Course Classification of Online Learning Platform Based on Sentence-Bert Model

Jiaze He[1], Qian Lu[1](✉), Ying Tong[2], and Yiyang Chen[3]

[1] Electric Power Research Institute, State Grid Jiangsu Electric Power CO. LTD, Nanjing, PR China
hzj@js.sgcc.com.cn
[2] School of Communication Engineering, Nanjing Institute of Technology, Nanjing, PR China
[3] MSc Computing and Information Technology, University of St Andrews, St Andrews, UK
chenyiyang2019@udirecter.com

Abstract. A large number of online learning platforms and the explosive growth of the types and quantity of resources on the platform greatly increase the difficulty of learning content selection. Traditional language search uses resource ranking to build index, which cannot meet the needs of professional and accurate search. An intelligent search method is proposed in this paper using Bert language model for pre-training to improve the learning and reasoning ability of the machine. Based on Sentence-Bert model and the concise and effective twin network (Siamese), the sentence vector features are generated to complete the downstream search task. Finally, experiments on the online learning resources of MOOC and the State Grid were analyzed on Google Colab platform, and results show that achieve rapid keyword matching of relevant courseware names.

Keywords: Natural language processing · BERT · Short text · Semantic search · Sentence embedding

1 Introduction

The number of resources of online learning platforms is growing rapidly with the rapid development of the Internet. At the same time, people's demand for more intelligent retrieval of course information on learning platforms is also increasing day by day [1]. With the explosive spread of information, there is an urgent need for effective information screening methods. While searching for relevant information quickly and accurately, discover hidden and higher-value information from the data. In this case, various intelligent search technologies, especially artificial intelligence technologies have been vigorously developed and widely used [2].

Natural language processing in intelligent search is dedicated to the ability of machines to understand and generate human languages. The ultimate goal is to make computers or machines as intelligent as humans in understanding languages [3]. Natural language processing has two research branches, one is based on grammatical rules, and the other is based on probability. Probability-based research methods With the popularity

© Springer Nature Singapore Pte Ltd. 2021
Y. Tian et al. (Eds.): ICBDS 2020, CCIS 1415, pp. 612–623, 2021.
https://doi.org/10.1007/978-981-16-3150-4_50

of finite state machines and empirical methods, coupled with the improvement of computer storage capacity and computing speed, natural language processing has expanded from a few application fields such as machine translation earlier to more fields, Such as information extraction and information retrieval. Various processing techniques based on different rules have also been integrated and used by researchers. The continuous creation of a variety of corpora based on statistics, examples and rules has injected a lot of vitality into the research of natural language processing [4]. Although my country's research on natural language processing started late, the current gap with international standards has been narrowing. Corpora and knowledge bases corresponding to Chinese are constantly being built, and advanced research results on semantic segmentation and syntactic analysis have also been developed. It keeps emerging [5].

Compared with English text classification, the research of Chinese text classification started late, but the speed of development is extremely fast. For decades, many domestic scholars have proposed many excellent classification algorithms when studying Chinese classification, which has laid a solid foundation for the research and development of Chinese text classification. For example, Li Xiaoli and Liu Jimin of the Institute of Computing Technology of the Chinese Academy of Sciences applied the conceptual reasoning network to text classification [8], resulting in a text recall rate of 94.2% and an accuracy rate of close to 99.4%. Literature [9] proposed a hypertext coordination classifier based on the study of KNN, Bayesian and document similarity, with an accuracy rate of close to 80%; Literature [10] studied the text classification of independent languages, and used vocabulary and category The amount of mutual information is the scoring function, considering single classification and multi-classification, so that the recall rate is 88.87%; Literature [11] combines word weights with classification algorithms, which are implemented in closed test experiments based on VSM The classification accuracy rate reaches 97%.

In 2018, the release of the BERT model [6] is considered to be the beginning of a new era in the field of natural language processing (NLP). It broke the records of many tasks in the field of natural language, and showed power in all major tasks of NLP. Gesture. In recent years, the research on Chinese text classification based on the BERT model has received extensive attention from scholars. Hu Chuntao [12] used the transfer learning strategy to apply the model to public opinion text classification tasks, Yao Liang [13] used BERT and domain-specific corpus to classify TCM clinical records, Zhang XH [14] used BERT entity extraction to extract the concept of breast cancer and its attributes; Jwa H [15] proposed exBAKE, which uses the BERT model to analyze the relationship between news headlines and content to detect fake news text. Some studies have also begun to implement pre-training models based on Chinese literature. For example, Wang Yingjie et al. [16] based on the pre-training process of BERT, mainly based on Chinese Encyclopedia, constructed a pre-trained language representation model for scientific and technological text analysis for classification experiments.

Considering that when the BERT model calculates semantic similarity, it needs to enter two sentences into the model at the same time for information exchange, which causes a lot of computational overhead. In this paper, the Sentence-BERT language model is used for pre-training, combined with the concise and effective Siamese network (Siamese), to complete the keyword matching of the generated sentence vector features.

Experimental simulations are carried out on MOOC online learning resources and State Grid online learning resources. The experimental results show that the Sentence-BERT model is better than the BERT model in matching speed and matching accuracy.

2 The Sentence-BERT Model

The Sentence-BERT model [7] is improved based on the BERT (Bidirectional Encoder Representations from Transformer) model. Although BERT and its enhanced models have achieved good results in the regression tasks of sentence pairs such as various sentence classification tasks and text semantic similarity, the excessive overhead restricts their actual application scenarios. In addition, although the BERT model can directly map the sentence vector to the vector space, and then generate a vector that can represent the semantics of the sentence through some other processing, the actual use effect is not ideal, and its own structure makes it similar to the semantics. Unsupervised degree tasks such as degree search and clustering lack applicability. Literature [7] is based on the improvement of the BERT network model and proposes the Sentence-BERT (SBERT) network structure, which uses the twin network or triple network structure to complement the advantages of BERT, so that the generated sentence embedding vector can better represent the semantics Features, and can be applied to large-scale semantic similarity comparison, clustering, and semantic information retrieval.

2.1 Sentence Vector Generation Strategy

The Sentence-BERT model defines three strategies for obtaining sentence vectors, which are mean pooling, maximum pooling and CLS vectors. 1) Mean pooling: all word vectors in the sentence are averaged, and the mean vector is used as the sentence vector of the whole sentence; 2) Maximum pooling: all word vectors in the sentence are subjected to maximum value operation, and the maximum value vector is used as The sentence vector of the whole sentence; 3) directly call the “CLS” mark in BERT as the vector representation of the sentence. Literature [7] gives the experimental results of using three sentence vector generation strategies, as shown in Table 1. It can be seen that the results of the mean strategy are the best on different data sets. Therefore, this article chooses the mean strategy to obtain the feature vector of the courseware name.

Table 1. Experimental comparison of three pooling strategies

Pooling strategy	Natural language inference data set	STS benchmark test set
MEAN	80.78	87.44
MAX	79.07	69.92
CLS	79.80	86.62

2.2 Model Objective Function

Compared with the BERT model, the Sentence-BERT model uses a twin network or a triplet network to update the initial weight parameters of the model, so as to achieve the purpose of the generated sentence embedding vector with semantics. According to different tasks, different objective functions are set.

Classification Objective Function
As can be seen from Fig. 1, the classification objective function inserts sentences passing through the Bert Model and the pooled layer into vectors **u** and **v** and the vector difference between them |**u** − **v**|, splicing them into a vector, and then multiplying it by a trainable weight parameter $W_t \in R^{3n \times k}$, where n is the dimension of the sentence vector and k is the category number. Cross-entropy loss function is used in training optimization. The classification objective function is defined as:

$$o = soft\max(W_t(u, v, |u - v|)) \tag{1}$$

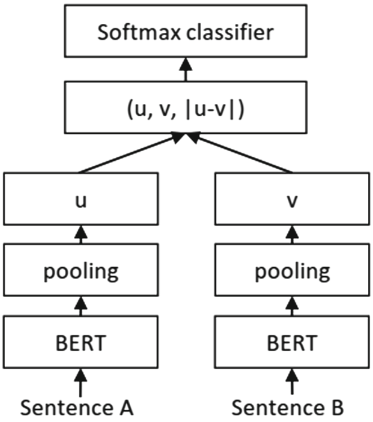


Fig. 1. Flow chart of classification objective function

Regression Objective Function
Calculate the cosine similarity of the embedding vector sum of two sentences, and the calculation structure is shown in Fig. 2. The mean square error loss function is used during training optimization.

2.3 Pre-training and Fine-Tuning of the Model

The SBERT model uses the joint data set ALLNLI, which is composed of two data sets, SNLI and MultiNLI, during pre-training. Among them, SNLI has 570,000 artificially labeled sentence pairs, and the tags are divided into three types: opposition, support and neutral; MultiNLI is an upgraded version of SNLI, which has 430,000 sentence pairs,

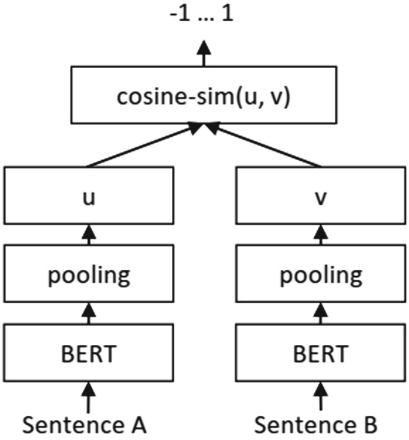


Fig. 2. Flow chart of regression objective function

mainly including spoken and written texts. The format and labels of the two data sets are uniform.

In the experiment, for each iteration, 3 types of Softmax classification objective functions are used to fine-tune SBERT. Each batch size is set to 16, and the Adam optimizer with a learning rate of 2 is used for optimization. The sentence vector generation strategy defaults to the mean strategy.

3 Method of Generating Sentence Embedding Based on SBERT

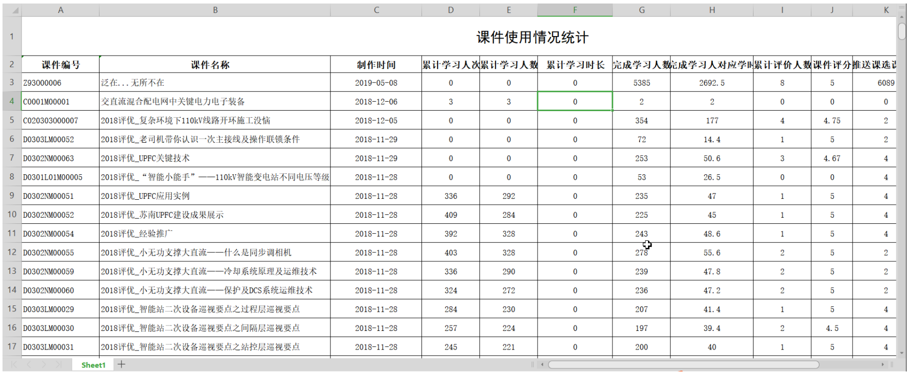
First, use an existing pre-training model such as BERT to fine-tune the data set using natural language inference and instantiate it, and map the tag vector of the sentence to the embedding layer of BERT for output. Then, for the output access pooling layer, such as the mean strategy or multiple pooling combination strategies, the sentence embedding vector is pooled. In the actual operation process, the sentence vector converter is formed by two modules, word_embedding_model and pooling_model. Each sentence is first passed through the word_embedding_model module, and then the sentence embedding vector with a fixed length is output through the pooling_model module.

Subsequently, we can specify a training data loader, use the NLIDataReader module to read the AllNLI data set, and generate a data loader suitable for training the sentence converter model. To calculate the training loss, Softmax can be used for normalized classification. These generated sentence embedding vectors can be applied to a variety of downstream tasks such as text clustering and semantic similarity analysis.

Among them, we can also specify a verification set to evaluate the sentence embedding model. The validation set can be used for testing on some invisible data. In this experiment, the validation set of the STS benchmark data set is used to evaluate the model.

4 The Experimental Results and Analysis

In this paper, data collection is carried out on the training platform of State Grid and MOOC network of China, and the training course data and MOOC online course data are obtained. In the training course data of State Grid, each course contains 13 data contents such as courseware number, courseware name, production time, and cumulative number of learners. There are a total of 1214 items of statistical course names. MOOC online courses include part of the catalogue of online MOOCs of Chinese universities from 2016 to 2018, with a total of 1352 course name information. For the training course files of State Grid, this article only needs the name information of the courseware. First use the script to take out the name information of the courseware and export it to a new text file named wenben1.txt. As for the MOOC online course catalog file, it can be directly imported into the new text corpus, named wenben2.txt. Part of the course catalog of the two data sets is shown in Figs. 3 and 4, respectively.



课件使用情况统计

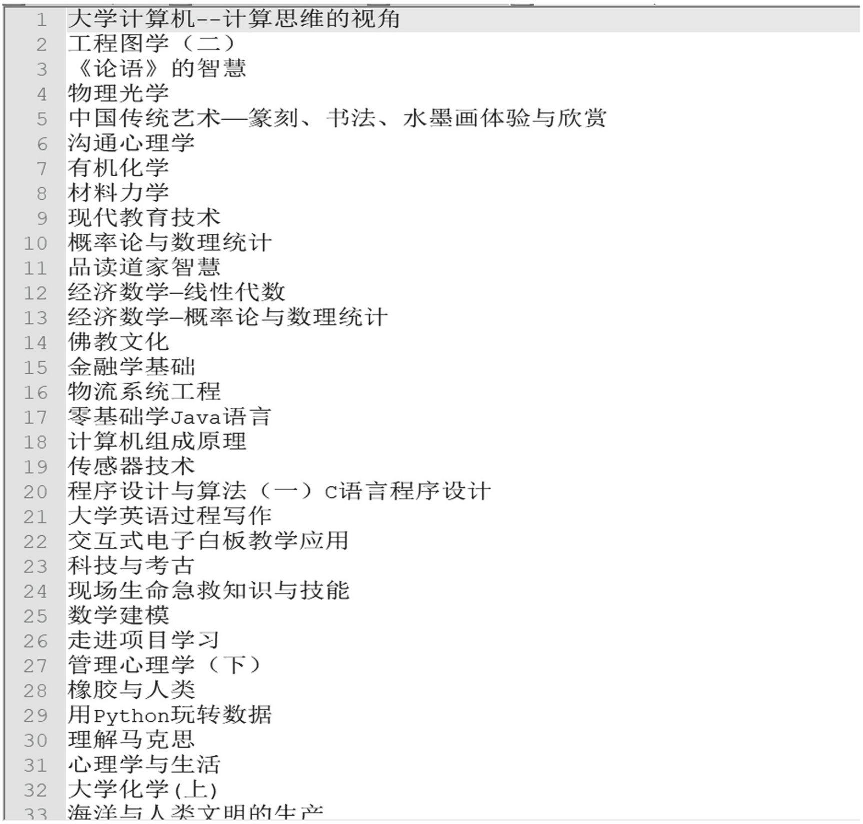
课件编号	课件名称	制作时间	累计学习人次	累计学习人数	累计学习时长	完成学习人数	完成学习人对应学时	累计评价人数	课件评分	推送课选课
Z93000006	泛在...无所不在	2019-05-08	0	0	0	5385	2692.5	8	5	6089
C0001M00001	交直流混合配电网中关键电力电子装备	2018-12-06	3	3	0	2	2	0	0	0
C020303000007	2018评优_复杂环境下110kV线路开环施工没恼	2018-12-05	0	0	0	354	177	4	4.75	2
D0303LM00052	2018评优_老司机带你认识一次主接线及操作联锁条件	2018-11-29	0	0	0	72	14.4	1	5	2
D0302NM00063	2018评优_UPFC关键技术	2018-11-29	0	0	0	253	50.6	3	4.67	4
D0301L01M00005	2018评优_"智能小能手"——110kV智能变电站不同电压等级	2018-11-28	0	0	0	53	26.5	0	0	4
D0302NM00051	2018评优_UPFC应用实例	2018-11-28	336	292	0	235	47	1	5	4
D0302NM00052	2018评优_苏南UPFC建设成果展示	2018-11-28	409	284	0	225	45	1	5	4
D0302NM00054	2018评优_经验推广	2018-11-28	392	328	0	243	48.6	1	5	4
D0302NM00055	2018评优_小无功支撑大直流——什么是同步调相机	2018-11-28	403	328	0	278	55.6	2	5	2
D0302NM00059	2018评优_小无功支撑大直流——冷却系统原理及运维技术	2018-11-28	336	290	0	239	47.8	2	5	2
D0302NM00060	2018评优_小无功支撑大直流——保护及DCS系统运维技术	2018-11-28	324	272	0	236	47.2	2	5	2
D0303LM00029	2018评优_智能站二次设备巡视要点之过程层巡视要点	2018-11-28	284	230	0	207	41.4	1	5	4
D0303LM00030	2018评优_智能站二次设备巡视要点之间隔层巡视要点	2018-11-28	257	224	0	197	39.4	2	4.5	4
D0303LM00031	2018评优_智能站二次设备巡视要点之站控层巡视要点	2018-11-28	245	221	0	200	40	1	5	4

Fig. 3. National grid training course statistics table

The classification of course names in this article belongs to short text analysis. Here, two analysis methods, K-means clustering and semantic search, are used to analyze and research the National Grid Courseware Name Text Data Set and MOOC Online Course Name Text Data Set to further illustrate the effectiveness of the Sentence-BERT model. The hardware experiment environment in this article is built and processed with the help of the Google Colab online computing platform. The software adopts transformers 2.8.0 and above, and installs tqdm and torch1.0.1 and above.

4.1 K-means Text Clustering Experiment Results and Analysis

First, calculate the sentence embedding vector for each course name short sentence in the National Grid courseware text database (wenben1.txt) and MOOC online course text database (wenben2.txt); then, use the first sentence in the python3 package library. The tripartite machine learning module sklearn performs K-means clustering (K is the number of self-clusters, which can be set to different values). Among them, the sentence

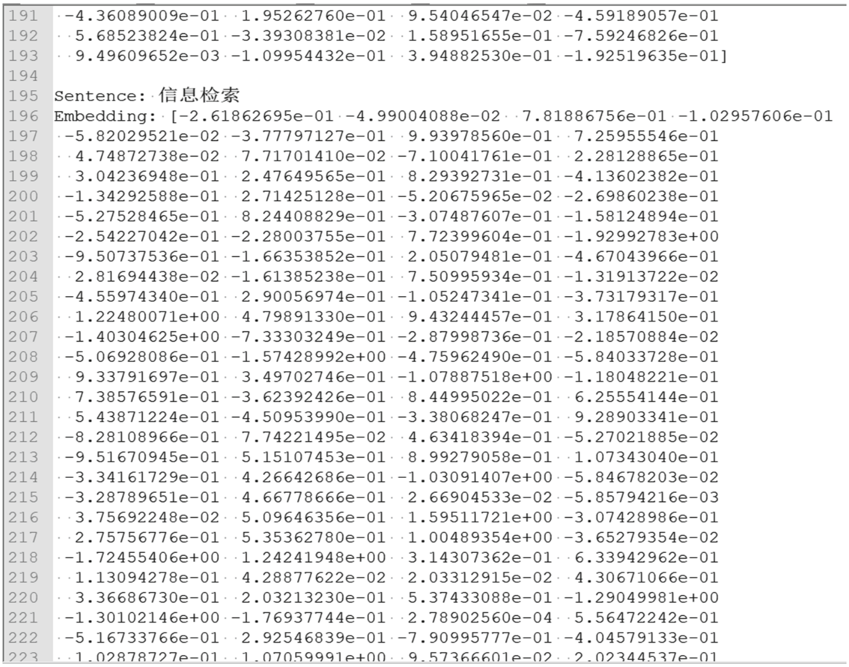


大学计算机--计算思维的视角
工程图学（二）
《论语》的智慧
物理光学
中国传统艺术—篆刻、书法、水墨画体验与欣赏
沟通心理学
有机化学
材料力学
现代教育技术
概率论与数理统计
品读道家智慧
经济数学–线性代数
经济数学–概率论与数理统计
佛教文化
金融学基础
物流系统工程
零基础学Java语言
计算机组成原理
传感器技术
程序设计与算法（一）C语言程序设计
大学英语过程写作
交互式电子白板教学应用
科技与考古
现场生命急救知识与技能
数学建模
走进项目学习
管理心理学（下）
橡胶与人类
用Python玩转数据
理解马克思
心理学与生活
大学化学(上)
海洋与人类文明的生产

Fig. 4. Statistics of MOOC online courses

embedding vector is generated by using a trained Sentence Transformer model, and a specific Numpy array containing 768-dimensional embedding is generated corresponding to each sentence information, as shown in Fig. 5.

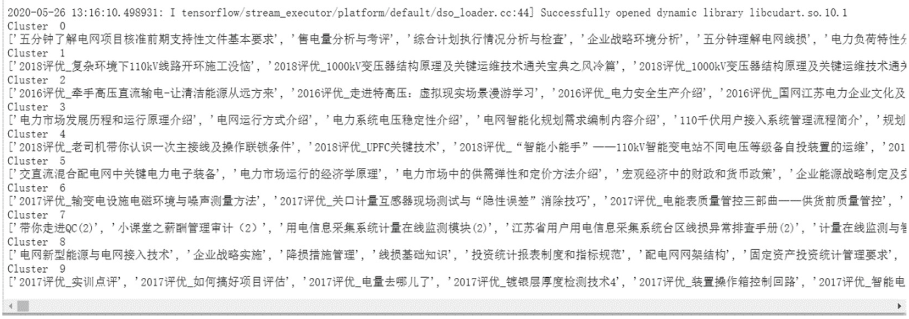
For the national grid courseware text database and MOOC online course text database, K values were set to 10, 20, and 50 respectively for cluster analysis and comparison experiments. Some experimental results are shown in the figure below. It can be seen from the experimental results that the larger the value K of K-means text clustering is selected, the finer the division of each group of clusters, but too large a value will also make the grouping confused. tendency. This is because the basis of clustering depends entirely on the sentence embedding vector generated above. If the amount of training data in the previous period is not very sufficient, the information will not be fully reflected in the embedded information, and ultimately result in insufficient differentiation density (Figs. 6, 7 and 8).



```
 -4.36089009e-01  1.95262760e-01  9.54046547e-02 -4.59189057e-01
  5.68523824e-01 -3.39308381e-02  1.58951655e-01 -7.59246826e-01
  9.49609652e-03 -1.09954432e-01  3.94882530e-01 -1.92519635e-01]

Sentence: 信息检索
Embedding: [-2.61862695e-01 -4.99004088e-02  7.81886756e-01 -1.02957606e-01
 -5.82029521e-02 -3.77797127e-01  9.93978560e-01  7.25955546e-01
  4.74872738e-02  7.71701410e-02 -7.10041761e-01  2.28128865e-01
  3.04236948e-01  2.47649565e-01  8.29392731e-01 -4.13602382e-01
 -1.34292588e-01  2.71425128e-01 -5.20675965e-02 -2.69860238e-01
 -5.27528465e-01  8.24408829e-01 -3.07487607e-01 -1.58124894e-01
 -2.54227042e-01 -2.28003755e-01  7.72399604e-01 -1.92992783e+00
 -9.50737536e-01 -1.66353852e-01  2.05079481e-01 -4.67043966e-01
  2.81694438e-02 -1.61385238e-01  7.50995934e-01 -1.31913722e-02
 -4.55974340e-01  2.90056974e-01 -1.05247341e-01 -3.73179317e-01
  1.22480071e+00  4.79891330e-01  9.43244457e-01  3.17864150e-01
 -1.40304625e+00 -7.33303249e-01 -2.87998736e-01 -2.18570884e-02
 -5.06928086e-01 -1.57428992e+00 -4.75962490e-01 -5.84033728e-01
  9.33791697e-01  3.49702746e-01 -1.07887518e+00 -1.18048221e-01
  7.38576591e-01 -3.62392426e-01  8.44995022e-01  6.25554144e-01
  5.43871224e-01 -4.50953990e-01 -3.38068247e-01  9.28903341e-01
 -8.28108966e-01  7.74221495e-02  4.63418394e-01 -5.27021885e-02
 -9.51670945e-01  5.15107453e-01  8.99279058e-01  1.07343040e-01
 -3.34161729e-01  4.26642686e-01 -1.03091407e+00 -5.84678203e-02
 -3.28789651e-01  4.66778666e-01  2.66904533e-02 -5.85794216e-03
  3.75692248e-02  5.09646356e-01  1.59511721e+00 -3.07428986e-01
  2.75756776e-01  5.35362780e-01  1.00489354e+00 -3.65279354e-02
 -1.72455406e+00  1.24241948e+00  3.14307362e-01  6.33942962e-01
  1.13094278e-01  4.28877622e-02  2.03312915e-02  4.30671066e-01
  3.36686730e-01  2.03213230e-01  5.37433088e-01 -1.29049981e+00
 -1.30102146e+00 -1.76937744e-01  2.78902560e-04  5.56472242e-01
 -5.16733766e-01  2.92546839e-01 -7.90995777e-01 -4.04579133e-01
  1.02878727e-01  1.07059991e+00  9.57366601e-02  2.02344537e-01
```

Fig. 5. The generated sentence embedding vector array

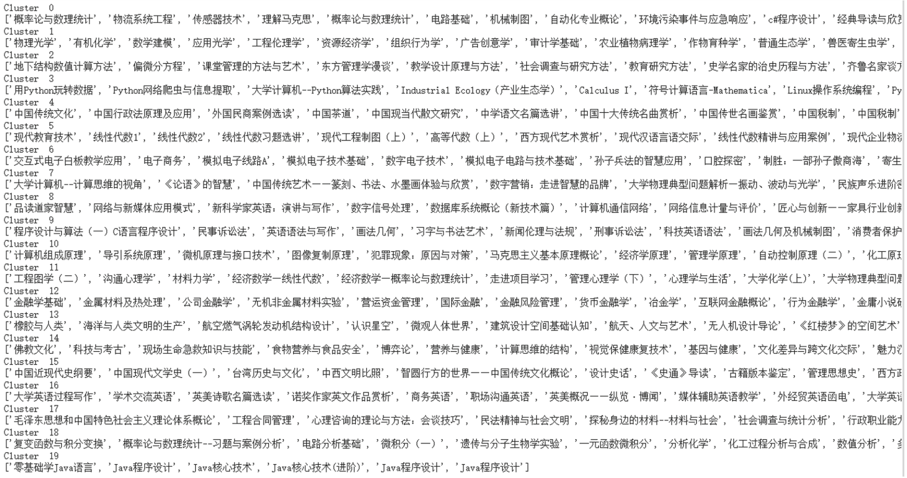


```
2020-05-26 13:16:10.498931: I tensorflow/stream_executor/platform/default/dso_loader.cc:44] Successfully opened dynamic library libcudart.so.10.1
Cluster  0
['五分钟了解电网项目核准前期支持性文件基本要求', '售电量分析与考评', '综合计划执行情况分析与检查', '企业战略环境分析', '五分钟理解电网线损', '电力负荷特性
Cluster  1
['2018评优_复杂环境下110kV线路开环施工没恼', '2018评优_1000kV变压器结构原理及关键运维技术通关宝典之风冷篇', '2018评优_1000kV变压器结构原理及关键运维技术通
Cluster  2
['2016评优_牵手高压直流输电-让清洁能源从远方来', '2016评优_走进特高压：虚拟现实场景漫游学习', '2016评优_电力安全生产介绍', '2016评优_国网江苏电力企业文化及
Cluster  3
['电力市场发展历程和运行原理介绍', '电网运行方式介绍', '电力系统电压稳定性介绍', '电网智能化规划需求编制内容介绍', '110千伏用户接入系统管理流程简介', '规划
Cluster  4
['2018评优_老司机带你认识一次主接线及操作联锁条件', '2018评优_UPFC关键技术', '2018评优_“智能小能手”——110kV智能变电站不同电压等级备自投装置的运维', '201
Cluster  5
['交直流混合配电网中关键电力电子装备', '电力市场运行的经济学原理', '电力市场中的供需弹性和定价方法介绍', '宏观经济中的财政和货币政策', '企业能源战略制定及
Cluster  6
['2017评优_输变电设施电磁环境与噪声测量方法', '2017评优_关口计量互感器现场测试与“隐性误差”消除技巧', '2017评优_电能表质量管控三部曲——供货前质量管控', '
Cluster  7
['带你走进QC(2)', '小课堂之薪酬管理审计（2）', '用电信息采集系统计量在线监测模块(2)', '江苏省用户用电信息采集系统台区线损异常排查手册(2)', '计量在线监测与
Cluster  8
['电网新型能源与电网接入技术', '企业战略实施', '降损措施管理', '线损基础知识', '投资统计报表制度和指标规范', '配电网网架结构', '固定资产投资统计管理要求',
Cluster  9
['2017评优_实训点评', '2017评优_如何搞好项目评估', '2017评优_电量去哪儿了', '2017评优_镀银层厚度检测技术4', '2017评优_装置操作箱控制回路', '2017评优_智能电
```

Fig. 6. 10 clustering results of the National Grid courseware text library

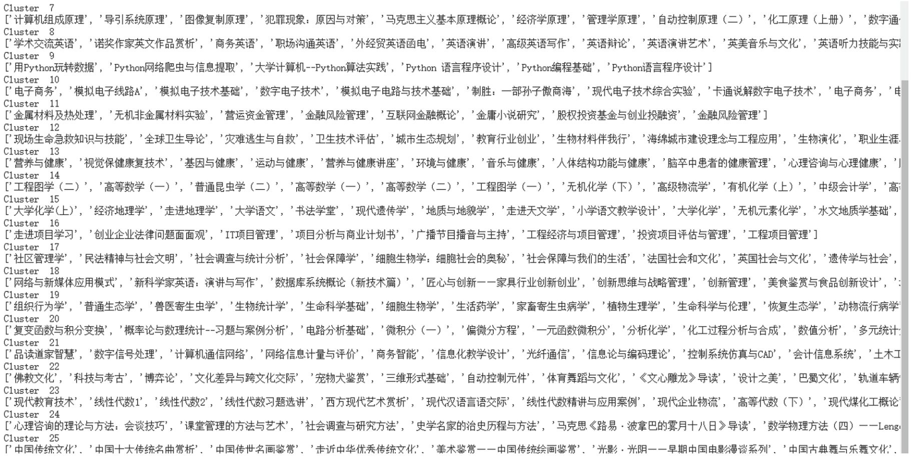
4.2 Sentence-BERT Semantic Search Experimental Results and Analysis

Semantic search is the task of finding sentences that are similar to a given sentence. As with cluster analysis, all sentences in the Corpus are embedded with the corresponding sentences, then the input query sentences are embedded with the same method, the Scipy toolkit in Python is used to search the Numpy array of the Corpus to retrieve the content most similar to the query statement, and display the first 8 results.



```
Cluster  0
['概率论与数理统计', '物流系统工程', '传感器技术', '理解马克思', '概率论与数理统计', '电路基础', '机械制图', '自动化专业概论', '环境污染事件与应急响应', 'c#程序设计', '经典导读与欣
Cluster  1
['物理光学', '有机化学', '数学建模', '应用光学', '工程伦理学', '资源经济学', '组织行为学', '广告创意学', '审计学基础', '农业植物病理学', '作物育种学', '普通生态学', '兽医寄生虫学',
Cluster  2
['地下结构数值计算方法', '偏微分方程', '课堂管理的方法与艺术', '东方管理学漫谈', '教学设计原理与方法', '社会调查与研究方法', '教育研究方法', '史学名家的治史历程与方法', '齐鲁名家谈
Cluster  3
['用Python玩转数据', 'Python网络爬虫与信息提取', '大学计算机--Python算法实践', 'Industrial Ecology（产业生态学）', 'Calculus I', '符号计算语言-Mathematica', 'Linux操作系统编程', 'Py
Cluster  4
['中国传统文化', '中国行政法原理及应用', '外国民商案例选读', '中国茶道', '中国现当代散文研究', '中学语文名篇选讲', '中国十大传统名曲赏析', '中国传世名画鉴赏', '中国税制', '中国税制'
Cluster  5
['现代教育技术', '线性代数1', '线性代数2', '线性代数习题选讲', '现代工程制图（上）', '高等代数（上）', '西方现代艺术赏析', '现代汉语言语交际', '线性代数精讲与应用案例', '现代企业物
Cluster  6
['交互式电子白板教学应用', '电子商务', '模拟电子线路A', '模拟电子技术基础', '数字电子技术', '模拟电子电路与技术基础', '孙子兵法的智慧应用', '口腔探密', '制胜：一部孙子傲商海', '寄生
Cluster  7
['大学计算机--计算思维的视角', '《论语》的智慧', '中国传统艺术——篆刻、书法、水墨画体验与欣赏', '数字营销：走进智慧的品牌', '大学物理典型问题解析—振动、波动与光学', '民族声乐进阶
Cluster  8
['品读道家智慧', '网络与新媒体应用模式', '新科学家英语：演讲与写作', '数字信号处理', '数据库系统概论（新技术篇）', '计算机通信网络', '网络信息计量与评价', '匠心与创新——家具行业创
Cluster  9
['程序设计与算法（一）C语言程序设计', '民事诉讼法', '英语语法与写作', '画法几何', '习字与书法艺术', '新闻伦理与法规', '刑事诉讼法', '科技英语语法', '画法几何及机械制图', '消费者保护
Cluster  10
['计算机组成原理', '导引系统原理', '微机原理与接口技术', '图像复制原理', '犯罪现象：原因与对策', '马克思主义基本原理概论', '经济学原理', '管理学原理', '自动控制原理（二）', '化工原
Cluster  11
['工程图学（二）', '沟通心理学', '材料力学', '经济数学—线性代数', '经济数学—概率论与数理统计', '走进项目学习', '管理心理学（下）', '心理学与生活', '大学化学(上)', '大学物理典型问
Cluster  12
['金融学基础', '金属材料及热处理', '公司金融学', '无机非金属材料实验', '营运资金管理', '国际金融', '金融风险管理', '货币金融学', '冶金学', '互联网金融概论', '行为金融学', '金庸小说研
Cluster  13
['橡胶与人类', '海洋与人类文明的生产', '航空燃气涡轮发动机结构设计', '认识星空', '微观人体世界', '建筑设计空间基础认知', '航天、人文与艺术', '无人机设计导论', '《红楼梦》的空间艺术'
Cluster  14
['佛教文化', '科技与考古', '现场生命急救知识与技能', '食物营养与食品安全', '博弈论', '营养与健康', '计算思维的结构', '视觉保健康复技术', '基因与健康', '文化差异与跨文化交际', '魅力
Cluster  15
['中国近现代史纲要', '中国现代文学史（一）', '台湾历史与文化', '中西文明比照', '智圆行方的世界——中国传统文化概论', '设计史话', '《史通》导读', '古籍版本鉴定', '管理思想史', '西方
Cluster  16
['大学英语过程写作', '学术交流英语', '英美诗歌名篇选读', '诺奖作家英文作品赏析', '商务英语', '职场沟通英语', '英美概况——纵览·博闻', '媒体辅助英语教学', '外经贸英语函电', '大学英
Cluster  17
['毛泽东思想和中国特色社会主义理论体系概论', '工程合同管理', '心理咨询的理论与方法：会谈技巧', '民法精神与社会文明', '探秘身边的材料--材料与社会', '社会调查与统计分析', '行政职业能
Cluster  18
['复变函数与积分变换', '概率论与数理统计--习题与案例分析', '电路分析基础', '微积分（一）', '遗传与分子生物学实验', '一元函数微积分', '分析化学', '化工过程分析与合成', '数值分析', '
Cluster  19
['零基础学Java语言', 'Java程序设计', 'Java核心技术', 'Java核心技术(进阶)', 'Java程序设计', 'Java程序设计']
```

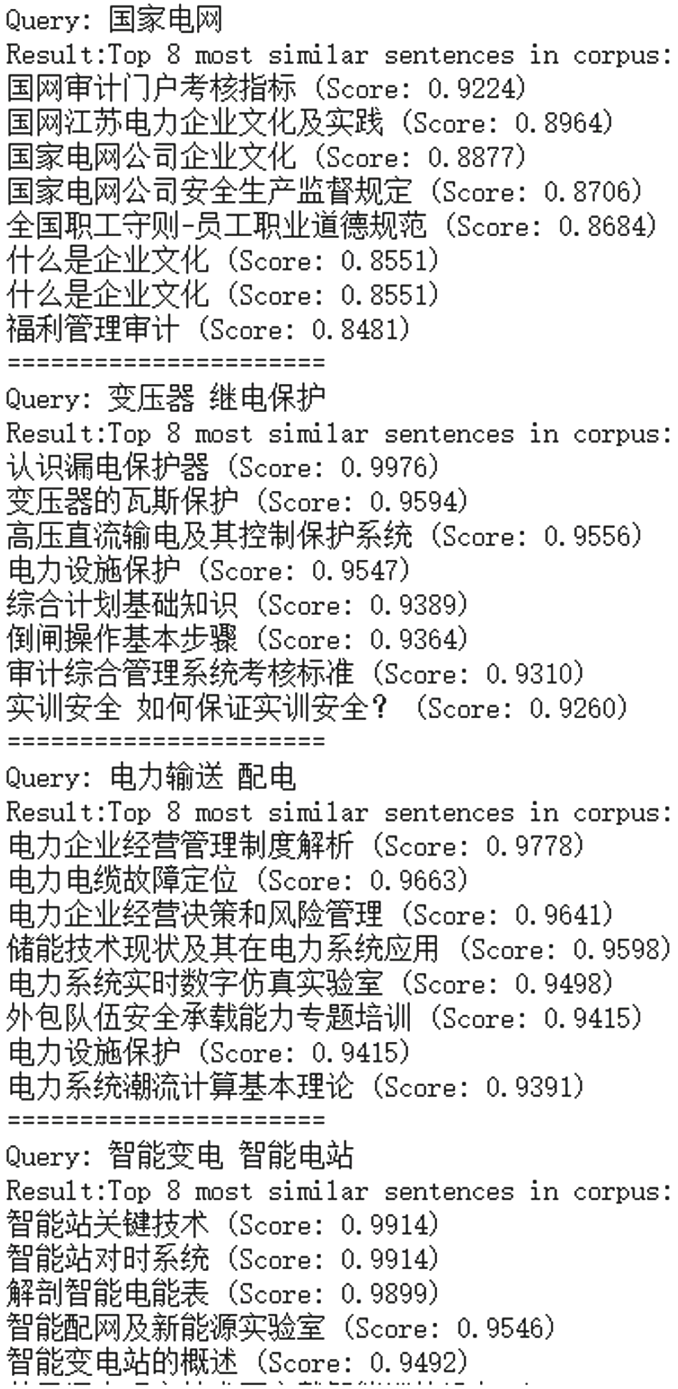
Fig. 7. 20 clustering results of MOOC online course text library



```
Cluster  7
['计算机组成原理', '导引系统原理', '图像复制原理', '犯罪现象：原因与对策', '马克思主义基本原理概论', '经济学原理', '管理学原理', '自动控制原理（二）', '化工原理（上册）', '数字通
Cluster  8
['学术交流英语', '诺奖作家英文作品赏析', '商务英语', '职场沟通英语', '外经贸英语函电', '英语演讲', '高级英语写作', '英语辩论', '英语演讲艺术', '英美音乐与文化', '英语听力技能与实
Cluster  9
['用Python玩转数据', 'Python网络爬虫与信息提取', '大学计算机--Python算法实践', 'Python 语言程序设计', 'Python编程基础', 'Python语言程序设计']
Cluster  10
['电子商务', '模拟电子线路A', '模拟电子技术基础', '数字电子技术', '模拟电子电路与技术基础', '制胜：一部孙子傲商海', '现代电子技术综合实验', '卡通说解数字电子技术', '电子商务', '电
Cluster  11
['金属材料及热处理', '无机非金属材料实验', '营运资金管理', '金融风险管理', '互联网金融概论', '金庸小说研究', '股权投资基金与创业投融资', '金融风险管理']
Cluster  12
['现场生命急救知识与技能', '全球卫生导论', '灾难逃生与自救', '卫生技术评估', '城市生态规划', '教育行业创业', '生物材料伴我行', '海绵城市建设理念与工程应用', '生物演化', '职业生涯
Cluster  13
['营养与健康', '视觉保健康复技术', '基因与健康', '运动与健康', '营养与健康讲座', '环境与健康', '音乐与健康', '人体结构功能与健康', '脑卒中患者的健康管理', '心理咨询与心理健康', '
Cluster  14
['工程图学（二）', '高等数学（一）', '普通昆虫学（二）', '高等数学（一）', '高等数学（二）', '工程图学（一）', '无机化学（下）', '高级物流学', '有机化学（上）', '中级会计学', '高
Cluster  15
['大学化学(上)', '经济地理学', '走进地理学', '大学语文', '书法学堂', '现代遗传学', '地质与地貌学', '走进天文学', '小学语文教学设计', '大学化学', '无机元素化学', '水文地质学基础',
Cluster  16
['走进项目学习', '创业企业法律问题面面观', 'IT项目管理', '项目分析与商业计划书', '广播节目播音与主持', '工程经济与项目管理', '投资项目评估与管理', '工程项目管理']
Cluster  17
['社区管理学', '民法精神与社会文明', '社会调查与统计分析', '社会保障学', '细胞生物学：细胞社会的奥秘', '社会保障与我们的生活', '法国社会和文化', '英国社会与文化', '遗传学与社会',
Cluster  18
['网络与新媒体应用模式', '新科学家英语：演讲与写作', '数据库系统概论（新技术篇）', '匠心与创新——家具行业创新创业', '创新思维与战略管理', '创新管理', '美食鉴赏与食品创新设计', '
Cluster  19
['组织行为学', '普通生态学', '兽医寄生虫学', '生物统计学', '生命科学基础', '细胞生物学', '生活药学', '家畜寄生虫病学', '植物生理学', '生命科学与伦理', '恢复生态学', '动物流行病学
Cluster  20
['复变函数与积分变换', '概率论与数理统计--习题与案例分析', '电路分析基础', '微积分（一）', '偏微分方程', '一元函数微积分', '分析化学', '化工过程分析与合成', '数值分析', '多元统计
Cluster  21
['品读道家智慧', '数字信号处理', '计算机通信网络', '网络信息计量与评价', '商务智能', '信息化教学设计', '光纤通信', '信息论与编码理论', '控制系统仿真与CAD', '会计信息系统', '土木工
Cluster  22
['佛教文化', '科技与考古', '博弈论', '文化差异与跨文化交际', '宠物犬鉴赏', '三维形式基础', '自动控制元件', '体育舞蹈与文化', '《文心雕龙》导读', '设计之美', '巴蜀文化', '轨道车辆
Cluster  23
['现代教育技术', '线性代数1', '线性代数2', '线性代数习题选讲', '西方现代艺术赏析', '现代汉语言语交际', '线性代数精讲与应用案例', '现代企业物流', '高等代数（下）', '现代煤化工概论
Cluster  24
['心理咨询的理论与方法：会谈技巧', '课堂管理的方法与艺术', '社会调查与研究方法', '史学名家的治史历程与方法', '马克思《路易·波拿巴的雾月十八日》导读', '数学物理方法（四）——Leng
Cluster  25
['中国传统文化', '中国十大传统名曲赏析', '中国传世名画鉴赏', '走近中华优秀传统文化', '美术鉴赏——中国传统绘画鉴赏', '光影·光阴——早期中国电影漫谈系列', '中国古典舞与乐舞文化',
```

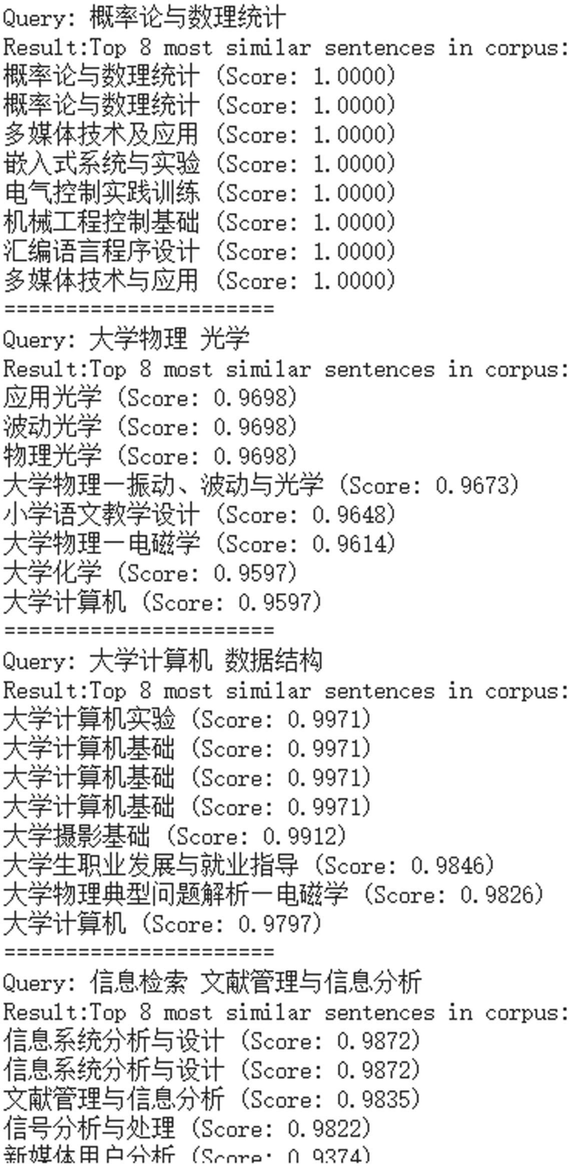
Fig. 8. 50 clustering results of MOOC online course text library

In the same way, semantic search experiments are carried out on the State Grid courseware text library and MOOC online course text library, and the results are shown in Figs. 9 and 10. It can be seen from the figure that, for each query sentence, whether it is a single text or a combination of multiple texts, when calculating the embedded vector of the query sentence, it is treated as a single line of text data for processing and calculation. Under each query sentence, the score is calculated according to the semantic similarity. The higher the score, the more similar the semantics of the query sentence of the retrieval text, and the 8 closest retrieval information texts are given according to the score.



```
Query: 国家电网
Result:Top 8 most similar sentences in corpus:
国网审计门户考核指标（Score: 0.9224）
国网江苏电力企业文化及实践（Score: 0.8964）
国家电网公司企业文化（Score: 0.8877）
国家电网公司安全生产监督规定（Score: 0.8706）
全国职工守则-员工职业道德规范（Score: 0.8684）
什么是企业文化（Score: 0.8551）
什么是企业文化（Score: 0.8551）
福利管理审计（Score: 0.8481）
======================
Query: 变压器 继电保护
Result:Top 8 most similar sentences in corpus:
认识漏电保护器（Score: 0.9976）
变压器的瓦斯保护（Score: 0.9594）
高压直流输电及其控制保护系统（Score: 0.9556）
电力设施保护（Score: 0.9547）
综合计划基础知识（Score: 0.9389）
倒闸操作基本步骤（Score: 0.9364）
审计综合管理系统考核标准（Score: 0.9310）
实训安全 如何保证实训安全？（Score: 0.9260）
======================
Query: 电力输送 配电
Result:Top 8 most similar sentences in corpus:
电力企业经营管理制度解析（Score: 0.9778）
电力电缆故障定位（Score: 0.9663）
电力企业经营决策和风险管理（Score: 0.9641）
储能技术现状及其在电力系统应用（Score: 0.9598）
电力系统实时数字仿真实验室（Score: 0.9498）
外包队伍安全承载能力专题培训（Score: 0.9415）
电力设施保护（Score: 0.9415）
电力系统潮流计算基本理论（Score: 0.9391）
======================
Query: 智能变电 智能电站
Result:Top 8 most similar sentences in corpus:
智能站关键技术（Score: 0.9914）
智能站对时系统（Score: 0.9914）
解剖智能电能表（Score: 0.9899）
智能配网及新能源实验室（Score: 0.9546）
智能变电站的概述（Score: 0.9492）
```

Fig. 9. Semantic search test results of the State Grid courseware text library



```
Query: 概率论与数理统计
Result:Top 8 most similar sentences in corpus:
概率论与数理统计（Score: 1.0000)
概率论与数理统计（Score: 1.0000)
多媒体技术及应用（Score: 1.0000)
嵌入式系统与实验（Score: 1.0000)
电气控制实践训练（Score: 1.0000)
机械工程控制基础（Score: 1.0000)
汇编语言程序设计（Score: 1.0000)
多媒体技术与应用（Score: 1.0000)
======================
Query: 大学物理 光学
Result:Top 8 most similar sentences in corpus:
应用光学（Score: 0.9698)
波动光学（Score: 0.9698)
物理光学（Score: 0.9698)
大学物理一振动、波动与光学（Score: 0.9673)
小学语文教学设计（Score: 0.9648)
大学物理一电磁学（Score: 0.9614)
大学化学（Score: 0.9597)
大学计算机（Score: 0.9597)
======================
Query: 大学计算机 数据结构
Result:Top 8 most similar sentences in corpus:
大学计算机实验（Score: 0.9971)
大学计算机基础（Score: 0.9971)
大学计算机基础（Score: 0.9971)
大学计算机基础（Score: 0.9971)
大学摄影基础（Score: 0.9912)
大学生职业发展与就业指导（Score: 0.9846)
大学物理典型问题解析一电磁学（Score: 0.9826)
大学计算机（Score: 0.9797)
======================
Query: 信息检索 文献管理与信息分析
Result:Top 8 most similar sentences in corpus:
信息系统分析与设计（Score: 0.9872)
信息系统分析与设计（Score: 0.9872)
文献管理与信息分析（Score: 0.9835)
信号分析与处理（Score: 0.9822)
新媒体用户分析（Score: 0.9374)
```

Fig. 10. Semantic search test results of MOOC online course text library

Comparing the two sets of semantic search test results, it is found that it is indeed possible to perform semantic matching calculations based on the sentence semantic vector rather than the character matching degree of the phrase in the sentence. However, as the experimental results show, the results of the semantic search are good or bad. The analysis believes that there are two possible reasons: first, the input training corpus text has a low degree of correlation with the text tested in the experiment, and the machine does not learn enough about the information contained in the sentence semantics; second, query and retrieval The amount of sentences in the text is small and the distribution is uneven, resulting in errors in semantic matching.

5 Conclusion

This paper presents an intelligent search method based on natural language processing technology. Sentence-BERT language model is used for pre-training to improve the learning and reasoning ability of the machine. Based on the feature of sentence vector generated by twin network and the embedded vector generated by twin network, clustering and semantic search are carried out respectively. In the Google Colab platform for the two tasks of the application of experimental analysis, achieved short-text intelligent search requirements.

References

1. Wang, Z.: Research and implementation of text categorization based on machine learning. Nanjing University of Posts and Telecommunications (2018). (Chinese)
2. Yang, T.: Application of artificial intelligence technology in information retrieval based on big data. Value Eng. **38**(10), 173–175 (2019). (Chinese)
3. Khurana, D., Koli, A., Khatter, K., et al.: Natural language processing: state of the art, current trends and challenges (2017)
4. Huikun, S.: A method for computing semantic similarity in information retrieval. J. Chizhou Univ. **30**(03), 26–29 (2016)
5. Kui, L.: Research on semantic similarity computation of short tex. Harbin Engineering University (2016). (Chinese)
6. Devlin, J., Chang, M. W., Lee, K., et al.: BERT: pre-training of deep bidirectional transformers for language understanding (2018)
7. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks (2019)
8. Xiaoli, L., Jimin, L., Zhongzhi, S.: Conceptual inference network and its application in text categorization. Comput. Res. Dev. **37**(9), 1032–1038 (2009). (Chinese)
9. Yan, F., Enhong, C., Qingyi, W., et al.: Performance Study of hypertext concordance classifier. Comput. Res. Dev. **37**(9), 1026–1031 (2009). (Chinese)
10. Xuanjing, H., Lide, W., Shizaki, Y., et al.: Text classification method independent of languages. Chin. J. Inf. **14**(6), 1–7 (2000). (Chinese)
11. Qian, D., Yongcheng, W., Huihui, Z., et al.: Word weight and classification algorithm in automatic text classification. Chin. J. Inf. **14**(3), 25–29 (2000). (Chinese)
12. Chuntao, H., Jinkang, Q., Jingmei, C., et al.: Application research of public opinion classification based on BERT model. Netw. Secur. Technol. Appl. **11**, 41–44 (2019)
13. Liang, Y., Zhe, J., Chengsheng, M., et al.: Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J. Am. Med. Inf. Assoc.: JAMIA **26**(12), 12–23 (2019)
14. Xiaohui, Z., Yaoyun, Z., Qin, Z., et al.: Extracting comprehensive clinical information for breast cancer using deep learning methods. Int. J. Med. Inf. **132**, 103985 (2019)
15. Jwa, H., Oh, D., Park, K., et al.: exBAKE: automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT). Appl. Sci.-Basel **9**(19), 32–46 (2019)
16. Yingjie, W., Bin, X., Ningbo, L.: A pre-trained technological language representation for Chinese technological text analysis. Comput. Eng. 12-08 (2019)