



# VIT<sup>®</sup>

## Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

NAME	:	SYED AYAZ IMAM ANANYA POKHRIYAL
REG.NO	:	18BCE0660 18BCI0195
SUBJECT CODE	:	CSE2004
SUBJECT TITLE	:	DATABASE MANAGEMENT SYSTEM
GUIDED BY	:	GOVINDA K

### SLAY ALLERGENS

**Repository and Suggestion Algorithm for Packaged Items that  
Trigger Reaction**

**ABSTRACT:**

Food allergies caused by allergens such as gluten, shellfish, nuts, soy, dairy are growing at an exponential rate impacting at least 3% of the world. In developed countries, allergen-free food is easily accessible due to greater medical and patient awareness, greater availability of substitutes and financial will to maintain an allergen-free diet. This is not the case in developing countries like India- to tackle this issue of inaccessibility; we endeavour to create a database that catalogs allergen-free food. In this database, we shall achieve this using two machine learning algorithms namely Cosine Similarity Method and Correlation Method to create a Recommendation System.

**KEYWORDS:**

Allergen Free Food, Recommendation Systems, Machine Learning, Database Management.

**INTRODUCTION:**

Allergies can be defined as an overreaction of body due to some substances which are considered otherwise harmless. Common allergic disorders are allergic rhinitis, asthma, anaphylaxis, certain drugs, food items, insect allergies, and eczema. The best way to treat an allergy is to avoid the foods which contain the allergens a person is triggered by as even the smallest quantity of an allergen can produce a severe response in sensitive people.

Allergen-free food, specifically gluten-free food is grossly under produced in India. According to the Institute of Agri-Business Management India has a potential of 2,347-kilotonnes of gluten-free foods; yet only 7.55-kilo tonnes were produced in 2016. Experts have estimated that the market share of gluten-free products in India is only 0.5-2% of the global produce even though millions of Indians suffer from Celiac Disease.

Due to huge technological advances in proteomics and genomics and current upgradements in data analysis, we have large resources available on the internet for basic research on allergy. There are several allergy related databases developed by organizations and government bodies. These databases serve different purposes as they differ in level of applications.

This document emphasizes in making allergen-free food available to more and more patients suffering from hyper-allergy. With greater availability of diet-sensitive food will come to a greater sense of awareness. The organization is as follows:

***Literature Review***

***Proposed Methodology***

***Codes***

***Results and Discussion***

***Conclusion***

***References***

## **LITERATURE REVIEW:**

1. Kiran Kadam, Rajiv Karbhel et al: The authors of this paper created a relational database, known as “AllerBase” which catalogs protein allergens and related data from existing bioinformatics resources along with published literature. AllerBase is a knowledgebase which has allergy data manually curated and brought to a single platform. [1]
2. Muskoko MM. Statistics corner: This article provides insight into the use of concepts of correlation for medical research and it highlights some of its misuses as well. [4]
3. Masih, Jolly, Rajkumar, et al: The study enables manufacturers to bring gluten-free foods to the mainstream market and to make it more affordable for all the sections of the society. The study focuses on India and the USA since both of the nations have a high potential for gluten-free foods. [5]
4. Palasik JM, Goss FR, Lai KH, et al: Using a Medical Text Extraction, Reasoning and mapping system, food items that cause strong sensitivity are grouped after study. The authors concluded that new strategies are needed to standardize food sensitivity concepts and to improve documentation. [6]
5. Wang, Jing et al: This article aims to integrate the various methods of predicting allergens. It compared three allergen prediction algorithms namely Sequence-Based, Motif-Based and Support Vector Machine on the basis of well-defined parameters and concluded that the SVM method was the most efficient. [3]

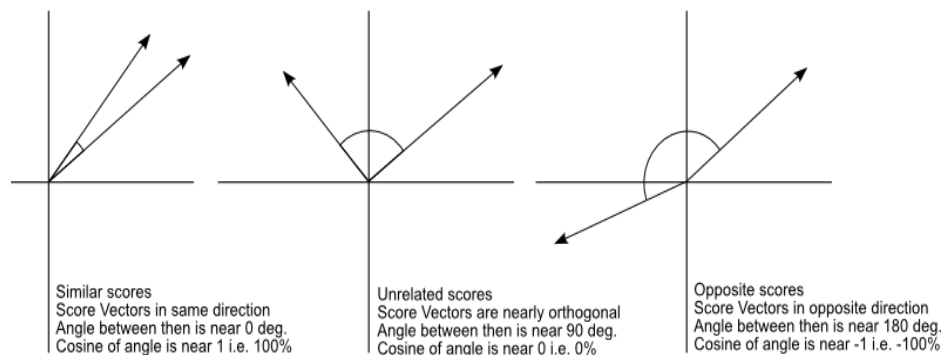
Smartphones nowadays are very powerful and can be used as allergen analysis and detection systems due to their powerful processors, optical sensors and IOT capability. In recent years there has been tremendous rise in machine learning solution to bio-

related problems. K-Nearest-Neighbours or KNN algorithm was one of the first method for allergy prediction. Bayesian Classifier has also been used for this purpose.

### PROPOSED METHODOLOGY:

The data for allergen-free food has been taken from <https://www.kaggle.com/>. All food items are listed in order of their ingredients and are approved by the FDA. This database helps us create recommendations, on the basis of sales, popularity and overall rating. Recommendation engines suggest data using various algorithms and then recommend the most relevant items to users. **We have achieved this using two algorithms:**

**Cosine Similarity Method:** The algorithm uses cosine rules to determine the similarity between the parameters and thus resulting in a score in between 0 to 1 which determines the similarity. This algorithm is very advantageous because if two products are very distant from each other by Euclidean distance, the angle between them can be smaller. According to cosine similarity algorithm, the products/documents smaller the angle, higher the similarity.



**Fig 1a. Graphical Representation of Cosine Similarity**

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

**Fig 1.b Mathematical Formula to Calculate Cosine Similarity**

**Correlation Method:** Correlation is the measure of how strong one variable is

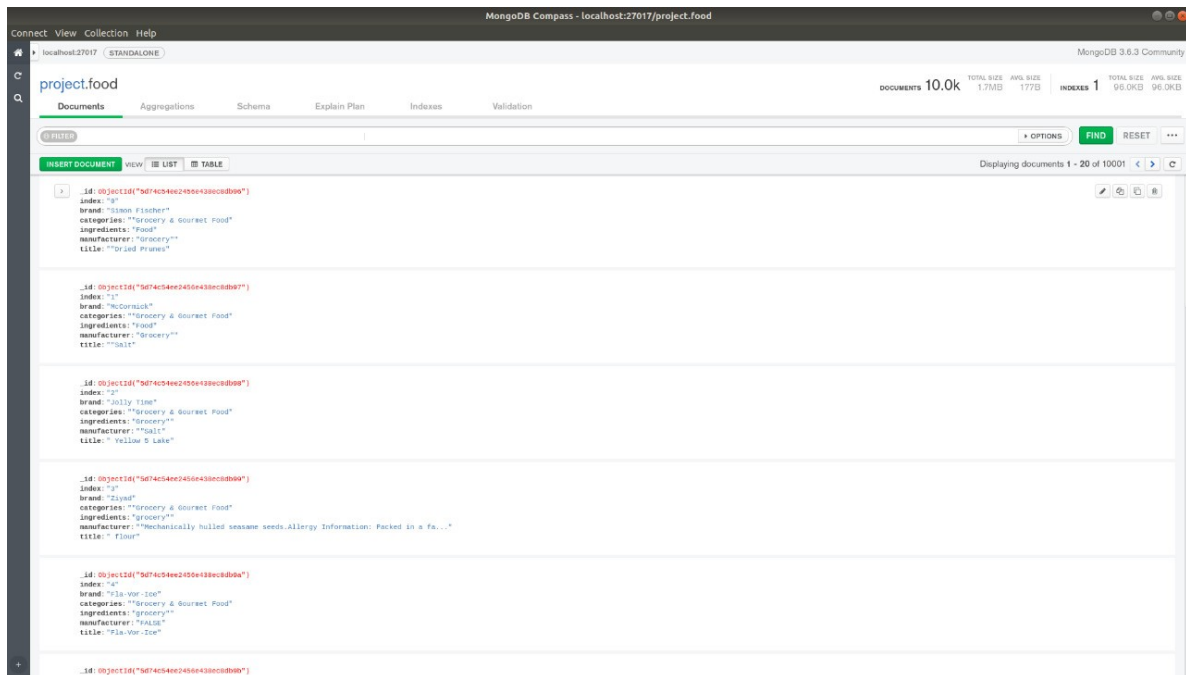
dependent on the other. The pearson's Correlation formula given below give a value in range of -1 to +1. Higher the value achieved, more the dependence of one variable on the other. The value of +1 means that x and y lie at a line with slope greater than zero where as a value of -1 means x and y lie on negative slope. Zero value indiates no correlation between the variables x and y.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

**Fig 2. Pearson's Correlation Formula**

The DBMS used to create a repository of allergen-free food is Mongo DB Compass. It is an open-Source, cross-platform, NoSQL document database that provides the advantages of stupendous performance, availability and scalability.

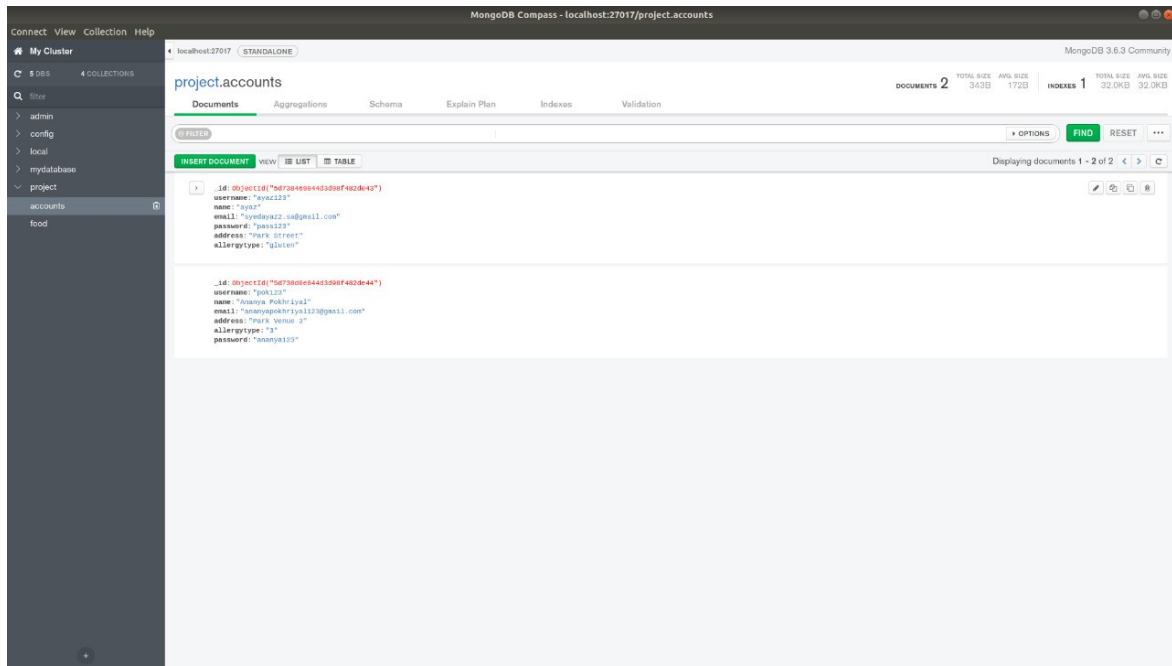
**Product** database is as follows:



**Fig 3. Product Database**

As we can see, each Product has Index, Brand, Category, Ingredients, Manufacturer and Title as its field.

Database for the **user accounts** is as follows:



**Fig 4. Account Database**

As we can see, each User has Username, Name, Email, Address, Allergy Type, and Password as its field.

**CODES:**

**main.py**

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Sat Oct 26 17:03:31 2019

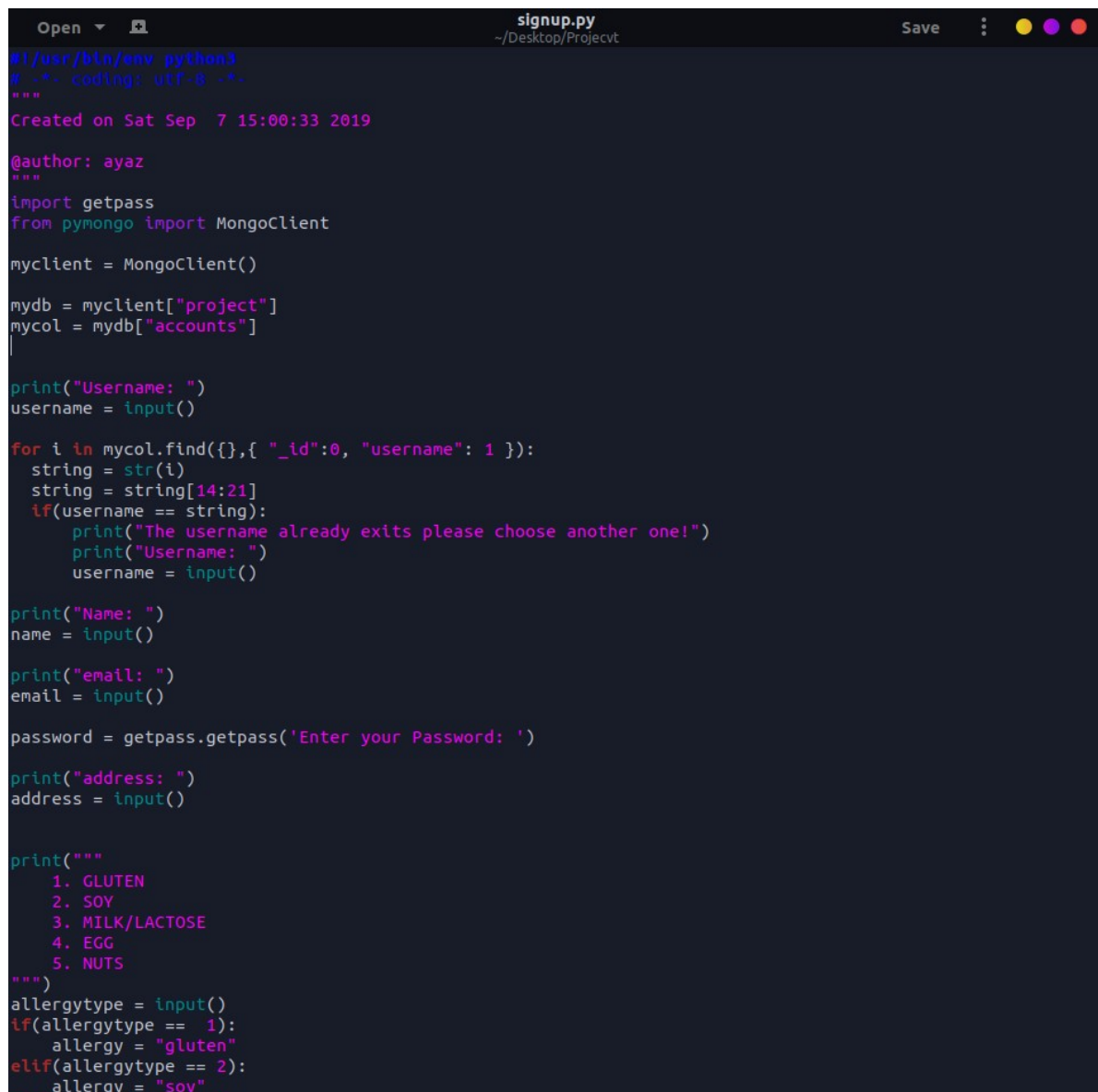
@author: ayaz
"""

print("Welcome to SlayAllergens! Your Diet Recommendation Helper!")

inp = int(input("""Press:\n
1.Sign In\n
2.Sign Up\n"""))
if inp == 2:
    import signup
if inp == 1:
    import signin
```

**Fig 5. Terminal/Welcome Page for User**

## signup.py



```
Open  signup.py  Save  ~/Desktop/Projectv
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Sat Sep  7 15:00:33 2019

@author: ayaz
"""
import getpass
from pymongo import MongoClient

myclient = MongoClient()

mydb = myclient["project"]
mycol = mydb["accounts"]

print("Username: ")
username = input()

for i in mycol.find({}, {"_id":0, "username": 1 }):
    string = str(i)
    string = string[14:21]
    if(username == string):
        print("The username already exists please choose another one!")
        print("Username: ")
        username = input()

print("Name: ")
name = input()

print("email: ")
email = input()

password = getpass.getpass('Enter your Password: ')

print("address: ")
address = input()

print("""
1. GLUTEN
2. SOY
3. MILK/LACTOSE
4. EGG
5. NUTS
""")
allergytype = input()
if(allergytype == 1):
    allergy = "gluten"
elif(allergytype == 2):
    allergy = "soy"
```

Fig 6. Sign Up Code

## signin.py

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Sat Oct 26 17:13:10 2019

@author: ayaz
"""

import getpass
from pymongo import MongoClient

myclient = MongoClient()

mydb = myclient["project"]
mycol = mydb["accounts"]

us = ps = logged = 0
username = input("Please Enter your Username: \n")

password = getpass.getpass("Enter your Password: \n")

query1 = {"username": username}
query2 = {"password": password}

doc1 = mycol.find(query1)
doc2 = mycol.find(query2)

for i in doc1:
    us = 1
for j in doc2:
    ps = 1

if us == 1 and ps == 1:
    print("""Logged In!\n
          Begin Your Search....."\n""")
    inp = int(input("1. Algorithm 1\n2. Algorithm 2\n"))
    logged = 1
else:
    print("Wrong Credentials")

if inp == 1:
    import food_cosinesim
else:
    import food_correlation
```

Fig 7. Sign In Verification



## ALGO 1: food\_cosinesim.py

```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity

def get_title_from_index(index):
    return df[df.index == index]["title"].values[0]

def get_index_from_title(title):
    return df[df.title == title]["index"].values[0]

df = pd.read_csv("editedIngredients.csv")

features = ['ingredients', 'brand', 'categories', 'manufacturer']

for feature in features:
    df[feature] = df[feature].fillna('')

def combine_features(row):
    return row['ingredients'] + " " + row['brand'] + " " + row['categories'] + " " + row['manufacturer']

df["combined_features"] = df.apply(combine_features, axis = 1)

cv = CountVectorizer()
count_matrix=cv.fit_transform(df["combined_features"])

cosine_sim = cosine_similarity(count_matrix)
print("\n\n")
print("Search Here: ")
food_user_likes = input()

food_index = get_index_from_title(food_user_likes)
similar_food = list(enumerate(cosine_sim[food_index]))

sorted_similar_food = sorted(similar_food, key = lambda x:x[1] ,reverse=True)

i = 0
for food in sorted_similar_food:
    print(get_title_from_index(food[0]))
    i = i+1
    if i>50:
        break
```

Fig 8. Cosine Similarity Algorithm

## ALGO 2: food\_correlation.py

```
Open  food_correlation.py  Save  ~/Desktop/Projectv
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Sat Oct 19 13:43:58 2019

@author: ayaz
"""

import pandas as pd
import pymongo
import numpy as np
from pymongo import MongoClient

client = MongoClient()

#getting the data
db = client.project
collection = db.filefood
df = pd.DataFrame(list(collection.find()))
del df['_id']
df.columns = ['rating', 'item_id', 'user_id']

#Importing food and thier respective IDs
collection = db.newwingredient
food_titles = pd.DataFrame(list(collection.find()))
del food_titles['_id']

#Merging df and food_titles by item id
data = pd.merge(df, food_titles, on = 'item_id')

#Converting the values into numeric
data['rating'] = pd.to_numeric(data['rating'])
data['item_id'] = pd.to_numeric(data['item_id'])
data['user_id'] = pd.to_numeric(data['user_id'])

#Calculate mean ratings and count of all food Products
data.groupby('title')['rating'].mean().sort_values(ascending = False)
data.groupby('title')['rating'].count().sort_values(ascending=False)

#Creating dataframe with 'rating' count values
ratings = pd.DataFrame(data.groupby('title')['rating'].mean())
ratings['num of ratings'] = pd.DataFrame(data.groupby('title')['rating'].count())

'''#VISUALIZING IMPORTS
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('white')
%matplotlib inline

# plot graph of 'num of ratings column'
plt.figure(figsize =(10, 4))
ratings['num of ratings'].hist(bins = 70)
```

Fig 9A: Matrix Correlation Algorithm

```

# plot graph of 'ratings' column
plt.figure(figsize =(10, 4))
ratings['rating'].hist(bins = 70)'''

#Sorting values according to the 'num of rating column'
ratings.sort_values('num of ratings', ascending = False).head(10)
foodmat = data.pivot_table(index ='user_id',
                           columns ='title', values ='rating')

# analysing correlation with similar food

item = input("Enter the name of the food item: \n")
item_user_ratings = foodmat[item]

# analysing correlation with similar food
similar_to_item = foodmat.corrwith(item_user_ratings)

# Similar food as of cinnamon
corr_item = pd.DataFrame(similar_to_item, columns=['Correlation'])
corr_item.dropna(inplace = True)

corr_item = corr_item.join(ratings['num of ratings'])
corr_item[corr_item['num of ratings']>100].sort_values('Correlation', ascending = False).head()

corr_item = corr_item.sort_values('Correlation', ascending = False)

#Printing the output
print(corr_item.iloc[:,0].head(25))

```

**Fig 9B: Matrix Correlation Algorithm**

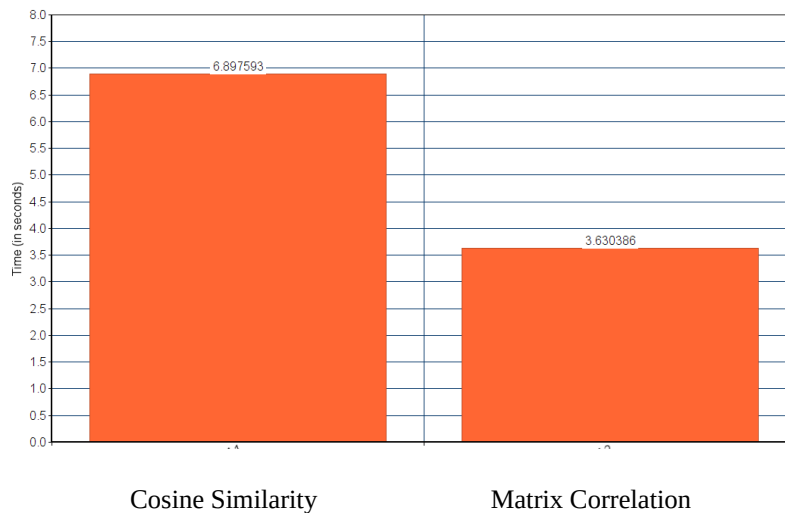
## RESULT AND DISCUSSION:

We compared the two algorithms used to generate product recommendations on the basis of the time consumed by each to suggest allergen free food. To test the same, we repeated the search for five different food items. The results obtained are as follows:

a) Corn Nuts Ranch Bag:

Runtime of Cosine similarity Algorithm is: 6.89753 s

Runtime of Matrix correlation Algorithm is: 3.630386 s

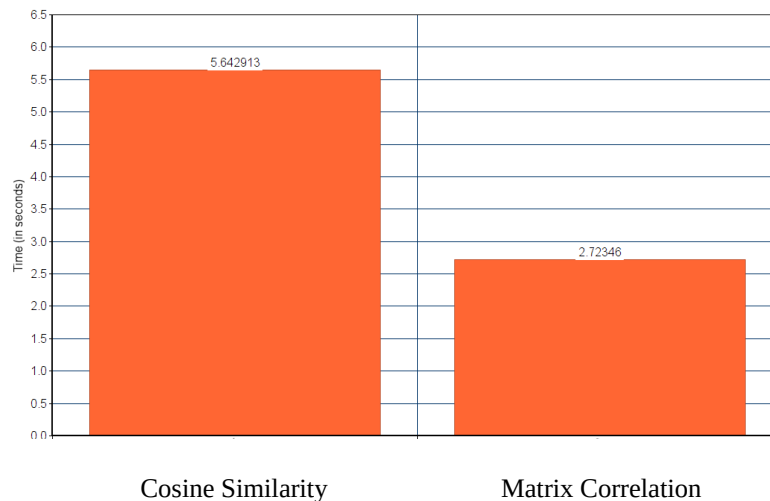


**Fig 10. Graph comparing Consine Similarity Method and Correlation Method with Sample Item Corn Nuts Ranch Bag**

b) Jolly Time Popcorn

Runtime of Cosine similarity Algorithm is: 5.642913 s

Runtime of Matrix correlation Algorithm is: 2.72346 s

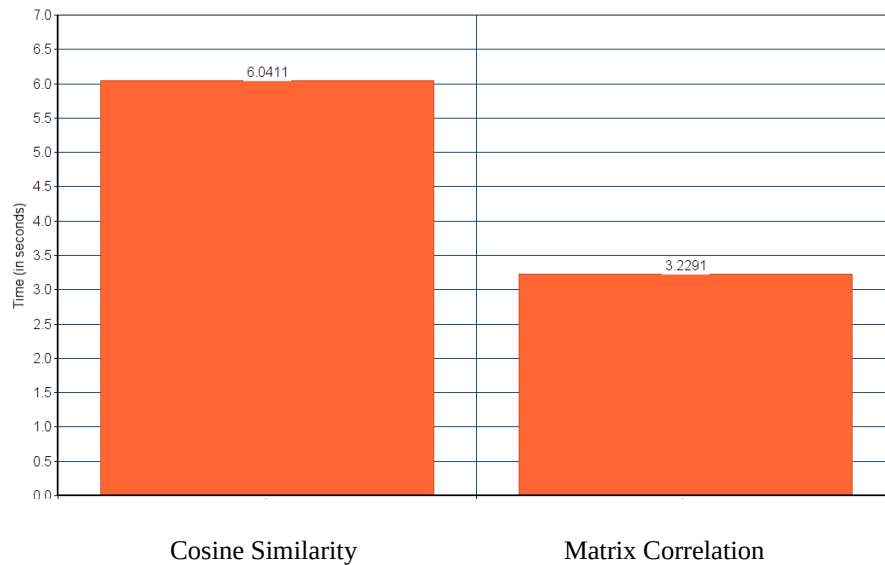


**Fig 11. Graph comparing Consine Similarity Method and Correlation Method with Sample Item Jolly Time Popcorn**

c) Roasted Turkey Gravy

Runtime of Cosine similarity Algorithm is: 6.0411 s

Runtime of Matrix correlation Algorithm is: 3.2291 s

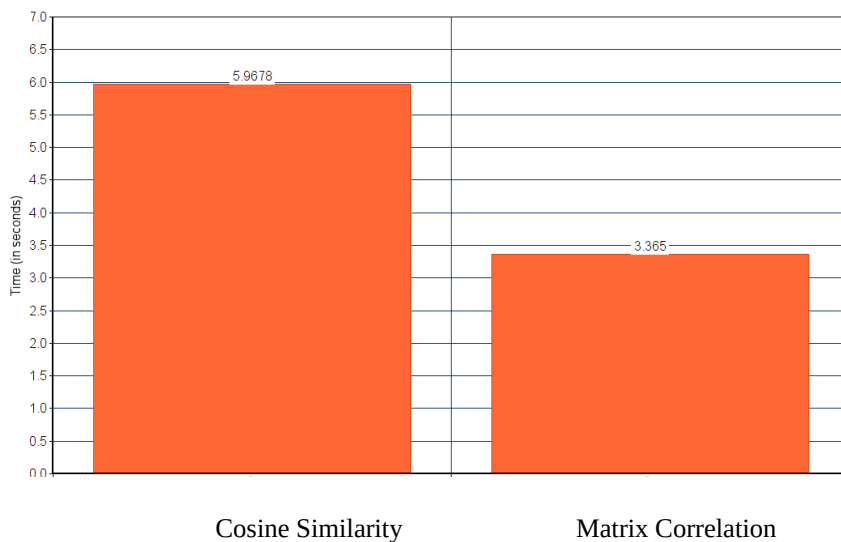


**Fig 12. Graph comparing Consine Similarity Method and Correlation Method with Sample Roasted Turkey Gravy**

d) Simply Organic Seasoning

Runtime of Cosine similarity Algorithm is: 5.9678 s

Runtime of Matrix correlation Algorithm is: 3.365 s

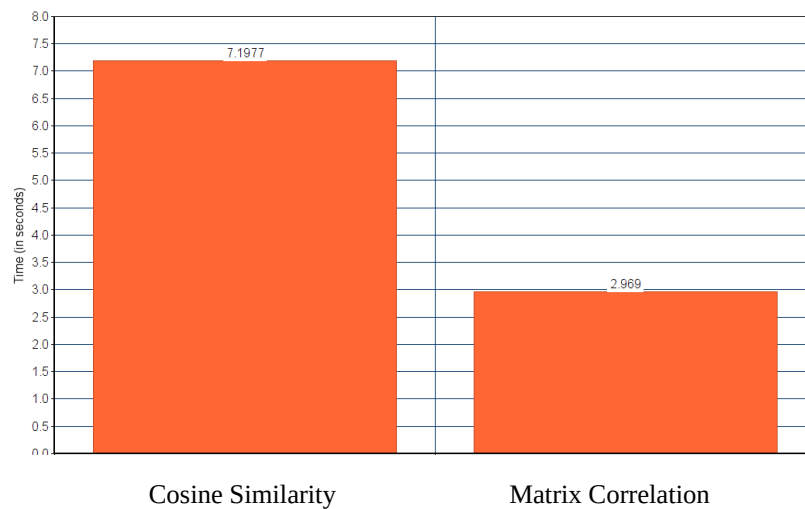


**Fig 13. Graph comparing Consine Similarity Method and Correlation Method with Sample Item Simply Organic Seasoning**

e) M&b Curry

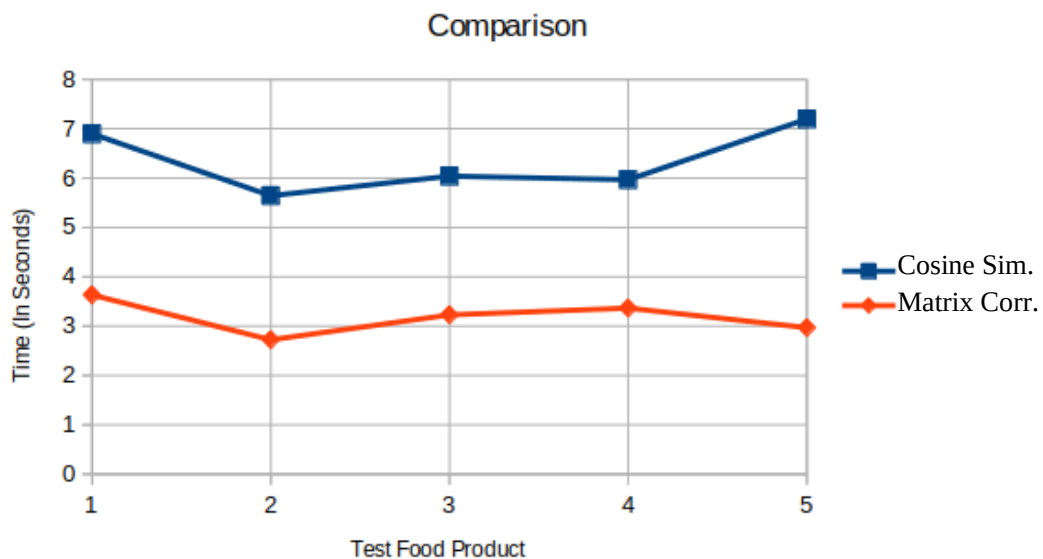
Runtime of Cosine similarity Algorithm is: 7.1977 s

Runtime of Matrix correlation Algorithm is: 2.969 s



**Fig 14. Graph comparing Consine Similarity Method and Correlation Method with Sample Item M&b Curry**

The final results are tabulated as follows:



**Fig 15. Cumulative Runtime Comparison**

From the above graphs it can be observed that the Matrix Correlation Algorithm performs better compared to Cosine Similarity Algorithm in terms of time complexity. This is because:

- Matrix Correlation Algorithm computes on correlation on numerics columns which are compluted as time complexity  $O(1)$ .

- Cosine Similarity Algorithm has a time complexity of  $O(n)$ .

Though Matrix Correlation Algorithm runs much faster compared to Cosine Similarity Algorithm, we get more viable output using the Cosine Similarity Algorithm generates recommendation which are similar to the food item we seek for. Matrix Correlation Algorithm generates recommendation based on the user ratings. Hence we might not get the recommendation which we seek for.

In modern e-commerce websites, a hybrid of both algorithms is used to recommend products to the users. This ensures similar and popular product recommendations to the user.

## CONCLUSION:

The aim of our project was to create a repository of allergen-free food in order to create a Recommendation System based on various allergens for the same. The project allows users to create a user profile and list their allergens or sign in if in-case they have already created a profile. Then an option to use the Algorithm-1 Cosine Similarity Method or Algorithm-2 Correlation method is given to them. The system tests two algorithms to use for its Recommendation System and finds that the Correlation Method takes markedly less time to give search results. Such a Recommendation System can be implemented within E-Commerce websites and Food Delivery Apps to make Allergen-Free Food more accessible to consumers. This shall, overall, create more awareness about allergens, make people more conscious about the food that they are eating and more sensitive to the needs of their own health and body.

## REFERENCES:

- [1] Kiran Kadan, Rajiv Karbal, V. K. Jayaaman, Sangeia Sawant, Urmila Kulkarni-Kale, AllerBase: comprehensive allergen knowledge base, *Database*, Volume 2017, 2017, bax066, <https://doi.org/10.1093/database/bax066>
- [2] Frank Eisenhaber, Unix interfaces, Kleisli, buccandin structure, etc. — The heroic beginning of bioinformatics in Singapore, *Journal of Bioinformatics and Computational Biology*, 10.1142/S0219720014710024
- [3] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly Media.
- [4] Mukaka MM. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24(3):69–71.
- [5] Masih, Jolly & Rajkumar, Rajasekaran & Matharu, Paviter & Sharma, Amita. (2019). Market Capturing and Business Expansion Strategy for Gluten-Free Foods in India and USA Using PESTEL Model. *Agricultural Sciences*. 10. 202-213. 10.4236/as.2019.102017.

- [6] Plasek JM, Goss FR, Lai KH, et al. Food entries in a large allergy data repository. *J Am Med Inform Assoc*. 2016;23(e1):e79–e87. doi:10.1093/jamia/ocv128
- [7] United States Food and Drug administration. <http://www.fda.gov/food/labelingnutrition/FoodAllergensLabeling/GuidanceComplianceRegulatoryInformation> Food Allergen Labeling and Consumer Protection Act of 2004 (Public Law 108-282, Title II).
- [8] American College of Allergy, Asthma, and Immunology. Food allergy: a practice parameter. *Ann Allergy, Asthma, Immunol*. 2006;96(3 Suppl 2):S1–S68
- [9] Substance Registration System Unique Ingredient Identifier (UNII). <http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm>
- [10] Salo PM, Arbes SJ, Jr, Jaramillo R, et al. Prevalence of allergic sensitization in the United States: result from the National Health and Nutrition Examination Survey (NHANES) 2005-2006. *J Allergy Clin Immunol*. 2014;134(2):350–359.