Statistical Natural Language Processing
Brandeis University
Fall 2015

PA1: Gender Classification

Huilin Gang
September 24, 2015

1 Naive Bayes

I implemented Multinomial Naive Bayes with BagOfWords represents each document in my classifier. My best performance model is 71.9% using NGram for $[x_{i-2}, xi-1, x_{i+1}, xi+1]$, $\alpha = 0.1$ and training set = 3000. I experimented with different size of training set and tested the performance of my model. The accuracy looks like below:

Table 1.1: Model Accuracy with different size of training set(with $\alpha = 0.1$)

2000	2500	3000
0.682	0.665	0.706

Then I experimented with different smooth method. My smoothing method looks like $P(x \mid y) = \frac{c(x,y)+\alpha}{c(y)+\alpha s}$. Let s be the number of features in the document lable y, and α be different number and test the performance of model.

Table 1.2: Model Accuracy with different α

$\alpha = 0.01$	$\alpha = 0.02$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 1$	$\alpha = 2$
0.698	0.702	0.706	0.703	0.698	0.694

2 Feature Engineering

1. Bag of Word

   (a) Bag of Words
   For bag of word, I tried to eliminate all stop words and also the one with outlier frequency, but none of these works efficient in my model. I also tried exclude all punctuation in document, it also works just fine.

   Table 2.1: Model Accuracy with different stemming method)

Without stop words	Without punctuation	Without outlier
0.694	0.700	0.612

   (b) Bag of Stemmed Words

   I implemented NLTK porter, lancaster and RSLP stemmer with $\alpha = 0.1$ and training set 3000. The result looks like below:

   Table 2.2: Model Accuracy with different stemming method)

porter stemmer	lancaster stemmer	RSLP stemmer
0.703	0.681	0.693

   (c) Bag of Tokenized Words
   I implemented NLTK Tokenized The performance for tokenized bag of words is 0.702

2. NGram
   I tried NGram in sequence $[x_{i-1}, x_i]$, $[x_{i-1}, x, x_{i+1}]$, $[x_{i-2}, xi-1, x_{i+1}, xi+1]$

Table 2.3: Model Accuracy with different stemming method)

$[x_{i-1}, x_i]$	$[x_{i-1}, x, x_{i+1}]$	$[x_{i-2}, xi-1, x_{i+1}, xi+1]$
0.707	0.707	0.714

3. FMeasure

I tried tagging and categorizing document in NLTK, thus I can get F measure for each document. But since I don't have much time on training and testing, I haven't completed building model with FMeasure as variables. Also, it's time consuming to finish tagging.

3 Conclusion

From what I learn in Mukherjee and Bing paper, I know there are some fancy methods for feature engineering like POS. I didn't have enough time to implement them in my model. For my further plan, I would like to use them to improve my model performance.