# Chapter 1

# Introduction

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

## Dissertation Background:

In the era of rapid development of the Internet and smartphones, searching restaurants or specific food will not be a complicated stuff anymore. For example, with the Google Maps, the customers could find restaurant or food they are interested in with a simpler and more accurate way. At the same time, this kind of applications also brings benefits to restaurant operators, to better promote their business [4]. After discovering such a business opportunity, an increasing number of restaurant operators begin to advertise the link of the menu on some informational applications or websites, such as Google Maps or TripAdvisor [5]， producing a large amount data about the restaurant menu. That means we could explore more information from such a menu rather than only the name of a dish. Therefore, using data mining techniques [2] on menu datasets may obtain more meaningful information, which would be an indication of the regional differences in the UK.

## Research Aim and Research Focus:

This research aims to mining menu data from "Fish & Chips" shops to discovery the distinctions regarding the content expression among different regions, thereby revealing the regional differences in the UK. ‘Fish & Chips’ is one of the most famous food in the UK and there are more than 1,000 ‘Fish & Chip’ shops in this country [3]. In this project, we will use the websites provided by some of these ‘Fish & Chip’ shops, obtaining the raw HTML data and then focusing on employing data cleaning, mining, and visualisation techniques to find the content with regional features, which contribute to achieving the goal of the project. For example, 'Haggis' is a traditional food in Scotland and widely distributed, while rarely seen in England. According to the methodologies applied in this research, we could provide evidence that 'Haggis' is loved by the Scottish people and it is a regional dish in Scotland.

## Research Methods:

In terms of data crawling, the dissertation will illustrate the selection of data sources and methods for crawling data from ‘Fish & Chip’ 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, HTMLPaser and Natural Language Processing (NLP). Considering the data mining procedure of the research, the data visualisation technics to 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 national content) of the extracted data sets, the project employs machine learning methods, such as decision tree and regression classifier to generate the regionality result. Specifically, this research is an iterative process and includes four rounds of evaluation and improvement since the entire study is an exploratory process that there are no existing criteria to verify the rationality of the method selection and the correctness of the results. In fact, 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. Therefore, regional features and regional results are derived from the constant attempts, evaluations, and improvements during the project. In this case, further research may use or update each of the methods and steps in the previous exploration. Besides, each iteration will also evaluate the results to identify problems and propose improvements for the next iteration.

## Value of the Research:

The research links seemingly unrelated menu information to regional differences of the UK through exploring regional content from the messy menu dataset. In addition, some features of regional content in terms of the geographical distribution are discovered. Furthermore, 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.

## 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 are all related to methodologies, findings, evaluations, and improvements, which composes the four iterations of the project. Chapter 3 describes the first iteration, presenting the procedures of obtaining regional results of single independent word by one kind of feature. Chapter 4 is related to the second iteration, which describes the application of decision tree to get regional results of the single word. This chapter uses two types of decision tree algorithms and the makes comparison between the two algorithms. Chapter 5 covers the third round iteration, which introduces the logistics 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.

**Introduction references**

[1] Fuller, Michael. (2015). Big Data: New science, new challenges, new dialogical opportunities. *Fuller , M 2015 , ' Big Data: New Science, New Challenges, New Dialogical Opportunities ' Zygon: Journal of Religion and Science , Vol 50 , No. 3 , Pp. 568-582 . DOI: 10.1111/zygo.12187,* Fuller , M 2015 , ' Big Data: New science, new challenges, new dialogical opportunities ' Zygon: Journal of Religion and Science , vol 50 , no. 3 , pp. 568-582 . DOI: 10.1111/zygo.12187.

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

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

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

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

# Chapter 2

# Background

## Web crawling

Before crawling data from websites of ‘Fish & Chips’ shops, the dissertation compared the data source which includes food delivery websites such as Just-Eat [3] and independent ‘Fish & Chips’ shops’ websites. The advantages of using food delivery websites as data source is that it is convenient to search "Fish & Chips" shops in each city of the UK by postcode. In addition, each shop which is searched out is available to crawl data directly. Unlike when using independent websites that many links are not available, or the content of the website is PDF which cannot be crawled. Thus, the dissertation originally planned to use the food delivery website as the data source. However, sites like just-eat have a clear description that crawling data is not allowed. Therefore, using the food delivery website to crawl data directly may against legal, and this fatal flaw directly leads to the unavailability of this method. As a consequence, the dissertation planes to find independent websites of "Fish & Chips" shops from the Google Maps and uses them as the data source to crawl data. Web crawling is the process which collects information from target websites [1]. The dissertation uses Python module urllib2 which can be used to simulate browser behaviour to download web pages and handle request errors [2] to get the full website HTML source code of ‘Fish & Chips’ shops.

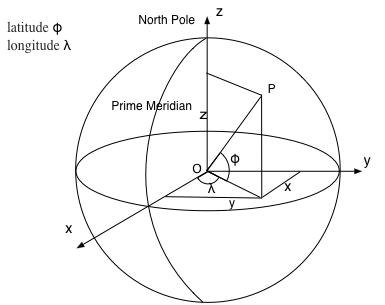
## HTML data cleaning technics:

Data cleaning is used for improving the quality of data which is used as data for subsequent data processing through detecting inconsistencies and removing errors [4]. 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, noun phrases and 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 costumers 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 classifying words that represent the same meaning into the same category (mainly focuses on the identification and classification of singular and plural nouns with the same meaning). Fortunately, the regular expression, HTMLPaser and NLP can help the project to achieve the data cleaning goal. HTMLParser is an open source, fast and robust HTML parsing tool for extracting and cleaning content of HTML [5, 6]. It can customize HTML tag content extraction based on user requirements [7]. In this project, in this project, the HTMLParser mainly plays the role of data extraction and filtering. Since the data source used in the project is an independent website, the HTML structure of most websites is different (small parts of the website structure are the same because they are developed by the same company). Therefore, HTMLParser plays a huge effect that it does not pay attention to the structure of the website design, only pay attention to the name of the HTML tag, such as ‘div’ and ‘script’. 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 are different, 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 [8] to filter the result of the HTMLParser. 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 a research about using computer to understand and manipulate natural text or speech to process tasks [9]. 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 [10, 11]. 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 to singular forms.

## Geographic data visualisation methods:

Because one of the target of the project is to explore geographically distributed features to represent regionality, the project uses Cartesian coordinate systems for geolocation calculations and Matplotlib for data visualisation. The core calculation in this project is the coordinates of the central point which is the centre of all shops which contain a specific content. The importance of the central point is that all other features of the project are derived from it. The project uses a set of coordinates containing this content to calculate the point. The algorithm used in this project for calculating the center point of a series of coordinates through combining geographical coordinate system with Cartesian coordinate systems which regards the Earth as a sphere (Fig.1). This combination is also known as ECEF ("earth-centered, earth-fixed") [12]. In Cartesian coordinates, earth is a sphere centered at the origin [13]. The z axis points to the north pole. The x, y axis 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 [14]. 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. 1 geographical coordinate system with a cartesian coordinate systems**

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

= ,

= ,

= ,

The centroid of these points is the average of the sum of :

() = ,

The coordinate of the centroid can be expressed as:

= ,

=

After the project gets the coordinates of central point and other features such as radius, 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 [15]. 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 [16]. Basemap provides a possibility that the project can draw Matplotlib plot over the real word map [17]. 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.

## Data mining and analysis with machine learning methods:

1 Decision tree: ID3 algorithm, Cart algorithm.

1. Logistic regression

Because this project is an exploratory project, 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 [18]. In the early stage of the project, the discovery of features is based on observing the visualization results of the geographic information. After the project finds some features, the project tries two machine learning methods to judge the regional content. One of them is the decision tree and other is logistics regression.

## 1.4.1 Decision Tree

Decision tree is mainly used for classification and prediction of models [19] and the project uses 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 use training dataset to create the tree and then use the tree to classify the test dataset [20].

ID3 algorithm constructs decision tree by selecting most useful features. 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 [21] 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:

Information Gain: a represents a feature.

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

1. Loading training dataset. The training dataset has 42 sample data. In this dataset, regional words are marked as 1 and widely distributed words and words which have few shops are marked as 0.
2. Calculating the Entropy.
3. Data segmentation based on optimal segmentation feature.
4. Selecting the best segmentation feature based on the maximum information gain.
5. Recursively building a decision tree.
6. Sample classification.

Considering the limitation of the ID3 algorithm that the ID3 algorithm can only deal with discrete values [22] so that the feature values must be classified based on numerical variables. Thus, the project should first observe the feature value to find the criteria and then mark each training data according to this criteria. This will cause the division point not to be the optimal choice. As a result, the project planes to use an algorithm to find the optimal divide point in the training dataset to generate the decision tree and compare the results. Fortunately, in Python, the Sklearn package provides method to generate the decision tree, including classification tree and regression tree. In this project, the classification tree is more suitable for the project, because the target of the decision tree is binary. The algorithm used in Sklearn package is an optimised version of the CART algorithm [23].

Cart algorithm uses binary recursive partitioning procedure to split datasets [24]. In classification tree, Cart algorithm uses Gini index as a property to determine partitioning [25]. 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 [26]. The following formulas shows 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 , then the Gini index is defined as:

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

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:

Cart algorithm implementation steps:

1. Using each feature A in the sample D and each possible value of A (A>=a and A<a) to divide the sample into two parts and calculate the Gini (D, A).
2. 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.
3. Recursive call (1) (2)

## 1.4.2 Logistic Regression

The project studies the classification problem, so the dependent variable of the model is classification variable (0 and 1) and the independent and dependent variables of the model are nonlinear. As a result, logistic regression model is more suitable for this project and the project selected the logistic regression model of the sklearn package as the classifier.

Logistic regression is well suited to describe the relationship which is expressed as probability between classification results and one or more classifications [27]. 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 [28]. The following model is the model of logistics regression:

denotes the vector of feature variables, and denotes the associated binary output. represents the weight vector. The logistic regression has model:

Logistic loss function:

Average logistic loss:

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 [29]. Thus, when there are many features, overfitting will become a problem of the model unless the training set is ample [30]. In order to solve this problem, L1 and L2 regularization were used.

L1:

Lasso (L1) penalty encourages the sum of the absolute values of the to be small [30]. It uses sparsity to fit model with many features [31]. 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 [32].

L2:

L2 penalty encourages the sum of the squares of the to be small [30]. It will reduce the regression coefficient but will not be zero [32]. Thus, if each feature has an effect on the classification, L2 penalty is more suitable.

**Background Reference**

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

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

[3] https://www.just-eat.co.uk

[4] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, *23*(4), 3-13.

[5] Oswald, D., Raha, S., Macfarlane, I., & Walters, D. (2006). HTMLParser.

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

[7] 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.

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

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

[10] 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.

[11] Madnani, N. (2007). Getting started on natural language processing with Python. *Crossroads*, *13*(4), 5-5.

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

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

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

[15] 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).

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

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

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

[19] 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.

[20] 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).

[21] 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*.

[22] 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.

[23] <http://scikit-learn.org/stable/modules/tree.html>

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

# Chapter 3

# Iteration 1

8 to 10 pages

Statement: aim to know independent words distribution through map and according to the map and ratio (the number of shops whose distance is less than 20000 meters from the center point/ total shop number) to find features of regional words.

* 1. **Methodology:**

1. Data obtain: Decomposing the content in HTML into independent words.
2. Data cleaning: Cleaning independent words (Special symbol filtering, uppercase conversion to lowercase, using NLP method to analyse part of speech to complete noun singular and plural combination).
3. Data visualisation:

* Mark points on the map, calculate central point of all shops, outlier points identify and filter, draw radius.
* Draw the ratio trend.
  1. **Findings:**

Find some regional words based on ratio and map.

**3.3 Evaluation:**

It is imperfect to rely solely on the ratio feature, and it requires more features. Give examples.

* 1. **Improvement:**

1. Observing the data set and according to the coordinates information to find more features, such as ‘city number’, ‘proportion’, ‘average distance’, ‘shop numbers’.
2. Decide to use decision tree to classify the words.

# Chapter 4

# Iteration 2

Around 8 pages

Statement: aim to use decision tree to classify the independent words and compare result of two decision tree algorithms

* 1. **Methodology:**

1. Generate different training sets that match the two algorithms based on the observations in iteration 1.
2. Generate two kinds of trees and visualise two trees.
   1. **Findings:**
3. Which words are judged as regional words.

(2) Differences between two algorithms.

**4.3 Evaluation:**

(1) Based on experiences.

(2) Difficulties of evaluation

* 1. **Improvement:**

1. Find HTML context of the words and analyse why these words are judged as regional words.
2. Can consider noun phrases and word pair as analyse target.
3. Can consider to use classification method in sklearn package.

# Chapter 5

# Iteration 3

Around 8 pages

Statement: aim to use logistic regression classifier to classify independent single word.

* 1. **Methodology:**

1. Use logistic regression model to classify independent single word.
   1. **Findings:**
2. Independent words findings.
3. In logistic regression, mainly describe the impact of features and the selection of features.
   1. **Evaluation:**
4. Independent words use context, compare the result between logistic classifier (L1) and logistic classifier (L2).
   1. **Improvement**

# Chapter 6

# Iteration 4

1 or 2 pages

Statement: aim to use noun phrases and word pairs as dataset to find regionality information in the menu.

# Chapter 7

# Conclusion

1 or 2 pages

# Chapter 8

# Future Work