Sentiment Analysis Politeness in Sentences

Text Analysis Final Project

Tokenization and counts

- **tokenizing** strings and giving an integer id for each possible token, for instance by using white-spaces and punctuation as token separators.
- **counting** the occurrences of tokens in each document.
- Use CountVectorizer

Training data set

- Split the total 4000 sentences as
  75% train data and 25% test data.

Tf–idf Term Weighting

- In a large text corpus, some words will be very present (e.g. “the”, “a”, “is” in English) hence carrying very little meaningful information about the actual contents of the document.
- If we were to feed the direct count data directly to a classifier those very frequent terms would shadow the frequencies of rarer yet more interesting terms.
- In order to re-weight the count features into floating point values suitable for usage by a classifier it is very common to use the tf–idf transform.

Results for count tf-idf for X-train data

- Unigram:(3000, 7762)
- Bigram:(3000, 46040)
- Trigram:(3000, 103825)

Supervised learning

- 1. Naïve Bayes
- 2. Support Vector Machines
- 3. Logistic Regression
- 4. Random Forest

Results with no chaining the parameters

- Naïve Bayes:

0.502

- SVM :

0.56

- Logistic Regression

0.585

- Random forest:

0.539

- Naïve Bayes with stop words and bigram:

0.506

- SVM with stop words and bigram:

0.563

- Logistic Regression with bigram

0.602

- Random forest with bigram

0.558

Find the best parameters

- Use the GridSearch in sklearn to search of the best parameters on a grid of possible values.
- Logistic Regression
- SVM
- All model use k-fold validation (k=10)

Logistic Regression

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

As an optimization problem, binary class L2 penalized logistic regression minimizes the above cost function

Seach best C
Use unigram,bigram and trigram

- clf__C: 10
- tfidf__use_idf: True
- vect__ngram_range: (1, 3)

Logistic Regression results on 1000 test data

- F1 score: 0.5915724607485247
- accuracy: 0.615
- precision: 0.6156845811100721
- recall: 0.615

  1 0 -1
- Array 1([[59, 157, 10],

  0 [35, 412, 54],

  -1[12, 117, 144]])

SVM

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0, i = 1, ..., n$$

clf__alpha: 100
tfidf__use_idf: True
vect__ngram_range: (1, 1)

Seach best C
Use unigram,bigram and trigram

SVM

- f1 score: 0.5227198240496134
- accuracy: 0.588
- precision: 0.6054327525493768
- recall: 0.588
- array([[20, 194, 12],
  [13, 465, 23],
  [5, 165, 103]]

Embedding different train model

- Equal weighted sum the value of the class from different train model (NB, LR, SVM,RF)
- If the weighted value is higher than 0 then polite
- If the weighted value is lower than 0 then impolite
- f1 score: 0.598136844704617
- accuracy: 0.617
- precision: 0.614370244461421
- recall: 0.617

Final test on the test data set

- Use Logistic Regression with trigram and C = 10
- array([[28, 44,   8],
- [11, 145, 27],
- [4, 38, 48]])
- f1 score: 0.6113427341350652
- accuracy: 0.6260623229461756
- precision: 0.626162858795047
- recall: 0.6260623229461756