Chip Shop Map of the UK

Jiahao Cao

August 23, 2018

MSc in High Performance Computing

The University of Edinburgh

Year of Presentation: 2018

Abstract

In the Internet era, an increasing number of restaurants are promoting their restaurants via the Internet. This has produced huge amounts of website data and these data are very messy. In the past, these restaurant data may only help diners to order food. However, in the eyes of data workers in the era of big data, there is a lot of potentially valuable information in these restaurant data. 'Fish & Chips' is one of the most famous dishes in the UK, and the dissertation focuses on 'Fish & Chips' shops' menu data to explore regional differences in the UK through the distribution of menu content.

The dissertation describes the entire exploration process from data acquisition to obtaining regional results. Specifically, it contains data source selection, methods for retrieving data from websites of 'Fish & Chips' shops, the procedure of cleaning website dataset, visualisation of the geographical distribution of menu content, the process of exploring regional features and machine learning methods used for classifying regional and non-regional content. The project finally found some features of regional content, regional dishes, regional expressions, and some speculations about regional results.

This is an innovative and exploratory study that there are no existing ideas or methods can be referred to and all ideas and methods require the dissertation to explore and evaluate. Thus, the methods, algorithms, and findings used in this dissertation may provide some reference for similar research in the future.

Contents

Chapter 1 Introduction....1

1.1 Dissertation background....1
1.2 Research aim and research focus....2
1.3 Research methods....2
1.4 Value of the research....3
1.5 Structure of the dissertation....3

Chapter 2 Background Theory....5

2.1 Web crawling....5
2.2 HTML data cleaning techniques....5
2.2.1 HTMLParser....6
2.2.2 Natural Language Processing (NLP)....7
2.3 Geographic data visualisation....8
2.3.1 Central point calculation algorithm....8
2.3.2 Visualisation with Matplotlib....9
2.4 Data mining with machine learning methods....10
2.4.1 Decision tree....10
2.4.2 Logistic regression....12

Chapter 3 Iteration 1....14

3.1 Methodology....14
3.1.1 Data acquisition....14
3.1.2 Data cleaning....15
3.1.3 Data visualisation....17
3.2 Findings....19
3.2.1 Regional words findings....19
3.2.2 Non-regional words findings....21
3.3 Evaluation and improvement....24
3.4 Summary and future work....28

Chapter 4 Iteration 2....29

4.1 Methodology....29
4.1.1 Training dataset....29
4.2 Findings....30
4.3 Evaluation and improvement....30
4.3.1 Cart algorithm findings and evaluations....32
4.4 Summary and future work....33

Chapter 5 Iteration 3....34

5.1 Methodology....34
5.2 Findings....35
5.3 Evaluation....35
5.4 Summary and Future Work....37

Chapter 6 Iteration 4 38
6.1 Noun phrase 38
6.1.1 Noun phrase methodology 38
6.1.2 Noun phrase decision tree findings 39
6.1.3 Noun phrase decision tree evaluation 40
6.1.4 Noun phrase logistic regression findings 40
6.1.5 Noun phrase logistic regression evaluation 40
6.2 Word pair 41
6.2.1 Word pair methodology 41
6.2.2 Word pair decision tree findings 41
6.2.3 Word pair decision tree evaluation 43
6.2.4 Word pair logistic regression findings 43
6.2.5 Word pair logistic regression evaluation 43
6.3 Summary 44
Chapter 7 Conclusion 45
Chapter 8 Future Work 47
8.1 Finding more 'Fish & Chips' shop menus 47
8.2 In-depth study of algorithms and parameters of classification models 47
8.3 Synonym detection 47
8.4 Extend the project to a wider range 48
Appendix A Logistic regression probability results 49
References 57

List of Tables

Table 1: Regional words with ‘ratio’ 24

Table 2: Non-regional words with ‘ratio’ 25

Table 3: Regional words with ‘average distance’ 26

Table 4: Non-regional words with ‘average distance’ 27

Table 5: Regional words with ‘proportion’.....27

Table 6: Non-regional words with ‘proportion’.....28

Table 7: Comparison of regularization choices for independent word when selecting ‘proportion’, ‘ratio’ and ‘average distance’ as features.....35

Table 8: Comparison of regularization choices for noun phrase.....40

Table 9: Comparison of penalty choices for word pair when selecting, ‘proportion’, ‘ratio’, ‘city number’ and ‘average distance’ as features 43

List of Figures

Figure 1: Workflow Diagram 3

Figure 2: HTML source code with HTML style of one of ‘Fish & Chips’ shops’ websites 7

Figure 3: HTML source code of one of ‘Fish & Chips’ shops’ websites.....7

Figure 4: geographical coordinate system with a cartesian coordinate systems 8

Figure 5: Haggis distribution range 9

Figure 6: Data operation flow chart of Map script and Reduce script.....15

Figure 7: The result of Reduce script.....17

Figure 8: Conjecture map of regional distribution words 18

Figure 9: Conjecture map of non-regional distribution words 18

Figure 10: ‘haggis’ distribution (95%) 19

Figure 11: The number of ‘haggis’ shops varies with distance.....19

Figure 12: ‘bru’ distribution (95%) 20

Figure 13: The number of ‘bru’ shops varies with distance 20

Figure 14: ‘naan’ distribution (95%) 20

Figure 15: The number of ‘naan’ shops varies with distance.....20

Figure 16: ‘roe’ distribution (95%).....20

Figure 17: The number of ‘roe’ shops varies with distance 20

Figure 18: ‘supper’ distribution (95%) 21

Figure 19: The number of ‘supper’ shops varies with distance.....21

Figure 20: ‘pakora’ distribution (95%).....21

Figure 21: The number of ‘pakora’ shops varies with distance.....21

Figure 22: ‘chip’ distribution (95%).....22

Figure 23: The number of ‘chip’ shops varies with distance 22

Figure 24: ‘sausage’ distribution (95%) 22

Figure 25: The number of ‘sausage’ shops varies with distance.....22

Figure 26: ‘supreme’ distribution (95%) 22
Figure 27: The number of ‘supreme’ shops varies with distance 22
Figure 28: ‘gift’ distribution (95%) 23
Figure 29: The number of ‘gift’ shops varies with distance 23
Figure 30: ‘soup’ distribution (95%) 23
Figure 31: The number of ‘soup’ shops varies with distance 23
Figure 32: ‘daily’ distribution (95%) 23
Figure 33: The number of ‘daily’ shops varies with distance 23
Figure 34: ‘massala’ distribution (95%) 25
Figure 35: The number of ‘massala’ shops varies with distance 25
Figure 36: ‘funghi’ distribution (95%) 26
Figure 37: The number of ‘funghi’ shops varies with distance 26
Figure 38: ID3 algorithm decision tree 30
Figure 39: Cart algorithm decision tree 32
Figure 40: Noun phrase decision tree 39
Figure 41: Word pair decision tree 42

Acknowledgements

First of all, I would like to thank my supervisors, Ally Hume and Jane Kennedy for their patient counselling throughout my dissertation. When I met problems, they always gave me timely constructive advice and these suggestions made me learn a lot during the dissertation. Sometimes they would explain to me very patiently in order to let me understand the problem because my native language is not English. For my writing of the dissertation, they gave me a lot of feedback which really helped me to improve a lot.

Besides, I would like to thank all the teachers who taught me knowledge during my postgraduate studies.

In addition, I want to thank my junior high school teacher, Ms. Ma, who is the most important teacher in my life. She made me develop the good habit of studying hard and taught me to be a responsible person. Without her, I might not have the chance to come to Edinburgh to study.

Finally, I want to thank my parents for their unconditional support for my studies and always give me encouragement and love. They are the most important part of my life, thanks for everything they have done for my growth.

Chapter 1

Introduction

This chapter provides a general overview of the dissertation, including dissertation background, the aim of this dissertation, the methods used in this dissertation, study value, and main structure of this research.

1.1 Dissertation background

In the era of big data, a lot of messy data is generated from many fields, such as industrial field, business field, and research field. Messy data is a kind of data which cannot provide clearly interpretable information directly [1]. However, there is plenty amount of potentially valuable information in messy data. Thus, processing messy data is a valuable activity that the messy data can be converted into structured data to facilitate analysis [2] to help engineers or statisticians get more valuable information. This project mainly focuses on processing messy data which generated in the catering field.

In the era of rapid development of the Internet and smartphones, searching restaurants or specific food has become easier. For example, with Google, customers could find the type of restaurant or food they are interested in more simply and more accurately. At the same time, restaurant operators can use these techniques to better expand their business [3]. On discovering this business opportunity, an increasing number of restaurant operators are advertising the link to their menu on some information applications and websites, such as Google and TripAdvisor [4]. This has resulted in a large amount of restaurant information such as menu and restaurant description, being publicly available in a digital form. In terms of ordinary diners, the content in restaurants' websites can only provide them with a reference for catering. However, in the eyes of data workers, a lot of potentially valuable information can be mined from these restaurants' data. That means some valuable findings such as regional differences in a country or region can be found if using data mining techniques [5] on restaurants' websites. This dissertation will introduce a method to get, process very messy restaurants' websites data, mining regional features and classifying regional content from this data to reveal the regional differences.

1.2 Research aim and research focus

The project hopes to find unknown regional dishes, but the project does not wish to manually find menu items through using the dictionary. Thus, the aim of this dissertation is to mine menu data from 'Fish & Chips' shops and reveal regional differences in the UK based on the geographical distribution of the content of the menu data. For example, 'Haggis' is a traditional food in Scotland and widely distributed, while rarely seen in England. According to the methodologies applied in this dissertation, we could provide evidence that 'Haggis' is loved by the Scottish people and it is a regional dish in Scotland. 'Fish & Chips' is one of the most famous foods in the UK and there are more than 1,000 'Fish & Chips' shops in this country [6]. To achieve the project aim we will obtain the raw HTML data from the websites of some of these 'Fish & Chips' shops, and then employ data cleaning, mining, and visualisation techniques to find the content associated with particular regions.

1.3 Research methods

In terms of data crawling, the dissertation will illustrate the selection of data sources and methods for crawling data from 'Fish & Chips' shops' websites in the UK. The data cleaning procedure focuses on extracting and cleaning text content which is used for exploring regionality from the website HTML content, such as single independent words, noun phrases, and word pairs. The methods used for extracting and cleaning HTML content is the combination of Regular Expressions, HTMLParser and Natural Language Processing (NLP) (will be detailed in 2.2). Considering the data mining procedure of the research, data visualisation technologies will be applied to mine the regional features based on the geographical distribution of the extracted content. In terms of the classification (regional content and non-regional content) of the extracted datasets, the project employs machine learning methods, such as decision trees and regression classifiers (will be detailed in 2.4) to generate the regionality result. Specifically, this research is an iterative process and includes four rounds of evaluation and improvement (showed by Fig. 1) since the entire study is an exploratory process. Excepting for the first iteration, each of the remaining iterations depends on the result of the last iteration. This means the project knows what to proceed next only after getting and evaluating the results of each iteration. Besides, there are no existing criteria to verify the rationality of the method selection and the correctness of the results. Thus, only after each iteration has finished, the project can know whether the choice of method is reasonable and whether the result is correct. In addition, the features that could be used for reflecting regionality of the text is unknown and the evaluation of regionality content is based on the evaluator's experiences to some extent. For example, the evaluator knows the 'Irn Bru' is a Scottish drink, so when this phrase is judged to be regional, the evaluator can assume the decision is correct. Therefore, regional features and regional results are derived from each iteration, including exploration, evaluations, and improvements. In each iteration, the project may use or update the methods in the previous iteration. Besides, each iteration will also evaluate the results to identify problems and propose improvements for the next iteration.

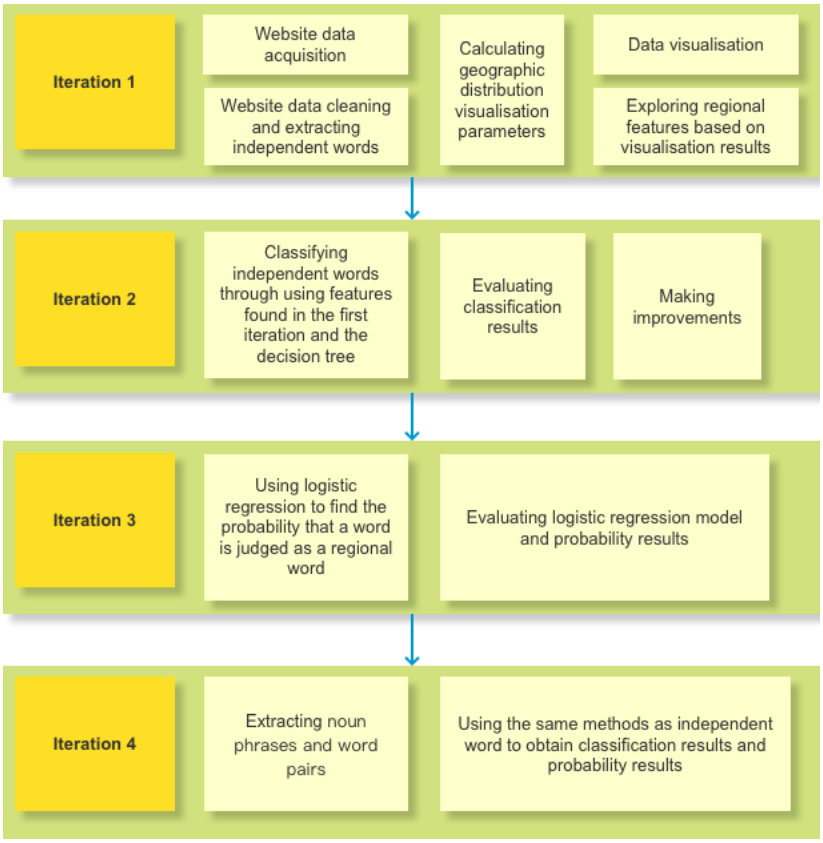


Figure 1: Workflow Diagram

1.4 Value of the research

The research is an innovative study that links seemingly unrelated menu information to regional differences of the UK through exploring regional content from the messy menu dataset. Thus, the biggest challenge for this project is that there are no existing ideas or methods to refer to and all ideas and methods require the project to explore and evaluate. Fortunately, the project succeeded in finding a solution to explore regional content and reflect regional differences, including ideas for finding regional features and methods for regional words classification. The methods and algorithms used in this project are universal, and they can also be used to find regional differences in other countries or used in similar studies. As a result, the project laid the foundation for similar subsequent researches. For example, if a study wants to know regional differences or consumption habits in other countries through restaurant data, it can refer to the methods and research processes used in this project.

1.5 Structure of the dissertation

The structure of this dissertation is organised as follows: Chapter 2 covers background knowledge, which mainly illustrates the main techniques and algorithms used in this research. Chapter 3, Chapter 4, Chapter 5 and Chapter 6 introduces the details of each iteration in the project, including methodologies, findings, evaluations, and improvements and summary and future work. Chapter 3 describes the first iteration,

presenting the procedures of exploring features of regional independent single words. Chapter 4 is related to the second iteration, which describes the application of a decision tree to get regional results of the single word. This chapter uses two types of decision tree algorithms and the makes a comparison between the two algorithms. Chapter 5 covers the third-round iteration, which introduces the logistic regression to obtain the probability that the single independent word is judged as a regional word. This chapter focuses on evaluating the importance of the selected features and the threshold regarding probability, in order to identify the number of probabilities exceeds the threshold, which would be judged as regional word. Chapter 6 is about the fourth iteration, which demonstrates the results of using the other two kinds of datasets (noun phrases and word pairs). Chapter 7 is a conclusion about this research. Chapter 8 provides a description of the improvement could be included in future work and also introduces the limitations, and recommendations.

Chapter 2

Background Theory

This chapter focuses on describing the techniques and methods used in the dissertation. Section 2.1 states the selection of data source and the definition of the web crawling. Section 2.2 presents an overview of data cleaning procedure and techniques used in the project. Section 2.3 provides the description of core algorithm and techniques used for geographic data visualisation. Section 2.4 aims to illustrate the data mining process in the project and some machine learning methods used for classifying regional content.

2.1 Web crawling

Before crawling data from websites of 'Fish & Chips' shops, the dissertation selected data sources that included food delivery websites such as Just-Eat [7] and independent 'Fish & Chips' shops' websites. The advantages of using food delivery websites as the data source are that it is convenient to search 'Fish & Chips' shops in each city in the UK by postcode. In addition, each shop which is searched out from food delivery websites is available to crawl data directly that the shop has a valid link and the page of that link has menu content. Whereas when searching on independent websites it may be the case that the desired content (e.g. menu) is only available in a PDF and hence cannot be crawled. Further on independent websites it may be the case that the URLs provide on the site are broken and hence also cannot be crawled. Thus, the dissertation originally planned to use the food delivery website as a data source. However, sites like Just-Eat clearly state that direct crawling of their website data is not allowed [8]. Therefore, using such a food delivery website to crawl data directly in this dissertation project may be illegal. This fatal flaw meant this method could not be used. As a consequence, Google has been used to find independent websites of 'Fish & Chips' shops and the data from these independent websites has then been used as the data source.

Web crawling is the mechanism by which information has been collected from target websites [9]. Specifically, the Python module urllib2 has been used to simulate browser behaviour to download web pages and handle request errors [10] to get the full website HTML data of 'Fish & Chips' shops.

2.2 HTML data cleaning techniques

Data cleaning is used for improving the quality of data which is used for subsequent processing through detecting inconsistencies and removing errors [11]. In this project, the dataset required to be cleaned is HTML data. The goal of data cleaning in this project is to obtain independent words (such as 'haggis'), noun phrases (such as 'monday supper deal haggis') and word pairs (such as 'salad with haggis' can be

divided into ‘salad with’ and ‘with haggis’ word pairs) with shop coordinates from HTML datasets and city dataset which contains coordinates.

In the web-based dataset, there is a lot of content that is not required by this project, such as name, attributes of HTML tags, script code, and special symbols. The project only focuses on information which the user can see on the page rather than the implementation details of the page. However, in terms of content which customers can see, there is a lot of redundancy, such as the singular and plural of the same noun all represent the same word. Therefore, the project should not only filter useless content in the HTML data but also classify words that represent the same meaning such as ‘chip’ and ‘chips’ into the same category (mainly focuses on the classification of singular and plural nouns with the same meaning). Fortunately, the regular expression, HTMLParser, and NLP can help the project to achieve the data cleaning goal.

2.2.1 HTMLParser

HTMLParser is an open source, fast and robust HTML parsing tool for extracting and cleaning the content of HTML [12]. It can customize HTML tag content extraction based on user requirements [13]. In this project, the HTMLParser mainly responsible for data extraction and filtering. The data source used in the project are independent websites that the HTML structure of most websites is different in Fig. 2 and Fig. 3 shows (small parts of the website structure are the same because they are developed by the same company). Fig. 2 shows part of the HTML source code of one of the websites that this website writes all the CSS styles on the page. Fig. 3 shows part of the HTML source code of another website that this page introduces some JavaScript code between the HTML element tags. In addition, these two websites are completely different in the HTML structure of the menu. Therefore, in order to extract content from HTML source code with different structures, the HTMLParser plays an important role. It mainly concerns about the name of the HTML tag such as ‘div’ and ‘script’ rather than the structure of the website design. As a consequence, the project can easily filter absolutely useless content based on the tag name, such as the content in the ‘script’ tag and extract potentially valuable content from the remaining tags. However, because the design styles of different web pages, the extracted data may also contain special symbols such as field trailing space symbol that will interfere with the cleaning result. Thus, the project also uses the regular expression which is a source language which can locate specific character strings in text [14] to filter the result of the HTMLParser.

Figure 2: HTML source code with HTML style of one of 'Fish & Chips' shops' websites

Figure 3: HTML source code of one of 'Fish & Chips' shops' websites

2.2.2 Natural Language Processing (NLP)

In order to solve the problem of data redundancy in the extracted content, the project uses method of semantic recognition in NLP. Natural Language Processing (NLP) is using computer to understand and manipulate natural text or speech to process tasks [15]. This project mainly wants to change the singular and plural nouns of the same root into singular nouns and the Natural Language Toolkit (nltk) can provide the solution. nltk is an open source tool written by Python with collection of modules and corpora [16]. nltk determines the part of speech of a word based on its corpus and the identification method has been encapsulated which the project can use directly to identify plural nouns and convert them into singular forms. However, in English some words can be both plural nouns and verbs and the nltk will treat all words as nouns and converts them into singular. Fortunately, this project does not care whether the word being converted is a verb or a noun. It only cares whether the word is distributed regionally.

2.3 Geographic data visualisation

One of the aims of the project is to explore geographically distributed features to represent regionality, so the project uses Cartesian coordinate systems for geolocation calculations and Matplotlib for data visualisation.

2.3.1 Central point calculation algorithm

The core calculation in this project is the set of coordinates' (the method of obtaining it will be described in detail in 3.1.2) central point which is the centre of all shops which contain a specific content. The importance of the central point is that most regional features found in this project are derived from it. The project uses a set of geographic coordinates containing this content to calculate the centroid. The algorithm used in this project for calculating the central point through combining geographical coordinate system with Cartesian coordinate systems which regard the Earth as a sphere (Fig. 4). This combination is also known as ECEF ("earth-centered, earth-fixed") [17]. In Cartesian coordinates, the earth is a sphere centered at the origin [18]. The z-axis points to the north pole. The x, y-axes are on the equatorial plane that the x-axis passes through the equator and the prime meridian and the y-axis points to the equator at 90 degrees east [19]. However, the coordinates of the central point obtained by using this algorithm in the project are not accurate since the algorithm regards the earth as a sphere rather than ellipse which is the shape of the earth itself. Fortunately, the requirement of the accuracy of the coordinates of the central point in this project is not high, because this project is concerned with the distribution of content.

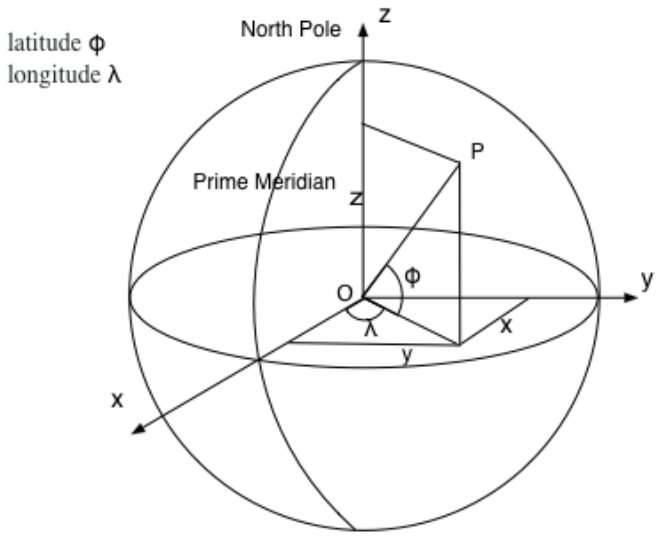


Figure 4: geographical coordinate system with a cartesian coordinate systems

As Fig. 4 shows, point P represents a geographical coordinate with latitude ϕ and longitude λ. A series of coordinates can be represented as latitude ϕ_i, longitude λ_i (i = 1...,n). Thus, in Cartesian coordinate systems, the coordinates of the three directions can be expressed as:

(1) $$x_i = r \cos \phi_i \cos \lambda_i$$

(2) $$y_i = r \cos \phi_i \sin \lambda_i$$

(3) $$z_i = r \sin \phi_i$$

The centroid of these points is the average of the sum of x_i, y_i, z_i:

(4) $$(\bar{x}, \bar{y}, \bar{z}) = \frac{1}{n} \sum (x_i, y_i, z_i)$$

The coordinate of the centroid can be expressed as:

(5) $$\bar{\phi} = atan2(\bar{z}, \sqrt{\bar{x}^2 + \bar{y}^2})$$

(6) $$\bar{\lambda} = atan2(\bar{x}, \bar{y})$$

2.3.2 Visualisation with Matplotlib

After the project gets the coordinates of central point and other features such as radius (details will be described in 3.1.3), the project hopes to display the distribution of a specific content on the UK map to observe the distribution of the content. The visualisation tool selected by the project is Matplotlib package of Python which is an open source portable Python plotting package used in scientific, engineering and financial fields [20]. It can implement complex data visualisation processes with simple encapsulated methods. Because of the convenience of this tool, most data visualisation processes in the project are achieved by it. Specifically, in this project, Matplotlib package mainly completes the visualising of geographic information distribution and the line graphs. In terms of the visualisation of geographic information distribution, the project uses one of the Matplotlib toolkit named Basemap [21]. Basemap provides a possibility that the project can draw Matplotlib plot over the real-world map [22]. This indicates the Basemap replaces the bottom canvas of the Matplotlib, so it can implement the goal of plotting other graphics such as radius and circumference curve on the map. The following figure (Fig. 5) is an example of using Basemap to visualise the distribution of 'haggis' (details will be stated in 3.1.3).

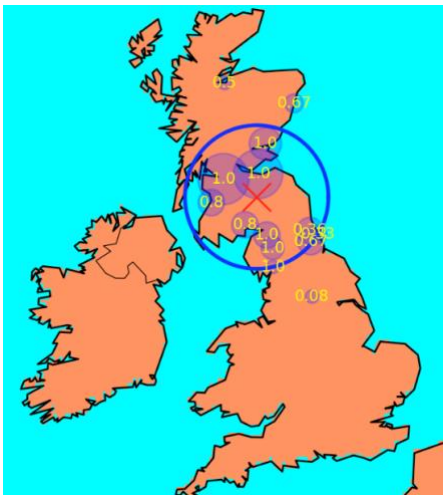


Figure 5: Haggis distribution range

2.4 Data mining with machine learning methods

This project is an exploratory project, therefore, at the beginning of the project, there is no clear definition of regional features. As a result, the discovery of features and obtaining regional results are the process of data mining. Data mining is a process to extract patterns which represent useful information from massive dataset [23]. In the early stage of the project, the discovery features went through two phases. The first is to find some content widely distributed on the map and some regionally distributed content based on the results of the geographical distribution of content such as Fig. 5. The second phase is to find the commonalities of these widely distributed content and regional content. The commonalities can be regarded as the features (regional or non-regional) of the content (will be described in 3.2 and 3.3). After the project used the above method to get more features, the project tried two machine learning methods to classify the regional content. One of them is the decision tree and the other is logistic regression and both of them belong to the method of supervised learning. Supervised learning means that the training data has both features and labels. Through training, the machine can find the connection between the features and the labels by itself and can judge the labels when facing data with only features without labels [24]. In this project, the training data set includes widely distributed content and regional content. They were judged and added based on developer's experiences and the results of the project.

2.4.1 Decision tree

A Decision tree is mainly used for classification and prediction of models [25] and the project used a decision tree to classify regional content and widely distributed content. There are two algorithms used in this project. One is the ID3 algorithm and the other is the Cart algorithm. Both algorithm uses training dataset to create the tree and then use the tree to classify the test dataset [26]. The reason why the ID3 algorithm was chosen by the project is that the features obtained by the project for the first time were categorical rather than continuous. Thus, the project uses the ID3 algorithm which uses categorical data to generate the decision tree [27] to classify the content to regional or not. However, there were some disadvantages to use the ID3 algorithm (details in 4.3) and the project tried another decision tree algorithm – Cart algorithm.

2.4.1.1 ID3 algorithm

The ID3 algorithm constructs a decision tree by selecting most useful features [28]. These features can make the classification of data set more effective. Thus, the project requires an algorithm to measure the suitability of features and select features. The Entropy can measure the impurity of training dataset [29] that the greater the entropy, the more complex the information. As a consequence, the project can use the information gain which is the amount of entropy lost by adding a feature to select representative features.

Entropy: X represents the collection of features, $X = \{x_1,\ x_2,\ x_3, \dots,\ x_n\}$, $p(x_i)$ is the probability of occurrence of x_i.

(7) $$H(X) = -\sum_{i=1}^{n} p(x_i)\, log_2\, p(x_i)$$

Information Gain: a represents a feature.

(8) $$IG(X|a) = H(X) - H(X|a)$$

The decision tree construction process of ID3 algorithm is divided into the following steps:

- Loading training dataset and treating the dataset as the first node.
- Splitting the dataset by each feature and calculating the Entropy and Information Gain based on the splitting result.
- Selecting feature which has maximum Information Gain as optimal segmentation feature.
- According to the optimal segmentation feature split dataset into two nodes.
- Repeat 2-4 steps for each newly acquired node to recursively build the tree
- Sample classification.

2.4.1.2 Cart algorithm

Cart algorithm uses a binary recursive partitioning procedure to split datasets [30]. In classification tree, Cart algorithm uses the Gini index as a property to determine to partition [31]. The Gini index indicates the uncertainty of the sample. The larger the Gini index, the greater the uncertainty of the sample set which means the probability of the sample belongs to a class is low. In terms of each feature, the Cart algorithm will traverse all possible splitting methods and select the feature which has minimum Gini index as the division criteria [32]. The following formulas show the calculating of the Gini index.

Assuming that there is a K class, the probability that the sample point belongs to the K class is p_k, then the Gini index is defined as:

(9) $$Gini(p) = 1 - \sum_{k=1}^{K} {p_k}^2$$

Assuming that C_k be the subset of samples belonging to the k class in D, then the Gini index is:

(10) $$Gini(D) = 1 - \sum_{k=1}^{K} (\frac{|C_k|}{|D|})^2$$

Assuming that feature A, divide the sample D, into two data subsets $D1$, and $D2$, then the Gini index of the sample D, under the feature A is:

(11) $$Gini(D, A) = \frac{|D_1|}{D} Gini(D_1) + \frac{|D_2|}{D} Gini(D_2)$$

The steps to generate a decision tree using the Cart algorithm are as follows:

- Using each feature A in the sample D, and each possible value of A ($A \geq a$ and $A < a$) to divide the sample into two parts and calculate the Gini (D, A).

- Find the optimal segmentation feature which has the minimum Gini (D, A). Next, determining whether the splitting stop condition is satisfied. If not, output the optimal segmentation point.
- Recursive call 1 and 2.

In this project, the result of the decision tree was binary (details in 4.2 and 4.3), which means content is judged to be regional or non-regional. However, the project also wants to know how much probability content is judged as regional because the project wants to find more evidence to verify the classification results. Thus, the project tried another method - logistic regression.

2.4.2 Logistic regression

The project studies the classification problem, so the dependent variable of the model is the classification variable (0 or 1) and the independent and dependent variables of the model are nonlinear. As a result, the logistic regression model is suitable for this project and the project selected the logistic regression model of the Sklearn-learn package as the classifier to find the possibility that each piece of content is judged to be regional.

Logistic regression is well suited to describe the relationship which is expressed as probability between classification results and one or more classifications [33]. It can adapt to multiple classification results. In this project, logistic regression is used to calculate the probability of a binary event occurring under multiple independent features [34]. The following model is the model of logistics regression:

x denotes the vector of feature variables, and $b \in \{0,1\}$ denotes the associated binary output. w represents the weight vector and the w^T is the transposed matrix of w. $\sigma(.)$ is the sigmoid function. v is the intercept. The logistic regression has model:

(12) $$Prob(b|x;w) = \sigma(w^T x) = \frac{1}{1+exp(-b(w^T x+v))}$$

Logistic loss function (z is $b(w^T x + v)$:

(13) $$f(z) = log(1 + exp(-z))$$

Supposing the training dataset is $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})$ and the average logistic loss:

(14) $$l_{avg}(v,w) = (1/m)\sum_{i=1}^{m} f(b(w^T x))$$

Logistic regression problem:

Overfitting problem: in supervised learning when there are many input features, but only a small number of key features determine the classification target. That is, when the number of training set data is insufficient, the classification model may perform well on the training dataset but not well on the test dataset [35]. Thus, when there are many features, overfitting will become a problem of the model unless the training set is ample [36]. In order to solve this problem, L1 and L2 regularizations were used.

L1 regularization [35]:

$$l_{avg}(v, w) + \lambda \|w\|_1 = (1/m) \sum_{i=1}^{m} f(b(w^T x)) + \lambda \sum_{i=1}^{n} |w_i| \quad (15)$$

Lasso (L1) penalty encourages the sum of the absolute values of the w_i to be small [36]. It uses sparsity to fit model with many features [37]. The sparsity means that L1 penalty will automatically filter some features that have less impact on classification. L1 penalty achieves the filtering by reducing the regression coefficient to 0 and slightly reducing other regression coefficients [38].

L2 regularization [35]:

$$l_{avg}(v, w) + \lambda \|w\|_2^2 = (1/m) \sum_{i=1}^{m} f(b(w^T x)) + \lambda \sum_{i=1}^{n} {w_i}^2 \quad (16)$$

L2 penalty encourages the sum of the squares of the w_i to be small [36]. It will reduce the regression coefficient but will not be zero [38]. Thus, L2 penalty will weaken the dominant classification feature and enhance the influence of other features. If each feature has an effect on the classification, L2 penalty is more suitable.

In this project, using logistic regression and its L1 and L2 regularizations can not only help the project to obtain probabilities but also help the project to see the impact of each feature on the classification results.

Chapter 3

Iteration 1

In iteration one, the following tasks: data acquisition, data cleaning and data visualisation were done by the project. The data acquisition procedure is mainly responsible for retrieving and storing menu websites. The data cleaning process is primarily responsible for obtaining independent single words from retrieved HTML files. The data visualisation phrase is divided into two parts. One of them is geographical map visualisation and another is trend visualisation (details in 3.1.3). The aim of the iteration one is to find regional features of regional words through observing the result of some known regional words' geographical distribution map and trend graph. In addition, iteration one details the process of discovering these features and makes plans for the next iteration.

3.1 Methodology

Iteration one mainly focuses on four aspects, which are web data acquisition, HTML data cleaning, geographic data and ratio trend visualisation and exploration of features.

3.1.1 Data acquisition

After the project identified the data source is independent 'Fish & Chips' shops' websites, the project started to find URLs of these websites through searching on the Google and other food recommendation websites. The method for searching websites is first finding the city, then searching for 'Fish & Chips' and get websites URLs from the searching result. At the beginning of the project, the project collected websites of 'Fish & Chips' shops in twenty-one cities which have a large population size in the UK, such as London, Manchester, and Glasgow. However, the project found that these collected shops are concentrated in the north-central (e.g. Edinburgh, Glasgow) and south-central (e.g. Manchester, Sheffield), with few shops in the north (e.g. Dundee, Inverness), southwest (e.g. Plymouth) and central regions (e.g. Newcastle). In order to solve the problem of uneven distribution of shops, the project added seventeen shops which distributed in northern, central, and southwestern cities.

The initial goal of the project is to obtain a collection of shops that their distribution can cover all parts of the UK. However, in the UK, although there are a lot of 'Fish & Chips' shops, not every 'Fish & Chips' shop offers a menu on its website. Besides, as 2.1 described, some of these 'Fish & Chips' websites cannot be retrieved. As a consequence, the project finally collected two hundred and forty available websites of the 'Fish & Chips' shop. The distribution of the shops which contain these websites basically covers most parts of the UK. However, although the distribution of shops covers most cities, the number of shops in each city still shows bias. That means most of the shops are concentrated in large population cities, and other cities with sparse populations such as

Inverness and Carlisle have fewer shops. Thus, this imbalance will be reflected when the project visualises the shops' geographic data in 3.1.3. Each ‘Fish & Chips’ shop collected by the project will be assigned Id, city name, and the URL and this information are stored in a CSV format file. Id is used to uniquely identify the shop and the city name is used to find the coordinates of the city where the shop is located (cities with their coordinates are stored in another file created by the project).

After the project completes the shop collection, the project wrote a Python script that uses urllib2 module to retrieve HTML data from the collected websites. The urllib2 module provides a way to simulate a browser to send HTTP requests to a website. This method avoids the problem of some websites’ denying access due to the detection of abnormal access. Besides, the script uses a file which has the coordinates of different cities and finds the geographic coordinates of each shop. (In this project, the coordinates of the shops in the same city are the coordinates of the city).

The script generates a file for every shop, and each file stores the HTML source code of a shop, and the file name is the Id of the shop (e.g. 1.csv). In addition, the script also generates a mapping file which contains shop Id, name of HTML source code file and coordinates of that shop. The main role of this mapping file is to associate the shop with its website HTML file and coordinates. As a result, after the data acquisition procedure, the project connected each shops’ HTML content with its coordinates.

3.1.2 Data cleaning

The entire process of data cleaning in this project was a process similar to Map-Reduce that the Map method generates a series of intermediate key and value pairs and the Reduce method merges the results of the Map method according to the same key value [39]. The reason why data cleaning procedure is similar to the Map-Reduce is that this project extracted the content from the HTML source code in the form of key (menu word)-value (shop id and a set of coordinates) and merge the content with the same key. Two Python scripts were used in the project to achieve the Map and Reduce processes to complete the data cleaning procedure. The Map script was responsible for extracting independent single words with their shop id (its role will be described in the Reduce script) and shop coordinates. The Reduce script was mainly responsible for merging the words output by the Map script. It added the coordinates of the same word but from different shops to the coordinate set of that word. The Fig. 6 is an example of the extraction operation of the Map script and the merging operation of the Reduce script.

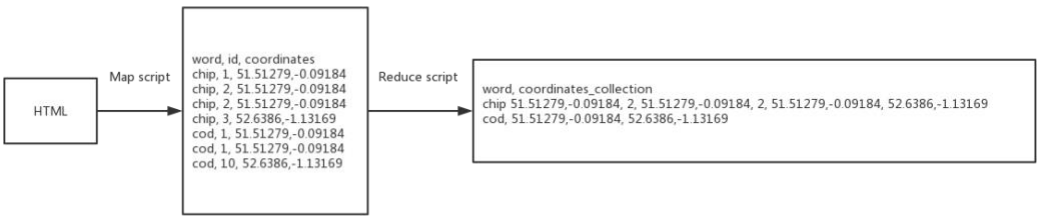


Figure 6: Data operation flow chart of Map script and Reduce script

The Map script mainly focused on filtering useless HTML content, extracting the content in the HTML tag, splitting the content into separate independent words and converting all plural nouns to singular. Firstly, the script read the mapping file generated in 3.1.1 and sequentially read the HTML files retrieved by the project and hand HTML content to the HTMLParser for processing. The HTMLParser firstly filtered the tags with the content in them that project did not need. The useless tags names were defined by the project and in this project, useless tags were 'script', 'style', 'link', 'head', 'a' and 'title', since most of the content in these tags was HTML code. Next, the HTMLParser sequentially recognized the names of other tags with their content, but this project only focused on the content. In order to extract independent single words, the project used regular expressions and *split()* function in the function which HTMLParser handles the contents of the HTML tag. The project firstly used regular expressions to filter numbers and process special symbols such as '.', '+' and '-', and then used the segmentation method to split the content into separate independent words and converted them to lowercase.

The project originally wanted to convert plural nouns into singular nouns after getting the independent single words. However, the ASCII encoded special spaces such as '\xc2\xa0' appeared in some of independent single words and the regular expressions could not recognize them. The reason for appearing these special spaces is the inconsistency between the character set encoding of some websites and the compiler character set encoding. As a consequence, if the project directly converts the part of speech, plural nouns with special spaces will be treated as proprietary singular nouns. Further, the project found that these words with special spaces only displayed the ASCII code of the special space after added to the list. Thus, the project first added all the results of HTMLParser to the output list. Each row of the output list included shop Id, independent single word (may contain special spaces) and the shop coordinates. Next, the project converted output array to string, and the ASCII code of special spaces existed as ASCII characters, and then the project used the regular expression to filter these characters. After filtering the special spaces, the project reconverted the string of the output list to a list and used as the output of the Map script.

Before the Reduce script processing the output data of the Map script, the Unix *sort()* method was used to sort the independent single word column of the output list of the Map script. This reduced the workload of the Reduce script that when processing each row, the Reduce script does not have to determine whether the word in current row appear in the words that have appeared before. The script only needs to determine whether the word in the current row is the same as the word in the previous row. If the same, add the coordinates of the current word to the previous coordinate set. If they are different, a new coordinate set is created for the current word, and the current coordinate is added to the new set. Besides, in this project, each word was only allowed to appear once in one shop. That means in terms of same words with the same shop id which is one of the output value of Map script, only one set of coordinates could be added to the word coordinates collection. The reason for this is because the coordinates of the word added to the collection can represent the coordinates of all the same words in a shop.

The output of the Reduce script is a CSV format file which each row is an independent single word with its coordinates collection as Fig. 7 shows. As a result, this file can be

used for calculating the central point of each word's distribution and the data cleaning procedure has finished.

swansea	51.62125,-3.94490
courgette	52.9536,-1.15047
greenish	51.51279,-0.09184
delightfully	53.79648,-1.54785
harriet	53.48095,-2.23743
stating	50.15201,-5.06654
vincenzo	55.95206,-3.19648
fritter	51.51279,-0.09184 51.51279,-0.09184 51.51279,-0.09184 53.7446,-0.33525 53.7446,-0.33525 53.7446,-0.33525 53.7446,-0.33525 52.6386,-1.13169 52
enjoy	51.51279,-0.09184 51.51279,-0.09184 53.7446,-0.33525 53.7446,-0.33525 52.6386,-1.13169 52.9536,-1.15047 52.9536,-1.15047 53.79648,-1.54785 53
farmhouse	52.6386,-1.13169 52.9536,-1.15047 52.9536,-1.15047 53.38297,-1.4659 52.40656,-1.51217 52.40656,-1.51217 54.77560,-1.58374 52.48142,-1.89983 5
bradfordwho	53.79391,-1.75206
specially	52.6386,-1.13169 52.9536,-1.15047 53.41058,-2.97794 57.47908,-4.22398 52.20765,0.13200 52.63220,1.28925
fudgey	55.95206,-3.19648
speci	52.9536,-1.15047 52.9536,-1.15047
pollock	53.48095,-2.23743
scorcher	52.48142,-1.89983
almond	51.51279,-0.09184 52.6386,-1.13169 53.79648,-1.54785 54.89228,-2.93206 56.46913,-2.97489 51.48,-3.18 55.45862,-4.62849
bacon	52.6386,-1.13169 52.6386,-1.13169 52.9536,-1.15047 52.9536,-1.15047 52.9536,-1.15047 52.40656,-1.51217 52.40656,-1.51217 53.79648,-1.54785 53
direct	54.97328,-1.61396
chef	52.6386,-1.13169 53.79648,-1.54785 53.79391,-1.75206 53.41058,-2.97794 53.41058,-2.97794 53.41058,-2.97794 51.48,-3.18 51.48,-3.18 55.95206,-3
elegant	51.51279,-0.09184
tortillla	53.79648,-1.54785
street	52.6386,-1.13169 54.90465,-1.38222 53.38297,-1.4659 53.79648,-1.54785 54.77560,-1.58374 53.79391,-1.75206 52.48142,-1.89983 53.00415,-2.1853
safe	53.79391,-1.75206 54.58333,-5.93333

Figure 7: The result of Reduce script

3.1.3 Data visualisation

In iteration one, the data visualisation of this project was divided into two parts. One of them was geographic maps data visualisation, and another was trend of the number of shops which contain a specific word increasing with distance (meter) from the central point. The project wanted to find the features of regional words through the results of these two kinds of data visualisation results. In terms of the geographic maps, the project considered this to be the most intuitive way to see if a word is a regional word. The reason why the project wanted to visualise the trend was that trend graph was a conjecture of the project for the distribution trend of words with regional features. The Fig. 8 and Fig. 9 are the project's conjecture maps for regional words and non-regional words. The black arrow indicates the distance from the central point. The circles indicate the distribution ranges of the word as the distance from the central point increases. The red circles in these two graphs have same size. The blue pointer represents the cities of words distribution and in each city, words may be distributed in many shops. The project thought that the regional words might be clustered together close to their centroid. Thus, if a word is a regional word, as the distance from the central point becomes larger, the number of shops will increase to a certain value and then no longer grow. Besides, when the distance from the central point begins to increase, the number of shops of regional words may increase rapidly, and as the distance increases to a certain distance, the growth of the number of shops will slow down. However, if a word is a non-regional word, as the distance increases, the number of shop will continue to increase until the distance increases to a great distance. Further, when the distance has not increased much (e.g. one-third of the distance of the arrow in Fig. 9), the number of shops of non-regional words

would not increase to a lot relative to the total shop number. As a consequence, the project believed that 'ratio' which means the number of shops whose distance is less than a specific distance (radius of the red circle in Fig. 8 and Fig. 9) from the central point divided by total shop number could be considered as a feature of regional word.

Figure 8: Conjecture map of regional distribution words

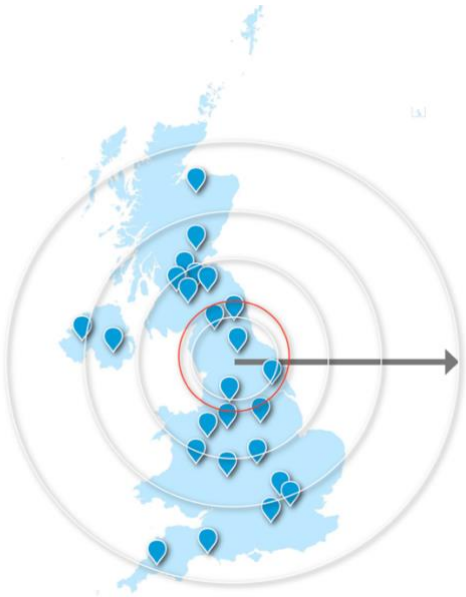


Figure 9: Conjecture map of non-regional distribution words

Before visualising the geographic maps, firstly, the central point of a word distribution was calculated by the project based on the coordinates set results of the word. Secondly, the radius of the word distribution was calculated. The method of calculating the radius is calculating the distance between all word's coordinates from the central point and find the largest distance as the radius. The distance between the central point and each coordinate point was derived from the Euclidean distance because the Basemap has converted the Earth's sphere into a plane. Next, the number of shops in the city which contain the word was calculated by the project and then the proportion of those shops in the total number of shops in the city was calculated. By this way, the geographic map could not only show the distribution of the word but also show the uneven distribution of the number of shops in each city. This can help users better understand the details of word distribution.

However, there is a problem that some outliers which means few shops are far from the central point will have a huge impact on the above parameters, especially on the radius that the radius will become very large because it covers all the shops. For example, if a word is distributed in the north of the UK, but there are one or two shops located in the southernmost part of the UK, the central point will be shifted south. Besides, the geographical map visualisation results of that word will have a large distribution range, because it contains the southernmost shops of that word. This will cause inconvenience to the developer to observe the word distribution range and affect the exploration for features. In addition, the large deviations of the central point will lead to deviations of

important features of the word, because most of regional features derived from the central point (details in 3.2 and 3.3).

In order to filter the outlier shops, the project firstly calculated the central point of shops which contain the word and then sorted the distances of all shops. Next, the project set a percentage that the project only took a percentage of the shops close to the central point and then recalculated the central points and other parameters. The project used 'haggis' and 'bru' as examples (the project knew these two words are regional word in advance) to adjust the filtering percentage to observe changes in the distribution range and finally decides to keep 95% of the shops. The project found that selecting 95% shops could filter out almost all outliers that have a huge impact on the results. In addition, this percentage can retain all normal distribution shops whose distance from the central point is within a reasonable range. The Fig. 5 is an example of the distribution of 'haggis' which contain 95% shops. In Fig. 5, 'X' represents the central point of the distribution of 'haggis'. The bold blue line represents the circumference of the distribution. The size of many blue small circles in the figure represents the number of shops which contain 'haggis' in each city. The decimal in each blue circle means the number of shops in the city that contain the word as a percentage of the total number of shops in the city.

3.2 Findings

In iteration one, the project obtained the features of some regional word by comparing the results of pre-known regional words with the results of pre-known non-regional words. The following diagrams show geographic map visualisation and the trend visualisation results of regional words and non-regional words.

3.2.1 Regional words findings

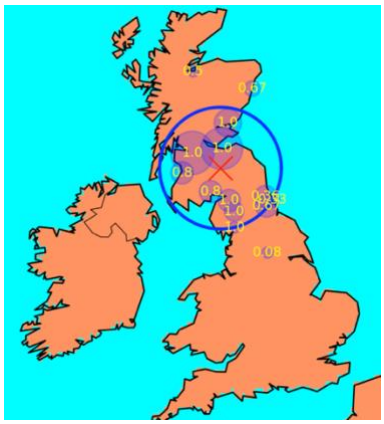


Figure 10: 'haggis' distribution (95%)

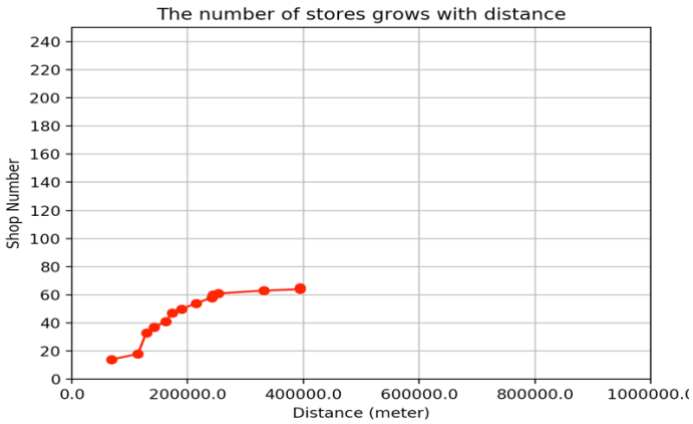


Figure 11: The number of 'haggis' shops varies with distance

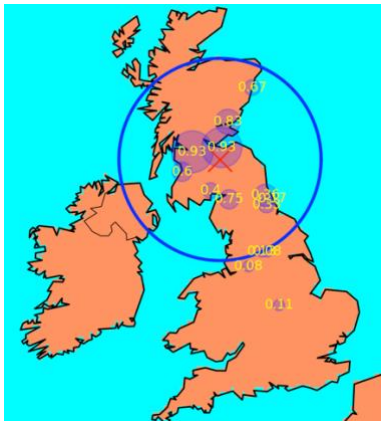


Figure 12: 'bru' distribution (95%)

Figure 13: The number of 'bru' shops varies with distance

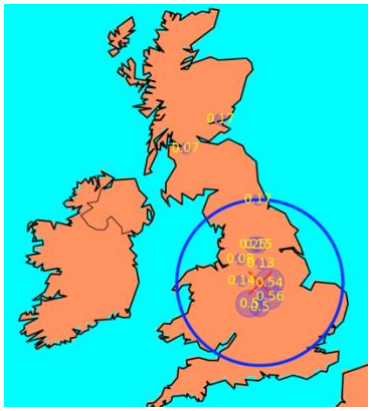


Figure 14: 'naan' distribution (95%)

Figure 15: The number of 'naan' shops varies with distance

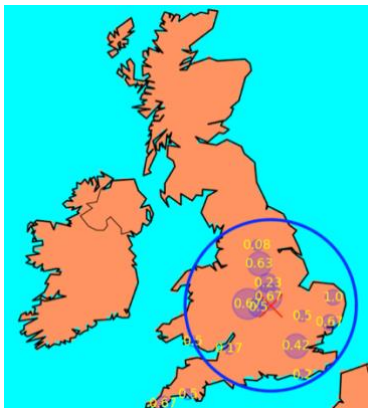


Figure 16: 'roe' distribution (95%)

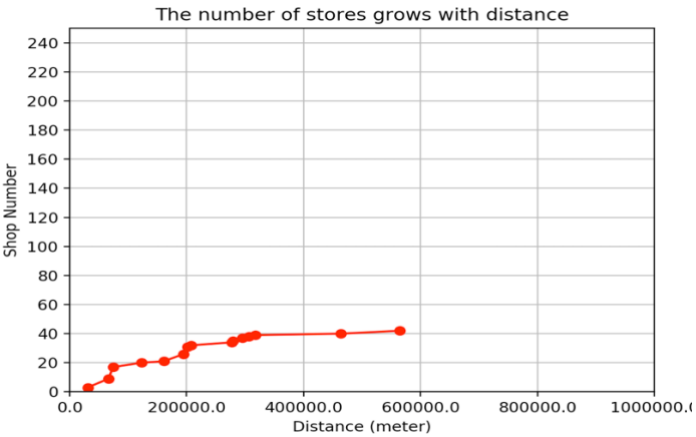


Figure 17: The number of 'roe' shops varies with distance

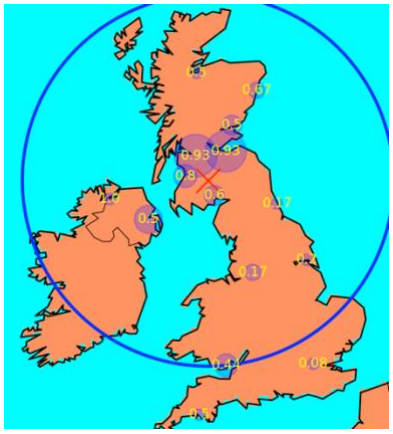


Figure 18: 'supper' distribution (95%)

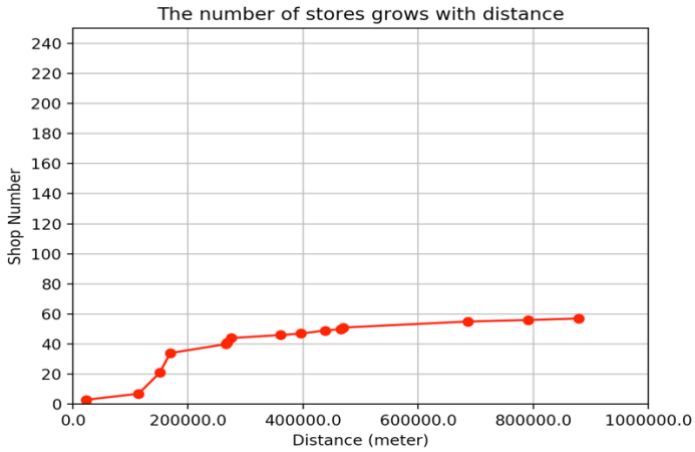


Figure 19: The number of 'supper' shops varies with distance

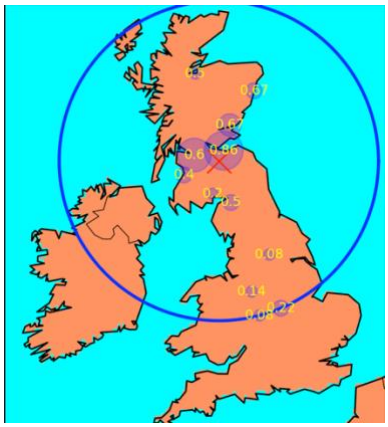
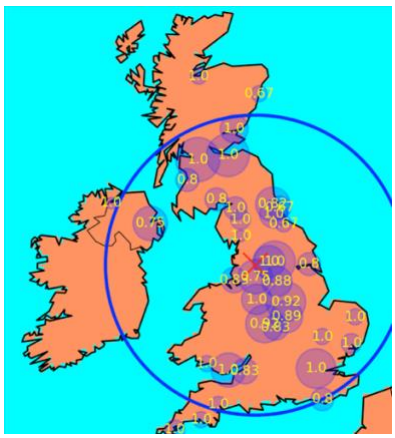


Figure 20: 'pakora' distribution (95%)

Figure 21: The number of 'pakora' shops varies with distance

3.2.2 Non-regional words findings



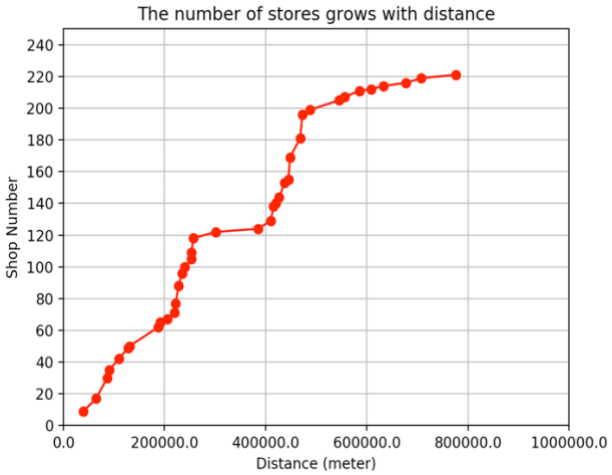


Figure 22: ‘chip’ distribution (95%)

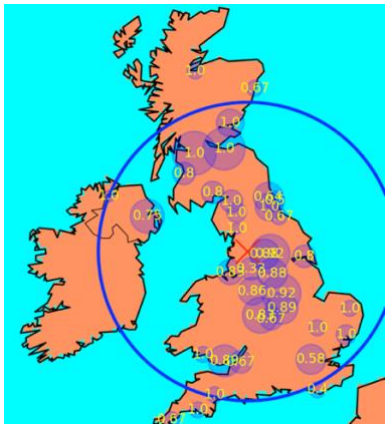


Figure 23: The number of ‘chip’ shops varies with distance

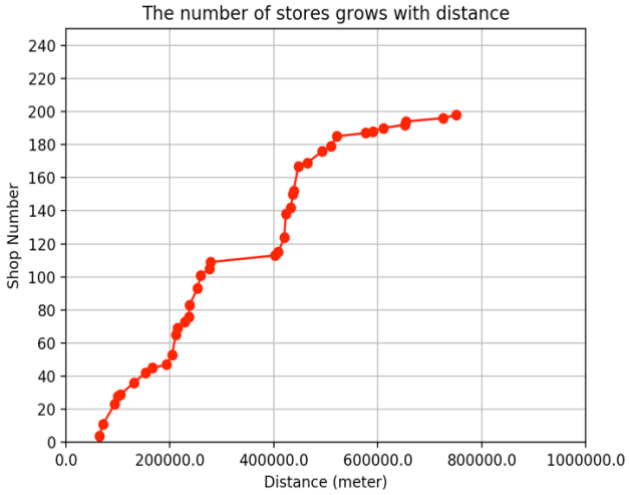


Figure 24: ‘sausage’ distribution (95%)

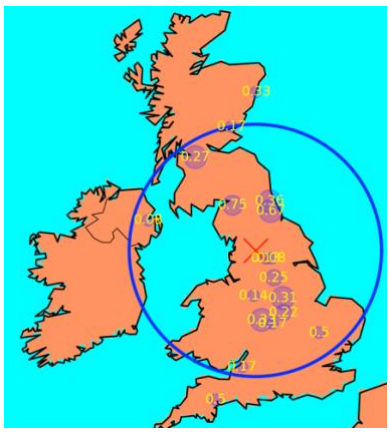


Figure 25: The number of ‘sausage’ shops varies with distance

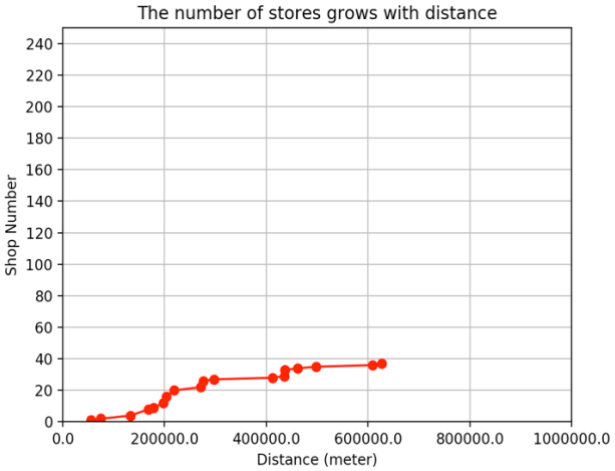


Figure 26: ‘supreme’ distribution (95%)

Figure 27: The number of ‘supreme’ shops varies with distance

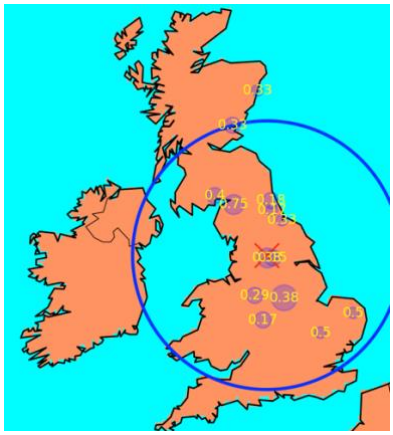


Figure 28: 'gift' distribution (95%)

Figure 29: The number of 'gift' shops varies with distance

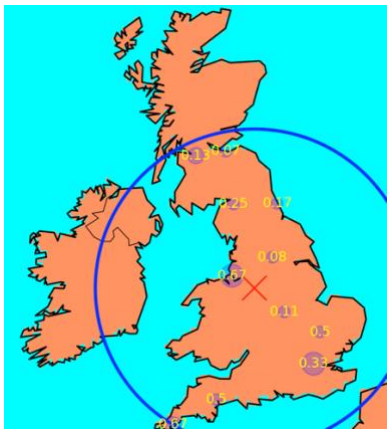


Figure 30: 'soup' distribution (95%)

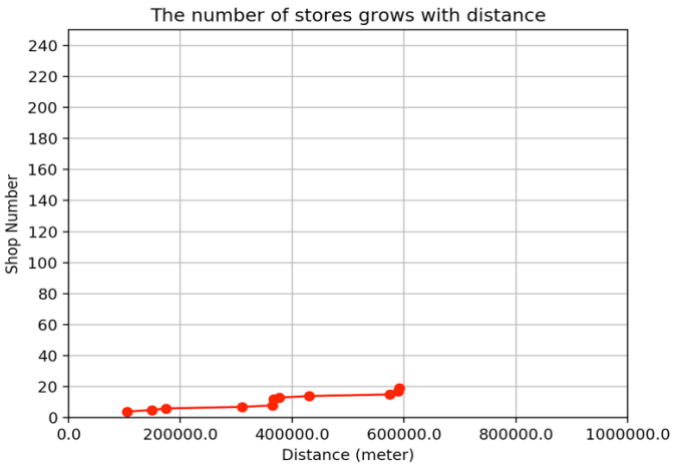


Figure 31: The number of 'soup' shops varies with distance

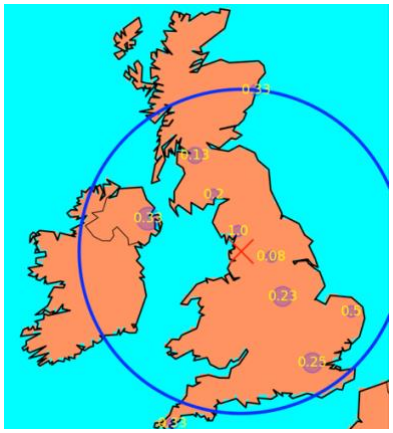


Figure 32: 'daily' distribution (95%)

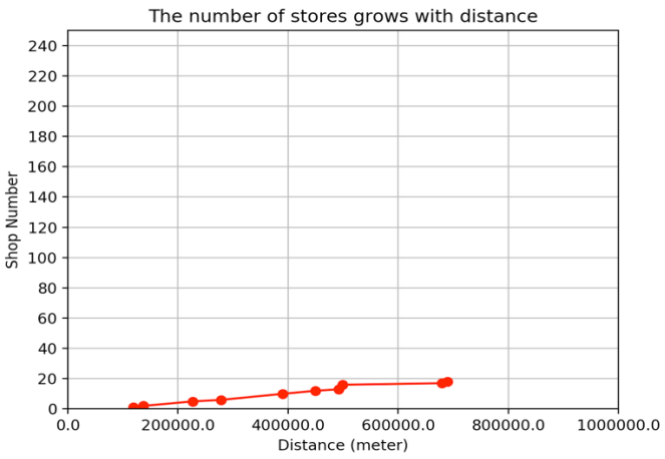


Figure 33: The number of 'daily' shops varies with distance

The above figure shows three types of non-regional words. Fig. 22 – Fig. 25 show the distribution and trend of non-regional words which distributed in almost every shop in every city.

Fig. 26 – Fig. 29 show words which distributed in almost every city, but the number of shops contained that word in each city is not a lot.

Fig. 30 - Fig. 33 show words which not distributes in many cities, but widely distributed throughout the UK.

According to these non-regional words findings, the project found that the growth of the number of shops will not be too fast when the distance start to grow. This is more certain that 'ratio' is a feature of regional words rather than non-regional words.

3.3 Evaluation and improvement

Through comparing the findings of regional words and non-regional words, the project found that the number of regional words' shops increases greatly within 200,000 meters from the central point. After 200,000 meters, the growth trend shows a slowdown. However, there is no such rule for the trend of the number of non-regional words' shops. Thus, this confirms the hypothesis of the project that regional words show a rapid growth trend within a certain distance and in this project, the certain distance is 200,000 meters. Next, the project calculated the 'ratio' of above regional words and non-regional words based on the distance threshold of 200,000 meters and the following two tables are the 'ratio' result.

Regional words	Ratio
haggis	77%
bru	75%
naan	90%
roe	65%
supper	60%
pakora	79%

Table 1: Regional words with 'ratio'

Non-regional words	Ratio
chip	29%
sausage	24%
supreme	32%
gift	32%

soup	32%
daily	11%

Table 2: Non-regional words with 'ratio'

According to Table 1 and Table 2, the 'ratio' of all non-regional words is very low. However, in the result of regional words, the 'ratio' of 'supper' is only '60%' and although 'supper' is densely distributed near the Edinburgh, it is widely distributed in many other cities. Thus, the project raised doubts about the regionality of 'supper' and the project did not want to use 60% as the threshold of 'ratio'. From the distribution of 'roe', 'roe' shows a strong regionality relative to 'supper', so, the project set the threshold of 'roe' to 65%.

However, there is a problem that most of words with few shops (less than ten shops) are distributed within 200,000 meters. Besides, the 'ratio' of many of these words is 1. For example, Fig. 34 and Fig. 35 show the distribution and trend of 'massala' which just has four shops and the 'ratio' is 1. In Fig. 34, 'massala' looks like a regional word and according to Fig. 35, 'massala' has the feature of regional word. However, the distribution sample of 'massala' is really too small that the project cannot directly determine that 'massala' is a regional word. As a consequence, the project decided that for words containing only ten or less stores, the project treated them as non-regional words.

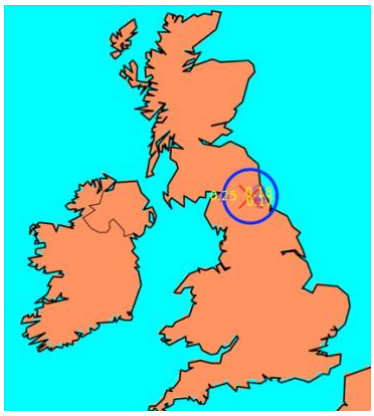


Figure 34: 'massala' distribution (95%)

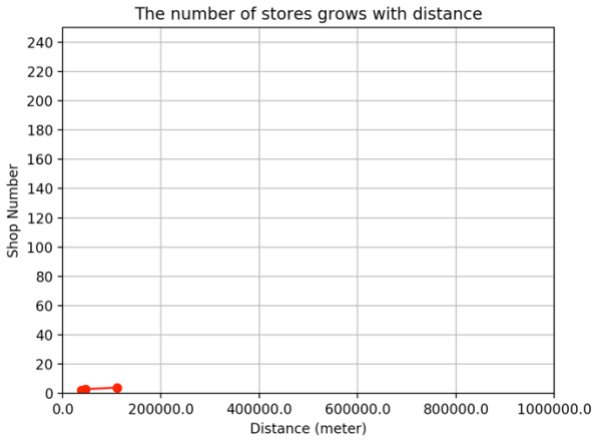


Figure 35: The number of 'massala' shops varies with distance

Through the above findings, it can be explained that 'ratio' > 65% can be regarded as one of features of regional word, but it is not enough to only rely on 'ratio' to judge all regional words such as 'funghi'. According to Fig. 36 and Fig. 37 which are the distribution and trend of 'funghi', the project found that the number of shops which contain 'funghi' rises smoothly and the 'ratio' of 'funghi' is just 50%. However, as Fig. 36 shows, there are many shops that contain this word in a small area and it looks like a regional word. Thus, if 'funghi' is only judged by 'ratio', it must be treated as a non-regional word. As a consequence, the project requires to discover more features to make a more accurate decision.

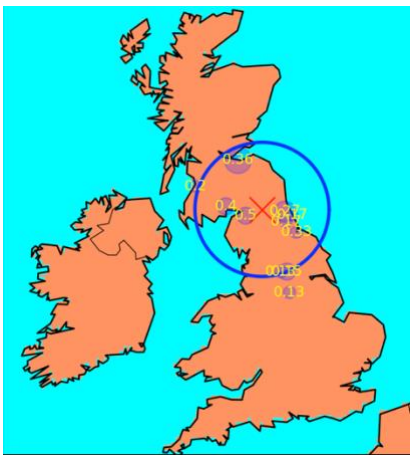


Figure 36: 'funghi' distribution (95%)

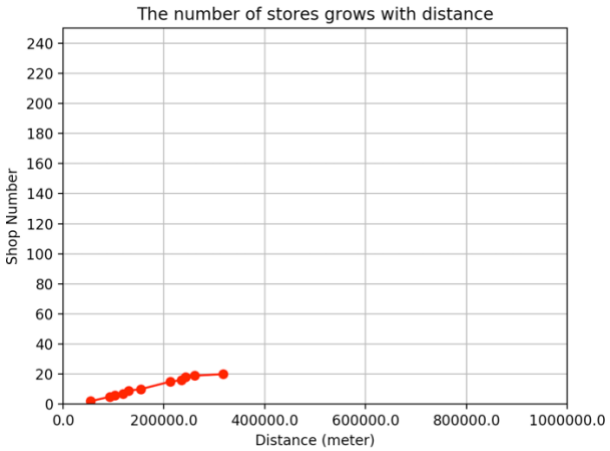


Figure 37: The number of 'funghi' shops varies with distance

According to the results of 'funghi', the project first thought of calculating the average distance (mean distance) of the shop which contains the word from the central point. This is because the small average distance from the central point means that the word distribution range will not be very large. The result of regional and non-regional words result is showed in the following table.

Regional words	Average distance
haggis	216926.14
bru	266813.33
naan	214933.34
roe	239904.54
supper	387295.02
pakora	306999.72
funghi	176748.19

Table 3: Regional words with 'average distance'

Non-regional words	Average distance
chip	351005.86
sausage	353703.24
supreme	308121.37
gift	262870.84
soup	361040.31

daily	389090.17

Table 4: Non-regional words with 'average distance'

According to the Table 3 and Table 4, the 'average distance' of most of regional words is less than 300,000 meters and most of non-regional words' average distance is larger than 300,000. Thus, the project initially decided to use 'average distance' < 300,000 as a feature of regional words.

In the exploring regional features process of the project, the project is inspired by the phenomenon that the closer to the central point, the denser the distribution of the shops. Thus, the project decided to use the median distance of cities which contain the word to try to discover new features. Firstly, the distances from the central point of all cities which contain the word were calculated and sorted by the project. Next, the median distance ('median') was found by the project and then the project found cities with distances below the median and calculated the number of shops included in those cities ('num_less'). Besides, the project also found cities with a distance larger or equal to the median distance and calculated the number of shops in those cities ('num_above'). By comparing the 'median', 'num_less' and 'num_above' separately of regional and non-regional words, the project did not find any features to distinguish regional and non-regional. However, the project found that 'proportion' which means 'num_less' divided by the total shop number of the word could distinguish regional and non-regional and the following table is the result.

Regional words	**Proportion**
haggis	76.92%
bru	83.01%
naan	74.19%
roe	67.80%
supper	77.19%
pakora	78.94%
funghi	45%

Table 5: Regional words with 'proportion'

Non-regional words	**Proportion**
chip	55.20%
sausage	55.05%
supreme	59.46%
gift	57.14%

soup	42.10%
daily	55.56%

Table 6: Non-regional words with 'proportion'

According to Table 5 and Table 6, the project found that excepting 'funghi', the 'proportion' of all regional words larger than 67%, and the 'proportion' of all non-regional words lower than 60%. As a consequence, the project defined 'proportion' > 67% as a feature of regional word.

In addition, the project also conjectured that if a word is distributed in many cities, it may be a regional word. In order to verify this thought, the project calculated the values of the above five features for all words and stored them in a CSV format file (word, ratio, proportion, average distance, city number, the number of shops). Through this file, the project found that when a word's 'city number' > 19, the project can directly conclude that it is a non-regional word because the words of 'city number' > 19 have low 'ratio', low 'proportion' and large 'average distance'. However, 'pasty' which distributed in eighteen cities shows regionality. Thus, the project defined 'city number' < 19 as one of the features of regional words.

3.4 Summary and future work

Through observing and comparing geographical maps and trends of known regional and non-regional words, iteration one derives five features ('ratio' > 65%, 'shop number' > 10, 'average distance' < 300,000 meters, 'proportion' > 67%, 'city number' < 19) of regional words. The project wants to use these features to make regional judgement for all the separated independent words. However, the project cannot use all these features in one conditional statement to judge regional words, because some words such as 'massala' only satisfy some of these features. As a result, the project can only divide a data set into two parts by selecting one feature at a time, and then divide the result of the division through another feature. However, the project cannot judge whether each division is the optimal division which means the currently selected feature can maximize the distinction between regional and non-regional words. Fortunately, the ID3 algorithm in decision tree can help the project to solve this problem that ID3 algorithm uses Entropy to select feature to achieve optimal division. Thus, in the next iteration, the project will use a decision tree to use these features to get the regional words result.

Chapter 4

Iteration 2

The aim of the iteration two is to use the decision tree and features found in iteration one to classify the independent single words. In this iteration, the following tasks will be done by the project: generating the training set for the decision tree, using ID3 algorithm to generate the decision tree, evaluating the ID3 algorithm and the regional result of the ID3 decision tree. In addition, after evaluating the ID3 algorithm, this iteration introduces another decision tree algorithm (Cart algorithm) and evaluates the algorithm and regional result and makes recommendations for the next iteration.

4.1 Methodology

In this section, the project focuses on the generation process of the training data set and the ID3 algorithm.

4.1.1 Training dataset

The training dataset is generated by the project, containing some known regional words and non-regional words and the project initially wanted to use 20% of the total data (5289 words) as the training dataset. According to the 'word with features' file generated from the iteration one, the project found that only one-fifth of the total words' shops number is more than ten and among these words, most words show obviously non-regional features. For example, non-regional words 'drink' and 'bean' have low 'ratio' and 'proportion', and very large 'average distance'. Thus, finding a sufficient number of non-regional words as training data is easy. However, it is difficult to find a sufficient number of regional words by observing the data set in the 'word with features' file. The reason is that finding regional words does not only require to observe the geographic maps visualisation results but also is influenced by the subjective criteria of the developer. In the process of generating a training set for the first time, the subjective criteria of the developer become the biggest obstacle to select regional word, because there are many words are commonly used in English, such as 'securely' and 'instantly' that present some of the regional features. The project though that although these words have some regional words features, they do not have all the features. As a consequence, the project did not add these words with their features into the training dataset. In the end, the project just defined twenty-five regional words, most of them were dish names and others were place names such as 'Yorkshire'. Besides, there were seventy non-regional words were added to the training dataset, then the project used this dataset to generate the ID3 algorithm decision tree.

4.2 Findings

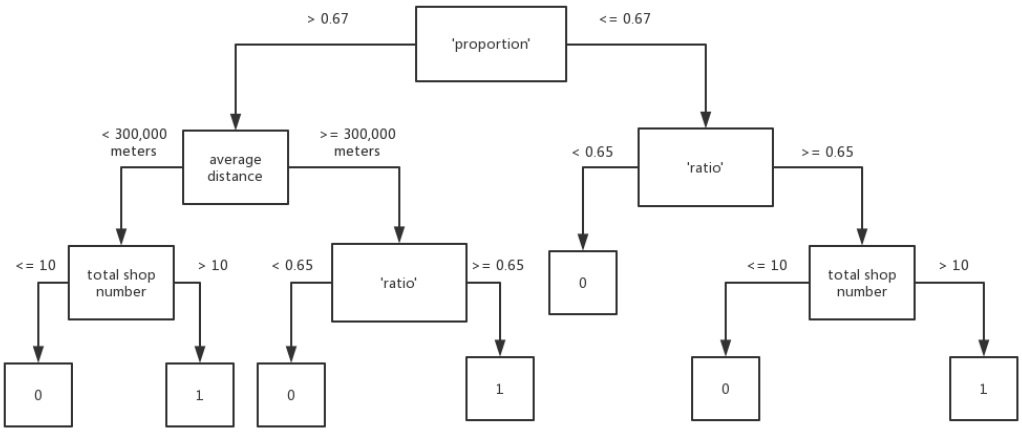


Figure 38: ID3 algorithm decision tree

Fig. 38 is the decision tree result of ID3 algorithm, where '0' represents non-regional words and '1' represents regional words. There are 54 following independent words are judged as regional words. In these words, 'haggis', 'irn', 'bru', 'pakora', 'pasty', 'roe', 'yorkshire', 'kiev', 'inferno', 'crunch', 'macaroni', 'naan', 'pattie', 'kidney', 'spaghetti', 'balty', 'haagen', 'dazs', 'cob', 'parmesan', 'plaice', 'pukka', 'savoury', 'sury', 'rump' were used in training dataset. 'skate', 'bolognese', 'hamburger', 'rib', 'carbonara', 'meaty', 'cucumber', 'guava', 'pattie', 'burdock', 'splash', 'dandelion', 'scallop', 'keema', 'samosa', 'give', 'smokey', 'cornish', 'bit', 'quattro', 'passion', 'facebook', 'chosen', 'value', 'securely', 'instantly', 'rock', 'shot', 'bull' were new founded.

4.3 Evaluation and improvement

Although the project got regional word, in addition to the dish words, many of these words are commonly used in English such as 'bit', 'give' and 'value'. In order to understand why these words were judged as regional words, the project decided to find the context of these words to figure out the usage of these words on the website. To achieve this, the project wrote a Python script whose input are these regional words. This script is responsible for searching the context of these regional words in all the HTML files obtained in the iteration one and generates a file of these words and their context. Thus, the project can according to this file to find the reason why the words were judged as regional words.

According to the word context file, the project found that the following words represent a kind of dish or part of the name of the dish: 'haggis', 'irn', 'bru', 'kiev', 'inferno', 'crunch', 'skate', 'bolognese', 'macaroni', 'naan', 'hamburger', 'plaice', 'rib', 'kidney', 'spaghetti', 'carbonara', 'pasty', 'roe', 'balty', 'meaty', 'cucumber', 'cob', 'guava', 'pakora', 'pukka', 'savoury', 'pattie', 'burdock', 'parmesan', 'splash', 'dandelion', 'scallop', 'keema', 'samosa', 'sury', 'rump', 'dazs', 'rock', 'shot', 'haagen', 'bull', 'macaroni', 'cornish', 'quattro', 'passion'. It is worth mentioning that in these regional

words, ‘irn-bru’ always uses together which represents a Scottish drink; ‘haagen’ and ‘dazs’ represent regional words because in that area, Haagen-Dazs maybe has more trade links with the merchants; ‘rock’ always used with ‘eel’. ‘rock eel’ represents a kind of fish; ‘shot’ always used with ‘hot’ and ‘hot shot’ represents a kind of dishes; ‘bull’ always used with ‘red’ and ‘red bull’ is a drink and the reason why ‘bull’ is regionally distributed maybe same as ‘haagen’; ‘cornish’ is always used with ‘pasty’ and ‘cornish pasty’ is a dish; ‘macaroni’ is always used with ‘cheese’ and ‘macaroni cheese’ is a dish; ‘quattro’ is always used with ‘stagioni’ and ‘quattro stagioni’ is a kind of pizza; ‘passion’ represents fruit or dish. Maybe in that area this kind of fruit is famous or selling well.

In addition to the regional dish words mentioned above, there are some words which are commonly used in English were judged as regional words and the project has found the reason for this based on these words’ context. ‘securely’ is mostly used with ‘with’ and ‘pay securely online’. Besides, the project found that when ‘securely’ used with ‘with’, all websites that use this usage have the same style. Similarly, all websites which have the usage of ‘pay securely online’ have the same style. This may be because the website of shops in the area was developed by the same company. As a consequence, ‘securely’ appears regionally. Similarly, the reason why ‘chosen’ (always used with ‘flavour’) and ‘instantly’ (always used with ‘chip shop takeaway - order online instantly’) were judged as regional words is same as ‘securely’. In terms of ‘Yorkshire’, most of ‘Yorkshire’ represent a place named ‘Yorkshire’. Therefore, ‘Yorkshire’ is a regional word that represents a place name. ‘give’ is used as a verb. Maybe in that area, people are used to expressing their own dishes in this way, such as give the best taste. ‘smokey’ is used as an adjective, usually in conjunction with a ‘bbq’ or ‘sausage’. ‘bit’ is usually used as a degree adverb. ‘value’ is always used with ‘box’ or ‘meal’ which represent dishes. In terms of ‘facebook’, the project found that some websites provide a Facebook account and shops which contain ‘facebook’ are distributed in a small region.

According to the results of the ID3 algorithm, although the project successfully determined the words’ regionality according to the regional features, the project also found a problem. This problem is that the threshold of each feature may not accurate because these thresholds are defined by the developer by observing a limited amount of data in iteration one. As a consequence, the project wants to use an algorithm to define the feature threshold automatically. However, the project cannot use ID3 algorithm to find the feature threshold, because the limitation of the ID3 algorithm is that it can only deal with discrete values [40]. That means the feature values must be classified based on numerical variables and the project has to mark each training data’s feature to numerical variables. For example, the ‘average distance’ of ‘haggis’ is 216,926 meters and the project will mark it as ‘<300,000’. Fortunately, Python provides a toolkit (Sklearn-learn) which integrates a variety of machine learning algorithms for supervisory and unsupervised problems [41]. This toolkit can help the project to achieve the goal of finding thresholds automatically. As a result, the project used the Sklearn-learn package decision tree algorithm which uses a kind of optimized Cart algorithm [42] to generate the decision tree, including a classification tree and regression tree. In this project, the classification tree is more suitable, because the target of the decision tree is binary.

4.3.1 Cart algorithm findings and evaluations

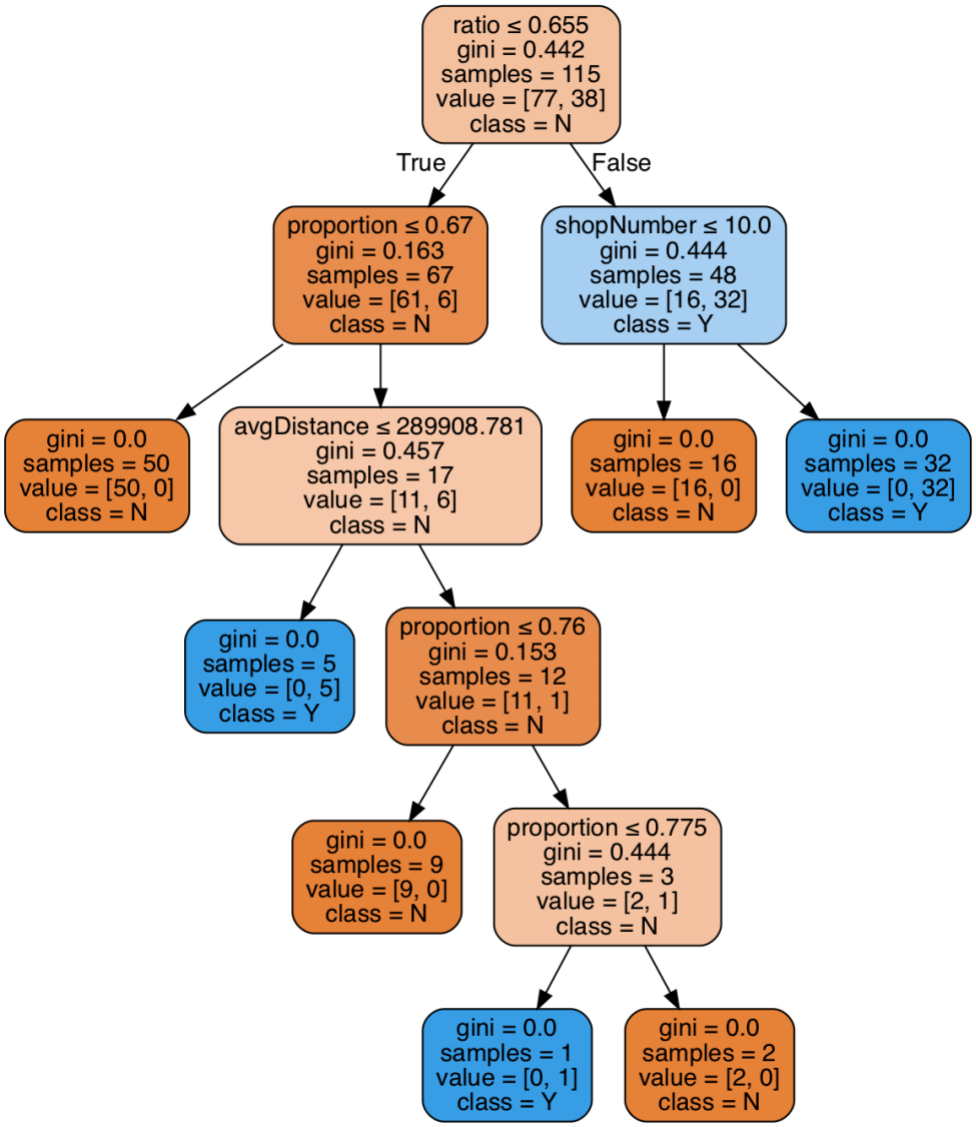


Figure 39: Cart algorithm decision tree

The Fig. 39 is the decision tree result of using Cart algorithm. Before generating the Cart algorithm decision tree, the project added some regional words to the training dataset based on the results of the ID3 algorithm. Although the threshold of feature is not accurate, the results of the ID3 algorithm classification were verified by the project to be regional. Thus, the project added some new regional words from the regional result of the previous classification. The purpose of this is to get a tree that can classify regional words more accurately.

In terms of the classification result of the Cart algorithm, Although the Cart algorithm generates more tree branches and the branching conditions become more precise, the classification results are not quite different from the ID3 algorithm. There are 56 words

were judged as regional words that in these 56 words, excepting ‘mince’, 55 are in the result of the ID3 algorithm. ‘mince’ is a word which widely distributed near Edinburgh. However, when using the ID3 algorithm, because of its ‘proportion’ larger than 67%, but its ‘ratio’ lower than 65%, so it was judged as a non-regional word.

4.4 Summary and future work

In this iteration, the project first obtained regional words. Besides, the project got the reason why the word is judged as a regional word based on the context in which the word appears. Further, the project found that ‘city number’ did not have any impact on the decision tree, therefore ‘city number’ is a useless feature for the project. However, the project wants to continue to explore the possibility of each classification result such as how likely is ‘haggis’ to be classified as regional content. Unfortunately, the project cannot get the probabilities through using the decision tree, because in this project, the result of the decision tree is binary. As a consequence, in the next iteration, the project wants to use a regression classifier of Sklearn-learn package to get the probability that a particular content is in a category.

Chapter 5

Iteration 3

The aim of iteration three is using logistic regression to find the probability that a word is judged as a regional word. During the process of iteration three, the project also found other valuable information, such as the impact of each feature on the classification results, impact of using L1 and L2 on classification results, and some newly discovered regional words. The iteration three divided into four sections: methodology, findings, evaluation and summary and future work. The methodology part focuses on describing the improvement of the training dataset and the changes in the selection of features. The findings part mainly discusses the logistic regression model results and probability results. The evaluation part concentrates on evaluating the findings. The summary and future work part is responsible for summarizing iteration three and arrange the plan for iteration four.

5.1 Methodology

The project wrote a Python script to achieve logistic regression. The training dataset used in the regression is an improved dataset which was added more results generated in the decision tree. Besides, the project used the features excepting ‘city number’ and ’shop number’ for the logistic regression. Considering ‘city number’, it has been confirmed by the decision tree that it had no effect on the classification. In terms of ‘shop number’, the reason for not using it as a feature is because of the first logistic regression classification result of the project. For the first classification, the project used all other features except ‘city number’ and found that logistic regression cannot directly classify word with ‘shop number’ less than ten into non-regional word like the decision tree. In contrast, the logistic regression set the regional probability of these words very high which means these words have a high probability of being regional words. In logistic regression, the probability that a word belongs to a category is affected by the weight of the features. In this project, the weight of ‘shop number’ is smaller than other features and almost closes to zero. As a consequence, words with ten or fewer shops may be given a high probability. In order to filter words with ‘shop number’ less than ten, the project filtered them in the script and did not use them as a training set for generating the logistic regression model. Besides, the project also did not make regional judgments on words with less than ten shops, because the project only cares about regional words, and the filtering of these words have no effect on regional words. In the end, the project decided to filter the words with less than ten shops instead of using the ‘shop number’ as a feature.

The project initially wanted to use logistic regression to obtain the probability that the word is judged as a regional word. However, some other findings were discovered during the classification procedure, such as the differences between using L1 and L2 regularizations.

5.2 Findings

The following table is generated using ‘proportion’, ‘ratio’ and ‘average distance’ as features. It compares the differences of using L1 and L2 regularizations. The classification rate means the amount of data in the training dataset correctly classified by the model divided by the total training set data amount. The percentage threshold means that if the regional probability of a word is higher than this threshold, the word will be judged as a regional word and it was obtained by observing the probability results. The project observed every 10% interval from high probability to low probability. If there are more than five regional words (decided in the second iteration) in an interval, the lower bound of this interval will be considered the threshold.

Regularization	**L1**	**L2**
Classification rate	83.13%	53.57%
Percentage threshold	50%	Cannot tell
Coefficients (‘proportion’, ‘ratio’ and ‘average distance’)	1.50840108e+00 4.48321094e+00 -1.51522389e-05	3.44111057e-12 8.63444565e-12 -1.77054477e-06
The number of regional words	59	Cannot tell
Regional words with probability	See full list of regional words in Appendix A-A.1	Meaningless

Table 7: Comparison of regularization choices for independent word when selecting ‘proportion’, ‘ratio’ and ‘average distance’ as features

The regional result of the words’ probability (greater than 50%) when using L1 regularization is showed in Appendix A.1. Compared to the results of the decision tree, some words were newly defined as regional words (Probability is greater than 50%). They are ‘premium’, ‘spam’, ‘greek’, ‘telephone’, ‘munchie’, ‘serving’, ‘cookie’, ‘farm’, ‘under’. However, there are six regional words defined in the decision tree were given low probability, such as ‘bolognese’, ‘plaice’, ‘kidney’, ’rump’, ‘bull’, ‘mince’.

Unfortunately, when using L2 regularization, the regional probability of all words is less than 50%, which means the logistic model cannot judge the regionality for each word. As a result, in Table 7, the project cannot tell the threshold and the number of regional words.

5.3 Evaluation

According to the result of Table 7, using different regularizations resulted in very large differences in classification rates. When using L1 regularization, L1 regularization reduced the weight of features which have less impact on classification. Through the

result of coefficients, the project knows that among the three features selected by the project, ‘ratio’ has the greatest impact on the results of the classification, followed by ‘average distance’ and ‘proportion’. The reason why the coefficient of ‘average distance’ closed to zero is not because the L1 regularization considered this property not important and reduced it to zero. It is because values of the ‘average distance’ are large, such as 300,000, so the model reduced the coefficient of ‘average distance’ to make the probability output of the logistic regression model in zero to one. In addition, the logistic regression model regarded as ‘average distance’ as an important feature.

The project speculated that the reason for appearing above new regional words was because of the ‘average distance’. In order to verify this speculation, the project observed the features of these new regional words and found that the ‘average distance’ of these words was small and ‘ratio’ and ‘proportion’ were low. Thus, in the decision tree, these words were judged as non-regional words because of the low ‘ratio’ and ‘proportion’. However, in the logistic regression model, these words were judged as regional words because their ‘average distance’ is small which means their shops are distributed in small areas. In terms of words which were judged as regional words by decision tree but given low probability by the logistic regression model, the project divides these words into two categories to analyse. The first category contains ‘bolognese’, ‘plaice’, ‘kidney’ and ‘mince’. The ‘ratio’ of these words is very low and the ‘average distance’ is not very small, but their ‘proportion’ is high. However, in the logistic regression model, ‘proportion’ has the lowest impact on the results. As a consequence, the probability of these words is less than 50% but close to 50%. The second category includes ‘bull’ and ’rump’. They have high ‘ratio’ and high ‘probability’, but very large ‘average distance’. Thus, the logistic regression model gave them a probability of close to 50% but lower than 50%.

When using L2 regularization, the logistic regression model reduced the coefficients of all features to close to zero. This means the model just used the ‘average distance’ feature to judge the regional words. The project speculates that L2 regularization wanted to fit all these three features, but it overfitted the feature of ‘average distance’, causing the other two parameters to be close to zero. As a consequence, the regional probability of all words was lower than 50% and the model could not make a decision. This is the reason why classification rates show large differences when using different regularizations.

The project also found the context of the new appearing regional words and found the following findings: ‘greek’ always used with ‘salad’ or ‘pizza’ and represent a kind of dish. Almost all ‘spam’ is used with ‘fritter’ and ‘spam fritter’ is a kind of dish. ‘premiums’ are often used with a kind of ingredients such as salmon and cod or a kind of sauce such as tartare sauce. The reason that ‘telephones’ have a high regional probability is that all shop websites with ‘telephone’ in that region are developed by the same company and each website has ‘telephone orders welcome’. ‘munchie’ is always used with ‘box’ that the ‘munchie box’ is a kind of fast-food in Scotland. ‘serving’ is mostly concentrated in the central England region and often used with ‘quality’ to show good ingredients. This may be a regional usage. The reason for the regionality of ‘cookie’ may be due to the good sales of cookies in that area. All ‘farm’ is used in conjunction with ‘house’ and ‘farm house’ are a regional brand of pizza. There are many usages of

‘under’, but the most appearing is ‘under 12’, and shops contain this usage can be discounted for people under 12 years old and they are concentrated in central England.

5.4 Summary and Future Work

Iteration three got the probability of the regional words. Besides, through the logistic regression model, the project discovered some new regional words and analysed the reasons for these words to be judged as regional words. Further, through analysing the logistic model coefficients, the project knew that the “ratio” has the greatest impact on the classification, followed by ‘average distance’ and ‘proportion’. In addition, the project also speculated the reason why L1 and L2 classification rates show a large difference. However, for independent single words, there are many words shows regionality because they are used in conjunction with other words. Thus, the project decides to use other types of content such as noun phrases and word pairs to find regionality of the content.

Chapter 6

Iteration 4

The aim of this iteration is to use noun phrases and word pairs as dataset to find content regionality. This iteration is mainly divided into two parts. The first one is noun phrase section and the second one is word pair section and both of them include methodology, findings, and evaluation. In the summary at the end of this chapter, the project compared the classification results of the three types of content (independent word, noun phrase and word pair).

6.1 Noun phrase

This section mainly describes the method of generating the noun phrases classification results and the evaluation of the result.

6.1.1 Noun phrase methodology

The project firstly needs to extract the noun phrases from the HTML files. The methods used to implement this is same as the methods used for extracting independent single words. However, it is different that the project extracted noun phrases by using spaCy which is an open source NLP toolkit [43] and it has encapsulated method that can directly identify noun phrases in paragraphs or sentences. Besides, in terms of noun phrases, the project does not need to convert noun plural to singular. Specifically, the project created a new script that was built on the original script which is used for extracting the independent words and the project made some changes to it. After the HTMLParser recognized the contents of the HTML tag and filtered special symbols, the project used the encapsulated function to extract noun phrases. However, the project found that the noun phrase extracted by spaCy sometimes contained some verbs and adjectives. As a consequence, the project continues to use the part of speech recognition method in spaCy to filter verbs and adjectives in noun phrases in this step. For example, ‘large chicken burger’ and ‘small chicken burger’ are the same dish but different size. Thus, the project needs to remove the adjectives ‘large’ and ‘small’ to ensure that the two noun phrases can be classified into one category. The noun phrases generated by the script contain both noun phrases and independent words because some of the HTML tag content is an independent word. Next, the project uses the same steps as the independent word to generate regional results. It is worth to mention that when generating the noun phrase training dataset, the project referred to the result of the independent words. This means the project priority found regional noun phrases from phrases which contain regional words such as ‘diet bru’ and ‘vegetable pakora’ because these noun phrases are more likely to be regional. In addition, the project also used geography map visualisation results and trends graph to verify regionality of the above noun phrases. The noun phrase training dataset contains twenty-four regional noun phrases and thirty-three non-regional

noun phrases. After the project was ready for the training dataset, the project firstly used ‘proportion’, ‘ratio’, ‘average distance’, ‘shop number’ and ‘city number’ as features to generate a decision tree. After discovering the ‘city number’ was also useless for noun phrase classification (detail in 6.1.2), the project used ‘proportion’, ‘ratio’ and ‘average distance’ features to obtain the logistic regression results.

6.1.2 Noun phrase decision tree findings

The following diagram is the result of noun phrase decision tree, which used Cart algorithm.

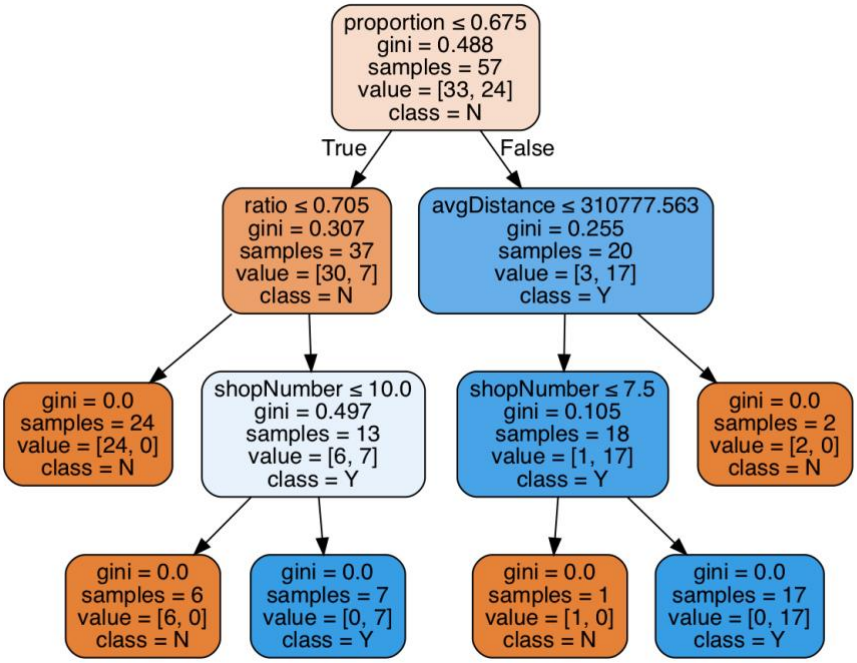


Figure 40: Noun phrase decision tree

There are 52 noun phrases were judged as regional phrases:

‘kidney pie’, ‘cod roe’, ‘pattie’, ‘bru’, ‘cheeseburger quarter pounder’, ’vegetable pakora’, ‘funghi’, ‘chicken pakora’, ‘haggis’, ‘pasty’, ‘pie supper’, ‘pudding supper’, ‘macaroni cheese’, ‘inferno’, ‘naan’, ‘hamburger’, ‘pakora’, ‘spaghetti’, ‘bull’, ‘diet bru’, ‘suey roll’, ‘roe’, ‘mince’ were in the training dataset.

‘rice’, ‘pop’, ‘pasties’, ‘fish chips’, ‘cookies’, ‘supper’, ‘chicken meat’, ‘shot’, ‘king rib’, ‘pasta’, ‘pollo’, ‘dandelion’, ‘cheeseburger half pounder’, ‘sausage supper’, ‘burdock’, ‘beef onion pie’, ‘bit’, ‘potato pie’, ‘cheese’, ‘ale’, ‘chip roll’, ‘cheese tomato’, ‘pizza supper’, ‘chicken breast supper’, ‘chip shop takeaway order’, ‘diet coke ltr’, ‘chicken leg supper’, ‘pineapple ring’, ‘pizza crunch’, ‘hamburger supper’ were newly discovered.

6.1.3 Noun phrase decision tree evaluation

According to the findings of the noun phrase, some phrases such as ‘vegetable pakora’ and ‘cod roe’ appeared in the appearing context of the independent regional words. However, some words such as ‘rock’ and ‘skate’ appeared in the results of independent word were not exited in the above noun phrases. This is because there are many types of phrases that appear with these independent words, and the number of shops of each noun phrase less than ten shops. Thus, these phrases were classified into non-regional phrases. For example, ‘rock’ appeared in the results of independent word and often used as ‘rock eel’, but the shop number of ‘rock eel’ was less than ten times. Thus, ‘rock’ disappeared in the results of noun phrase. In addition, there are some new words were judged as regional words, such as ‘pop’ and ‘pollo’ and the project found the reason for their appearing. Taking ‘pop’ as an example that ‘pop’ is concentrated in the central of the UK (around Manchester), representing a kind of drink. However, it also appears in noun phrases such as ‘bottle pop’ and ‘pop pepsi’. Thus, the ‘proportion’ of ‘pop’ in noun phrase form is higher than ‘pop’ in independent word form. As a consequence, ‘pop’ meets the conditions of regional word. In terms of phrase ‘chip shop takeaway order’, the reason why it was judged as a regional phrase is that the web pages which contain this phrase has the same style and was developed by the same company.

6.1.4 Noun phrase logistic regression findings

The following table is the comparison between using L1 and L2 regularizations of the logistic regression model of noun phrase.

Regularization	**L1**	**L2**
Classification rate	89.74%.	89.74%.
Percentage threshold	50%	50%
Coefficient (‘proportion’, ‘ratio’ and ‘average distance’)	1.43506897e+00 4.21097118e+00 -1.42088335e-05	1.27526880e+00 1.74406725e+00 -1.25081676e-05
The number of regional noun phrases	64	66
Regional noun phrases with probability	See full list of regional noun phrases in Appendix A-A.2	See full list of regional noun phrases in Appendix A-A.3

Table 8: Comparison of regularization choices for noun phrase

6.1.5 Noun phrase logistic regression evaluation

According to the Table 8, the project found that in terms of the classification result, there is not much difference between using L1 penalty and L2 regularizations. After the project

analysed the training dataset of the noun phrases, the project found that using any of the three features (‘ratio’, ‘proportion’ and ‘average distance’) can successfully classify most noun phrases though ‘ratio’ still has the biggest impact on the results. Thus, when using L2 regularization, the coefficients of the three features are almost the same.

In addition, compared to the result of the decision tree, when using logistic regression, some new phrases appeared. The project speculates that the reason for the appears of new phrases in logistic regression is because the imperfect training datasets lead to insufficient feature threshold. Thus, some phrases such as ‘spam fritter’ whose ‘proportion’ and ‘ratio’ are low, but it also has low ‘average distance’ are judged as non-regional phrases in the noun phrase decision tree. Besides, the project found that the reason why these noun phrases were judged as regional phrases is same with reasons described above (include regional words’ reason and regional noun phrases’ reason).

6.2 Word pair

This section mainly describes the method of generating the word pair classification results and the evaluation for the result.

6.2.1 Word pair methodology

The script used to generate the word pair is built based on the script that generates the noun phrase. The change in the script is that after the HTMLParser recognized the content of the tag and filtered special symbols, the script split the content into word pairs by space. For example, ‘I am XXX’ will be split into ‘I am’ and ‘am XXX’. The result of word pair script contains independent words, noun phrases, and word pairs because some of the content itself is a word or noun phrase. The rest of the word pair script is the same as the noun phrase.

6.2.2 Word pair decision tree findings

The word pair decision tree used ‘ratio’, ‘proportion’, ‘average distance’, ‘shop number’ and ‘city number’ as features.

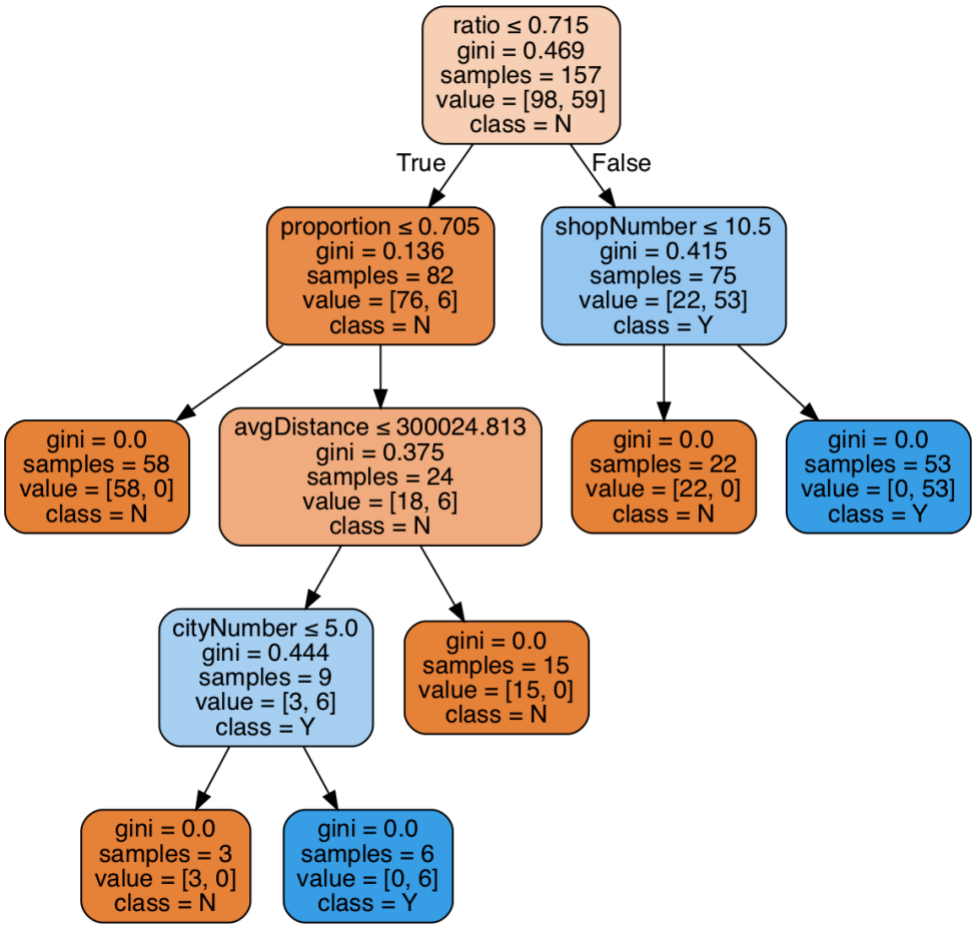


Figure 41: Word pair decision tree

There are 117 word pairs were judge as regional word pairs. (black pudding, in pitta, cod roe, irn bru, sausage large, fish chips, peas small, or gravy, mince pie, chicken meat, or curry, special fish, calzone, cornish pasty, chicken pakora, pasties, large sausage, large haddock, mixed meat, vegetable pakora, order online, fish supper, haggis, smoked sausage, cheese tomato, peas curry, sausage supper, south fried, bolognese, mini fish, large peas, kebab in, chicken chips, half pizza, red bull, king rib, salad or, scallop, wrap meal, of juice, cockles, salt, meat in, baked potatoes, hot shot, donner in, coke ml, bru l, coke ltr, chop suey, potato pie, al funghi, all steak, rubicon guava, baked potato, naan bread, smokey sausage, scampi supper, sausage single, bolognese sauce, mixed pakora, pie supper, bites and, burger single, spaghetti bolognese, pudding supper, cheese pattie, chip roll, chip steak, pizza supper, haagen dazs, pizza single, fried pizza, bru ltr, macaroni cheese, breast supper, haggis supper, white pudding, chicken balti, scampi single, burger supper, kebab pizza, rib supper, fanta ml, diet irn, in naan, pie single, nuggets supper, nuggets single, chicken supper, bru ml, fish butty, leg single, fish single, half fried, rib single, chicken single, online now, takeaway order, skate, suey roll, breast single, online instantly, pudding single, leg supper, securely with, pineapple ring, steak supper, pakora

vegetable, shop takeaway, spicy haggis, steak single, hamburger supper, haggis single, pizza crunch, rump steak, hamburger single).

6.2.3 Word pair decision tree evaluation

In these word pairs, many of them contain regional independent word, such as 'securely with' and 'rubicon guava'. Besides, many of them appeared in the result of noun phrases such as 'cheese tomato' and 'vegetable pakora'. Further, some word pairs like 'steak single', 'coke ml' and 'half pizza' maybe reflect regional marketing habits that in these regional, residents may like to mark the drink in millilitres, or they will sell the pizza in half. In addition, according to the word pair decision tree, the project finds that the 'city number' feature influenced the classification result. Thus, in word pair logistic regression model, the project would use this feature to generate the model.

6.2.4 Word pair logistic regression findings

Regularization	**L1**	**L2**
Classification rate	88.77%	74.48%
Percentage threshold	Cannot tell	Cannot tell
Coefficient ('proportion', 'ratio', 'city number' and 'average distance')	1.83013399e+00 6.22791230e+00 -2.69416502e-01 -7.79706487e-06	1.06066688e-01 1.80860335e-01 -1.26984839e-01 5.89637008e-06
The number of regional words	Cannot tell	Cannot tell
Regional words with probability	meaningless	meaningless

Table 9: Comparison of penalty choices for word pair when selecting, 'proportion', 'ratio', 'city number' and 'average distance' as features

There are 200 word pairs whose regional probability exceed 50% when using L1 regularization and when using L2 regularization, there are 684 word pairs' regional probability exceed 50%.

6.2.5 Word pair logistic regression evaluation

According to the logistic regression result of word pair, although the classification rate was high when using L1 regularization, in the 200 word pair results with a regional probability greater than 50%, the project could not find the regional threshold. The project found that in these 200 word pairs, there are many word pairs that are not regional word pairs through geographic visualization results and trend graphs. Besides, through

analysing these non-regional word pairs, the project found that these word pairs have the same characteristics that they have high ‘ratio’, large ‘average distance’ and low ‘proportion’. By observing the word pair training dataset, the project knew that the ‘ratio’ of all regional word pairs in the training set is higher than 59%, but the other three features cannot tell a certain threshold. In addition, according to the coefficient of each feature when using L1 regularization, the project found that ‘ratio’ has the most impact on the results and the impact of ‘average distance’ is small. Thus, this led to a bad classification result. However, the project believes that the root cause for the poor classification of the model is because the training dataset samples are not enough.

When using L2 regularization, ‘average distance’ has the greatest impact on the results of the classification. As a result, this is why there are more non-regional word pairs were identified as regional word pairs.

Based on the above description, the project thinks that the classification result of the word pair obtained by using the logistic regression is meaningless. As a consequence, the project only made a discussion of the result and did not show the results details.

6.3 Summary

By comparing the classification results of these three data types, using independent words for categorization can find most regional words, but requires the project to continue to find their usage context which is used to determine why they were judged as regional word. When using noun phrases to classify, projects can directly find the reason why they are classified as regional noun phrases from most of the results, because most of noun phrases are dish names. When using word pair dataset, there are more regional classification results (generated by decision tree) than the results of independent words and noun phrases. The main reason for this is because after the project split the content, the number of word pairs generated by the project was between the number of independent words and noun phrases. Besides, some words that were filtered in the process of generating independent words and noun phrases appeared in the word pairs. For example, ‘ml’ and ‘ltr’ will be filtered in the extracting process of generating independent word and noun phrases, but they were kept in the word pairs. Thus, when these words used with regional independent words such as ‘bru’, ‘bru ml’ and ‘bru ltr’ were judged as regional word pairs. In addition, because of the imperfection of the word pair training dataset, the classification effect of the logistic regression model was not good that the results of the regression model have no reference value.

Chapter 7

Conclusion

The biggest success of the project is that the project successfully explored ways to link 'Fish & Chips' shop menu information to regional differences of the UK. Through four iterations of exploration, the project successfully explored potential regional information in restaurant websites. The first iteration found methods to acquire data and clean data, including data source selection, data retrieving, extracting independent words from HTML files and finally obtained independent words with their coordinates. Besides, by visualising the distribution of words and trend of the number of shops increasing with distance (meter) from the central point, the project identified some regional words and successfully identified the features of regional words. Based on these features, the project explored classification methods for regional words. The second iteration successfully classified words through using the decision tree and this was the first time for the project to see regional results. However, in terms of independent words, it was difficult for the project to directly explain why they showed regionality. Thus, the project identified the context in which these words appeared and successfully explained the regional reasons for the words. Through the exploration of the third iteration, the project successfully found the probability of words presenting regionality and the impact of each feature on the classification results. After successfully obtaining the regional results of independent words, the project explored the regionality of noun phrases and word pairs in the fourth iteration.

Through evaluating the regional results of the project, the project found some regional differences. The first one is the most obvious difference among all the differences which is the differences in eating habits. For example, residence in Scotland like to eat haggis and drink irn bru and the people of England like to eat cod roe and naan. The second one is the difference in shopping habits. For example, in London and the cities around it, there are many 'Fish & Chips' restaurants selling Haagen-Dazs ice cream. This may mean that people in that area like to buy Haagen-Dazs when they consume in the 'Fish & Chips' shops. The third is the difference in sales habits of 'Fish & Chips' shop. For example, in Manchester and the surrounding cities, many 'Fish & Chips' shops offer discounts for children under the age of twelve, and there are many shops that sell half of the dish such as 'half pizza'. The last difference is reflected in language usage such as Scots like to use 'supper' to describe dinner.

However, the project also has limitations. The first one is the imperfect training datasets which occurred in three types of content. This may lead to deviations in the classification results obtained by the project or even incorrect classification, such as the logistic regression results of the word pair. The second limitation is that limited time leads to limited analysis of results that the project thought there are other reasons can be used to explain why logistic regression model could not classify regional and non-regional content very well. Besides, the project wants to study more parameters of the

classification model to optimize the classification model, but because of the time, the project did not do it. The third limitation is that cities selected by the project distribute unevenly in some areas such as region near the Newcastle. The project found that in that area, several selected cities are densely distributed, and this may cause some bias to regional results.

Although the project has these limitations, the project initially explored feasible methods for this study that this project answered where to obtain data, how to retrieve data, how to clean data, how to mine features from data and how to classify data. These methods can be referenced or reused by subsequent researchers of similar studies. In addition, the regional results gained by this project can be used as evaluation references by subsequent similar studies.

Chapter 8

Future Work

For this project, there are still many things that can be done in the future.

8.1 Finding more 'Fish & Chips' shop menus

This project only used Google and TripAdvisor to search for 'Fish & Chips' shops, but the project believes that there are many other websites can be used to search for more menus of 'Fish & Chips' shops. Besides, the project can find a way to extract information from some PDF menu pages. By these way, the project can add more content samples to the dataset and do so will bring two benefits. The first benefit is that the project may find more regional content, including new emerging content (not exist in the result of regional or non-regional in this project) and content that was originally judged to be non-regional (content whose shop number less than ten) into regional content. The second benefit is that more samples will make the regional or non-regional features of the content more distinct. This means there will be more regional content concentrated at the central point and there will be more non-regional content distributed in a wider area from the central point. As a consequence, the training dataset can be improved, and the classification model can be generated more accurate by more accurate coefficients. Finally, the project can gain more accurate classification results.

8.2 In-depth study of algorithms and parameters of classification models

Because of the limited time, the project only tried two kinds of classification model and only analysed a few parameters of the model. In terms of the decision tree, the project can try to use more parameters to build the tree. For example, the project can set a minimum sample size that if a branch has at least samples of this size, the branch can be divided, rather than building a huge tree to satisfy all training samples. Considering logistic regression, the project can visualise the regression model, so the project can better know how the model divides the dataset. This can help to analyse the result of the logistic regression model. In addition, the project can analyse more logistic regression parameters such as learning rate and loss function optimization method to better explain the probability results obtained by logistic regression in this project.

8.3 Synonym detection

In the future, the project can try to use a new method to merge the same content that this method will determine the degree of similarity between the content. If the degree of similarity is high, the similar content will be divided into one category and has the same

content name. For example, ‘chips’ and ‘chip’ may have a high probability of similarity. Thus, the project can merge them as ‘chip’. By this way, the project can avoid content of the same meaning being divided into different contents due to string mismatch.

8.4 Extend the project to a wider range

In the future, the project can try to collect data and mine information from other kinds of restaurants rather than only focuses on ‘Fish & Chips’ shops. This can increase the amount of data in the content sample so that the content sample can better reflect its features, which is beneficial to improve the accuracy of the classification. Further, the project can use same methods to mine restaurant data in other countries not just only in the UK to explore other countries’ regional differences. Moreover, the project can mine other information from the menu data, such as predicting the taste of local people, according to the frequency of dishes or sauces on the menu. By this way, the project can give reasonable advice to businessmen who want to open a restaurant or sell ingredients locally.

Appendix A

Logistic regression probability results

A.1 independent word classification result (L1 penalty with ‘ratio’, ‘proportion’, ‘average distance’ features).

Word	Non-regional & Regional
cob	[0.011876565040122133, 0.9881234349598779]
dazs	[0.03565785374686514, 0.9643421462531349]
instantly	[0.03880521047969243, 0.9611947895203076]
haagen	[0.045216138886797874, 0.9547838611132021]
hamburger	[0.045310140513992514, 0.9546898594860075]
yorkshire	[0.06726020112458264, 0.9327397988754174]
carbonara	[0.0748655490961001, 0.9251344509038999]
inferno	[0.07820954149680193, 0.9217904585031981]
burdock	[0.08061048740063548, 0.9193895125993645]
dandelion	[0.08061048740063548, 0.9193895125993645]
splash	[0.08855926617249821, 0.9114407338275018]
parmesan	[0.10952897901931857, 0.8904710209806814]
quattro	[0.1148709886937076, 0.8851290113062924]
pattie	[0.11607358231246456, 0.8839264176875354]
naan	[0.12525699647456212, 0.8747430035254379]
rock	[0.12849691511617833, 0.8715030848838217]
give	[0.14043323848824363, 0.8595667615117564]
keema	[0.15252997882774477, 0.8474700211722552]
macaroni	[0.18026996474752066, 0.8197300352524793]
stagioni	[0.18552435704470105, 0.814475642955299]

cornish	[0.18743374731899054, 0.8125662526810095]
meaty	[0.1885092770891743, 0.8114907229108257]
skate	[0.19105596269953462, 0.8089440373004654]
haggis	[0.20066987838408545, 0.7993301216159145]
guava	[0.21053335042016352, 0.7894666495798365]
smokey	[0.2188211503446117, 0.7811788496553883]
scallop	[0.2374227627360933, 0.7625772372639067]
passion	[0.24220789410029497, 0.757792105899705]
shot	[0.24768428446102064, 0.7523157155389794]
rib	[0.24915028186127286, 0.7508497181387271]
crunch	[0.2595798061837933, 0.7404201938162067]
kiev	[0.29097630707112465, 0.7090236929288753]
pukka	[0.29878720611735254, 0.7012127938826475]
suey	[0.3070038321413041, 0.6929961678586959]
bit	[0.3167334820875807, 0.6832665179124193]
under	[0.3168042925668779, 0.6831957074331221]
farm	[0.3227256066834283, 0.6772743933165717]
balty	[0.3348070496159057, 0.6651929503840943]
value	[0.3501170228726148, 0.6498829771273852]
bru	[0.35124217479327335, 0.6487578252067266]
irn	[0.35124217479327335, 0.6487578252067266]
pasty	[0.35460931105113624, 0.6453906889488638]
cucumber	[0.3625614270875258, 0.6374385729124742]
securely	[0.38772979177115596, 0.612270208228844]
cookie	[0.39519298167919703, 0.604807018320803]
spaghetti	[0.3957820821304686, 0.6042179178695314]

facebook	[0.42024142622620053, 0.5797585737737995]
serving	[0.4248062839850639, 0.5751937160149361]
roe	[0.42622824594002295, 0.573771754059977]
munchie	[0.4367089175920039, 0.5632910824079961]
funghi	[0.4385420913404726, 0.5614579086595274]
telephone	[0.4650031510322067, 0.5349968489677933]
chosen	[0.4743260958093295, 0.5256739041906705]
pakora	[0.47695369015475175, 0.5230463098452482]
greek	[0.4793352062612427, 0.5206647937387573]
spam	[0.4824534401241375, 0.5175465598758625]
premium	[0.4826230587021527, 0.5173769412978473]
samosa	[0.49091441238010813, 0.5090855876198919]

A.2 noun phrase classification result (L1 penalty with 'ratio', 'proportion', 'average distance' features).

Noun phrase	Non-regional & Regional
cheese pattie	[0.013309049172268272, 0.9866909508277317]
hamburger supper	[0.0186291426474251, 0.9813708573525749]
pizza crunch	[0.022596261386979788, 0.9774037386130202]
potato pie	[0.023568146763607967, 0.976431853236392]
inferno	[0.03220485137533258, 0.9677951486246674]
naan	[0.03268870363025178, 0.9673112963697482]
pineapple ring	[0.034481232740873224, 0.9655187672591268]
pudding supper	[0.03671211066273117, 0.9632878893372688]
pizza supper	[0.037933276927266446, 0.9620667230727336]
macaroni cheese	[0.038477080449522916, 0.9615229195504771]
diet bru	[0.041033740318725975, 0.958966259681274]

chip shop takeaway order	[0.049003760243348515, 0.9509962397566515]
diet coke ltr	[0.05278428446365713, 0.9472157155363429]
hamburger	[0.05353617178504022, 0.9464638282149598]
pattie	[0.06093147368861507, 0.9390685263113849]
chip roll	[0.06283895877899259, 0.9371610412210074]
dandelion	[0.06420796357829794, 0.9357920364217021]
chicken meat	[0.06710259810974517, 0.9328974018902548]
suey roll	[0.07105552105199131, 0.9289444789480087]
king rib	[0.07476984265165865, 0.9252301573483414]
bit	[0.07763404564334098, 0.922365954356659]
pollo	[0.08065191355388068, 0.9193480864461193]
pie supper	[0.08337077826346695, 0.916629221736533]
shot	[0.09245513281940321, 0.9075448671805968]
burdock	[0.11106739501985041, 0.8889326049801496]
cheese tomato	[0.1334193658886207, 0.8665806341113793]
beef onion pie	[0.1360885812031546, 0.8639114187968454]
haggis	[0.14537848581225887, 0.8546215141877411]
funghi	[0.15603603785303422, 0.8439639621469658]
chicken breast supper	[0.16533691272400008, 0.8346630872759999]
pasty	[0.16887846051290278, 0.8311215394870972]
supper	[0.1828044881275689, 0.8171955118724311]
fish chips	[0.19210831009980167, 0.8078916899001983]
cheeseburger half pounder	[0.20097439977528253, 0.7990256002247175]
cheeseburger quarter pounder	[0.20097439977528253, 0.7990256002247175]
chicken pakora	[0.2495477396698309, 0.7504522603301691]
cookies	[0.2592799667107383, 0.7407200332892617]

bull	[0.26622326820247466, 0.7337767317975253]
bru	[0.2797324130622031, 0.7202675869377969]
chicken nuggets meal	[0.297673913638301, 0.702326086361699]
nuggets	[0.31071290604648216, 0.6892870939535178]
vegetable pakora	[0.3141644368669968, 0.6858355631330032]
pasties	[0.32791435320038853, 0.6720856467996115]
spaghetti	[0.3284135989436908, 0.6715864010563092]
chips salad	[0.33646537799041176, 0.6635346220095882]
scallops	[0.3597985894131531, 0.6402014105868469]
pop	[0.3661374073969268, 0.6338625926030732]
pakora	[0.38093837612429626, 0.6190616238757037]
site	[0.3822906709066207, 0.6177093290933793]
pasta	[0.3924413678971085, 0.6075586321028915]
kebab wrap	[0.4060773055076441, 0.5939226944923559]
inch margherita	[0.4082409512546662, 0.5917590487453338]
sausage supper	[0.4371353690721911, 0.5628646309278089]
piece	[0.44206894811490416, 0.5579310518850958]
inch bread	[0.4448106111373529, 0.5551893888626471]
mince	[0.4451640822512588, 0.5548359177487412]
pie chips	[0.45505646034034897, 0.544943539659651]
cod roe	[0.46053626058973207, 0.5394637394102679]
spring roll	[0.4624339277638466, 0.5375660722361534]
facebook	[0.4765961922514401, 0.5234038077485599]
spam fritter	[0.4841356602295611, 0.5158643397704389]
kidney	[0.4919224144884715, 0.5080775855115285]

A.3 noun phrase classification result (L2 penalty with ‘ratio’, ‘proportion’, ‘average distance’ features).

Noun phrase	Non-regional & Regional
cheese pattie	[0.04302663367975612, 0.9569733663202439]
hamburger supper	[0.05669634713763139, 0.9433036528623686]
pizza crunch	[0.06500484200373025, 0.9349951579962698]
potato pie	[0.06658411254621566, 0.9334158874537843]
inferno	[0.07750836744789169, 0.9224916325521083]
pineapple ring	[0.08145369004390668, 0.9185463099560933]
naan	[0.08670872839951937, 0.9132912716004806]
macaroni cheese	[0.09432914146692861, 0.9056708585330714]
pudding supper	[0.09525936198881202, 0.904740638011188]
pizza supper	[0.09820495677690055, 0.9017950432230994]
diet bru	[0.09986031241828819, 0.9001396875817118]
chip shop takeaway order	[0.11084386246459865, 0.8891561375354013]
diet coke ltr	[0.11520124734668147, 0.8847987526533185]
suey roll	[0.11605131791091994, 0.8839486820890801]
pattie	[0.12619217081624445, 0.8738078291837555]
hamburger	[0.1293061165498678, 0.8706938834501322]
chicken meat	[0.13293857088732897, 0.867061429112671]
dandelion	[0.13319255416754494, 0.8668074458324551]
chip roll	[0.13874956035593355, 0.8612504396440664]
bit	[0.14303300507885575, 0.8569669949211443]
pollo	[0.15360521749090306, 0.8463947825090969]
king rib	[0.1536955211457186, 0.8463044788542814]
shot	[0.1593261813458673, 0.8406738186541327]
burdock	[0.16458133161698252, 0.8354186683830175]
pie supper	[0.1776010926627819, 0.8223989073372181]

funghi	[0.19483683998675805, 0.805163160013242]
cheese tomato	[0.20707470207423562, 0.7929252979257644]
beef onion pie	[0.20907048034755127, 0.7909295196524487]
pasty	[0.22080400831990743, 0.7791959916800926]
haggis	[0.22384433767294343, 0.7761556623270566]
cheeseburger half pounder	[0.24157381600573946, 0.7584261839942605]
cheeseburger quarter pounder	[0.24157381600573946, 0.7584261839942605]
chicken breast supper	[0.2433056468951732, 0.7566943531048268]
fish chips	[0.2505011589087093, 0.7494988410912907]
supper	[0.2820484701165006, 0.7179515298834994]
cookies	[0.2926068844345514, 0.7073931155654486]
bull	[0.3199562641957926, 0.6800437358042074]
chicken nuggets meal	[0.32125571887427606, 0.6787442811257239]
chicken pakora	[0.32258630530369103, 0.677413694696309]
nuggets	[0.33817842687597777, 0.6618215731240222]
pasties	[0.3434913363840276, 0.6565086636159724]
chips salad	[0.3437544174739232, 0.6562455825260768]
scallops	[0.37133107972519297, 0.628668920274807]
pop	[0.37566114070951995, 0.62433885929048]
site	[0.3828808722957264, 0.6171191277042736]
bru	[0.3870878662375089, 0.6129121337624911]
inch bread	[0.3949447590422569, 0.6050552409577431]
inch margherita	[0.40518983399044395, 0.594810166009556]
kebab wrap	[0.4092167362183604, 0.5907832637816396]
spaghetti	[0.4148651966000986, 0.5851348033999014]
vegetable pakora	[0.4220166362053063, 0.5779833637946937]

pie chips	[0.42219252756608205, 0.577807472433918]
sausage supper	[0.43141384413907924, 0.5685861558609208]
cod roe	[0.4401889430287208, 0.5598110569712792]
cod fish	[0.4441861002797133, 0.5558138997202867]
rubicon mango	[0.4468756085262042, 0.5531243914737958]
spam fritter	[0.4510542222928562, 0.5489457777071438]
fish bites	[0.45304200897496705, 0.546957991025033]
spring roll	[0.454370446988012, 0.545629553011988]
piece	[0.4562983102715691, 0.5437016897284309]
facebook	[0.4754315709258996, 0.5245684290741004]
pakora	[0.47801416041298306, 0.5219858395870169]
mince	[0.47887974808592726, 0.5211202519140727]
kidney	[0.47997648893148603, 0.520023511068514]
inch meat feast	[0.4899211474645704, 0.5100788525354296]
pasta	[0.49921432925156173, 0.5007856707484383]

References

[1] Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1-23.

[2] Hafley, W. L., & Lewis, J. S. (1963). Analyzing messy data. *Industrial & Engineering Chemistry*, *55*(4), 37-39.

[3] Vasumita S Adarsh. (2013, December 26). TastyKhana launches Google map feature for website.(Internet). *The Economic Times*, p. The Economic Times, Dec 26, 2013.

[4] O'Connor, P. (2010). Managing a hotel's image on TripAdvisor. *Journal of Hospitality Marketing & Management*, *19*(7), 754-772.

[5] Hand, D. J. (2007). Principles of data mining. *Drug safety*, *30*(7), 621-622.

[6] Fish and chips. (2010). *Nutrition & Food Science*, *40*(6), 157-165.

[7] *Just Eat.* Online at https://www.just-eat.co.uk/ (referenced 19/08/2018).

[8] *JUST EAT WEBSITE TERMS AND CONDITIONS.* Online at https://www.just-eat.co.uk/termsandconditions (referenced 19/08/2018).

[9] Castillo, C. (2005, June). Effective web crawling. In *Acm sigir forum* (Vol. 39, No. 1, pp. 55-56). Acm.

[10] Goerzen, J. (2004). Web Client Access. In *Foundations of Python Network Programming* (pp. 113-126). Apress, Berkeley, CA.

[11] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current

[12] LI, W., & HUANG, Y. (2007). Web information extraction based on HtmlPaser [J]. *Ordnance Industry Automation*, *7*, 024.

[13] Lin, S., & Hu, Y. (2010, July). An approach of extracting web information based on htmlparser. In *Information Technology and Computer Science (ITCS), 2010 Second International Conference on* (pp. 284-287). IEEE.

[14] Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Communications of the ACM*, *11*(6), 419-422.

[15] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

[16] Bird, S., & Loper, E. (2004, July). NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions* (p. 31). Association for Computational Linguistics.

[17] Zhu, J. (1994). Conversion of Earth-centered Earth-fixed coordinates to geodetic coordinates. *IEEE Transactions on Aerospace and Electronic Systems*, *30*(3), 957-961.

[18] Clynch, J. R. (2006). Earth coordinates. *Electronic Documentation, February*.

[19] Montenbruck, O., Gill, E., & Terzibaschian, T. (2000). Note on the BIRD ACS Reference Frames.

[20] Barrett, P., Hunter, J., Miller, J. T., Hsu, J. C., & Greenfield, P. (2005, December). matplotlib--A Portable Python Plotting Package. In *Astronomical data analysis software and systems XIV* (Vol. 347, p. 91).

[21] Whitaker, J. (2011). The Matplotlib Basemap Toolkit User's Guide. *Matplotlib Basemap Toolkit documentation, February*.

[22] Tosi, S. (2009). *Matplotlib for Python developers*. Packt Publishing Ltd.

[23] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[24] Guo, & Koelsch. (2015). The effects of supervised learning on event-related potential correlates of music-syntactic processing. *Brain Research,1626*, 232-246.

[25] Jin, C., De-Lin, L., & Fen-Xiang, M. (2009, July). An improved ID3 decision tree algorithm. In *Computer Science & Education, 2009. ICCSE'09. 4th International Conference on*(pp. 127-130). IEEE.

[26] Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2).

[27] Umanol, M., Okamoto, H., Hatono, I., Tamura, H. I. R. O. Y. U. K. I., Kawachi, F., Umedzu, S., & Kinoshita, J. (1994, June). Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems. In *Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the Third IEEE Conference on* (pp. 2113-2118). IEEE.

[28] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

[29] Peng, W., Chen, J., & Zhou, H. (2009). An implementation of ID3-decision tree learning algorithm. *From web. arch. usyd. edu. au/wpeng/DecisionTree2. pdf Retrieved date: May*, *13*.

[30] Steinberg, D., & Colla, P. (2009). CART: classification and regression trees. *The top ten algorithms in data mining*, *9*, 179.

[31] Rutkowski, L., Pietruczuk, L., Duda, P., & Jaworski, M. (2013). Decision trees for mining data streams based on the McDiarmid's bound. *IEEE Transactions on Knowledge and Data Engineering*, *25*(6), 1272-1279.

[32] Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications*, *33*(1), 1-5.

[33] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3-14.

[34] Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1-2), 167-179.

[35] Koh, K., Kim, S. J., & Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, *8*(Jul), 1519-1555.

[36] Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning* (p. 78). ACM.

[37] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273-282.

[38] Goeman, J., Meijer, R., & Chaturvedi, N. (2012). L1 and L2 penalized regression models. *cran. r-project. or*.

[39] Dean, J., & Ghemawat, S. (2010). MapReduce: a flexible data processing tool. *Communications of the ACM*, *53*(1), 72-77.

[40] Xu, L., Liu, G., & Chen, Z. (2012, December). Research on optimization model of threshold setting for half-rate based on the decision tree algorithm. In *Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on*(pp. 316-320). IEEE.

[41] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

[42] *Decision Trees*. Online at http://scikit-learn.org/stable/modules/tree.html (referenced 19/08/2018).

[43] Wester, A., Øvrelid, L., Velldal, E., & Hammer, H. L. (2016). Threat detection in online discussions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 66-71).