Information Retrieval Homework, Report III

R08725002 資管碩二 黃文鴻

1. Environment

```
VSCode
```

2. Programming Language

```
Python 3.8
```

3. Usage

```
pip3 install nltk
python3 ./pa3.py
```

4. Introduction

`In chi_square():` , chi square function for features selection

`In tokenize_data():` , tokenize document with porter stemmer and remove punctuation exclude hyphens

`In read_training_data(document_token):` , use tokenized dataset to generate training data and training label, training data is group by document id

`In MultipleNBClassifier():`

`read_model()` , get pre-trained model in previous training

`fit(x, y, k_features = float('nan'), method = 'chi_square', save = False):` , training model by data and label, which has following options:

- k_features: select only k features(token) for predict
- method: methods for features selection

- save: save pre-trained model or not

in this function, I implement pseudo-code below: ``` TrainMultinomialNB(C, D) V <- ExtractVocabulary(D) N <- CountDocs(D)

for each c in C do Nc <- CountDocsInClass(D, c) prior[c] <- Nc / N textc <- ConcatenateTextOfAllDocsInClass(D, c)

```
for each t in V
do
    Tct <- CountTokensOfTerm(textc, t)
for each t in V
do
    condprob[t][c] <- (Tct+1) / ∑(Tct'+1)
```

return V, prior, condprob ```

`predict_proba(W):` , show probabilities for a document in classes (W are tokens in the document)

`predict(W)` , show predict class for a document (W are tokens in the document)

`_concatenate_text_of_all_docs_in_class(x, y, c)` , flatten tokens in class c documents

`_select_features(x, y, c, k_features, method):` , select 500 / n_class features with default chi square to make an order as following pseudo code

```
SelectFeatures(D, c, k)
V <- ExtractVocabuliary(D)
L <- []
for each t in V
do
    A(t,c) <- ComputeFeatureUtility(D,t,c)
    Append(L, <t, A(t,c)>)
return FeaturesWithLargestValues(L,k)
```

`_compute_feature_utility(x, y, t, c, method = 'chi_square'):` , compute feature for selection with chi square function by default including following features selection methods:

- Chi square
- Likelihood ratios
- Expected mutual information

`_extract_vocabulary(docs):` , extract training data into 1-D list