# COMP9318

# Data WareHouse and Data Mining

# project report

Z3492986
Guohao Yi

## Introduction

This report describes the strategy used to train a classifier to perform Name Entity Recognition (NER) given the appropriate training data, from feature extractions to performance improvement by using the logistic regression method in python 3. This classifier is specific for classifying if a given word is a title.

## Feature analysis

A title is given to describe the position of certain characters in an organization, which is usually given to show what duties and responsibilities that those people need to take. The followings are the features I have found for determining if a given word is a title.

**Token**: This is the word itself. If this word is found to occur as a title, it has very high chance that it will be a title again. If this is the only  feature, it will introduce the problem of overfitting.

**POS**: This is the part of speech tag given as input. If the title is only a word, the part of speech feature will takes an important role in evaluating whether it's a title or not, so this should be included in the feature tags.

**Next**: This is the next word of the given word. For title with composite words, considering the next word and previous word is a reasonable approach to summarize the pattern of a title occurring in some text material,

**Prev**: Same reason as Next feature

**Posfix2**: the last 2 characters of the given word. A considerable number of word that ends with "er' or 'or' has a very high chance that it's a title.

**Posfix3**: the last3 characters of the given word. Though it has a similar reason as Posfix2 feature, it is included because it might be useful for filtering some ot the words that are not title but end with simiar pattern as Posfix2 like word such as 'minor'

**Len**: length of the word. This is included to eliminate some  very short or very long word, which normally related to if it's a title or not

**Isnoun**:Indicate if this is a noun. This is similar to POS but might increase the weight of POS because it's similar to POS

**Isconjuntset**:Indicate if this is in a subset that is very unlikely be the POS of the title. This is initially included for filtering some conjunction word. Actually it filter more than conjunction word because some more type of POS is included in the filter list.

**Isupper**: Indicate if all the characters are uppercase, which is used to improve the chance that some of the titles such as "CEO" would be included

**Commonabrev**: Indicate if the word is in a list of common abbreviation of titles

# Improvement of Performance

At the very beginning of doing this project, only token is included in the feature set, but as it will have problem of overfitting and might not correctly reflect the properties of title,  so extra features as the above session is added to improve the performance and reduce overfitting. Features are considered in a way that try building the model with added features and observe if the outputs are improved or not. Then WordNetlemmatizer is used to change nouns from plural to singular to enhance data extraction accuracy.Finally after considering proper properties that play important roles in reflecting the features of words that are title, ten-fold cross-validation is used to further reduce the overfitting problem. For each feature above, one of each is excluded every time and maybe none of them are excluded, and logistic model is built with the remaining features. Training data is divided into many groups and large part the training data is used to train each model and the remaining data participate in the testing and the score is added together get each model has its total score. The model that gets the highest score would be actual one that used to extract the features and train the actual classifier. K-folds cross-validation is an effective method that decrease overfitting and may improve performance.

# Conclusion

Name Entity Recognition is a very important topic in data mining and the classifier implemented in this project is satisfying but not very perfect as the limitation of background knowledges and experience in title properties.