Online Chinese News Classification System based on Python Reptile and Convolutional Neural Network

Zhe Zhang[1], Genyao Zhang[1], Chi Zhao[1], PeikeWang[2]

1College of Mathematics and Computer Science, Yan'an University, Yan'an, 716000, China
2College of Computer Engineering, Huaihai Institute of Technology, Huaihai, 222000, China

Abstract: With the advent of the era of big data, the amount of data has shown an exponential growth. A series of news retrieval platforms represented by Netease News have new content all the time. In order to allow users to more directly understand the current hot spots, the system usesTHUCNews data setsfor training, including a total of 14 categories, 830,000 articles. Using TensorFlow to train a character-level convolutional neural network with a single convolutional layer, a 92.24% accuracy rate was achieved on the validation set. Combined with the Python crawler, it eventually achieved automatic crawling and classification of news information, presented to the user in a concise manner.

Keywords: Crawler; Convolutional neural network; Python

1. Introduction

In the current era of information explosion, various kinds of information are uploaded daily from various media. In order to allow users to more intuitively understand the current hot news in the society, people have used crawlers to crawl and display news titles and content, but the news leaderboard only gives the news headlines, and does not give specific news categories. Even the classification of news content pages is classified by people subjectivity. There is also an article corresponding to multiple categories, so how to objectively carry out the news correctly classification becomes an important research issue in natural language processing. Therefore, it was can't achieved that use the simple adoption of crawler technology, so this paper regularly uses crawler technology to crawl the top 20 hits with the highest click rate from the Netease news rankings, put them into a trained neural network, these news will be classified and stored in the database by the neural network.

2. Related Research Work

Kim[1] used CNN[2][3] to classify sentences based on word2vec[4], and conducted comparative experiments on seven sets of datasets to prove a single layer of convolutional nerves. CNN is effective in text categorization tasks, it also shows that using unsupervised learning methods to pre-train word vectors is an important part of natural language processing (NLP). Xiang Zhang, Junbo Zhao, and Yann LeCun[5] provided empirical research on text classification for character-level convolutional neural networks. Using some large-scale datasets, the traditional model and the deep learning model were compared to find that the character-level convolutional neural network is an effective method for text classification.

This paper uses THUCNews[6][7] to construct vocabulary based on the training set and then converts the input text into a fixed-length id sequence by using the word2id operation. These a fixed-length id sequence as neural networks' input. It was adopted that character-level convolutional neural network model. The model achieves good results by adjusting different hyperparameters.

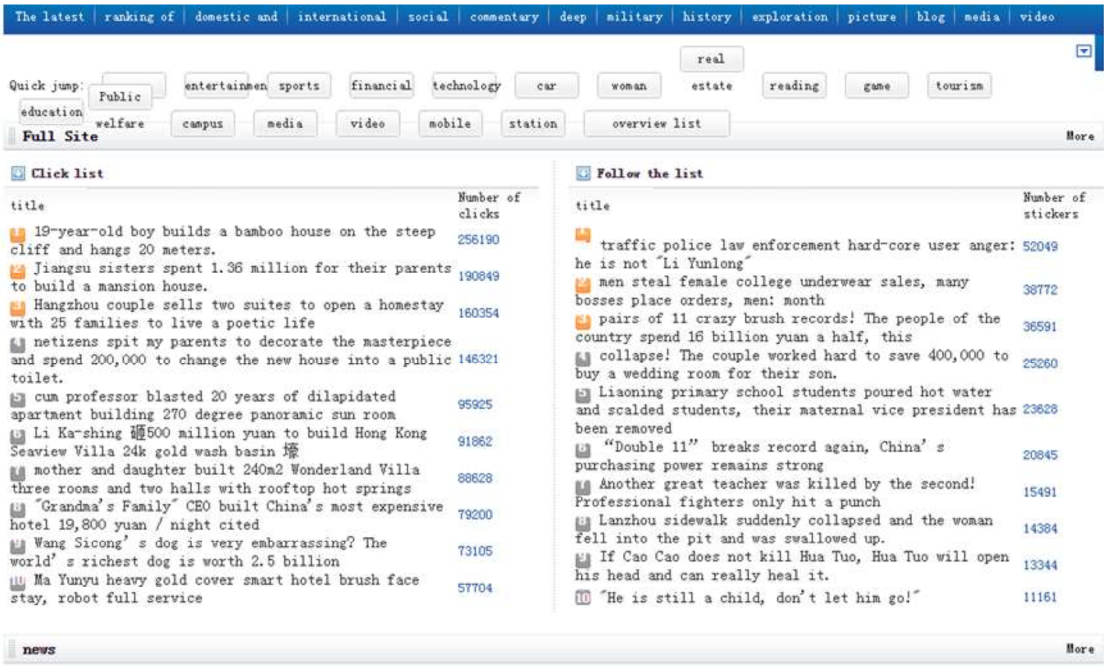
3. System Design and Implementation

In order to realize news online classification, this paper uses crawler technology to automatically crawl and store the required data, and these data are input into the neural network after preprocessing, and then the results are stored in the database, and displayed on the front-end of the website by friendly means.

3.1. Model design

The crawler is an automated program that requests websites and extracts data. This paper analyzes and compares the rankings of various news portals, and finally selects the Netease News rankings with regular contents, less bad information, and fewer advertisements. Crawl every hour and select the top 20 hot news with the highest click rate, as shown in Figure 1.

Analyzing the landing page: http://news.163.com/special/0001386F/rank_news.html, the page leaderboard is not d ynamically generated by JavaScript, so the page can be p rocessed directly after crawling, using the BeautifulSoup + requests library can be achieved, crawler process show n in Figure 2. After the above steps, a simple Netease news crawler is implemented.



Click list

	title	Number of clicks
1	19-year-old boy builds a bamboo house on the steep cliff and hangs 20 meters.	256190
2	Jiangsu sisters spent 1.36 million for their parents to build a mansion house.	190849
3	Hangzhou couple sells two suites to open a homestay with 25 families to live a poetic life	160354
4	netizens spit my parents to decorate the masterpiece and spend 200,000 to change the new house into a public toilet.	146321
5	cum professor blasted 20 years of dilapidated apartment building 270 degree panoramic sun room	95925
6	Li Ka-shing 砸500 million yuan to build Hong Kong Seaview Villa 24k gold wash basin 壕	91862
7	mother and daughter built 240m2 Wonderland Villa three rooms and two halls with rooftop hot springs	88628
8	"Grandma's Family" CEO built China's most expensive hotel 19,800 yuan / night cited	79200
9	Wang Sicong' s dog is very embarrassing? The world' s richest dog is worth 2.5 billion	73105
10	Ma Yunyu heavy gold cover smart hotel brush face stay, robot full service	57704

Follow the list

	title	Number of stickers
1	traffic police law enforcement hard-core user anger: he is not "Li Yunlong"	52049
2	men steal female college underwear sales, many bosses place orders, men: month	38772
3	pairs of 11 crazy brush records! The people of the country spend 16 billion yuan a half, this	36591
4	collapse! The couple worked hard to save 400,000 to buy a wedding room for their son.	25260
5	Liaoning primary school students poured hot water and scalded students, their maternal vice president has been removed	23628
6	"Double 11" breaks record again, China' s purchasing power remains strong	20845
7	Another great teacher was killed by the second! Professional fighters only hit a punch	15491
8	Lanzhou sidewalk suddenly collapsed and the woman fell into the pit and was swallowed up.	14384
9	If Cao Cao does not kill Hua Tuo, Hua Tuo will open his head and can really heal it.	13344
10	"He is still a child, don't let him go!"	11161

Figure 1. Netease news ranking

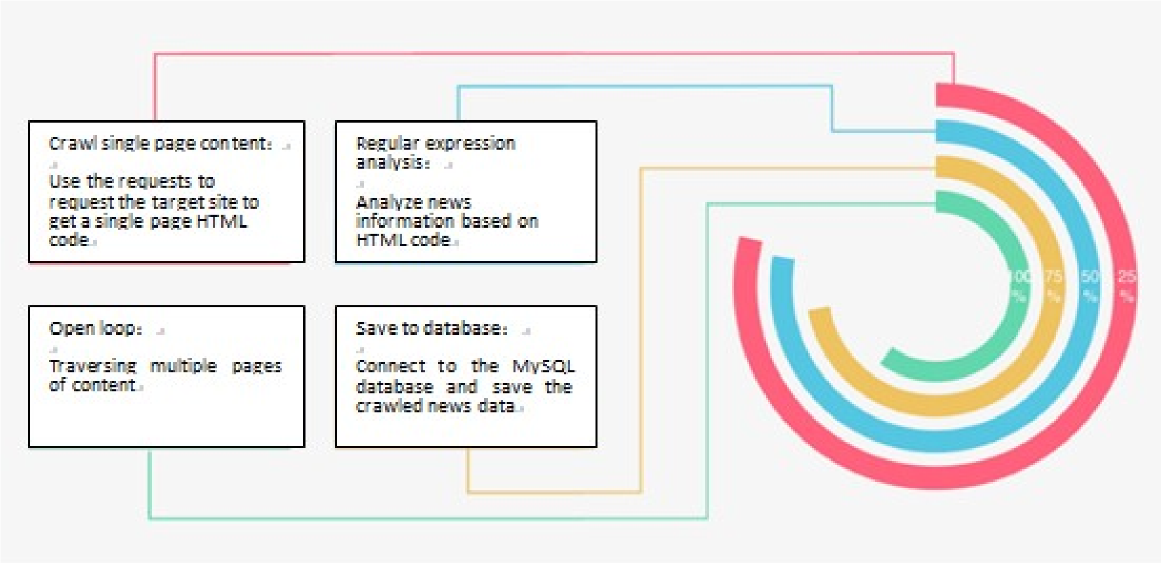

Crawl single page content:
Use the requests to request the target site to get a single page HTML code
Regular expression analysis:
Analyze news information based on HTML code
Open loop:
Traversing multiple pages of content
Save to database:
Connect to the MySQL database and save the crawled news data
100%
75%
50%
25%

Figure 2. Crawler process

The model architecture adopted in this paper is shown in Figure 3. The first-layer word embedding layer (Embedding) [8] converts a one-dimensional sequence into a two-dimensional vector, and accesses a convolutional layer composed of multiple filters. The convolutional layer is used as a feature extraction. The extraction layer extracts local features through filters and generates feature maps through convolution kernel function operations and outputs them to Max Pooling. The max pooling layer belongs to the feature mapping layer, which samples the feature maps generated by the convolutional layer and outputs the local optimal features. Immediately after the Fully Connected layer, the layer is added with dropout[9] regularization and relu activation functions. The last full connection layer and Softmax layer are used as classifiers, and finally map the output of multiple neurons to within the (0,1) interval, multiple classifications are performed.

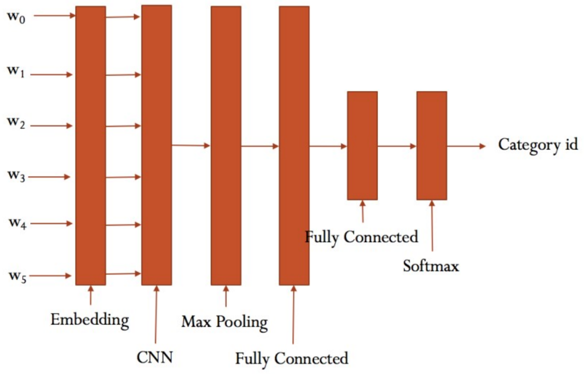


Figure 3. Model architecture

3.2. Data preprocessing and model training

This paper selects the THUCNews data sets generated by the Natural Language Processing Laboratory of Tsinghua University according to the historical data of the Sina News RSS subscription channel from 2005 to 2011. The data sets contains about 830,000 news documents (2.19 GB) and 14 candidate classification categories: Finance, lottery, real estate, stocks, home, education, technology, society, fashion, politics, sports, constellations, games, entertainment.

First,preprocessing the data sets, and 830,000 independent documents are integrated into three files, which are training set, verification set and test set. The proportion of the data sets is 70%, 15%, 15%. The specific division is shown in Table 2 After preprocessing, the data in the format of Table 1 is obtained.

Table 1. Pre-processed Data

Data	Shape	Data	Shape
x_train	[585247, 600]	y_train	[585247, 14]
x_val	[125403, 600]	y_val	[125403, 14]
x_test	[125425, 600]	y_test	[125425, 14]

Table 2. Data Set Partitioning

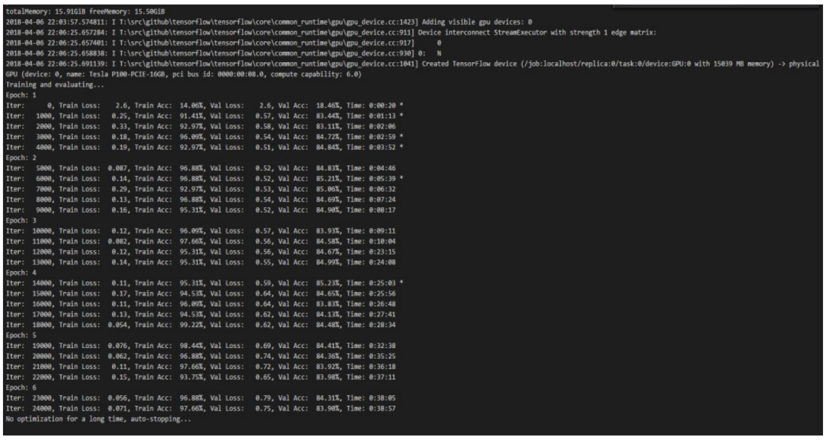
Data set partition				
Classification	**Training set**	**Verification set**	**Test set**	**Total**
Finance	25968	5564	5566	37098
Stock	108078	23159	23161	154398
Technology	114050	24439	24440	162929
Society	33594	7627	7628	50849
Game	17061	3655	3657	24373
Constellation	2504	536	538	3578
Politics	44160	9462	9464	63086
Fashion	9357	2005	2006	13368
Education	29355	6290	6291	41936
Property	14035	3007	3008	20050
Lottery ticket	5311	1138	1139	7588
Household	22810	4887	4889	32586
Entertainment	64842	13894	13896	92632
Sports	92122	19740	19742	131604

Total	585247	125403	125425	836075

The 10-fold cross-validation method was used during the time of training.Set the training to be suspended in advance if the model does not improve over 10,000 rounds. After continuous improvement, the following parameters are finally selected.

embedding_dim = 64 #Word vector dimension
seq_length = 600 #Sequence length
num_classes = 14 #Number of categories
num_filters = 512 #Number of convolution kernels
kernel_size = 9 #Convolution kernel size
vocab_size = 6000 #Glossary size
hidden_dim = 128 #Fully connected neurons
dropout_keep_prob = 0.6 #dropoutRetention ratio
learning_rate = 1e-3 #Learning rate
batch_size = 128 #Training size per batch
num_epochs = 10 #Total iteration round
print_per_batch = 1000 #Output results every few rounds
save_per_batch = 10 #How much of each round into tensorboard

The experimental results are shown in Figure 4 The highest accuracy of 85.23% was achieved on the verification set and stopped only after 6 rounds.



```
totalMemory: 15.91GiB freeMemory: 15.50GiB
2018-04-06 22:03:57.574811: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:1423] Adding visible gpu devices: 0
2018-04-06 22:06:25.657284: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:911] Device interconnect StreamExecutor with strength 1 edge matrix:
2018-04-06 22:06:25.657401: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:917]      0
2018-04-06 22:06:25.658838: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:930] 0:   N
2018-04-06 22:06:25.691139: I T:\src\github\tensorflow\tensorflow\core\common_runtime\gpu\gpu_device.cc:1041] Created TensorFlow device (/job:localhost/replica:0/task:0/device:GPU:0 with 15039 MB memory) -> physical
GPU (device: 0, name: Tesla P100-PCIE-16GB, pci bus id: 0000:00:08.0, compute capability: 6.0)
Training and evaluating...
Epoch: 1
Iter:      0, Train Loss:    2.6, Train Acc:  14.06%, Val Loss:    2.6, Val Acc:  18.46%, Time: 0:00:20 *
Iter:   1000, Train Loss:   0.25, Train Acc:  91.41%, Val Loss:   0.57, Val Acc:  83.44%, Time: 0:01:13 *
Iter:   2000, Train Loss:   0.33, Train Acc:  92.97%, Val Loss:   0.58, Val Acc:  83.11%, Time: 0:02:06
Iter:   3000, Train Loss:   0.18, Train Acc:  96.09%, Val Loss:   0.54, Val Acc:  84.72%, Time: 0:02:59 *
Iter:   4000, Train Loss:   0.19, Train Acc:  92.97%, Val Loss:   0.51, Val Acc:  84.84%, Time: 0:03:52 *
Epoch: 2
Iter:   5000, Train Loss:  0.087, Train Acc:  96.88%, Val Loss:   0.52, Val Acc:  84.83%, Time: 0:04:46
Iter:   6000, Train Loss:   0.14, Train Acc:  96.88%, Val Loss:   0.52, Val Acc:  85.21%, Time: 0:05:39 *
Iter:   7000, Train Loss:   0.29, Train Acc:  92.97%, Val Loss:   0.53, Val Acc:  85.06%, Time: 0:06:32
Iter:   8000, Train Loss:   0.13, Train Acc:  96.88%, Val Loss:   0.54, Val Acc:  84.69%, Time: 0:07:24
Iter:   9000, Train Loss:   0.16, Train Acc:  95.31%, Val Loss:   0.52, Val Acc:  84.90%, Time: 0:08:17
Epoch: 3
Iter:  10000, Train Loss:   0.12, Train Acc:  96.09%, Val Loss:   0.57, Val Acc:  83.93%, Time: 0:09:11
Iter:  11000, Train Loss:  0.082, Train Acc:  97.66%, Val Loss:   0.56, Val Acc:  84.58%, Time: 0:10:04
Iter:  12000, Train Loss:   0.12, Train Acc:  95.31%, Val Loss:   0.56, Val Acc:  84.67%, Time: 0:23:15
Iter:  13000, Train Loss:   0.14, Train Acc:  95.31%, Val Loss:   0.55, Val Acc:  84.99%, Time: 0:24:08
Epoch: 4
Iter:  14000, Train Loss:   0.11, Train Acc:  95.31%, Val Loss:   0.59, Val Acc:  85.23%, Time: 0:25:03 *
Iter:  15000, Train Loss:   0.17, Train Acc:  94.53%, Val Loss:   0.64, Val Acc:  84.65%, Time: 0:25:56
Iter:  16000, Train Loss:   0.11, Train Acc:  96.09%, Val Loss:   0.64, Val Acc:  83.83%, Time: 0:26:48
Iter:  17000, Train Loss:   0.13, Train Acc:  94.53%, Val Loss:   0.62, Val Acc:  84.13%, Time: 0:27:41
Iter:  18000, Train Loss:  0.054, Train Acc:  99.22%, Val Loss:   0.62, Val Acc:  84.48%, Time: 0:28:34
Epoch: 5
Iter:  19000, Train Loss:  0.076, Train Acc:  98.44%, Val Loss:   0.69, Val Acc:  84.41%, Time: 0:32:38
Iter:  20000, Train Loss:  0.062, Train Acc:  96.88%, Val Loss:   0.74, Val Acc:  84.36%, Time: 0:35:25
Iter:  21000, Train Loss:   0.11, Train Acc:  97.66%, Val Loss:   0.72, Val Acc:  83.92%, Time: 0:36:18
Iter:  22000, Train Loss:   0.15, Train Acc:  93.75%, Val Loss:   0.65, Val Acc:  83.98%, Time: 0:37:11
Epoch: 6
Iter:  23000, Train Loss:  0.056, Train Acc:  96.88%, Val Loss:   0.79, Val Acc:  84.31%, Time: 0:38:05
Iter:  24000, Train Loss:  0.071, Train Acc:  97.66%, Val Loss:   0.75, Val Acc:  83.90%, Time: 0:38:57
No optimization for a long time, auto-stopping...
```

Figure 4. Training results

3.3. Test of the model

By testing on the test sets, the following results were obtained: Test Loss: 0.28, Test Acc: 92.24%

The classification effect is shown in Table3.

Table 3. Classification Results

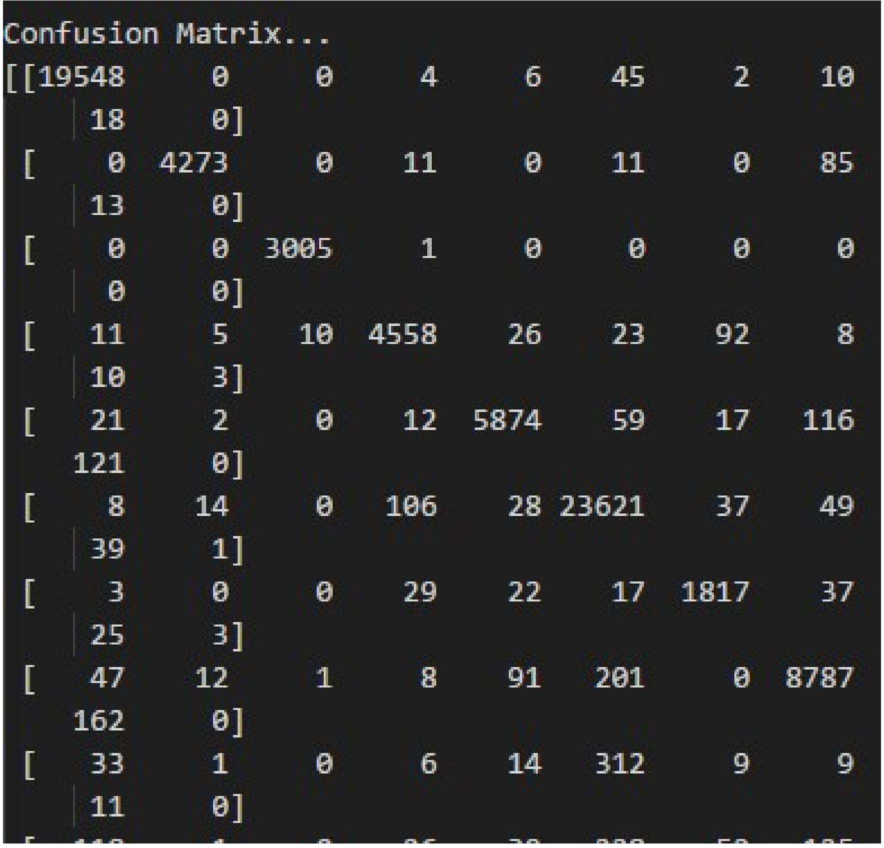
Classification	Precision	Recall	F1-score	Support
Sports	0.98	0.99	0.99	19741
Finance	0.83	0.77	0.8	5565
Property	1	1	1	3007
Household	0.94	0.93	0.94	4888
Education	0.93	0.93	0.93	6290
Technology	0.91	0.97	0.94	24439
Fashion	0.88	0.91	0.89	2005
Politics	0.84	0.93	0.88	9463
Game	0.9	0.87	0.88	3656
Entertainment	0.95	0.94	0.94	13895
Stock	0.93	0.87	0.9	23160
Lottery ticket	0.97	0.85	0.91	1138
Society	0.91	0.84	0.87	7627
Constellation	0.98	0.87	0.92	537
Avg / total	0.92	0.92	0.92	125411

The confusion matrix is shown in Figure 5.

The results show that the accuracy rate on the test sets reached 92.24%. Excluding the low F1-Score of finance and economics, the other classifications were all over 87%. From the confusion matrix, it can be seen that the classification effect is obvious.

3.4. System integration

Django is a large and comprehensive web application framework written in Python. This paper uses Django to integrate reptiles, convolutional neural network news classifiers, and web front ends. It sets the crawler to crawl once every whole point, the crawler runs over, and sends the crawled results to the neural network for prediction and stores the results. The effect is shown in Table 4.



```
Confusion Matrix...
[[19548     0     0     4     6    45     2    10
     18     0]
 [    0  4273     0    11     0    11     0    85
     13     0]
 [    0     0  3005     1     0     0     0     0
      0     0]
 [   11     5    10  4558    26    23    92     8
     10     3]
 [   21     2     0    12  5874    59    17   116
    121     0]
 [    8    14     0   106    28 23621    37    49
     39     1]
 [    3     0     0    29    22    17  1817    37
     25     3]
 [   47    12     1     8    91   201     0  8787
    162     0]
 [   33     1     0     6    14   312     9     9
     11     0]
```

Figure 5. Result confusion matrix

Table 4. System Display

News hotspot	Classification	Time
African students are fainting in Hubei heatstroke	Fashion	2018.06.30 14:00
The first prisoner of the 18th National Congress was full of corruption	Politics	2018.06.30 14:00
The man pretending to be a captain cheated, and lied that he would be the secretary of the municipal party committee.	Society	2018.06.30 14:00
Taxation education housing expenditure can be deducted before tax	Politics	2018.06.30 14:00
The daughter of the daughter of the school will receive 19 boss gifts.	Education	2018.06.30 14:00
Trump: Didn't say that I want to withdraw from the WTO, but the WTO is really unfair to us.	Entertainment	2018.06.30 14:00

4. Conclusion

In this paper, we have achievedthe online classification of Netease news clicks by Python crawler and character-level convolutional neural network, and obtained higher accuracy on the THUCNews dataset. compared with the 88.6% accuracy reached by THUCTC toolkit used the LibSVM and Liblinear classification algorithms by the Tsinghua University Natural Language Processing Laboratory, have been significantly improved. Therefore, character-level convolutional neural networks are better than traditional linear classifiers for Chinese text classification. However, the data sets used in this paper are relatively old, and it is believed that using crawler technology to crawl according to today's news classification and using the methods proposed in this project can achieve better results.

5. Acknowledgement

We gratefully acknowledge the Natural Language Proces sing Laboratory of Tsinghua University offers the THUC News for this research. The work described in this paper was fully supported by agrant from the Yan 'an Research Project (2017KG-01, 2017WZZ-04-01), and the NationalNatural Science Foundation of China (No.61761042).

References

[1] Kim Y. Convolutional Neural Networks for Sentence Classification. EprintArxiv. 2014.

[2] Lecun Y., Boser B., Denker J. S.. Backpropagation applied to handwritten zip code recognition. Neural Computation. 2014, 1, 541-551.

[3] Lécun Y., Bottou L., Bengio Y.. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998, 86, 2278-2324.

[4] Burges C. J. C., Bottou L., Welling M.. Advances in neural information processing systems 26: 27th Annual Conference on Neural Information Processing Systems 2013: December 5-10, Lake Tahoe, Nevada, USA. None. 2014.

[5] Zhang X., Zhao J., Lecun Y.. Character-level Convolutional Networks for Text Classification. 2015, 649-657.

[6] Li J. Y., Mao S.. A comparison and semi-quantitative analysis of words and character-bigrams as features in Chinese text categorization[C]// ACL 2006, International Conference on Computational Linguistics and, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July. DBLP. 2006.

[7] Li J., Sun M.. Scalable Term Selection for Text Categorization[C]// EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic. DBLP. 2007, 774-782.

[8] Lev G., Klein B., Wolf L.. In Defense of Word Embedding for Generic Text Representation[C]// International Conference on Applications of Natural Language to Information Systems. Springer, Cham. 2015, 35-50.

[9] Hinton G. E., Srivastava N., Krizhevsky A.. Improving neural networks by preventing co-adaptation of feature detectors. Computer Science. 2012, 3, 212-223.