Text Classification with Multinomial Naive Bayes

Eşref Özdemir

April 9, 2018

Contents

1 Training and Test Sets

Training and test sets are constructed using the parsing and tokenizing algorithms of the previous assignment. We again performed stop word removal, normalization and stemming steps. After tokenization and splitting the documents by their LEWIS_ID field, training set consisted of 5791 documents whereas test set consisted of 2300 documents.

2 50 Most Discriminating Words

In this section, we give the 50 most important word for each class. Top left word is the most important one, and the importance decreases as one goes from left to right and from top to bottom.

Table 1: Most discriminating words for earn

vs	ct	shr	net	said
qtr	to	rev	loss	4th
ha	div	at	profit	note
dividend	record	bui	31	avg
qtly	u.	prior	pct	market
not	agreement	would	acquir	year
mth	offer	agre	offici	had
sell	purchas	sai	exchang	bank
price	todai	about	export	1st
were	jan	wheat	tender	more

Table 2: Most discriminating words for acq

vs	ct	shr	said	acquir
net	qtr	rev	acquisit	to
stake	compani	bui	merger	year
loss	4th	ha	complet	offer
sell	share	record	note	common
div	corp	group	own	inc
sharehold	outstand	undisclos	purchas	agre
avg	approv	term	file	qtly
profit	bid	subsidiari	takeov	dividend
unit	control	transact	disclos	31

Table 3: Most discriminating words for money-fx

bank	dollar	currenc	monei	rate
market	central	treasuri	dealer	yen
ct	vs	england	japan	pari
inc	monetari	around	interven	intervent
shr	todai	shortag	net	u.k
at	fed	trade	bill	exchang
assist	corp	deficit	foreign	share
qtr	compani	stabil	econom	germani
sai	stg	band	rev	further
reserv	repurchas	against	nation	u.

Table 4: Most discriminating words for grain

wheat	agricultur	tonn	grain	corn
usda	export	crop	u.	depart
farmer	ct	vs	soybean	soviet
farm	inc	commod	maiz	barlei
net	shr	program	1986/87	qtr
rice	bushel	ec	feed	ussr
compani	share	winter	corp	shipment
cereal	rev	harvest	union	grower
to	sorghum	season	import	enhanc
subsidi	offici	china	acreag	hectar

Table 5: Most discriminating words for crude

oil	barrel	crude	bpd	opec
petroleum	energi	dai	vs	price
refineri	product	shr	ga	minist
drill	output	explor	ct	quota
gasolin	net	said	ecuador	to
sea	qtr	last	produc	state
countri	natur	rev	gulf	at
iran	iraq	import	pipelin	were
saudi	report	earthquak	inc	would
iranian	tanker	refin	sai	suppli

3 Metrics

3.1 Using All Words

Table 6: Micro averaged precision, recall and F1-score using all words

Metric	Score
Precision	0.9791
Recall	0.9791
F1-score	0.9791

Table 7: Macro averaged precision, recall and F1-score using all words

Metric	Score
Precision	0.9767
Recall	0.9814
F1-score	0.9767

Table 8: Unaveraged precision, recall and F1-score of each class using all words

	earn	money-fx	crude	grain	acq
Precision	0.9906	0.9568	0.9775	0.9932	0.9656
Recall	0.9723	0.9944	0.9667	0.9865	0.9873
F1-score	0.9656	0.9932	0.9775	0.9906	0.9568

3.2 Using Top 50 Words

Table 9: Micro averaged precision, recall and F1-score using top 50 words

Metric	Score
Precision	0.7504
Recall	0.7504
F1-score	0.7504

Table 10: Macro averaged precision, recall and F1-score using top 50 words

Metric	Score
Precision	0.7075
Recall	0.7968
F1-score	0.7075

Table 11: Unaveraged precision, recall and F1-score of each class using top 50 words

	earn	money-fx	crude	grain	acq
Precision	0.9932	0.9606	0.8013	0.4515	0.3308
Recall	0.8072	0.5845	0.7022	0.9122	0.9778
F1-score	0.3308	0.4515	0.8013	0.9932	0.9606

4 Screenshots

4.1 Constructing Datasets

```
eozd@asus_arch [14:47] [~/Downloads/cmpe493/assignment2] [master *]
-> % ./construct_datasets
Constructing train and test datasets...OK!
Writing train and test dataset files...OK!
5791 documents was indexed to construct the train dataset at train.txt
2300 documents was indexed to construct the test  dataset at test.txt
eozd@asus_arch [14:47] [~/Downloads/cmpe493/assignment2] [master *]
-> % ls -l {train,test}*
-rw-r--r-- 1 eozd eozd  995200 Apr  1 14:47 test.txt
-rw-r--r-- 1 eozd eozd 2683342 Apr  1 14:47 train.txt
eozd@asus_arch [14:47] [~/Downloads/cmpe493/assignment2] [master *]
-> %
```

4.2 Help Message

```
eozd@asus_arch [14:48] [~/Downloads/cmpe493/assignment2] [master *]
-> % ./classifier
usage: classifier [--fit train_set model_path [--num-features N]]
                  [--predict test_set model_path]

Fit a classifier using a training set; or predict the classes
of a test set using an already fitted model.

optional arguments:
  --fit train_set model_path     Fit a Naive Bayes classifier from given
                                 train_set and save the model to model_path.

  --num-features N               Number of features to use during training.
                                 Best N features are chosen using Mutual Information.
                                 If not given, all the words are used as features.

  --predict test_set model_path  Predict the classes of samples in test_set
                                 using an already fitted model in model_path
                                 and output the results to STDOUT.
eozd@asus_arch [14:48] [~/Downloads/cmpe493/assignment2] [master *]
-> %
```

4.3 Fitting with All Words

```
eozd@asus_arch [14:48] [~/Downloads/cmpe493/assignment2] [master *]
-> % ./classifier --fit train.txt model.txt
eozd@asus_arch [14:48] [~/Downloads/cmpe493/assignment2] [master *]
-> % ls -l model.txt
-rw-r--r-- 1 eozd eozd 651090 Apr  1 14:48 model.txt
eozd@asus_arch [14:48] [~/Downloads/cmpe493/assignment2] [master *]
-> %
```

4.4 Fitting with Top 50 Words

```
eozd@asus_arch [14:48] [~/Downloads/cmpe493/assignment2] [master *]
-> % ./classifier --fit train.txt model.txt --num-features 50
eozd@asus_arch [14:49] [~/Downloads/cmpe493/assignment2] [master *]
-> % ls -l model.txt
-rw-r--r-- 1 eozd eozd 3737 Apr  1 14:49 model.txt
eozd@asus_arch [14:49] [~/Downloads/cmpe493/assignment2] [master *]
-> %
```

4.5 Prediction

```
eozd@asus_arch [14:49] [~/Downloads/cmpe493/assignment2] [master *]
-> % ./classifier --predict test.txt model.txt > out 2> log
eozd@asus_arch [14:49] [~/Downloads/cmpe493/assignment2] [master *]
-> % head out
ID: 15273 | Test:      grain | Pred:      grain
ID: 18482 | Test:      grain | Pred:      grain
ID: 15217 | Test:        acq | Pred:        acq
ID: 16710 | Test:      crude | Pred:      crude
ID: 16012 | Test:      grain | Pred:      grain
ID: 16744 | Test:   money-fx | Pred:   money-fx
ID: 18313 | Test:        acq | Pred:        acq
ID: 15104 | Test:       earn | Pred:       earn
ID: 15327 | Test:       earn | Pred:       earn
ID: 17538 | Test:        acq | Pred:        acq
eozd@asus_arch [14:50] [~/Downloads/cmpe493/assignment2] [master *]
-> % head log
Micro Averaged Stats
--------------------
Precision:    0.9791
Recall:       0.9791
F1 score:     0.9791

Macro Averaged Stats
--------------------
Precision:    0.9767
Recall:       0.9814
eozd@asus_arch [14:50] [~/Downloads/cmpe493/assignment2] [master *]
-> %
```