# Chapter 1

# Introduction

This chapter provides a general description of the background, aim, methods, values, and structure of the dissertation.

## Dissertation Background:

In the era of rapid development of the Internet and smart phones, searching restaurants or specific food has become simpler and more accurate. For example, Google Maps not only provides customers with convenience to find the restaurants which have the food they want to eat but also allow restaurant operators to better promote their business [4]. As a consequence, an increasingly number of restaurant operators public the link of their menu on some of the informational intermediators such as Google Maps or TripAdvisor [5] for advertising. As a consequence, the volume of menu data of restaurants’ websites becomes huge, and the information contained in these menu data maybe more than just the name of the dish. Thus, using data mining techniques which aims to discover potential and unexpected information in the large datasets [2] on menu datasets may obtain valuable information. In this dissertation, this valuable information represents regional differences of the UK.

For example, users of Google Maps can search more than 500 ' Fish & Chips' shops in the UK on the Google Maps.

In the era of big data, the information on the menus of these restaurants is more than the name of the dish but the culture difference of the UK. This dissertation focuses on mining menu data of "Fish & Chips" shops to find the regional usage of the dish expression to demonstrate regional differences of the UK.

A big data set is likely to be comprised of data relating to a large number of parameters, collected with little “filtering” or standardization regarding its content or format[1].

## Research Aim and Research Focus:

The aim of this dissertation is mining menu data from "Fish & Chips" shops to find the regional usage of the dish expression to demonstrate regional differences of 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 the UK [3]. The dissertation will use websites provided by part of these ‘Fish & Chip’ shops which provide menu links to get the raw HTML data and then focuses on data cleaning, mining and visualisation technics to find the content with regional characters. This content can help the project find regional differences. For example, 'Haggis' is widely distributed in Scotland, and there are very few in England, so the project can decide 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 are used for exploring regionality from the website HTML content, such as single independent words, noun phrases and word pairs. The method 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 dissertation, the project uses data visualisation technics 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 dissertation uses machine learning methods, such as decision tree and regression classifier to generate the regionality result. Specifically, this research is an iterative process and this dissertation carried out 4 rounds of evaluation and improvement. This is because the entire study is an exploratory process that there is no existing standard to verify the rationality of the method selection and the correctness of the results. For example, the dissertation did not know which features can be used for reflecting regionality of the text and the evaluation of regionality content is also based on the evaluator’s experiences. Thus, regional features and regional results are derived from the constant attempts, evaluations and improvements of the project. This means in each round, the research may use or update each of the methods and steps mentioned above. Besides, each iteration will evaluate the results to identify problems and propose improvements for the next iteration.

## Value of the Research:

The dissertation links seemingly unrelated menu information to regional differences of the UK through exploring regional content from the messy menu dataset. In addition, the dissertation discovered some features of regional content in terms of the geographical distribution. Furthermore, the methods and algorithms used in this paper 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 Background, which mainly illustrates the main technics and algorithms used in the paper. Chapter 3, Chapter 4, Chapter 5 and Chapter 6 all contains mythologies, findings, evaluation and improvements. These chapters are four iterations of the dissertation. Chapter 3 Iteration 1, which mainly presents the procedure of obtaining regional results of single independent word by one kind of feature. Chapter 4 Iteration2, which mainly describes using decision tree to get regional results of the single word. This chapter used two kinds of decision tree algorithms and the dissertation compared these two algorithms. Chapter 5 Iteration 3, which introduces the logistics regression to obtain the possibility 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 of the probability which means deciding how many probabilities exceeds the threshold will be judged as regional word. Chapter 6 Iteration 4, which demonstrates the results of using other two kinds of datasets (noun phrases and word pairs). Chapter 7 Conclusion includes the summary of the dissertation. Chapter 8 Future Work, which covers where dissertation can be improved, limitations and recommendations. Chapter 9 References. Chapter 10 Appendix.

Introduces

Illustrates

Presents

Describes

Demonstrates

Covers

Includes

Contains

[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

8 pages about

## Web crawling

## HTML data cleaning technics:

1. HTMLPaser.
2. NLP.

## Map data visualisation methods:

1 BaseMap.

2 Cartesian algorithm to calculate the central point according to geographic coordinates.

## HTML data analysis with machine learning methods:

1 Decision tree: ID3 algorithm, Cart algorithm.

1. Logistic regression

# 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 noun phrases and word pairs as dataset to find regionality information in the menu. Besides, using logistic regression classifier to classify.

* 1. **Methodology:**

1. Data obtain: Decomposing the content in HTML into noun phrases and word pairs.
2. Data cleaning: Cleaning independent words (Special symbol filtering, uppercase conversion to lowercase).
3. Generate training dataset for noun phrases and word pairs.
4. Use Cart algorithm to generate decision tree and use logistic regression model to classify.
   1. **Findings:**
5. Independent words findings.
6. Noun phrases findings.
7. Word pairs findings.
8. In logistic regression, mainly describe the impact of features and the selection of features.
   1. **Evaluation:**
9. Independent words use context, compare the result between logistic classifier and the decision tree.
10. Noun phrases and word-pair can evaluate directly and compare the result between logistic classifier and the decision tree.
    1. **Improvement**

# Chapter 6

# Conclusion

1 or 2 pages

# Chapter 7

# Future work

1 or 2 pages