**IFT3335 TP2 Report**

**1. Experimental design**

To deal with the word meaning disambiguation problem, we aim to classify a word used in a context into the appropriate meaning class. The implementation process mainly includes data preprocessing, analysis of multiple classification algorithms, parameter sensitivity analysis, and draws some experimental conclusions and prospects for future optimization directions.

**2. Data preprocessing**

Data preprocessing mainly includes four parts: data denoising, feature extraction, feature selection and data set division.

**2.2 Feature Extraction**

Feature extraction includes two ways: feature extraction using grammatical part-of-speech granularity (gc) and feature extraction using word granularity (nw). The part-of-speech extraction method is to obtain the part-of-speech of each n adjacent words before and after the predicted word, such as VB VBG NNS IN, and the stem extraction method is to obtain the stems of each n adjacent words before and after the predicted word. Perform stem restoration, such as working->work.

**2.3 Feature Selection**

Since a word can appear several times in the text, and its importance may vary according to its frequency of occurrence. Therefore, consider extracting features from part-of-speech or stemming words by using CountVectorizer and TfidfVectorizer for feature selection, and using "frequency" to measure the importance of a word in the text. Among them, CountVectorizer uses the frequency of each word or part of speech in the entire training corpus as a feature, and TfidfVectorizer uses the word frequency-inverse text frequency of each word or part of speech in the entire corpus as its feature.

**2.4 Dataset division**

This dataset includes 2369 sentences, 6 semantics about "interest" or "interests". Detailed data can be found at: <http://www.d.umn.edu/~tpederse/data.html>, using the training set to train the model.

**3. Performance Analysis and Comparison of Classification Algorithms**

**3.1 decision\_tree:**

**3.2 random forest:**

**3.3 svm:**

**3.4 Naive Bayes:**

**3.5 mlp:**

**4. Parameter sensitivity analysis**

**4.1 Analysis of removing stop words**

Removing stop words includes removing punctuation marks (such as "/", ":", "\n", etc.), parts of speech ('/IN', '/DT', '/CC') and meaningless vocabulary (tool word list(stoplist).txt) has three parts, no stop words are removed, only punctuation marks and parts of speech are removed. Through experiments, if the words contained in tool word list(stoplist).txt are removed, the effect of all classifiers will be reduced. We think it is because tool word list(stoplist).txt contains a lot of words and names, Words related to the user's emotional color, whether to delete stop words depends largely on the tasks we are performing and the goals we want to achieve, and we are training a model that can recognize the meaning of "interest", and keep stop words make the classification effect better.

**(1) Remove stop words:**

**\\\**

**(2) Do not stop words:**

**\\\**

**4.2 Analysis of Feature Selection Method**

**TfidfVectorizer and CountVectorizer.**

CountVectorizer only considers the frequency of vocabulary in the text, which belongs to the bag of words model feature, while TfidfVectorizer not only considers the frequency of a certain vocabulary in the text, but also pays attention to the number of all texts containing this vocabulary. It can reduce the impact of high-frequency meaningless words and mine more meaningful features.

The experimental training parameters and the effects of various classifiers are as follows:

**(1) tf-idf**

**(2)count**

**4.3 Window size**

**\\\**

**4.4 Classifier parameters**

**4.4.1 Number of Network Layers**

**4.4.2 Optimizer**

**4.4.3 Activation function**

**5. Conclusions and expectations**

**\\\**