# Chapter 1

# Introduction

2 or 3 pages about

## Background:

Big Data, Fish & Chips, Data analysis, Data visualisation, Regional differences.

## Research focus:

Data crawling, Data cleaning, Data analysis, Data visualisation, Find regional differences.

## Overall Research Aim and Individual Research Objectives:

Overall aim is to find regionality from menu.

Specifically, this project is an iterative process:

1. Function selection.
2. Identify and find regional menu information.
3. Analyse the result.
4. Evaluate the result.

## Outline Research Methods and Timescales:

1. Data Source Selection
2. Obtain menu data
3. Clean HTML data
4. Iterative analysing data

## Value of the Research:

Illustrating regional differences in the UK through regional information of messy menu data.

## 1.6 Structure of the Dissertation

# 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 single 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 single words.
2. Data cleaning: Cleaning single 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 single 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 single 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. Single 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. Single 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