**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:

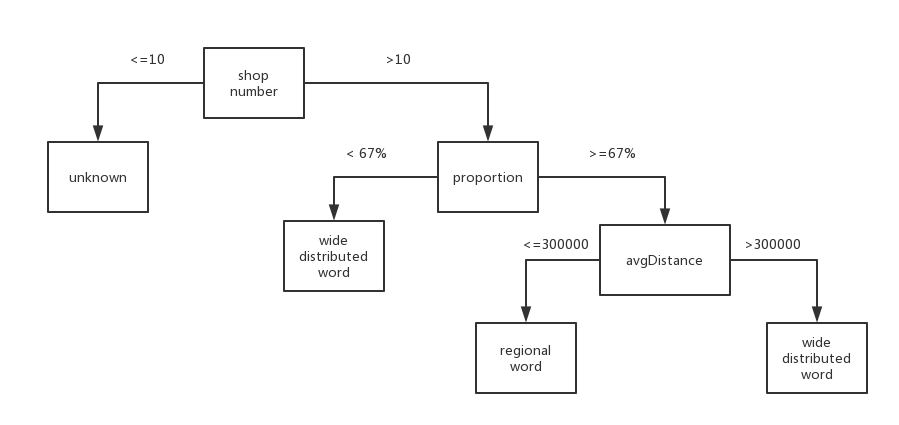
Information Gain: a represents a feature.

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

1. Loading training dataset. The training dataset has 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.

**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**

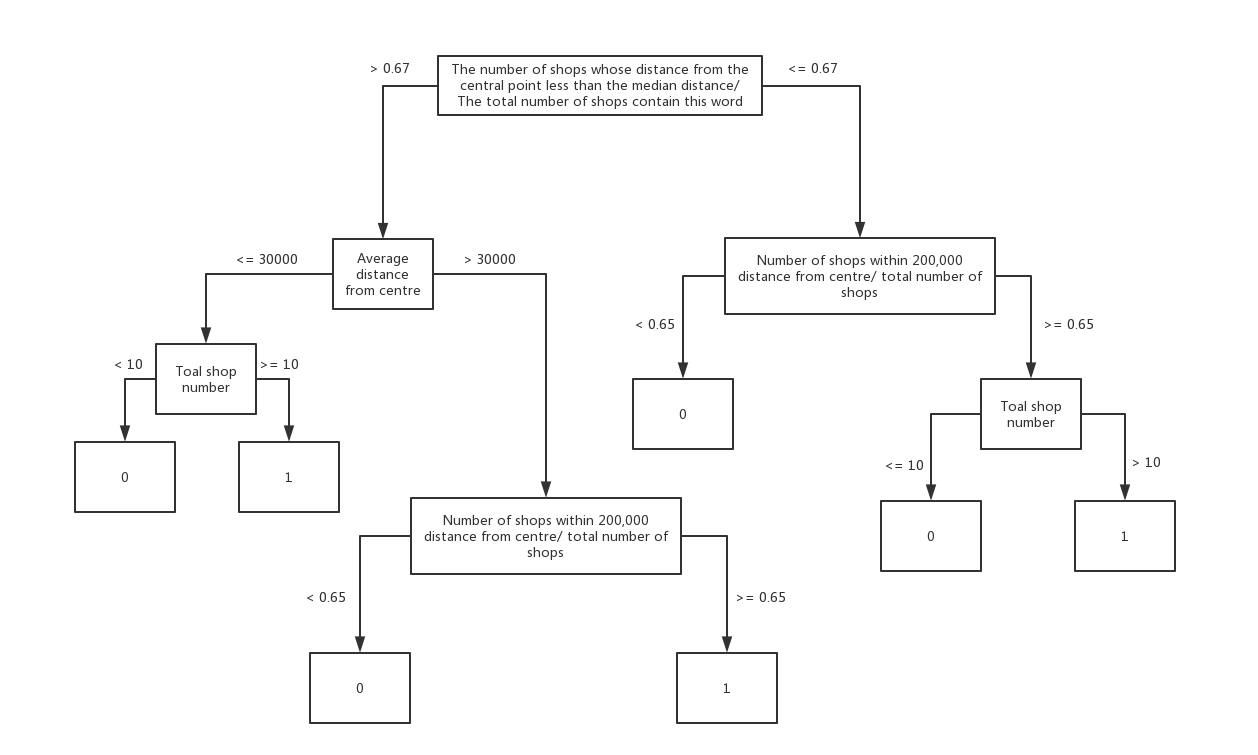
Evaluation is based on the context in which the words appear. In terms of words which are judged by decision tree as regional words, the project found out the context in which they appear in the web page. The content of the context is sentences which has the regional words. The project has generated 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).

**Add ratio feature into the decision tree**

The project has tried another method ratio which means the number of shops whose distance is less than 20000 meters from the center point/ total shop number to judge regional words. Thus, the project added ratio feature into the decision tree and according to the best feature selection algorithm of decision tree, the project generated the following decision tree (Fig. 2).



**Figure 2: Decision tree (added ratio feature)**

**Result**

There are 54 words are judged as regional words. Among them, the following words are newly appearing as regional words.

(meaty, give, cucumber, cob, smokey, guava, pakora, instantly, pukka, savoury, pattie, macaroni, burdock, parmesan, splash, dandelion, scallop, cornish, bit, stagioni

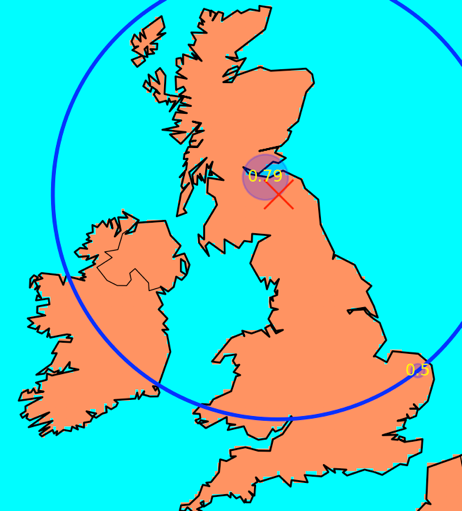
passion, facebook, keema, chosen, samosa, suey, rump, value.

**Evaluation**

1. ‘meaty’, ‘cucumber’, ‘cob’, ‘guava’, ‘pakora’, ‘pukka’, ‘savoury’, ‘pattie’, ‘burdock’, ‘parmesan’, ‘splash’, ‘dandelion’, ‘scallop’, ‘keema’, ‘samosa’, ‘sury’, ‘rump’
2. ‘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.
3. ‘smokey’ is used as an adjective, usually in conjunction with a ‘bbq’ or ‘sausage’.
4. ‘macaroni’’ is always used with ‘cheese’ and ‘macaroni cheese’ is a dish.
5. ‘cornish’ is always used with ‘pasty’ and ‘cornish pasty’ is a dish.
6. ‘bit’ is usually used as a degree adverb.
7. ‘quattro’ is always used with ‘stagioni’ and ‘quattro stagioni’ is a kind of pizza.
8. ‘passion’ represents fruit or dish. Maybe in that area this kind of fruit is famous or selling well.
9. ‘facebook’: The websites have a Facebook account, and it may be a coincidence that this word shows regionality.
10. ‘chosen’ is always used with ‘flavour’. All sites that use this usage have the same style in that area and these websites were developed by the same company. As a consequence, ‘chosen’ appears regionally.
11. ‘value’ is always used with ‘box’ or ‘meal’ which represent dishes.

Compare two decision trees:

Words such as ‘rump’ (Fig. 3) will be excluded if the project according to average distance to judge in the first tree.



**Figure 3: ‘rump’ distribution**

The proportion (The number of shops whose distance from the central point less than the median distance/ The total number of shops contain this word) of words such as ‘burdock’, ‘guava’ and ‘smokey’ is lower than 67. Thus, in the first tree, these words will be excluded. However, these words have have regional distribution characteristics.

As a consequence, adding ratio (Number of shops within 200,000 meters from centre/ The total number of shops) feature into the decision tree will make the result more accurate.

**Sklearn package**

Considering the limitation of the ID3 algorithm that the ID3 algorithm can only deal with discrete values [4] so that the feature values must be classified based on numerical variables. For example, in Fig. 1 and Fig. 2, the average distance feature was divided into two categories (>300000 meters and <300000 meters) and the average distance of each training data were marked into these two categories. Thus, the project should first observe the feature value to find the criteria and then mark each training data according to this criteria. As a result, the project planned to use continuous value in the training dataset directly 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. The algorithm used in Sklearn package is an optimised version of the CART algorithm [5].

**Cart algorithm (classification tree)**

Cart algorithm uses binary recursive partitioning procedure to split data sets [6]. In classification tree, Cart algorithm uses Gini index as a property to determine partitioning [7]. 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 [8]. 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:

**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*.
4. 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.
5. <http://scikit-learn.org/stable/modules/tree.html>
6. Steinberg, D., & Colla, P. (2009). CART: classification and regression trees. *The top ten algorithms in data mining*, *9*, 179.
7. 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.
8. 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.