

Reporting Date: 12th April 2024

TABLE OF CONTENTS

[ABSTRACT 2](#_Toc159599442)

[INTRODUCTION 3](#_Toc159599443)

[DATA PREPARETION 3](#_Toc159599444)

[EXPLORATORY DATA ANALYSIS 4](#_Toc159599445)

[CONCLUSION 10](#_Toc159599446)

[REFRENCES 11](#_Toc159599447)

# ABSTRACT

# 

# INTRODUCTION

AutoBasket is a company founded by Larry and Veronica Smiles with the purpose of simplifying the grocery shopping experience for individuals and families. Situated in the heart of Toronto, the company's headquarters pulsate with the energy of a city known for innovation and diversity. The company's app automates grocery and recipe lists, linking recipes to required products, and streamlining the shopping process. Today, AutoBasket assists households across the world to save time and energy on their weekly grocery runs, making it easier for busy families and individuals to get everything they need for a great meal.

AutoBasket has hired us as interns to focus on all areas of development for the organization and to help create solutions to common issues within the industry. Our role as interns is not only focused on the knowledge you have learned in school but also on developing your soft skills, including presentations, teamwork, and leadership. This internship will provide us with valuable real-world experience and an opportunity to contribute to the continued success of AutoBasket.

# COMPREHENSIVE PLAN FOR DATA COLLECTION

To comprehensively gather data related to Christmas recipes and food preferences in Canada from online forums and social media platforms, we have decided structured plan. Here is an outline of the plan covering data sources, data types, and methods for extracting and securely storing the data that we followed to comprehend our project objective.

## DATA SOURCES

For the immediate purpose, we have collected data from Online forums such as Reddit's food-related subreddits, cooking forums, and Canadian-specific culinary communities. Recipe websites like Good Housekeeping, Delish, AllRecipes Canada, and local Canadian cooking blogs.

To populate our dataset with varieties of data to prepare our platform to sustain for longer in the market we have also decided in future to collect the data form Social media platforms including Twitter, Facebook groups dedicated to food and cooking, and Instagram posts with relevant hashtags like #CanadianFood or #ChristmasRecipes.

## DATA TYPES

Since our project is about recommending the dishes, we have focused on the textual data. The data we have gathered somewhat commonly consists of the following titles in and around throughout the data Recipe titles, ingredients, instructions, user comments, and discussions related to Christmas recipes and food preferences. We have also collected the images so that later on we can use them to display in the front end with the name of the dishes.

## METHODS FOR EXTRACTION

In this sense, we are looking at the most comprehensive and the most efficient approach for this process to be convenient and can also be scalable when needed. The web scraping technique is essential for extracting information from different websites and forums. There is a scrapy toolkit in python used for efficient scraping. Nevertheless, we have been considering compliance to a scrapings website policies.

Employed by NLP methods, especially NER, we can obtain dish names, ingredients and other significant attributes from text data. NLTK (Natural Language Toolkit) packages have been taken advantage for this case.

## DATA PROCESSING AND STORAGE

After the Data Extraction, the second step has been about the cleaning of all the scrapped data. The pipeline includes the data cleaning process, which aims to eradicate noise, irrelevant information and duplicates from data extracted so that it is clear and clean. Using NLTK libraries for Named Entity Recognition (NER) to detect and obtain entity names, dishes, and other main class of phrases from the textual content.

Processed data is already stored in the data frame and later the data from the data done above will be stored in CSV format. A comprehensive plan has pushed us to collect accurate and appropriate data about Christmas recipe and food preferences from various online sources and to meet data safety, privacy and applicable policies and regulations.

# 

# DATA VISUALIZATION

# A diagram of a distribution of ingredients Description automatically generated

# CONCLUSION

# REFRENCES

1. (2024, 2 16). Retrieved from faker: https://faker.readthedocs.io/en/master/
2. *synthetic-data-generation-techniques*. (2024, 2 16). Retrieved from turing: https://www.turing.com/kb/synthetic-data-generation-techniques