Mitigating Hazards of Food Allergens

Mariam Hakobyan, Devavrat Tomar, Harshdeep Singh
EPFL

Objectives

OpenFood Facts dataset with food allergens features was processed to

1. Identify possible allergies that a given unlabeled food item may cause,
2. Recommend food items that have specific nutrients, without allergic risks.

Methods applied

- *Word2Vec* to group different allergens together for their further predictions based on the *food categories* present in the dataset,
- *Random Forest* model with average test accuracy of about **89.88**% for the task of predicting allergens,
- *Binary Search Tree based Vector Quantization* for efficient search of food items with specific nutrition and allergic limitations.

Introduction

Working on *Open Food Facts* dataset by 5000+ contributors with 600000+ products from 150 countries is not only an interesting data science task for analyzing ingredients, allergens and nutrition facts of food, but also challenging in terms of its incompleteness and multilingual data representations. In this project, the distribution of allergens and additives present in the food products were explored, *Word2Vec* was applied to cluster allergens into meaningful categories and classification on allergen categories based on ingredients and food categories. Binary search tree based fast vector query algorithm was implemented for recommending food items with allergic constraints.

Data preprocessing

Initially, there were more than 3000 unique allergens in the dataset. Because of the open source contribution, the allergens were represented in more than 15 languages, detected with the help of the **langdetect** library. Manual translation was the only way to make the text linguistically uniform. The allergens just made up of numbers and unwanted characters were removed from observations. Sentence representations were converted into meaning expressing words using **nltk** and taking only nouns and adjectives.

Food Additives: The purposes and effects

Many people are allergic and show adverse effects on consuming food with certain types of additives, added to food products for keeping the safety, taste and freshness.Purposes and effects of certain groups of additives are observed by clustering them based on the characterizing code level **e-xxx**, into 8 main categories.
We have observed that the most common additives in the dataset are acidity regulators which control the acidity level of the food to prevent from spoilage.
Brands like *Carrefour*, *Auchen* and *U-food* companies are the three top companies in terms of almost all additives' usage in their production.
Products with additives as common ingredients:

- Ice creams (gelatin) and candies mostly have *coloring* additives
- Condensed soups, chips contain most of the *flavour enhancers*
- Cookies and dough related products contain *anti-caking agents.*

Word2Vec Features

Since there were more than 1500 unique allergens we grouped them together according to the higher level categories they belong to. If the allergen is made up of multiple words such as 'milk solids', we find a vector for both 'milk' and 'solids'; add them together to get a resultant vector. The words are grouped then on the basis of KMeans clustering into 7 groups such as **acids**, **proteins**, **wheat/flour** etc.

Efficient Vector Query on Nutrients

To query the dataset for food items having similar nutrition content, one trivial algorithm is to find n nearest neighbours of the given nutrition vector by comparing it will all the food items. The time complexity for this trivial algorithm is $\mathcal{O}(n)$, which is too slow for querying around $600,000$ food items. Thus, we used Binary Search Vector Quantization **(BSVQ)** [1], [2] to store all food items into buckets as leaves of a BST. Vector Quantization is commonly used to quantize vectors into K codebook symbols (or clusters). Unlike K-means algorithm, Vector Quantization can be used to construct a BST for vectors as after every iteration, the number of clusters become twice and the algorithm ends in $\mathcal{O}(log(K))$ iterations.

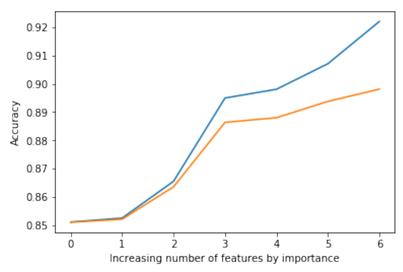

Figure 1: Accuracy increase as number of features increase

Predicting allergens from categories

Given categories for a product, we predict the allergens. We used 8300+ categories for this and used the Random Forest model since the training times are faster and it handles unbalanced data well. The model was trained on 58496 products and 89.88% accuracy was achieved. Also, the random forest model was used to observe which features were most important: **fishes**, **sugary snacks** were some of them.

Important Result

	Protein	Fat	Carbohydrate	Sugar	Salt	Sat. Fat	Allergen
Query	9.30	8.20	20.70	0.60	0.85	3.10	wheat/flour, fruit/nut
Results	9.00	6.10	23.90	3.70	1.11	2.80	seafood
	8.50	4.00	19.40	1.40	0.92	1.20	seafood, powder/protein
	10.10	4.60	22.10	4.60	1.00	1.00	seafood
	8.50	4.00	19.40	1.40	0.92	1.20	seafood, powder/protein

Table 1: Example - BSVQ query for $100g$ of food item

Conclusion

With only 10% of food items having allergen labels, in this project we trained a machine learning algorithm that can effectively predict possible allergens given the food categories. However, we think that ingredients of food items would give more suitable features for this type of classification problem but for the current dataset, the ingredient content of food items is very challenging to preprocess and extract useful features for training. The major challenges are non availability of ingredients text in one major language and sparse distribution of languages.

Results

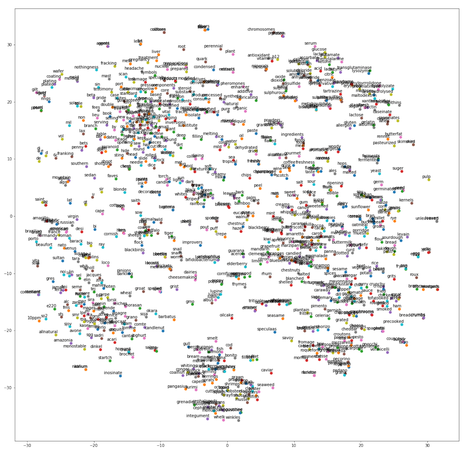

Figure 2: Word2Vec Allergens grouping

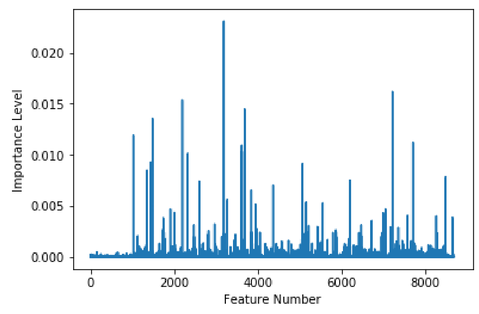

Figure 3: Important features by Random Forest

References

[1] A Lowry, S Hossain, and W Millar.
Binary search trees for vector quantisation.
pages 2205 – 2208, 05 1987.

[2] I. Katsavounidis, C. C.J. Kuo, and Zhen Zhang.
Fast tree-structured nearest neighbor encoding for vector quantization.
Trans. Img. Proc., 5(2):398–404, 1996.

Reach us at

- Repository: https://github.com/epfl-ada-2018/Project_dev_hak_deep