Assignment 2 Natural Language Processing

Perceptron article classification

Xuhong Guo 450489321

Abstract. In this assignment, we will implement an end-to-end document classification system.

Contents

1 Machine Learning Problem

The machine learning problem on this assignment is that, the program run out of memory and the extraction is too slow. For solving this problem, this assignment use a 26G memory computer and Pandas library in code. Pandas can iterate though large file lazily in IO rather than read the entire file[1].

2 Evaluate Averaged Perceptron Implementation

For a 10-fold cross-validated averaged perceptron algorithm training for 3 epochs on this data, the averaged perceptron implementation achieve a mean accuracy of 0.95.

		precision	recall	f1-score	support
1 fold	avg / total	0.97	0.97	0.97	328
2 fold	avg / total	0.93	0.92	0.93	328
3 fold	avg / total	0.95	0.95	0.95	328
4 fold	avg / total	0.95	0.95	0.95	328
5 fold	avg / total	0.95	0.95	0.95	328
6 fold	avg / total	0.96	0.96	0.96	328
7 fold	avg / total	0.96	0.95	0.96	328
8 fold	avg / total	0.95	0.94	0.94	328
9 fold	avg / total	0.94	0.93	0.94	328
10 fold	avg / total	0.94	0.93	0.94	328
		0.95	0.9466666667	0.949	

3 Baseline Set of Features

3.1 Introduce

This assignment use each word in the total vocabulary of the Wikipedia Article document set as features, which is the standard feature or baseline set of feature on this assignment.

3.2 Data description

The baseline set has 104317 features, and 2971 articles. The features is include each word in articles but punctuation.

3.3 Perceptron setting

The averaged perceptron algorithm training for 5 epochs on this data with 10-fold cross-validated.

3.4 Evaluation

3.4.1 accuracy

1 fold	accuracy	0.69
2 fold	accuracy	0.78
3 fold	accuracy	0.77
4 fold	accuracy	0.66
5 fold	accuracy	0.76
6 fold	accuracy	0.73
7 fold	accuracy	0.74
8 fold	accuracy	0.76
9 fold	accuracy	0.66
10 fold	accuracy	0.7
		0.725

3.4.2 micro-average

		precision	recall	f1-score	support
1 fold	avg / total	0.695	0.695	0.695	200
2 fold	avg / total	0.775	0.775	0.775	200
3 fold	avg / total	0.765	0.765	0.765	200
4 fold	avg / total	0.655	0.655	0.655	200
5 fold	avg / total	0.76	0.76	0.76	200
6 fold	avg / total	0.735	0.735	0.735	200
7 fold	avg / total	0.745	0.745	0.745	200
8 fold	avg / total	0.755	0.755	0.755	200
9 fold	avg / total	0.655	0.655	0.655	200
10 fold	avg / total	0.7	0.7	0.7	200
		0.724	0.7266666667	0.724	

3.4.3 macro-average

		precision	recall	f1-score	support
1 fold	avg / total	0.5014733806	0.4932280534	0.4743204658	200
2 fold	avg / total	0.5595891201	0.5715134921	0.5363834898	200
3 fold	avg / total	0.5432400932	0.5301063671	0.5248279069	200
4 fold	avg / total	0.412858162	0.3853067985	0.377047541	200
5 fold	avg / total	0.5047319347	0.5043025493	0.4942938881	200
6 fold	avg / total	0.5151676152	0.4708227714	0.4610420229	200
7 fold	avg / total	0.5775773036	0.5388349346	0.5338130371	200
8 fold	avg / total	0.4897496067	0.4477353267	0.446084	200
9 fold	avg / total	0.358469145	0.3680262	0.34595943	200
10 fold	avg / total	0.459832352	0.43311288	0.4388559998	200
		0.4922688713	0.4788751659	0.4632627781	

3.4.4 running time

Each fold spend 2 hours.

3.5 Analyzing Error and Shortcoming

The label Health, label Hospitality, and label Other has a 0 precision recall and f-value. After checked the number of those labels, find that the size of those labels is small than other labels. And the label Entertainment get 0.65 precision, 0.92 recall and 0.76 f-value which is more higher value than other labels. And the label Sports get 0.87 precision, 0.89 recall, and 0.88 f-value, After checked the number of size, label Entertainment and Sport has a higher size numbers. And the running time is too much longer.

3.6 Summary

This assignment base on the baseline set of feature, the accuracy is 72.5%, the micro-average is around 72%, macro-average is around 47%

4 Advance Unigrams Word Set of Features

4.1 Introduce

This assignment use semantics to selected the features for advance unigrams word set of features. The assignment selected the word type is Verb or Noun in each article, then found each Noun word's hypernyms as features of the article.

4.2 Reason

This assignment use the Noun and verb word as the feature is because, Noun and verb are able to show the meaning of a article. So the assignment try to use this two type of word as feature to represent articles. In the baseline set of feature, the adjective, adverb preposition word, maybe negatively effected the W value for classification. And Noun word have a hypernyms word, this assignment use the word hypernyms, which is not only decrease the number of useless features but also more meaningful for representation a article.

4.3 Data description

All article's Verb words and Noun word's hypernyms has 5608 features, and 2975 articles.

4.4 Perceptron setting

The averaged perceptron algorithm training for 5 epochs on this data with 10-fold cross-validated.

4.5 Evaluation

4.5.1 accuracy

1 fold	accuracy	0.66
2 fold	accuracy	0.75
3 fold	accuracy	0.76
4 fold	accuracy	0.69
5 fold	accuracy	0.79
6 fold	accuracy	0.72
7 fold	accuracy	0.75
8 fold	accuracy	0.76
9 fold	accuracy	0.69
10 fold	accuracy	0.71
		0.728

4.5.2 micro-average

		precision	recall	f1-score	support
1 fold	avg / total	0.66	0.66	0.66	200
2 fold	avg / total	0.75	0.75	0.75	200
3 fold	avg / total	0.76	0.76	0.76	200
4 fold	avg / total	0.695	0.695	0.695	200
5 fold	avg / total	0.79	0.79	0.79	200
6 fold	avg / total	0.72	0.72	0.72	200
7 fold	avg / total	0.75	0.75	0.75	200
8 fold	avg / total	0.755	0.755	0.755	200
9 fold	avg / total	0.69	0.69	0.69	200
10 fold	avg / total	0.715	0.715	0.715	200
		0.7285	0.73	0.7285	

4.5.3 macro-average

		precision	recall	f1-score	support
1 fold	avg / total	0.466978	0.452555	0.448327	200
2 fold	avg / total	0.523166899	0.52939169	0.5158013074	200
3 fold	avg / total	0.573292	0.61672192	0.584980429	200
4 fold	avg / total	0.454737	0.467966	0.456466676	200
5 fold	avg / total	0.5227831918	0.5545462562	0.5279583315	200
6 fold	avg / total	0.5216815857	0.48402047	0.483795017	200
7 fold	avg / total	0.62467906	0.6201995373	0.61961318	200
8 fold	avg / total	0.51728869	0.532531376	0.5117067156	200
9 fold	avg / total	0.4715510392	0.4803848004	0.4691811069	200
10 fold	avg / total	0.4256904498	0.44172671	0.4319673077	200
		0.5101847916	0.5264796722	0.5049797071	

4.5.4 running time

Each fold spend 20 minutes.

4.6 Analysing Error and Shortcoming

The label Health, label Hospitality, and label Other has a 0 precision recall and f-value. After checked the number of those labels, find that the size of those labels is small than other labels. And the label Entertainment get 0.75 precision, 0.92 recall and 0.83 f-value which is more higher value than other labels. And the label Sports get 0.94 precision, 0.89 recall, and 0.92 f-value, After checked the number of size, label Entertainment, Sports has a higher size numbers than other label.

4.7 Summary

This assignment base on the verb and noun word's hypernyms features data, the accuracy is 72.8%, micro-average is around 73%, macro-average is around 52%, each value is little higher than the value of baseline set. The more improve is the running time, which is more faster than baseline set. From the result, the Noun word's hypernyms and verb can represent the articles.

5 Bigrams Words Set of Features

5.1 Introduce

This assignment use bigrams words in the vocabulary of the Wikipedia article document set as features. Depend on the memory of computer is limited, and the number of the bigrams words is larger, so this assignment selected the bigrams words base on types of words. This types are the two words with (Adjective, Noun) or (Verb, Noun).

5.2 Reason

This assignment use bigrams words as feature, because bigrams words can be more meaningful for a specific article, so this assignment try to use the bigrams words of features to represent a article.

5.3 Data description

This bigrams words has 99688 features, and 2971 articles.

5.4 Perceptron setting

The averaged perceptron algorithm training for 5 epochs on this data with 10-fold cross-validated.

5.5 Evaluation

5.5.1 accuracy

1 fold	accuracy	0.54
2 fold	accuracy	0.57
3 fold	accuracy	0.5
4 fold	accuracy	0.54
5 fold	accuracy	0.55
6 fold	accuracy	0.53
7 fold	accuracy	0.54
8 fold	accuracy	0.57
9 fold	accuracy	0.53
10 fold	accuracy	0.53
		0.54

5.5.2 micro-average

		precision	recall	f1-score	support
1 fold	avg / total	0.5436241611	0.5436241611	0.5436241611	200
2 fold	avg / total	0.5738255034	0.5738255034	0.5738255034	200
3 fold	avg / total	0.49664429	0.49664429	0.49664429	200
4 fold	avg / total	0.543624161	0.543624161	0.543624161	200
5 fold	avg / total	0.546979865	0.546979865	0.546979865	200
6 fold	avg / total	0.5268456375	0.5268456375	0.5268456375	200
7 fold	avg / total	0.5402684564	0.5402684564	0.5402684564	200
8 fold	avg / total	0.570469798	0.570469798	0.570469798	200
9 fold	avg / total	0.5335570469	0.5335570469	0.5335570469	200
10 fold	avg / total	0.5268456375	0.5268456375	0.5268456375	200
		0.5402684557	0.5417598799	0.5402684557	

5.5.3 macro-average

		precision	recall	f1-score	support
1 fold	avg / total	0.3690803291	0.44439481	0.374537679	200
2 fold	avg / total	0.3572157177	0.361100837	0.350421539	200
3 fold	avg / total	0.379464996	0.35048284	0.355446702	200
4 fold	avg / total	0.393152409	0.33388136	0.343354269	200
5 fold	avg / total	0.3595122532	0.329472856	0.3405740116	200
6 fold	avg / total	0.4554322	0.3937066041	0.4131624812	200
7 fold	avg / total	0.3066306416	0.301570037	0.301771978	200
8 fold	avg / total	0.3535944251	0.3683065274	0.35428794	200
9 fold	avg / total	0.411169549	0.35529804	0.35788543	200
10 fold	avg / total	0.4554322	0.3937066041	0.4131624812	200
		0.3840684721	0.3598015457	0.3604604511	

5.5.4 running time

Each fold spend 2 hours.

5.6 Analysing Error and Shortcoming

The label Health, label Hospitality, and label Other has a 0 precision recall and f-value. After checked the number of those labels, find that the size of those labels is small than other labels. And the label Entertainment get 0.84 precision, 0.64 recall and 0.74 f-value which is more higher value than other labels. And the label Sports get 0.94 precision, 0.76 recall, and 0.84 f-value, After checked the number of size, label Entertainment, Sports has a higher size numbers than other label.

5.7 Summary

This assignment value base on the bigrams words set of features, accuracy is around 54%, micro-average is around 54%, macro-average is around 36%, so this result showed the two words with (Adjective, Noun) or (Verb, Noun) can not represent a article well.

6 Important Feature

6.1 Introduction

After this assignment finished the advance unigrams word in noun and verb, the assignment also test the noun, verb and adj each training and evaluating, the result is below.

6.2 Feature evaluation

6.2.1 Noun

NN					
		precision	recall	f1-score	support
average of 10 foder	macro	0.459	0.455	0.42	200
	micro	0.679	0.679	0.679	200

6.2.2 Verb

VB					
		precision	recall	f1-score	support
average of 10 foder	macro	0.226	0.283	0.26	200
	micro	0.491	0.491	0.491	200

6.2.3 Adj

ADJ					
		precision	recall	f1-score	support
average of 10 foder	macro	0.27	0.25	0.24	200
	micro	0.475	0.475	0.475	200

6.3 Summary

When the assignment selected the noun and verb word as the feature and making a training, the result get a same accuracy as the result of baseline features get. And then the assignment also evaluate each of type, the result is that, if the assignment use noun word as feature only, the macro-average is around 43%, the micro-average is around 67%, while if the assignment use the verb word as feature only, the macro-average is around 28%, the micro-average is around 49%, while if the assignment use the adj word as feature only, the macro-average is around 25%, the micro-average is around 47%. This comparison showed the noun and verb can represent is more important as feature. This assignment also use the other types of words except noun and verb to do a training, the result get 45% precision, 42% recall and 38%f-value. This result proved the noun and verb is more important features in this assignment.

7 Errors of the System

The category number of articles in this assignment are imbalanced, such as the health category,the hospitality category and the other category, they have little number of articles in this assignment. When training the dataset, those types of articles can not be learned well.

The system did not test the Adjective word or Adverb word, and when running the bigrams words, it also ignore the (Noun,Noun) words, because after selected the features of (Noun,Noun), the number of feature almost have 200,000, the computer memory can not afford it.

References

[1] Wikipedia, *Pandas (software) — wikipedia, the free encyclopedia* (2017), [Online; accessed 29-April-2017], https://en.wikipedia.org/w/index.php?title=Pandas_(software)&oldid=774897003