**Decision Tree:**

Decision tree is mainly used for classification and prediction of models [2] and the project uses decision tree to classify regional words and widely distributed words. The project uses ID3 algorithm to create the decision tree that the tree is created by the training dataset and use the tree to classify the test dataset [1]. ID3 algorithm constructs decision tree by selecting most useful features. These features can make the classification of data set more effective. Thus, the project needs an algorithm to measure the suitability of features and select features. The Entropy can measure the impurity of training dataset [3] 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:

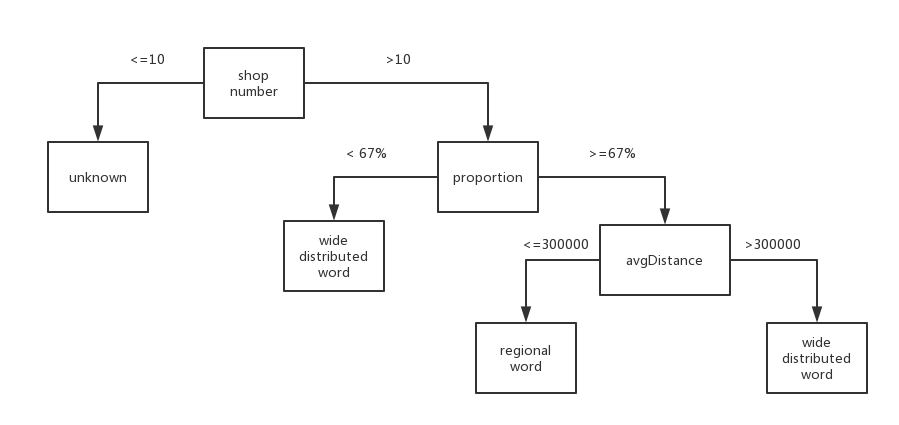
Information Gain: a represents a feature.

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

1. Loading training dataset. The training dataset has 39 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.

**Result**

The project generated the following decision tree based on the training data set and the information gain.



**Figure 1: Decision tree**

The number of regional words: 26.

The number of widely distributed words: 710.

The number of unknown words: 4552.

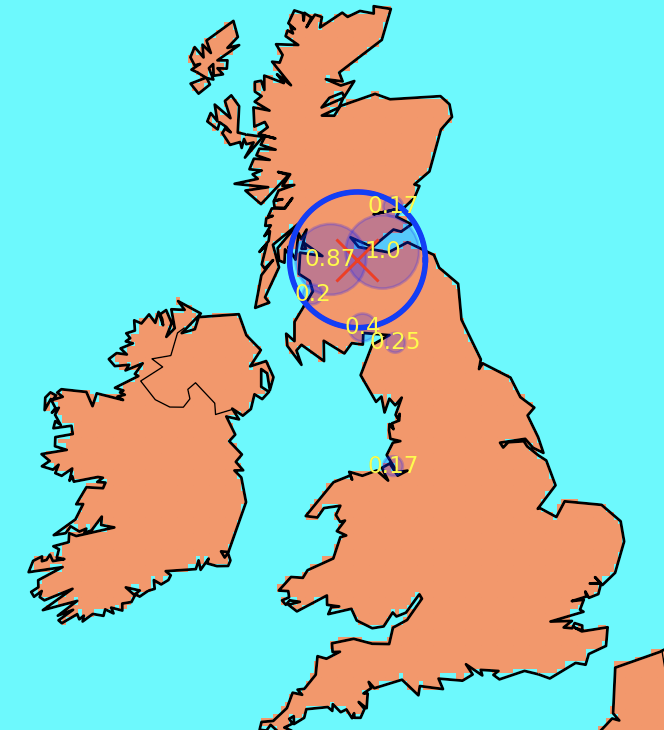
**Evaluation**

Evaluation based on experience. In these 26 regional words, there are 18 words are judged correctly and 8 words (securely, skate, instantly, rock, shot, hamburger, haagen, bull) should be widely distributed words. In these 8 words, ‘securely’, ‘skate’, ‘instantly’, ‘roc’, ‘haagen’, ‘bull’ appear at a frequency of 15 or less. The project believes that the low frequency of occurrence leads to misjudgement of these words. However, ‘shot’ and ‘hamburger’ appear more than 22 shops. Interestingly, ‘hamburger’ is widely distributed in Scotland.

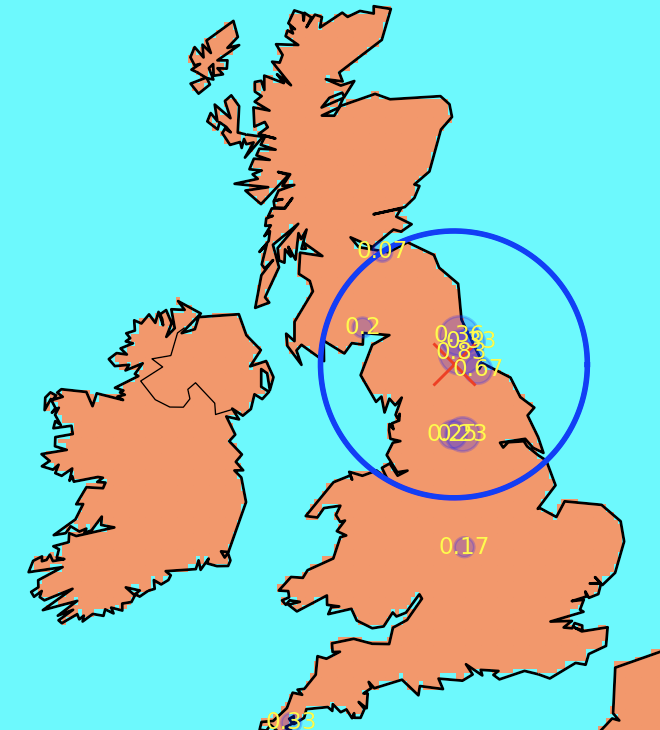
Evaluation is based on the context in which the words appear. In terms of words which are judged by decision tree as regional words, the project found out the context in which they appear in the web page. The content of the context is sentences which has the regional words. The project has generated a table which contains words, context (sentences) and the ids of the shop containing the words and the project will according to this table to find the reason that a word is judged as a regional word.

After analysing the result of words context, the project has the following findings:

1. ‘haggis’, ‘irn bru’, ‘kiev’, ‘inferno’, ‘crunch’, ‘skate’, ‘bolognese’, ‘macaroni’, ‘naan’, ‘hamburger’, ‘plaice’, ‘rib’, ‘kidney’, ‘spaghetti’, ‘carbonara’, ‘pasti’, ‘roe’, ‘balti’ represent a dish in the menu. Thus, maybe they are local dishes.
2. ‘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.
3. In terms of ‘yorkshire’, most of ‘yorkshire’ represent a place named ‘yorkshire’. Therefore, ‘yorkshire’ is a regional word that represents a place name.
4. All ‘instantly’ are used in this sentence (‘chip shop takeaway - order online instantly!’) and the websites with this sentence has the same design style. Thus, the reason why ‘instantly’ is judged as a regional vocabulary is the same as (2).
5. ‘rock’ always used with ‘eel’. ‘rock eel’ represents a kind of fish. Thus, the reason why the ‘rock’ is regionally distributed is because it represents a dish when used with ‘eel’.
6. ‘shot’ always used with ‘hot’ and ‘hot shot’ represents a kind of dishes. Thus, the reason why the ‘shot’ is regionally distributed is same as (5).
7. The project speculates that ‘haagen’ presents a regional distribution because in that area, Haagen-Dazs has more trade links with the merchants in that area.
8. ‘bull’ always used with ‘red’ and ‘red bull’ is a drink. The reason why ‘bull’ is regionally distributed maybe same as (7).



**Figure 2: The distribution of ‘hamburger’**

****

**Figure 3: The distribution of ‘shot’**

**References**

1. 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).
2. 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.
3. 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*.