Web scraping back of pack food label data from online retailer in the UK

Student: Maria Galazoula Student ID: 201075916
Supervisors: Prof Janet Cade, Dr Darren Greenwood, Dr Adam Martin and Prof Mark Birkin

Introduction

- New trends in diet with "free from" food items, such as vegan, dairy free, gluten free, etc (Adiamo et al, 2017; Walker et al, 2018).
- Nutritionists usually rely on third-party companies on accessing data.
- There is the need of updating food databases more efficiently and frequently (Baskaran and Ramanujan, 2018).
- Web scraping could provide the solution to this issue.

Methodology

- Due to time constraints, the focus will be Tesco, as the "Grocery search" API provides the product IDs (Figure 1).
- The product ID is also the unique ID of each link of the products.
- The products that were scraped were from the gluten free category.
- One function was created to automatically scrape and store all the information from the products (Figure 2).

Results and Future plans *

- Results were very promising for creating a food database efficiently using automated data extraction techniques, i.e. web scraping.
- The process is cheap and relatively fast.
- Future plans include getting data straight from the websites of other retailers.
- Instead of one function, creating multiple functions that scrape the product data.

Figure 1. APIs provided by Tesco.

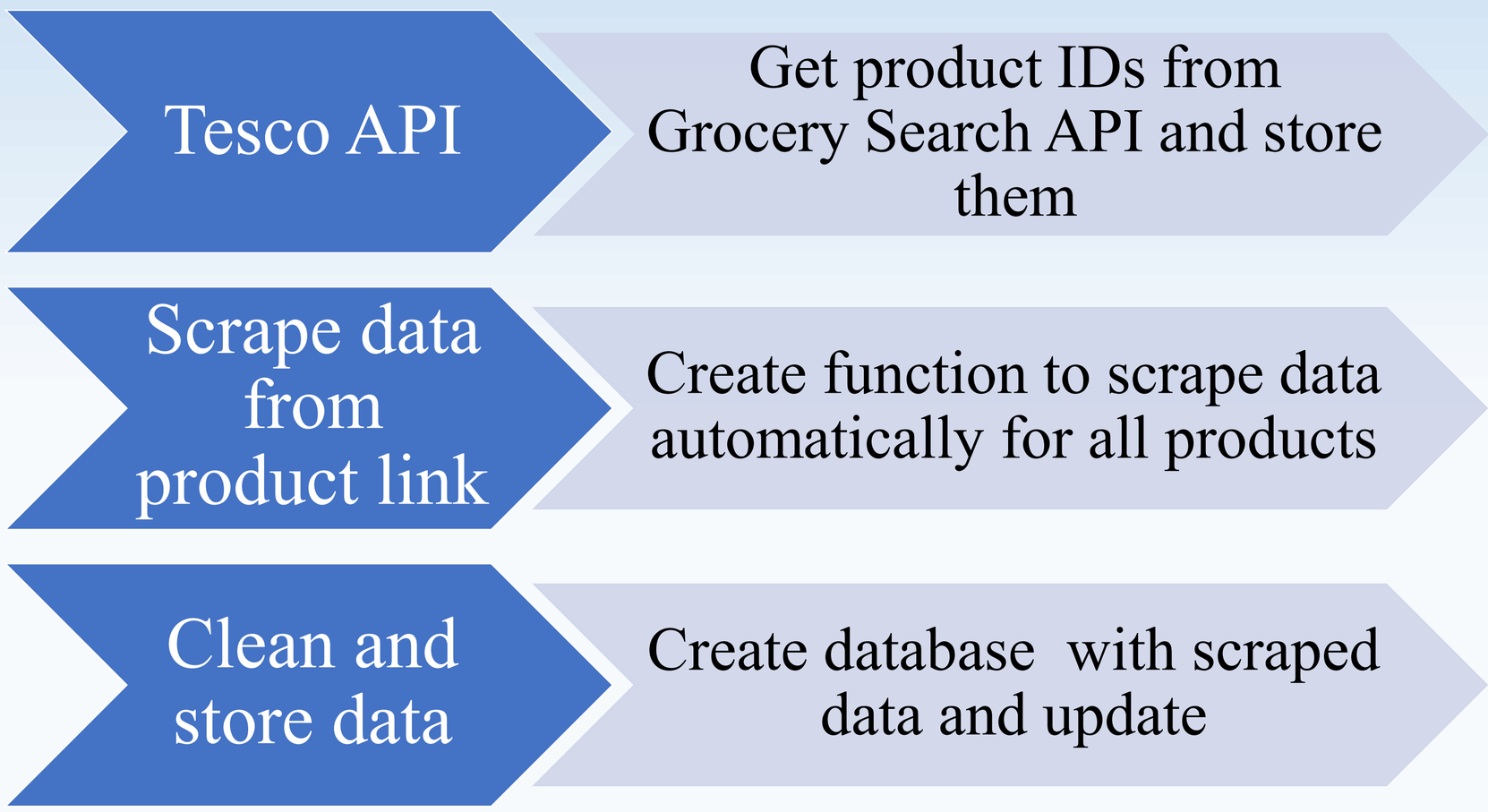

Figure 2. Work flow of web scraping data from online retailer.

Key References

Adiamo, O.Q., Fawale, O.S. and Olawoye, B. 2017. Recent Trends in the Formulation of Gluten-Free Sorghum Products. *Journal of Culinary Science & Technology*. **16**(4), pp.311–325.

Baskaran, U. and Ramanujam, K. 2018. Automated scraping of structured data records from health discussion forums using semantic analysis. *Informatics in Medicine.* **10**, pp.149–158.

Walker, A.J., Curtis, H.J., Bacon, S., Croker, R. and Goldacre, B. 2018. Trends, geographical variation and factors associated with prescribing of gluten-free foods in English primary care: a cross-sectional study. *BMJ Open*. **8**, p.21312.

*No data produced by this project was used by Dietary Assessment Ltd due to the Terms and Conditions of Tesco.

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
UNIVERSITY OF LEEDS
myfood24
Data Analytics & Society