**Logistic regression:**

Penalty: L1

C=1.0 Inverse of regularization strength.

Classification rate (The degree of fit between the model and the training set): 89.74%.

Noun phrases with a percentage greater than 50% are regional noun phrases.

Number of regional noun phrases: 65.

**Logistic regression:**

Penalty: L2

C=1.0 Inverse of regularization strength.

Classification rate: 89.74%.

Noun phrases with a percentage greater than 50% are regional noun phrases.

Number of regional noun phrases: 67 (‘inch meat feast’, ‘rubicon mango’).

**L2 logistic (Multinomial) regression:**

Penalty: L2

C=1.0 Inverse of regularization strength.

Classification rate: 61.53%.

This situation cannot be indicated that a phrase is regional, because the regional probability of all phrases is around 50%.

**Linear SVC:**

C=1.0 Inverse of regularization strength.

Classification rate: 79.48 %.

This situation cannot be indicated that a phrase which is larger than a certain probability is regional as many regional phrases such as ‘bru’, ‘mince’ and ‘vegetable pakora’ are determined as non-regional. At the same time, many non-regional phrases such as ‘email sms’, ‘half’ and ‘everything’ have more than 50% probability. As a consequence, regional phrases cannot be correctly classified by using Linear SVC.

**Conclusion**

By comparing four classifiers, logistic regression is the most suitable classifier for this project. Furthermore, the project will use L2 as the penalty parameter.

**Logistic regression:**

Logistic regression is well suited to describe the relationship which is expressed as probability between classification results and one or more classifications [5]. 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 [4].

Model:

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:

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 [6]. Thus, when there are many features, overfitting will become a problem of the model unless the training set is ample [3]. 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 [3]. It uses sparsity to fit model with many features [1]. 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 [2].

L2:

L2 penalty encourages the sum of the squares of the to be small [3]. It will reduce the regression coefficient but will not be zero [2]. Thus, if each feature has an effect on the classification, L2 penalty is more suitable.

After comparing the classification results of noun phrase, the project found that there is not much difference between using L1 penalty and L2 penalty. This is because that the project selected just three main features (‘proportion’, ‘ratio’ and ‘average distance’). The reason for choosing these three features instead of the five in the decision tree is that from the results of the decision tree, the impact of ‘city number’ feature on the classification results is negligible. Besides, the project found that if the project adds ‘city number’ and ‘shop number’ features, the model will overfit the training dataset that noun phrases with less than 10 shops will not be excluded. This is because in the logistic model, ‘shop number’ feature has the smaller weight than other features. Thus, some words that should not participate in classification are given regionality with high probability. In order to solve the above problems, the project ignored ‘city number’ feature and manually excluded words with less than 10 in the test dataset. In terms of the selection of penalty, the project decided to use ‘L2’ penalty, because the number of feature is few and the remaining three features are decisive.

**Result**

Single word

Noun phrase

Word-pair

**References**

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.