Allergen Filtered Recipe Searching: Research Project Proposal

By Billy Bonds Grant (220006428)

1. Introduction

The advent of data science has opened doors to innovative applications that are transforming lives. This proposal introduces a research project aiming to leverage data science to address a prevalent issue: dietary restrictions due to food allergies. The project, titled 'Allergen Filtered Recipe Searching', focuses on using web scraping techniques to develop a tool that offers personalised recipe suggestions for individuals with specific dietary restrictions, enhancing their culinary journey. This tool will filter out online recipes containing specific allergens, tailored to the user's needs.

The importance of this project lies in its potential to substantially improve the quality of life for people living with food allergies. It serves as a practical solution to the often challenging and time-consuming task of finding suitable recipes. Furthermore, it showcases how data science can be used to cater to unique and individual needs, particularly in the realm of food and nutrition.

This proposal will detail the project's research description, expected outcomes, necessary resources, potential constraints, as well as social, ethical, professional, and legal considerations. We will also delve into the proposed research questions, design and methods, initial literature search, research outline, and the approach to the tool's development.

2. Project Description

Food allergies present a significant obstacle in the pursuit of culinary diversity, a struggle experienced by numerous individuals worldwide. With an array of over 170 food substances known to trigger allergic reactions, locating appropriate recipes amidst the vast expanse of online resources is a daunting task. The proposed project aims to address this predicament by utilising data science methodologies to develop an efficient, user-friendly, web-based application that filters recipes according to specified allergen restrictions.

This project, while not driven by a specific workplace requirement or client request, targets a pervasive global issue. As per the World Allergy Organization, it is estimated that food allergies may affect between 220 and 250 million individuals worldwide. By aiming to reduce the time and stress associated with the manual filtration of recipes, the project proposes to improve the culinary experiences of these individuals.

The principal beneficiaries of the proposed project are individuals with allergen-induced dietary restrictions. Nevertheless, the ripple effect of such an application is extensive. Those who prepare meals for individuals with allergies, dieticians creating customised meal plans, and application

developers seeking to incorporate such functionality into their platforms could also reap benefits from the project.

The tangible deliverables of the proposed project encompass a web-based application capable of extracting recipes from a variety of online sources and filtering them according to designated allergen categories. With an emphasis on user experience, the application will be developed with an intuitive user interface that facilitates easy input of dietary restrictions and generates a customised list of suitable recipes. The efficacy of the application will be measured through an assessment of its accuracy and efficiency in sourcing allergen-free recipes. Thus, the project not only presents a practical solution to a prevalent issue but also enriches the existing knowledge on the applicability of web scraping techniques in solving real-world challenges.

3. Preliminary Literature Review

The literature review for this project will focus on two main areas: the understanding of food allergies and allergens, and the use of web scraping techniques for data acquisition.

3.1. Food Allergies and Allergens

Food allergies occur when the human immune system erroneously identifies specific proteins in food as harmful, initiating an allergic response. A select group of allergens - namely, milk, eggs, peanuts, tree nuts, wheat, soy, fish, and shellfish - account for a majority of food allergies. However, it is notable that the European Union recognises a total of 14 allergen groups, inclusive of cereals containing gluten, crustaceans, molluscs, mustard, sesame seeds, celery, and lupin (Turner et al., 2015).

The incidence of food allergies, akin to other forms of allergies, has demonstrated an escalating trend in recent decades, predominantly within developed nations. It is postulated that up to 20% of the populace claim to have food allergies. In the United Kingdom alone, it is approximated that two million individuals are grappling with a diagnosed food allergy (Prescott and Allen, 2011).

The underlying mechanisms of food allergies are intricate and remain partially understood, but they inherently involve a blend of genetic and environmental factors. Empirical evidence, such as that provided by Muthukumar et al. (2020), illustrates that specific food allergies, including peanut allergies, may be hereditary, thereby suggesting a genetic component. Concurrently, environmental elements, such as diet, timing of the introduction of particular foods to infants, and gut microbiota, also contribute to the onset of food allergies.

The implications of living with food allergies can profoundly impact an individual's lifestyle. Beyond the inherent risk of severe allergic reactions, these individuals must ascertain safe foods for consumption, a task that may prove burdensome when dining out or consuming pre-packaged foods, which may potentially contain concealed allergens (Leftwich et al., 2011).

Food allergy management primarily necessitates the avoidance of allergenic foods, thus requiring the precise labelling of food allergens in pre-packaged foods. This is mandated by law in numerous countries, including the United Kingdom, where the Food Information Regulations 2014 stipulate that food businesses must furnish information regarding the 14 recognised allergens (FSA, 2021).

However, such rules are not necessarily applicable to recipes located online, thereby exacerbating the difficulty faced by individuals with food allergies in their quest for suitable recipes. This accentuates the necessity for innovative solutions that can aid individuals with food allergies to safely and efficiently traverse the culinary landscape (Muthukumar et al., 2020).

3.2. Web Scraping Techniques

Web scraping, at its core, is the automated process of collecting large volumes of data from websites and storing it locally, typically arranged in a coherent format (Patel, 2020). This technique

is a cornerstone of data gathering from the web, especially when collecting data manually would be impractical due to the data's enormous volume (Munzert et al., 2015).

The execution of web scraping can be done through various methods. For tasks on the simpler side, software tools like import.io or ParseHub come in handy with user-friendly interfaces that help identify the data to be collected. But when it comes to complex tasks or if we're talking about a grand scale of data extraction, crafting a custom web scraper in programming languages like Python or Java offers the needed flexibility and control (Russell, 2018).

Web scraping usually follows three key steps: requesting the website server to download the page, analysing the web page to pinpoint the wanted data, and then pulling out and storing this data (Munzert et al., 2015). Patel (2020) details how these steps can range in complexity levels, all depending on the structure and size of the data at hand.

The first step involves connecting with the website server through HTTP or a web browser, then pulling the HTML content of the page. A common tool to perform this task is Python's requests library (Mitchell, 2015).

Next comes parsing the web page, essentially sifting through the HTML to find the required data. Here, Python's BeautifulSoup library shines with its functions that allow for navigating, searching, and tweaking the parse tree (Richardson, 2007).

The final step is the extraction of data. Once we find the needed elements within the parse tree, the data can be pulled out and stored in an organised format, like a CSV file or a database. This is the step where the durability of the scraper is put to the test, particularly when handling big data scales (Patel, 2020).

However, we should not forget that web scraping has its own set of ethical and legal considerations. Numerous websites have terms of service that limit or totally forbid web scraping. Also, certain laws, like the Computer Fraud and Abuse Act (CFAA) in the U.S. or the General Data Protection Regulation (GDPR) in the EU, impose regulations and limits on this practice (Krotov, 2017). We are aware of the importance of understanding and abiding by these laws and guidelines to ensure legal compliance (Patel, 2020).

3.3. Combining the Two Fields

The merging of web scraping techniques and food allergen identification provides an exciting opportunity for innovation. This intersection, rooted in both computer science and food science, allows for the creation of tools that can cater to specific dietary requirements by providing personalised recipe recommendations free from specified allergens (Chen, Y. et al., 2018).

The synergy between these two fields forms the basis of our research project. Our aim is to develop a web scraping tool that can analyse the ingredients listed in online recipes, identify potential allergens, and filter out those recipes that contain specified allergens. This tool can help those with food allergies to easily find suitable recipes, potentially improving their dietary diversity and quality of life (Wong, D., 2020).

Web scraping techniques are a crucial part of this project, as they provide a means to gather the necessary data from online recipes. The recipe data can then be analysed for the presence of the 14 allergen groups defined by the UK's Food Standards Agency (FSA, 2021). This fusion of web technology and food allergen knowledge is what allows our project to deliver value to those living with food allergies.

Furthermore, web scraping can be used to keep the database of recipes up-to-date, providing users with new, safe, and diverse meal options. It can help circumvent the barriers often faced by individuals with food allergies when it comes to meal planning and preparation, namely limited options and fear of cross-contamination (Netting, M. et al., 2017).

However, it's important to note the challenges and limitations in combining these two fields. The accuracy of allergen identification in the scraped recipes will be dependent on the clarity of the ingredient list and the knowledge base of allergens utilised. In addition, adherence to ethical and legal web scraping practices must be ensured (Krotov, V., 2017).

4. Research Questions, Design and Methodology

The primary goal of this research is to devise a user-friendly, web-based tool, leveraging web scraping methodologies to offer personalised recipe recommendations to individuals with particular food allergies. The primary research questions are:

I. What is the global prevalence of food allergies and how do these allergies influence the culinary experiences of the individuals affected?
II. What technological solutions are currently available for people with food allergies in search of suitable recipes, and what limitations do these solutions present?
III. How can we effectively utilise web scraping techniques to resolve the challenge of identifying recipes free of specific allergens?

The anticipated outcomes of the research are to gain a thorough understanding of food allergies, the existing technological aids, and the application of web scraping techniques, to ultimately create a tool that addresses the identified problem.

Given the nature of this project, the research does not require a hypothesis as the focus is not on verifying or falsifying a proposition, but rather on developing a pragmatic solution to a delineated problem.

Regarding the research design, this project employs an applied research approach, centred on addressing a specific problem and being inherently practical. It involves two fundamental stages: comprehending the problem through literature reviews and user interviews, followed by developing and testing the proposed solution.

The methodology involves a qualitative approach to understand the implications of food allergies on individuals' culinary experiences. This approach includes conducting a comprehensive literature review and potentially administering interviews or surveys with people affected by food allergies, dieticians, and other relevant stakeholders.

The tool's development will follow an iterative approach, harnessing web scraping techniques to gather recipes from various online sources, and employing machine learning techniques to accurately identify and exclude allergens from the recipes. The tool's effectiveness will be gauged based on its proficiency in filtering allergens and feedback from users regarding its usability.

It is crucial to note that while web scraping is a potent technique for data acquisition, it carries with it ethical considerations. This project will ensure compliance with the terms of service of websites, respect the limitations imposed by robots.txt files, and refrain from inundating servers with excessive requests within a short span.

The research design and methodology are well-aligned to address the research questions and achieve the research aim. This framework provides a holistic approach to understanding the problem, devising a solution, and assessing its effectiveness.

5. Resources and Constraints

To successfully execute this research project, various resources will be required:

Hardware and Software: A high-performance computing system with robust internet connectivity is paramount for implementing the web scraping operation, analysing data, and developing the tool. The project will leverage open-source software such as Python, which boasts

libraries like BeautifulSoup for web scraping and TensorFlow for machine learning purposes. Additionally, a web hosting service will be needed for deploying the web-based tool.

Datasets: An expansive set of food recipes will be crucial as the project's dataset. The data will be procured by web scraping popular culinary websites that offer a wide range of dishes. The dataset must be substantial to ensure that the developed tool can provide a diverse array of recommendations to the users.

Access to Participants: To evaluate the tool's effectiveness and usability, it is essential to gain access to individuals living with food allergies, who can offer valuable feedback. Partnerships with allergy support groups or dietitian networks could enable this access.

Despite meticulous planning, there are potential constraints that could impact the project:

Data Privacy and Legal Restrictions: Web scraping, despite being a potent technique, is bound by certain ethical and legal constraints. Some websites may prohibit scraping, which could restrict the variety of recipes collected. To circumvent this, the project will adhere strictly to the legal regulations and terms of service of each website.

Quality of Scraped Data: The quality and structure of the data obtained from disparate websites may differ, which could influence the consistency of the results. A solution to this issue is to implement rigorous data cleaning and preprocessing procedures.

Time Constraint: The time necessary for data collection, tool development, and testing could pose a challenge. A detailed project timeline will be established and adhered to, ensuring the project's completion within the stipulated timeframe.

Although this project necessitates specific resources and confronts potential constraints, with careful planning and adherence to ethical guidelines, it is anticipated to reach successful completion.

6. Social, Ethical, Professional, and Legal Considerations

The execution of this project invites a series of social, ethical, professional, and legal aspects that need meticulous consideration.

Social Considerations: This project is conceived with the intention of improving the gastronomic experiences of people suffering from food allergies, thereby addressing a considerable societal issue. It aims to establish a platform tailored to the dietary requirements of a significant portion of society, promoting inclusivity. However, it is crucial to ensure that the developed tool doesn't unintentionally exclude or discriminate against any individuals or groups. For instance, it should accommodate a broad range of cultural dietary preferences and restrictions, extending beyond allergens.

Ethical Considerations: The principal ethical aspect concerning this project involves the methodology of data collection. Web scraping can often be a contentious technique, as it involves the procurement and utilisation of data from websites. This project will comply with all ethical standards during the data scraping process. This compliance entails respecting the robots.txt files of websites, abstaining from overburdening servers with excessive requests, and utilising the data purely for research objectives, devoid of any commercial intentions.

Professional Considerations: This project will uphold the highest echelons of professional standards throughout its lifecycle. This commitment involves ensuring the integrity of the data and the functionality of the tool, preserving user privacy, and guaranteeing the tool's accessibility to users. The team will maintain transparency regarding the tool's capabilities and limitations and will actively respond to user feedback for continuous enhancement of the tool.

Legal Considerations: Web scraping is lawful when conducted in adherence to the law and the terms of service of websites. This project will conform stringently to these guidelines, thereby ensuring that any data procured does not violate copyright laws or infringe upon privacy laws.

Moreover, as the tool will offer recipe suggestions to individuals with food allergies, it is critical to clearly state that the tool operates as a recommendation system and does not substitute professional medical advice. Users should be strongly advised to consult with their healthcare provider or a certified dietitian prior to making substantial alterations to their diet.

To ensure these considerations are addressed effectively, they will undergo regular review throughout the project's course. The overarching aim is to develop a valuable tool for individuals with food allergies while maintaining the highest ethical, professional, and legal standards.

7. References

Brefeld, U. et al. (eds) (2019) Machine learning and data mining for sports analytics: 5th international workshop, MLSA 2018, co-located with ECML/PKDD 2018, Dublin, Ireland, September 10, 2018: proceedings. Cham: Springer (Lecture notes in computer science Lecture notes in artificial intelligence, 11330).

Brough, H.A. et al. (2014) 'Peanut allergy: Effect of environmental peanut exposure in children with filaggrin loss-of-function mutations', Journal of Allergy and Clinical Immunology, 134(4), pp. 867-875.e1. Available at: https://doi.org/10.1016/j.jaci.2014.08.011.

Chen, Y., Elenee Argentinis, J. and Weber, G. (2016) 'Ibm watson: how cognitive computing can be applied to big data challenges in life sciences research', Clinical Therapeutics, 38(4), pp. 688–701. Available at: https://doi.org/10.1016/j.clinthera.2015.12.001.

Keen, J. and Malby, R. (1992) 'Nursing power and practice in the united kingdom national health service', Journal of Advanced Nursing, 17(7), pp. 863–870. Available at: https://doi.org/10.1111/j.1365-2648.1992.tb02009.x.

Krotov, V. (2017) 'The Internet of Things and new business opportunities', Business Horizons, 60(6), pp. 831–841. Available at: https://doi.org/10.1016/j.bushor.2017.07.009.

Leftwich, J. et al. (2011) 'The challenges for nut-allergic consumers of eating out: The challenges for nut allergic consumers', Clinical & Experimental Allergy, 41(2), pp. 243–249. Available at: https://doi.org/10.1111/j.1365-2222.2010.03649.x.

Mitchell, R.E. (2018) Web scraping with Python: collecting more data from the modern web. Second edition. Sebastopol, CA: O'Reilly Media.

Munzert, S. (2014) Automated data collection with R: a practical guide to Web scraping and text mining. Chichester, West Sussex, United Kingdom: Wiley.

Muthukumar, J. et al. (2020) 'Food and food products associated with food allergy and food intolerance – An overview', Food Research International, 138, p. 109780. Available at: https://doi.org/10.1016/j.foodres.2020.109780.

Netting, M.J. et al. (2017) 'An australian consensus on infant feeding guidelines to prevent food allergy: outcomes from the australian infant feeding summit', The Journal of Allergy and Clinical Immunology: In Practice, 5(6), pp. 1617–1624. Available at: https://doi.org/10.1016/j.jaip.2017.03.013.

Patel, J.M. (2020) Getting structured data from the Internet: running web crawlers/scrapers on a big data production scale. New York, NY: Apress.

Prescott, S. and Allen, K.J. (2011) ‘Food allergy: Riding the second wave of the allergy epidemic: The food allergy epidemic’, Pediatric Allergy and Immunology, 22(2), pp. 155–160. Available at: https://doi.org/10.1111/j.1399-3038.2011.01145.x.

Richardson, L. (2007) 'Beautiful Soup Documentation', Crummy.

Russell, M.A. and Klassen, M. (2018) Mining the social web. Third edition. Beijing ; Boston ; Farnham ; Sebastopol ; Tokyo: O’Reilly.