Data Collection and Pre-Processing (DCPP)

DCPP Group Assignment

Term-1 Residency

Team:

Anuj Verma (12120050)
Kartikey Ajay Rai (12120014)
Krithika Nagendiran (12120015)
Navin Wadhwani (12120073)
Ratna Manedhar Punjala (12120024)

Table of Contents

1. Executive Summary

Problem Statement, Proposed Solution, Brief understanding of the challenges (1 page)

What is the problem? -

Why does one need a DATA SET? For example, in the current contemporary world – *looking for a recipe with specific ingredients, diet preferences, cuisine, occasion/festival* recipe with specific parameters is highly recommended and many follow a specific food preference. But it is not easy to find a recipe with a defined proportion of ingredients and its health effects, as there are so many recipes available on the internet. At the same time, it is astronomical task to filter out recipes with specific ingredient from those thousands of recipes out there. Hence, it becomes imperative that some smart applications be developed that can pull specific information about a recipe like a diet plan based on given ingredients or a famous delicacy based on the region and so on.

Indian cuisine reflects an 8000-year history of various groups and cultures interacting with the Indian subcontinent, leading to diversity of flavours and regional cuisines found in modern-day India. Later, trade with British and Portuguese influence added to the already diverse Indian cuisine. For all such smart applications the main ingredient is a SMART DATA SET.

What is the solution? –

Data plays a key role as enabler in assisting the resolution to a problem. A very well-defined database can further help to identify a quick solution with less complexity. In this assignment, the group of Data Collection students, present an algorithm that extracts the Indian food recipes and their details using techniques like web crawling and scraping and perform EDA to analyse the data, all using Python. This dataset can be further used to create a variety of FOOD RECIPE based applications. For example, a *Smart Chef* application can quickly recommend an Indian Cuisine for given set ingredients in a matter of few nanoseconds and likewise many other such applications can be built. All this possible given the INDIAN FOOD RECIPE database has that breadth and depth in its data content and structured to support such applications.

What are the Challenges? –

Some unique challenges encountered include:

- Identifying the seed source to begin data collection – Searching for recipe online would throw too many links and selecting the right link which would not only give a classical recipe but also ensure the methods are listed appropriately for any naïve chef to follow and achieve the outcome.
- Find the right web crawler to extract different recipe web links
- Identifying the best web scraping technique to extract useful information
- Crowdsourcing methods to further enhance the dataset
- EDA – exploratory data analysis to analyse the data collected

2. The chosen domain and Seed sources:

What analysis was conducted in this part? Why did you make those choices?

One of the best ways to approach a Data Collection problem is to choose a domain with a variegated attribute line.

As a team of distinct cultural composition and background experiences, everyone had a domain of choice to start the journey of data collection. As a beginning, we were given a set of 36 different domains. Everyone had a unique choice based on their preference but one thing that intrigued the team was the space of INDIAN FOOD RECIPES.

It was a general impression that, in today's mundane and busy world, food/diet/recipes have evolved, only to being more important than ever before. *Indian Food Recipe* is one such domain, which has thousands of years of history and overly complex with several variety of food types, cuisines, recipes, and ingredients that there is no one data type to bring them all under a single class.

One can draw a lot of variations in terms of data attributes like – ingredients, recipe, quantities, cooking time, taste and connects several other non-cooking aspects like culture, region, style, seasons, festivals and so on. Given this wide range of attributes, and common interest among the group members to explore the space, Team has finally decided to zero down on INDIAN FOOD RECIPES as our domain to conduct DCPP.

Once we zeroed on the domain, it was a daunting task to choose the right source for the data collection process. As a team we deliberated and looked up for all the prominent Indian Chef's and their work. We all agreed to look up to Padma Shri Chef Sanjeev Kapoor and was acknowledged as Government of India for his contribution and was awarded "National Award for Best Chef in India in 2017." We all have grown watching host popular TV show Khana Khazana. His easy demeanour and simple instructions make the most difficult recipe's easy to prepare.

3. Structured and unstructured sources:

What are the sources chosen and explain the reasoning behind the choices.

To begin the Data Cleaning Pre-Processing assignment, first critical step was to identify the starting point, the Seed source. Seed source had to be such that it can pull or connect to the many other web links/documents related to Indian food recipes. With that, team started off with some brain storming sessions and web trails and noted that many prominent chefs in India maintain their own Indian Cuisine website. Our obvious choice was to use the Structured data present on weblinks where information is stored in a database in a specific defined format like – Cuisine Type, Ingredients, Preparation Time, Preparation Method, User Review and so on.

Keeping the complexity of Unstructured data sources in view, and the purpose to make the best use of the DCPP techniques, team decided to use the prominent chef's websites as our seed source which presents data in structured format.

One of such prominent Indian Chef is Sanjeev Kapoor. He is a famous enthusiast among the food explorers and captured all his work on Indian Food into his master web link – www.sanjeevkapoor.com. This formed our seed source for data collection/exploration. There were similar other sources as well like – www.archanaskitchen.com, www.vahrehvah.com, www.yummyindiankitchen.com and so on.

4. Download/ crawl/ collect data from all the sources:

What are the methods used for this process? What are the challenges and how did you overcome them?

With the seed source fixed as weblink - www.sanjeevkapoor.com, the next challenge was to crawl the website to pull all the links to the various Indian food recipes and then pull the required data from these weblinks using web scraping techniques to create a dataset. However, retrieving data from weblinks was a challenging task as one had to deal with:

Crawling -

- Identifying and moving withing the website to gather all the links for recipes from seed source

Scraping –

- Web structure
- web formatting
- availability of the parameters
- and in general, fetching the required information from the browser

Translation –

- The dataset which was provided to us had Devanagari script in the dataset for many variables, and partial rows were translated to English. We tried to translate using Translator and google_trans tried to translate the Devanagari script to English Language. On researching google_trans has a known bug for JSON encode error which led us to move to Translator library.

To overcome issues with scraping, we used – Selenium library. This not only helps to move across different pages with the seed source but also provides capabilities to scrape through complex java scripts to extract data from the indexed webpages.
Team was able to pull close to 5000 weblinks for food recipes using the seed source. Next task was to perform web scraping on each recipe weblink to extract information for almost 25 attributes related to the recipe. These include:

- Ingredients
- Process
- Type of Meal
- Veg/Non-Veg/Vegan
- Equipment Required
- Cooking Duration
- ….

Refer Appendix-1 for the list of attributes.

Using Python web scraping library – Beautifulsoup, team was able to extract all the datapoints for various recipes that were present on the page.
Once the recipe weblinks were available as list, each link was loaded using WebDriver functionality, and information was extracted. This needed to view the source code which was viewable by right-clicking on the link and selecting INSPECT. This provides a TREE view with HTML codes. By searching specific TAG information within the HTML page,

several data points were extracted for the attributes. Often times, the tag information had some formatting which made it difficult to read the values of the attributes.

Example - While trying to scrape through a link to capture images for the recipes, many links returned with no images/no links. This was little confusing and then further checking these specific recipe pages, the images were assigned to a different tag. After multiple iterations, all the images for the recipes were obtained. However, it was also noted that, not all recipes have images and left them with a dummy image link.

5. Convert data from original sources to structured data fields:

Explain the process chosen and what are all the challenges taken?

Once the data was scraped (using BeautifulSoup library) a full set of unstructured data set was available. This needed transformation by applying formatting techniques to make it readable. Team utilized basic functionalities from Pandas library to format the data into structured data frame. Later *regex tool* was used to derive patterns from the data set and applied those patterns to generate a structured data frame.

6. Data cleaning/pre-processing as needed:

What data cleaning/ pre-processing techniques were taken up for this stage and why did you feel the requirement for it?

After fixing the data structure, next major step was to clean the data. This is a critical step in the process which would bring meaning to the structured data gathered. Several examples from the current analysis are presented below which defaults the importance of data cleaning at this stage of DCPP. However, the process of Data cleaning was a daunting task, as the data scraped through website was in many forms. For example, data was present in multiple languages i.e., English and Devanagari script, recipes were from different region i.e., Italian, Tex-Mex, etc., measurements were in different scales i.e., 1 tsp, 5 grams, 1 tablespoon, 15 grams, etc. To ensure the data is cleaned and in a standard structured format Pandas and NumPy libraries were used.

Below are a few steps performed during the data cleaning process:

- The first thing that needed a fix was - Total time (Prep Time + Cooking Time). Some of the dishes were in hours and others were in days. This step ensured that all the dishes are in the same time format.
- A list of all ingredients was available for each recipe but without a count. It is particularly important to know the number of ingredients which was calculated in this step.
- Based on the equipment used for the recipe, created a list of the utensils/equipment.
- Categorized the recipes based on their sugar content. Discovered that 1296 of them contained sugar and 3478 did not. An important detail for health-based applications.
- Categorized the recipes into vegetarian and non-vegetarian and found 731 were non-vegetarian while 4043 were vegetarian dishes.

- Categorized the recipes based on the presence of nuts as an ingredient, which resulted in a total of 106 dishes containing nuts.
- Categorized the recipe based on whether the dishes could be served hot or cold.
- Converted Devanagari script to English using google translate & Translator library in python. However, there was a data loss of 1196 recipes from the overall data scrapped as the library's failed to translate even after multiple attempts.
- Eventually, filtered out all non-Indian cuisines from the list and divided the data frame into categories based on the dishes that contained fruits and whether the fruits were seasonal or available all year.
- Likewise, there were several other clean-up activities performed on the data set to make it more meaningful.

7. Observations/ Insights and Analysis on the data collected

- Majority Indian recipes are Vegetarian and easy in preparation.
- Indian recipes inherit the cultural history from Ayurveda principles which relate to the human aspects of – Rajas, Tamas, and Satva. This is clearly observed in the recipes across India both in terms of preparation methods and ingredients.
- Lot of fusion recipes adapted from global cuisines and modified to suit the Indian pallet.
- Use of spices and condiments is predominant in Indian recipes.
- Milk and milk products form the major ingredients in the Indian classical desserts (Mithai).

8. Strategy to enhance the data with crowd sourcing methods:

What strategy did you implement to improve on the crowd sourcing methods and why?

Above methods to collect data include techniques that look at resources that are present ONLINE. This means that the sample data may be useful for developing applications that cater the needs of ONLINE population but not to cover the entire population that includes both ONLINE and OFFLINE customers/users. Indian Food Recipes is a domain that contain as much data online as there might be in offline sources.

So, to further refine the data set, one must adapt other methods of gathering the data from OFFLINE populations. Crowdsourcing is one of the best methods to gather such data. This could include methods like Surveys, interviews, feedbacks and so on.

For example, one can conduct an offline or online survey to gather information of the regional food delicacies, recipes, ingredients, spices information and get information of the existing food recipes as a customer review. These surveys can be done at several places like restaurants, hotels, fast food centres etc.

Strategic Initiatives for Crowd Sourcing to enhance the dataset:

1) Begin at Home – We are a group of 5 members, and all have minimum 500 friends each on Facebook, Instagram, and other social media channels. We would initiate a simple form to all our friends and ask them about sharing Ingredients Names in Local/regional languages which would help us enhance the data set and be able to make the recipes available in regional languages.
2) Feedback and Preference Capturing – We would also run a survey to capture preferences for food within our contacts and use the input to add more parameters and update the recipes to cater to the received preferences.
3) Reaching Industry Specialist – There are many Hotel Management & Catering Colleges in India and there are young students who are innovating and leaning. We would reach out to these colleges, ask for more recipes, and seek their input on the current recipes and share with them the dataset for their trial-and-error analysis.
4) Use of social media – In today's world where a lot of time is spent on social media, it is a channel we could not ignore. While we reached out our friends/contacts, here we would reach out to humans on a larger scale. We could replicate the simple survey to gather regional names of ingredients, regional recipes, fusion food recipes, etc.
5) Reach out Dietician – While food is an important aspect of Indians, today the trend is shifting towards healthy eating. It would not be possible to ignore aspect, we would reach out to the dietician's, health consultants, fitness trainers, etc. to gather dietary requirements for people performing special activities or sports or preference diets. This would enhance the dataset with recipes with diet needs and cater to the audience who are looking for dietary information for the food intake.

(Although there are several pros and cons associated with crowdsourcing methods, below we highlight only one of the major advantage and disadvantage of using this method.)

One of the big advantages of such activity is, many time the data points collected would not be influenced by the previous data which was the case with online resources where a user review could have been influenced by the other reviewers.

One major disadvantage of crowdsourcing method is – the repetitiveness in the data collection. Once the data is collected using a particular method, repeating it would require both time and money. While, in case of online resources, data can be easily updated very frequently from anywhere by just accessing the internet.

(Please note that, only the strategy to improve the data set has been suggested as part of crowdsourcing technique. But the current dataset does not include data from those methods as none of such methods feasible in the given context.)

9. References and Sources used for this Assignment

- www.stackoverflow.com – to learn and understand Selenium library
- https://github.com/ - GitHub for Beautiful Soup
- www.wikipedia.org/ -Wikipedia – to gather seasonal ingredients
- https://selenium-python.readthedocs.io/ - Selenium Documentation
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/ - BeautifulSoup Documentation

10. Appendix – 1 – Attribute List

Variables	Description	Data Type	ValueType
Recipe_Name	This variable is the name of the recipe	Categorical	Object
Main_Ingredients	List of Main Ingredients in the Recipe	Categorical	Object
Total_Time	This is the recipe (Prep time + Cooking time).	Categorical	Integer
Cuisine	This variable refers to the cuisine of various regions.	Categorical	Object
Ingredients_List	This variable represents the recipe's ingredient list.	Categorical	Object
Number_of_Ingredients	Total number of Ingredients in recipe	Categorical	Object
Prep_Time	Preparation time for a recipe	Categorical	Integer
Cook_Time	Cooking time for a recipe	Categorical	Integer
Serves	The number of people to whom the dishes can be served	Categorical	Integer
Level_of_Cooking	Indicates the level of cooking. Easy/Moderate/Difficult	Categorical	Object
Course	The recipes are divided into courses, such as main course/snacks.	Categorical	Object
Recipe_Link	Weblink to the Recipe		Link
Taste	Indicates taste of the recipe Sweet/Spicy/Tangy	Categorical	Object
Image_URL	Web address for the Recipe Image		Link
Recipe	Instructions for preparing the dishes	Categorical	Object
Equiment_Required	Equipment's needed to make the recipe	Categorical	Object
Contains_Sugar	Indicates whether the dish contains sugar or not	Categorical	Object
Veg/NonVeg	Tells us whether the recipe is vegetarian or not.	Categorical	Object
Contains_Nuts	Indicates whether the dishes contain nuts.	Categorical	Object
Served Hot/Cold/Normal	This variable indicates whether the recipe is served hot or cold.	Categorical	Object
Vegan/Non_Vegan	This variable indicates whether the recipe is vegan or not.	Categorical	Object
Region	Indicates whether the recipe is from the North or South or another region	Categorical	Object

Fruits_Present	Indicates whether the dish contains fruits.	Categorical	Object
Contains alcohol	Indicates whether alcohol is used in the recipe. Contain (1), if not (0)	Categorical	binary
Gluten Content	Indicates whether the recipe contains gluten.	Categorical	Object
Quick_Recipe	Indicates whether the recipe can be made quickly (1) or requires preplanning (0)	Categorical	binary
Seasonal	This variable represents the availability of ingredients based on the season.	Categorical	Object
Accompaniments	This variable determines whether or not the recipe includes Accompaniments.	Categorical	Object
Jain/Non-Jain	This variable classifies food as Jain or Non-Jain.	Categorical	Object
Diabetic Friendly	Indicates whether or not the recipe is suitable for diabetics.	Categorical	Object
Diet	Tells us what kind of diet recipe it is.	Categorical	Object