Title: Text Classification
Module: COM6513 Natural Language Processing
Name: Ziyuan XIA
Registration Number: 180216466

Description:

In this lab, a text classification system is implemented using python3 (unfortunately it is edited and compiled under Windows OS). This system is based on perceptron algorithm and bag of words representation. The perceptron algorithm takes 800 movie reviews of each class (positive/negative) as training data and 200 of each class to test. The bag of words representation is built with uni-gram, bi-gram and tri-gram for different feature types. Bi-gram and tri-gram models are expected to have better performance than uni-gram BOW.

Training is done in multiple passes (100 passes in this case), the training process of each pass for uni-gram, bigram and trigram are shown in figure 1.

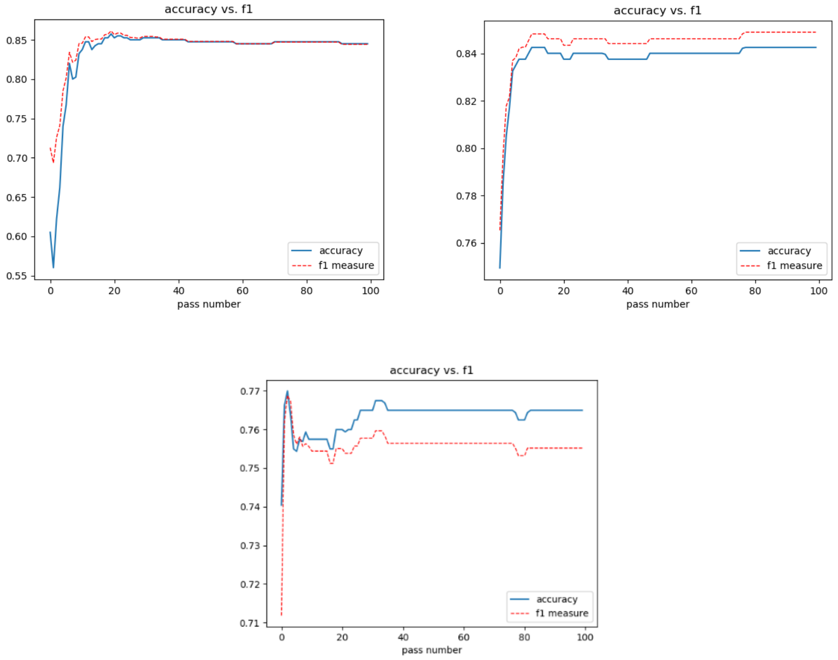


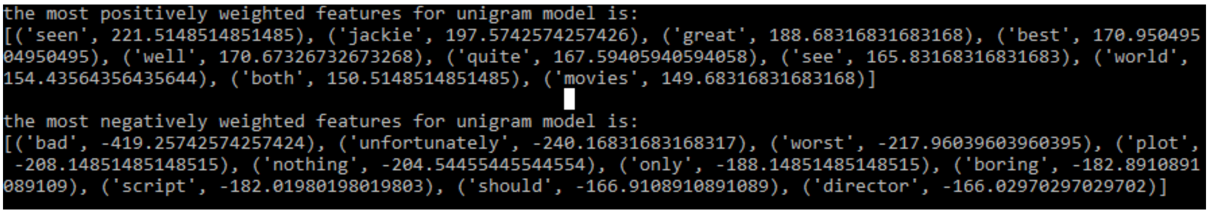
Figure 1. Training process (uni-gram: top left, bi-gram: top right, trigram: bottom)

The result takes form of system evaluation and top ten positively/negatively weighted features. Detailed information is shown in table 1.

Table 1. Result

	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Uni-gram	84.50	84.85	84.00	84.42
Bi-gram	84.25	81.57	88.50	84.89
Tri-gram	76.50	78.80	72.50	75.52

The top ten positively/negatively weighted features for uni-gram is shown in figure 2.



```
the most positively weighted features for unigram model is:
[('seen', 221.5148514851485), ('jackie', 197.5742574257426), ('great', 188.68316831683168), ('best', 170.950495
04950495), ('well', 170.67326732673268), ('quite', 167.59405940594058), ('see', 165.83168316831683), ('world',
154.43564356435644), ('both', 150.5148514851485), ('movies', 149.68316831683168)]

the most negatively weighted features for unigram model is:
[('bad', -419.25742574257424), ('unfortunately', -240.16831683168317), ('worst', -217.96039603960395), ('plot',
 -208.14851485148515), ('nothing', -204.54455445544554), ('only', -188.14851485148515), ('boring', -182.8910891
089109), ('script', -182.01980198019803), ('should', -166.9108910891089), ('director', -166.02970297029702)]
```

Figure 2. top ten features

Observation and discussion:

As it is shown in figure 1, the training process become stable after approximately 40 passes. However, the training process of bi-gram and tri-gram is not as stable as unigram. Additionally, the accuracy and f1 measure of bi-gram model and tri-gram model is lower than expected.

These phenomena could be caused by lack of training data. With the number of words in a feature increases, more features in test dataset is unseen and not useful for making prediction, more training data is required to address this problem. If there is enough training data, bi-gram model might performance better than uni-gram BOW because they take context in account.

As for the training process of tri-gram not stable and the results are not good, the reason could be that tri-gram models have too much complexity than the task required, and an overfitting problem has occurred. This assumption cannot be proved or disproved only if plenty of training data has been applied to these models.

From figure2, the top ten positively weighted features contain ‘great’, ‘well’ and ‘best’ and the top ten negatively weighted features contain ‘bad’, ‘unfortunately’, ‘worst’, ‘boring’. It is reasonable that these words have high weight on their polarity. However, applying this perceptron to reviews from another domain is not generalized well. Although the top ten words seem to be general, there are many other words maybe positively weighted in one domain but negatively weighted or even not making any sense in another. For example, delicious could be weighted highly positive in restaurant reviews but should not carry too much weight in laptop reviews. In addition, applying a perceptron to a different domain can meet a lot of unseen words which is not helpful for classifying polarity. To apply this classifier to a new domain, for example, laptop review, training data with features like price or functionality might be useful.