 944 , phosphorous content ranging from the 0 to 1366.4, total nitrogen content ranging from 0 to 38.6, extractable potassium content from the range 0 to 7342 , cation exchange capacity from the range 1 to 234. Each of these parameters provides valuable information about the samples, which can be used to gain insights into their composition, properties, and behaviour (LUCAS, 2015).

Programming and Libraries

Python: Python is a popular high-level programming language widely used to build a variety of apps, including web applications, scientific computing, data analysis, and artificial intelligence (Hao & Ho, 2019). Some of the well-established python libraries are numpy and pandas which are capable of handle large amount of data (Hao & Ho, 2019). We have utilized pandas and NumPy for analysing and manipulating the data in our project.

Data Set Analysis and Results

Initially, to build a cosine similarity we are going to apply the following steps:

1) Get the Land Use/Land Cover Area Frame Survey (LUCAS, 2015) dataset ready for use.

2) Perform text mining like deleting unnecessary features, adding new features to the data, and applying text preprocessing methods.

3) Next step would be to choose the correct model for the content-based recommendation system.

4) Selecting the embedding methods (Count Vectorization and TF-IDF).

5) Apply Cosine similarity based on the Count Vectorization and check the recommendations output.

6) Finally apply cosine similarity based on the TF-IDF and check the recommendations output.

This project used the Anaconda platform and Jupyter notebooks. It is an integrated development environment (IDE) used primarily by Python programmers for a wide range of essential tools for building machine learning models. I first installed the panda’s library to read the dataset. After uploading the dataset to the environment, create two variables 'X' and 'Y' to pass all characteristics of the dataset to 'X' except the 'target' column assigned to Y. bottom. After effectively reading and writing the variables, we choose an appropriate algorithm (in this case cosine similarity) to use for our agroforestry recommendation system.

We used Count Vectorization (Shahnoor et.al, 2017) to describe agroforestry practices, and associated information in a numerical format before applying cosine similarity to our agroforestry suggestion system. Each paragraph is converted into a vector of word counts by Count Vectorization, where each word represents a feature, and its count reflects the worth of that feature (Shahnoor et.al, 2017). This enables us to describe the paragraphs in a way that permits them to be readily compared using mathematical methods.

The input was then processed by the TFIDF Vectorizer. TFIDF stands for Term Frequency-Inverse Document Frequency and is a method of de-emphasizing words that appear frequently in documents and increasing the importance of words that appear infrequently (Shahnoor et.al, 2017). This is done to offer more weight to discriminative terms and less weight to non-discriminative words.

Finally, we calculated the cosine similarity between each pair of paragraphs using their count vectorize and TFIDF vectors. The input was then processed by the TFIDF Vectorizer. TFIDF stands for Term Frequency-Inverse Document Frequency and is a method of de-emphasizing words that appear frequently in documents and increasing the importance of words that appear infrequently (Shahnoor et.al, 2017). We thus found the topmost comparable data between each pair of vectors and recommended only the top three vector scores to the user by calculating the cosine similarity between all sets of content.

Results:

Below are the top three trees' recommendations using the TFIDF and Count Vectorization techniques calculated with the help of cosine similarity between the 2 vector pairs. A function named recommend was created with the logic of count vectorization and TF-IDF. Soil id was passed as an argument to the function to get the three topmost tree species recommendations.

Recommendation Based on Count Vectorization

```
def recommend(id):
    soil_id=df[df['sample_ID']==id].index[0]
    distances=similarity[soil_id]
    trees_list=sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[1:4]

    for i in trees_list:
        print(dfc.iloc[i[0]].Species_Name)
```
[77]

```
recommend(31035)
```
[80]

```
Irvingia wombolu
Calliandra calothyrsus
Acacia auriculiformis
```

Figure 2: Results of recommendations using Count Vectorization

Recommendation Based on TF-IDF

```
def recommend1(id):
    soil_id=df[df['sample_ID']==id].index[0]
    distances=similarity1[soil_id]
    trees_list=sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[1:4]

    for i in trees_list:
        print(dfc.iloc[i[0]].Species_Name)
```
[94]

```
recommend1(31035)
```
[95]

```
Ficus glumosa
Pentaclethra macrophylla
Acacia aneura
```

Figure 3: Results of recommendations using TF-IDF.

CHAPTER FIVE

DISCUSSION, CONCLUSION AND AREAS OF FURTHER STUDIES

Discussion

The research questions are:

1. What methods can be used to improve the accuracy and reliability of soil-based agroforestry tree species recommendation systems?
2. How can agroforestry tree species recommendation systems be tailored to the needs of different stakeholders, such as smallholder farmers or agribusinesses?
3. What will be the top three tree recommendations using natural language processing based on varying soil content?

What follows is the discussion of the findings and conclusion, followed by suggestions for areas for further study.

The findings and conclusion for each of the questions:

Q1. Various natural language processing techniques can significantly enhance the ability of soil-based agroforestry tree species recommendation systems to analyze and process large amounts of data. Cosine similarity, count vectorization, and TF-IDF are some of the techniques that can be used to achieve this (Shahnoor et.al, 2017). Count vectorization turns text into a matrix of token counts, and TF-IDF measures the importance of terms within a record (Shahnoor et.al, 2017). By improving the system's ability to analyze and process large volumes of data accurately, the recommendations provided by the system can become more effective and reliable, leading to better outcomes for soil-based agroforestry systems.

Q2. Using the similarity measures, cosine similarity (Alanzi et al., 2019) can recommend top varieties of trees on the topsoil survey data taken from European union database. To meet the requirements of small-scale farmers and stakeholders, cosine similarity approach can easily be customized to give recommendations using soil features like potential of hydrogen, percentage of nutrient content and texture of the soil. The system can enhance acceptance and aid in the success of agroforestry systems by offering recommendations that are pertinent and useful to the requirements of many stakeholders.

Q3. The application of natural language processing techniques namely cosine similarity measurement, count vectorization conversion, and Term frequency inverse document frequency assessment can be effectively leveraged in analyzing topsoil data, to determine ideal tree species selections based on soil specialties. With the recommendation systems ability to assess vector similarities as well as to transform textual material into token matrices while identifying significant vocabulary terms, this proposed system is a key resource that could substantially benefit farmers, foresters or any other stakeholders seeking guidance on ideal strategies for availing agroforestry system benefits.

Text based processing methods along with machine leaning can be leveraged in the recommendation systems, refining the accuracy and acceptance of recommendations in agroforestry tree species systems. The objective of this project to provide a top three tree recommendations has been achieved by incorporating cosine similarity, Term frequency inverse document frequency, and Count vectorization techniques, which are discussed in chapter 3.

Furthermore, the recommendation systems build for recommending tree species can be personalized as per the farmers soil test report and requirements.

Conclusion

The integration of text-based processing and machine learning methods has significant potential to refine agroforestry tree species recommendation systems. With a focus on accurately handling data within a given context these innovative approaches can generate reliable recommendations for optimal tree species.

The tree species recommendations given by the model can be tweaked as per the soil test report and requirements of the small holder farmers or agricultural businesses.

The recommendation system outputs the top three trees fulfilling the primary objective of the thesis project. Such a system can be a valuable tool for farmers, foresters, and other stakeholders involved in agroforestry systems which would help them to select the most suitable tree species for their land and increase the productivity and sustainability of their agroforestry systems.

Future work

Areas for further study need to improve the sustainability and productivity of agroforestry practices and enhance the ecosystem services they provide. Future studies could explore additional strategies for improving the accuracy and reliability of agroforestry tree species recommendation systems, such as incorporating feedback from stakeholders and farmers directly into the system's design and development. This could help to refine the system's recommendations and ensure they are more practical and relevant to the needs of different stakeholders.

APPENDIX A

CODE OF CRUCIAL PART

<u>Code</u>

```
#Import necessary Libraries
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from geopy.point import Point
import numpy as np
import pandas as pd

#Reading csv file
df=pd.read_csv('LUCAS_TOPSOIL_v1.csv')
dfc=pd.read_csv("Agroforestry_treespecies.csv")

#Based on Sand, Clay and Silt Features values creating a Soil Texture
def soiltextureclass(sand,clay,silt):
    textural_class=[]
    for i,j,k in zip(sand,clay,silt):
        if k + 1.5*j < 15:
            textural_class.append('sand')
        elif k + 1.5*j >= 15 and k + 2*i < 30:
            textural_class.append('loamy sand')
        elif (j >= 7 and j < 20 and i > 52 and k + 2*j >= 30) or (j < 7 and k < 50 and k + 2*j >= 30):
            textural_class.append('sandy loam')
        elif j >= 7 and j < 27 and k >= 28 and k < 50 and i <= 52:
            textural_class.append('loam')
        elif (k >= 50 and j >= 12 and j < 27) or (k >= 50 and k < 80 and j < 12):
            textural_class.append('silt loam')
#Based on pH in CaCl2 values creating a New Feature
def soilph(pH_in_CaCl2):
    soilphval=[]
    for i in pH_in_CaCl2:
        if i>5 and i<8:
            soilphval.append('Ideal soil pH range for plant growth')
        elif i>=4 and i<=5:
            soilphval.append('plants that are highly sensitive to acidity,such as some legumes and barley, are adversely affe
        elif i>=3.7 and i<4:
            soilphval.append('At a low pH, beneficial elements such as molybdenum (Mo), phosphorus (P), magnesium (Mg) and ca
        elif i<3.7:
            soilphval.append('No plant growth possible at this pH value as it is too acidic Applying sufficient lime to lift
        elif i>=8 and i<9:
            soilphval.append('when the pH(CaCl2) is greater than 7.5, calcium can tie up phosphorus, making it less available
        elif i>=9:
            soilphval.append('Alkaline soils are characterized by soil-solution pH greater than 7.0, which is caused by build
        else:
            soilphval.append('incorrect value')
    return soilphval

#Based on N (Nitrogen), P (phosphorus), K (Potassium) values creating a New Feature Nitrozen Content
def nitrogencontent(N):
    nitroper=[]
    for i in N:
        i=(i/1000)*100
        if i>=1.5:
            nitroper.append('Suitable level of nitrogen present in soil for most plant growth')
        elif i<1.5:
```

```python
#Based on N (Nitrogen), P (phosphorus), K (Potassium) values creating a New Feature Nitrozen Content
def nitrogencontent(N):
    nitroper=[]
    for i in N:
        i=(i/1000)*100
        if i>=1.5:
            nitroper.append('Suitable level of nitrogen present in soil for most plant growth')
        elif i<1.5:
            nitroper.append('Nitrogen level is low, nitrogen fixation required Nitrogen is the element responsible for lush g
    return nitroper

#Based on N (Nitrogen), P (phosphorus), K (Potassium) values creating Soil richness ratio
def soil_richness_ratio(N,P,K):
    soilrichnessratio=[]
    for i,j,k in zip(N,P,K):
        try:
            if(i*100)/j<14.5 and (i*100)/k<2.1:
                soilrichnessratio.append('N-limited site')
            elif(i*100)/j>14.5 and k/j>3.4:
                soilrichnessratio.append('P or P+N-limited sites')
            elif(i*100)/k>2.1 and k/j<3.4:
                soilrichnessratio.append('K- or K+N-limited sites')
            else:
                soilrichnessratio.append('No value')
        except ZeroDivisionError:
            soilrichnessratio.append(0)
    return soilrichnessratio
```

```python
Method 1: Count Vectorization
cv=CountVectorizer(max_features=500,stop_words='english')
cv1=CountVectorizer(max_features=302,stop_words='english')
vectors=cv.fit_transform(df['tags']).toarray()
new_vectors=cv1.fit_transform(dfc['Description']).toarray()
similarity=cosine_similarity(vectors,new_vectors)
recommend(31035)
recommend(13291)

#Method 2: TFIDF Vectorizer
tfidf_vectors=tfidf.fit_transform(tags).toarray()
tfidf_new_vectors=tfidf1.fit_transform(description).toarray()
similarity1=cosine_similarity(tfidf_vectors,tfidf_new_vectors)
recommend1(31035)
recommend1(13291)
```

REFERENCES

Anjuma. N and B. Srinivasu. (2021) “A Comparative Study on Classification Algorithms Using Different Feature Extraction And Vectorization Techniques For Text”. https://www.researchgate.net/publication/358118682

Alanzi, AbuZeina (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing, Journal of King Saud University-Journal of Computer and Information Sciences, 29 (2017), pp. 189 95 https://doi.org/10.1016/j.jksuci.2016.04.001

Alfirna. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, (2016). Cosine similarity to determine similarity measure: A study case in online essay assessment, 4thInternational Conference on Cyber and IT Service Management, Bandung,Indonesia, 2016, pp. 1-6, doi: 10.1109/CITSM.2016.7577578.

Bargali S.S, K. Padalia, K. Bargali “Effects of tree fostering on soil health and microbial biomass under different land use systems in the Central Himalayas”,Land Degrad. Dev., 30 (16) (2019), pp. 1984-1998 https://onlinelibrary.wiley.com/doi/full/10.1002/ldr.3394

Barrios E, Valencia V, Jonsson M, Brauman A, Hairiah K, Mortimer PE, Okubo S. 2018. Contribution of trees to the conservation of biodiversity and ecosystem services in agricultural landscapes. Int J Biodivers Sci Ecosyst Serv Manage. doi:10.1080/21513732.2017.1399167.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '10). Association for Computing Machinery, New York, NY, USA, 841–842. https://doi.org/10.1145/1835449.1835643

Bruno Trstenjak, Sasa Mikac, Dzenana Donko,KNN with TF-IDF based Framework for Text Categorization, Procedia Engineering,Volume 69,2014, Pages 1356-1364,ISSN 18777058, https://doi.org/10.1016/j.proeng.2014.03.129

Donghwa Kim, Deokseong Seo, Suhyoun Cho, Pilsung Kang,Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec, Information Sciences, Volume 477,2019,Pages 15 - 29,ISSN 0020-0255, https://doi.org/10.1016/j.ins.2018.10.006

Eswaran, H., Lal, R., & Reich, P. F. (2019). Land degradation: an overview. Response to land degradation, 20-35. https://www.taylorfrancis.com/chapters/edit/10.1201/9780429187957- 4/land-degradation-overview-eswaran-lal-reich

Fahad S, Chavan SB, Chichaghare AR, Uthappa AR, Kumar M, Kakade V, Pradhan A, Jinger D, Rawale G, Yadav DK, Kumar V, Farooq TH, Ali B, Sawant AV, Saud S, Chen S, Poczai P. (2022). Agroforestry Systems for Soil Health Improvement and Maintenance. Sustainability. 2022; 14(22):14877. https://www.mdpi.com/2071-1050/14/22/14877

Fkih F., (2022) “Similarity measures for collaborative filtering-based recommender systems: Review and experimental comparison”J. King Saud Univ. - Comput. Inform. Sci., 34 pp. 7645-7669, https://doi.org/10.1016/j.jksuci.2021.09.014

Frontiers in artificial intelligence, Semantics, and natural language processing in agriculture Frontiers, https://www.frontiersin.org/research-topics

Grammarly, writing assistance, apa, mla formatting, https://www.grammarly.com

Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, David Pinto (2014) Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model" Vol 18, No 3 (2014) https://doi.org/10.13053/CyS-18-3-2043

Haochen Chen, Syed Fahad Sultan, Yingtao Tian, Muhao Chen, and Steven Skiena. 2019. Fast and Accurate Network Embeddings via Very Sparse Random Projection. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM'19). https://doi.org/10.1145/3357384.3357879

Jaegul Choo, Hanseung Lee, Zhicheng Liu, John Stasko, and Haesun Park (2013) "An interactive visual testbed system for dimension reduction and clustering of large- scale high-dimensional data", Proc. SPIE 8654, Visualization and Data Analysis 2013, 865402; https://doi.org/10.1117/12.2007316

Islam.F, M. S. Arman, N. Jahan, M. H. Sammak, N. Tasnim and I. Mahmud, "Model and Popularity Based Recommendation System- A Collaborative Filtering Approach," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), https://ieeexplore.ieee.org/abstract/document/9984348

Iyabano.A, Klerkx.L &Leeuwis.C (2023) Why and how do farmers organizations get involved in the promotion of agroecological techniques? Insights from Burkina

Faso, Agroecology and Sustainable Food Systems, https://doi.org/10.1080/21683565.2023.2164881

Jawo T.O, Teutscherova.N, Mesele Negash, Kefyalew Sahle & Bohdan Lojka (2023) Smallholder coffee-based farmers' perception and their adaptation strategies of climate change and variability in South-Eastern Ethiopia, International Journal of Sustainable Development & World Ecology, https://doi.org/10.1080/13504509.2023.2167241

Jiwani, N., Gupta, K., Whig, P. (2023). Analysis of the Potential Impact of Omicron Crises Using NLTK (Natural Language Toolkit). In: Khanna, A., Gupta, D., Kansal, V., Fortino, G., Hassanien, A.E. (eds) Proceedings of Third Doctoral Symposium on Computational Intelligence . Lecture Notes in Networks and Systems, vol 479. Springer, Singapore. https://doi.org/10.1007/978-981-19-3148-2_37

Kumar.R and Sahula. V, (2021) Intelligent Approaches for Natural Language Processing for Indic IEEE International Symposium on Smart Electronic Systems (iSES), Jaipur, India, 2021, pp. 331-334, doi: 10.1109/iSES52644.2021.00084.

Kumar.R, M. P. Singh, P. Kumar and J. P. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), Avadi, India, 2015, pp. 138-145, doi: 10.1109/ICSTM.2015.7225403.

Lahitani A.R, A. E. Permanasari and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," 2016 4th International Conference on Cyber and IT Service Management, Bandung, Indonesia, 2016, pp. 1-6, doi: 10.1109/CITSM.2016.7577578.

Lovedeep Singh.(2022) “Clustering Text: A Comparison Between Available Text Vectorization Techniques”. In: Soft Computing and Signal Processing. Springer, 2022, https://link.springer.com/chapter/10.1007/978-981-16-1249

Magerman, T., Van Looy, B. & Song, X. Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. Scientometrics82, 289–306 (2010). https://doi.org/10.1007/s11192-009-00466

Meijer SS, Catacutan D, Ajayi OC, Sileshi GW, Nieuwenhuis M (2015) The role of knowledge, attitudes, and perceptions in the uptake of agricultural and agroforestry innovations among smallholder farmers in sub-Saharan Africa. Int J Agric Sustain 13(1):40–54 https://doi.org/10.1080/14735903.2014.912493

Nathalie Korboulewsky, Gabriel Perez, Matthieu Chauvat, (2016) How tree diversity affects soil fauna diversity: A review,Soil Biology and Biochemistry Volume 94, Pages 94-106,ISSN https://www.sciencedirect.com/science/article/pii/S0038071715004174

Orwa C, Mutua A, Kindt R, Jamnadass R, Simons A. 2009. Agroforestry Database: a tree reference and selection guide version 4.0. World Agroforestry Centre,Kenya.

https://www.worldagroforestry.org/output/agroforestree-database

Ospina, C. (2017). Climate and economic benefits of agroforestry systems. The Climate Institute, Washington. http://climate.org/wp-content/uploads/2017/03/Agroforestry-Article-3.6.17.pdf

Pantera, A., Mosquera-Losada, M.R., Herzog, (2021) Agroforestry and the environment. Agroforest Syst 95, 767–774 . https://doi.org/10.1007/s10457-021-00640-8

Patil.P, Panpatil.V, Shrikant Kokate,(2020)."Crop Prediction System using MachineLearningAlgorithms https://www.irjet.net/archives/V7/i2/IRJET-V7I2163.pdf

Patil, K., & Gunjal, A. (2022). Digital agriculture: Crop recommendation in India using machine learning models. Nucleation and Atmospheric Aerosols. https://doi.org/10.1063/5.0108469

Peipei Xia, Li Zhang, Fanzhang Li,Learning similarity with cosine similarity ensemble, Information Sciences, Volume 307,2015,Pages 39-52,ISSN 0020-0255, https://doi.org/10.1016/j.ins.2015.02.024

Priano ME, Fuse VS, Mestelan S, Berkovic AM, Guzmán SA, Gratton R, Juliarena MP (2018) “Afforested sites in a temperate grassland region: influence on soil properties and methane uptake”. Agrofor Syst. https://doi.org/10.1007/s10457-017-0104-7

Priyadharshini A, S. Chakraborty, A. Kumar and O. R. Pooniwala, "Intelligent Crop Recommendation System using Machine Learning," 2021 5th

International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 843-848, doi: 10.1109/ICCMC51019.2021.9418375

Ramos HMN, Vasconcelos SS, Kato OR, Castellani DC (2018) “Above- and belowground carbon stocks of two organic, agroforestry-based oil palm production systems in eastern Amazonia”. Agrofor Syst. https://doi.org/10.1007/s10457-017-0131-4

Ramos, N.C.; Gastauer, M.; de Almeida Campos Cordeiro, A.; Meira- Neto, J.A.A. “Environmental filtering of agroforestry systems reduces the risk of biological-invasion.” Agrofor.Syst. 2015, 89,279–289. https://link.springer.com/article/10.1007/s10457-014-9765-7

Roberto J. Bayardo, Yiming Ma, and Ramakrishnan Srikant. 2007. Scaling up all pairs similarity search. In Proceedings of the 16th international conference on World Wide Web. https://dl.acm.org/doi/10.1145/1242572.1242591

Rosati, A., Borek, R. & Canali, S. Agroforestry and organic agriculture. Agroforest Syst 95, 805–821 (2021). https://doi.org/10.1007/s10457-020-00559-6

Schwarz, M. Lobur and Y. Stekh, "Analysis of the effectiveness of similarity measures for recommender systems," 2017 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Lviv, Ukraine, 2017, pp. 275- https://ieeexplore.ieee.org/abstract/document/7916133

Shin, S., Park, M. S., Lee, H., Lee, S., Lee, H., Kim, T. H., & Kim, H. J. (2021).

Global Trends in Research on Wild-Simulated Ginseng: Quo Vadis Forests, 12(6), 664. MDPI AG. http://dx.doi.org/10.3390/f12060664

Singh R, Behera MD, Das P, Rizvi J, Dhyani SK, Biradar CM.(2022) Agroforestry Suitability for Planning Site-Specific Interventions Using Machine Learning Approaches. Sustainability. 14(9):5189. https://doi.org/10.3390/su14095189

Simranjit Singh, Singara Singh Kasana,(2022) Quantitative estimation of soil properties using hybrid features and RNN variants, Chemosphere, Volume 287, Part 1, 2022, 131889, ISSN 0045-6535, https://www.sciencedirect.com/science/article/pii/S0045653521023614

Srivastava, P.; Giri, N.; Mandal, D. (2019) 137 Cs technology for soil erosion and soil carbon distribution. Curr. Sci. 2019,116, 888–889. https://currentscience.ac.in/Volumes/116/06/0888.pdf

Sitikhu. P, K. Pahi, P. Thapa and S. Shakya, "A Comparison of Semantic Similarity Methods for Maximum Human Interpretability," 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 2019, pp. 1-4, https://doi.org/10.1016/j.ins.2015.02.024

Soons M.B, A.L. Brochet, E. Kleyheeg, A.J. Green Seed dispersal by dabbling ducks: an overlooked dispersal pathway for a broad spectrum of plant species J. Ecol., 104 (2016), https://doi.org/10.1111/1365-2745.12531

Sunita Sarawagi and Alok Kirpal. 2004. Efficient set joins on similarity predicates. In Proceedings of the 2004 ACM SIGMOD international conference on Management of data (SIGMOD '04). Association for Computing

Machinery, New York, NY, USA, 743–754.
https://doi.org/10.1145/1007568.1007652

Tao, Z., Liu, H., Fu, H., & Fu, Y. (2017). Image Cosegmentation via Saliency-Guided Constrained Clustering with Cosine Similarity. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).
https://doi.org/10.1609/aaai.v31i1.11203

Temgoua LF, Momo Solefack MC, Nguimdo Voufo V, Tagne Belibi C, Tanougong A (2018) Spatial and temporal dynamic of land-cover/land-use and carbon stocks in Eastern Cameroon: a case study of the teaching and research forest of the University of Dschang.
https://doi.org/10.1080/21580103.2018.1520743

Van der wolf, J., Jassgone, L., Gram, G., & Vaast, P. (2016). Turning local knowledge on agroforestry into an online decision support tool for tree selection in smallholder farms. Experimental Agriculture, 55(S1)
https://doi.org/10.1017/S001447971600017X

U.S. DEPARTMENT OF AGRICULTURE.Soil Health. USDA,
https://www.nrcs.usda.gov/wps/portal/nrcs/main/soils/health

U.S. DEPARTMENT OF AGRICULTURE."Soil Survey Manual." USDA, 2017
https://www.nrcs.usda.gov/sites/default/files/2022-09/The-Soil-Survey-Manual.pdf

Wendland. A, Marco Zenere, and Jörg Niemann. (2021) "Introduction to Text Classification: Impact of Stemming and Comparing TF-IDF and Count Vectorization as Feature Extraction Technique". In: Springer,
https://link.springer.com/chapter/10.1007/978-3-030-85521-5_19

Xin Ye, Hui Shen, Xiao Ma, Razvan Bunescu and Chang Liu, “From word embeddings to document similarities for improved information retrieval in software engineering”, in Proceedings of the International Conference on Software Engineering, Texas, USA, 14-22 May 2016, ISBN: 9781450339001, pp. 404–415, DOI: 10.1145/2884781.2884862 https://doi.org/10.1145/2884781.2884862

Yengwe J, Gebremikael MT, Buchan D, Lungu O, Neve SD (2018) “Effects of Faidherbia albida canopy and leaf litter on soil microbial communities and nitrogen mineralization in selected Zambian soils”. Agrofor Syst https://doi.org/10.1007/s10457-016-0063-4

Z. Doshi, S. Nadkarni, R. Agrawal and N. Shah (2018) AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms, Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, https://ieeexplore.ieee.org/abstract/document/8697349