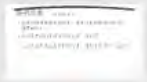
基于深度神经网络的科技文章多标签提取技术以及新知识发现

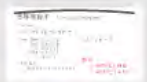
北京邮电大学 大数据与智能信息处理实验室

2018年11月

目录

- 研究背景
- 数据说明
- 多标签技术
- 新知识发现
- 应用场景讨论





基于深度神经网络的科技文章多标签提取技术以及新知识发现

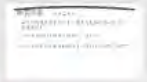
北京邮电大学 大数据与智能信息处理实验室

2018年11月

目录

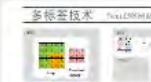
- 研究背景
- 数据说明
- 多标签技术
- 新知识发现
- 应用场景讨论



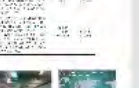


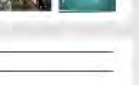










基于深度神经网络的科技文章多标签提取技术以及新知识发现

北京邮电大学 大数据与智能信息处理实验室

2018年11月

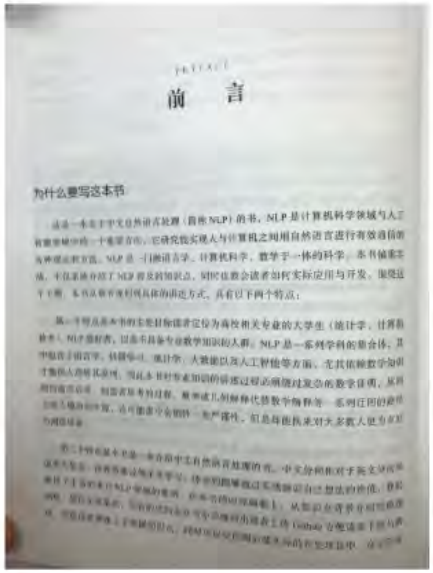
目录

目录

- 研究背景
- 数据说明
- 多标签技术
- 新知识发现
- 应用场景讨论

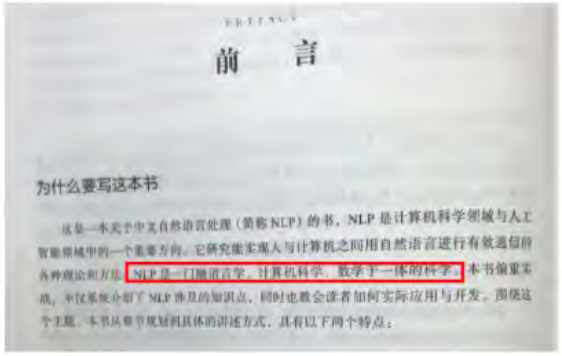
研究背景

随着信息的井喷式增长，我们接触到的不仅仅是已有传统科学分类。很多的新科学新知识伴随着不同学科的交叉应运而生。



前　言

为什么要写这本书



前　言

为什么要写这本书

这是一本关于中文自然语言处理（简称NLP）的书，NLP是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。NLP是一门融语言学、计算机科学、数学于一体的科学。本书偏重实战，全仅系统介绍了NLP涉及的知识点，同时也教会读者如何实际应用与开发。围绕这个主题，本书从章节规划到具体的讲述方式，具有以下两个特点：

分类与多标签？

PREFACE

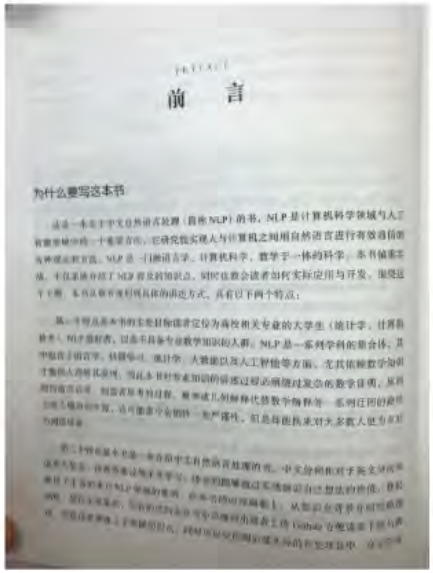
前　言

为什么要写这本书

这是一本关于中文自然语言处理（简称 NLP）的书，NLP 是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。NLP 是一门融语言学、计算机科学、数学于一体的科学。本书偏重实战，不仅系统介绍了 NLP 涉及的知识点，同时也教会读者如何实际应用与开发。围绕这个主题，本书从章节规划到具体的讲述方式，具有以下两个特点：

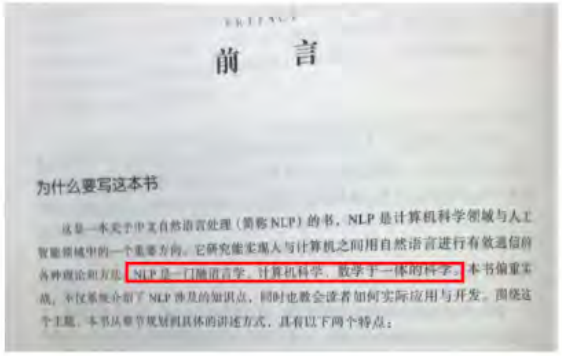
研究背景

随着信息的井喷式增长，我们接触到的不仅仅是已有传统科学分类。很多的新科学新知识伴随着不同学科的交叉应运而生。



前　言

为什么要写这本书



前　言

为什么要写这本书

这是一本关于中文自然语言处理（简称NLP）的书，NLP是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。NLP是一门融语言学、计算机科学、数学于一体的科学。本书偏重实战，全仅系统介绍了NLP涉及的知识点，同时也教会读者如何实际应用与开发。围绕这个主题，本书从章节规划到具体的讲述方式，具有以下两个特点：

分类与多标签？

研究背景　新知识与多标签

- 目前学科特点
  - 基础学科：覆盖全面，且基础知识完备。
  - 新生学科：对于基础学科的深入性研究及应用性探索，通常产生学科分支或融合各个学科，重新组合成新生学科。

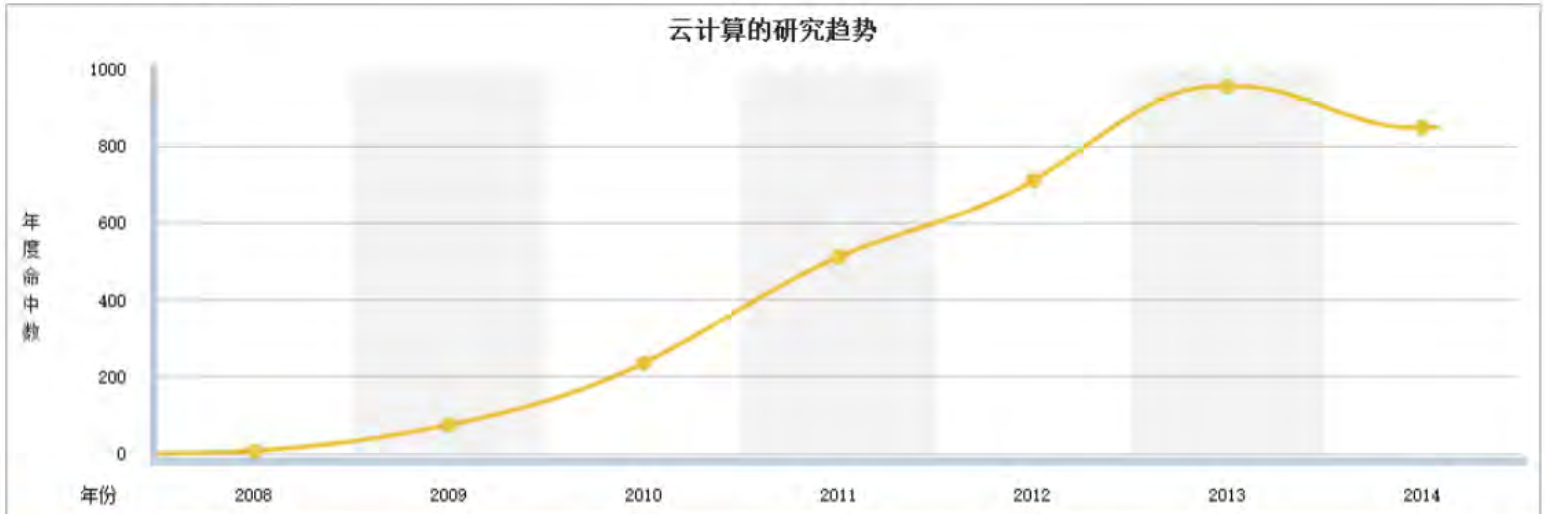
- 多标签、新知识及分类
  - 多标签：一个基础学科为一个标签，而多个标签组合为新生学科。
  - 新知识：学科交叉产生，过去没有过的多标签组合。
  - 分类：当某个新的多标签组合逐渐趋于稳定时，产生一个新的学科分类。

研究背景　　云计算为例

- 云计算领域兴起于2008年，由亚马逊推出的云计算服务开始诞生。
- 2010年至2011年被业界所认识，并采用。
- 2012年发展为比较成熟的技术，得到了业界的广泛使用。

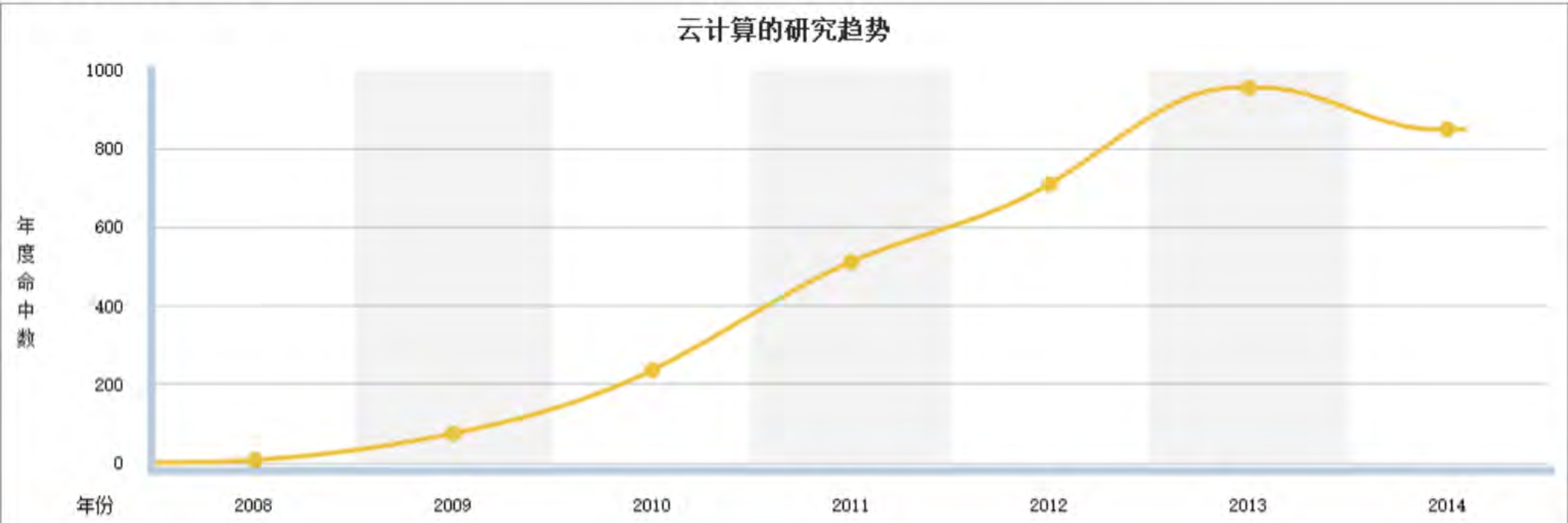
研究背景　云计算为例

- 新知识发展趋势：随着应用需求产生，伴随着相关期刊和科技文献的发表逐渐成熟。



云计算	16	165	594	1477	2265	3120	1129
热词	cloud computing(2)	saas(17)	图书馆(39)	cloud computing(158)	cloud computing(293)	cloud computing(832)	cloud computing(383)
	网络技术(2)	网格计算(15)	虚拟化(38)	虚拟化(103)	虚拟化(157)	虚拟化(188)	大数据(67)
	网格计算(2)	云安全(13)	saas(26)	物联网(90)	物联网(121)	物联网(164)	虚拟化(62)
	google api(1)	虚拟化(11)	云安全(24)	图书馆(57)	图书馆(81)	大数据(109)	物联网(59)
	安全(1)	cloud computing(10)	物联网(24)	云存储(53)	应用(77)	hadoop(108)	hadoop(54)

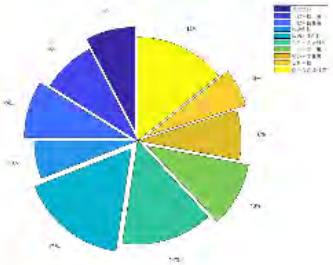
知识发展趋势：随着应用需求产生，伴随着相关期刊和科技文献
表逐渐成熟。



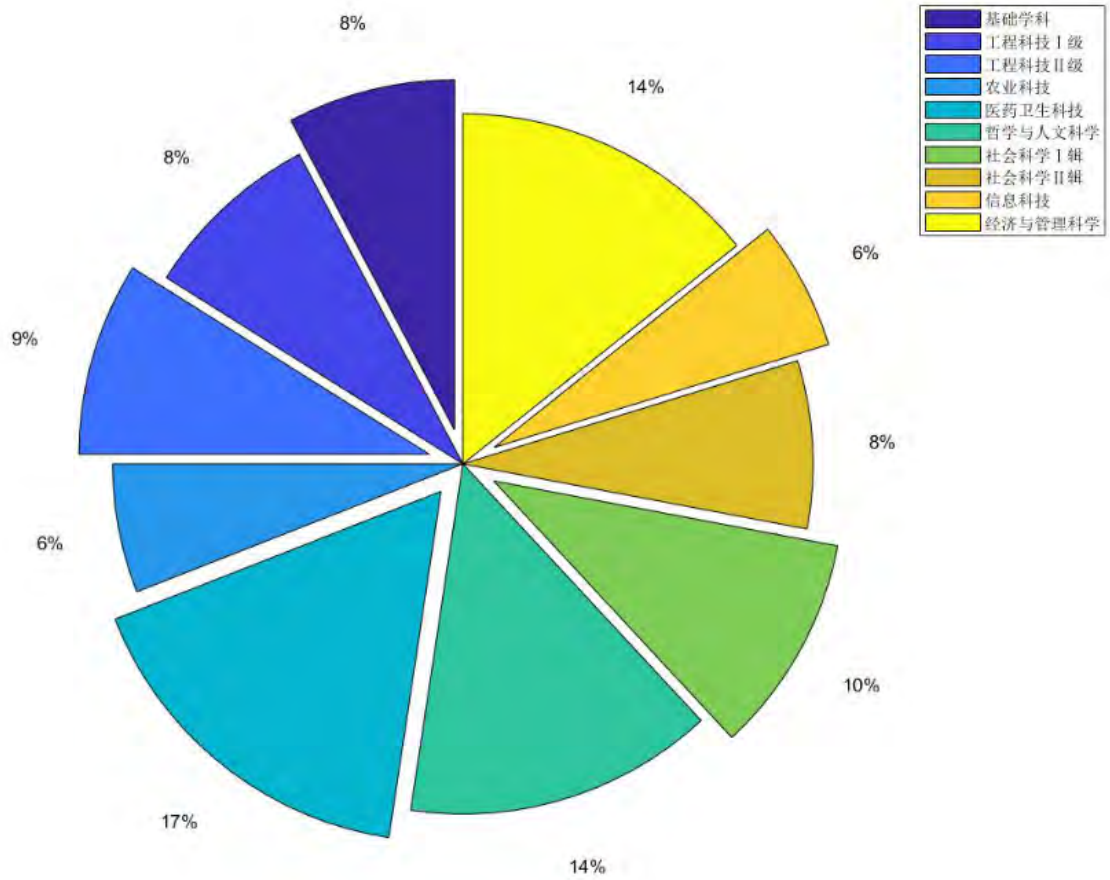
云计算	16	165	594	1477	2265	3120	1129
热词	cloud computing(2)	saas(17)	图书馆(39)	cloud computing(158)	cloud computing(293)	cloud computing(832)	cloud computing(383)
	网络技术(2)	网格计算(15)	虚拟化(38)	虚拟化(103)	虚拟化(157)	虚拟化(188)	大数据(67)
	网格计算(2)	云安全(13)	saas(26)	物联网(90)	物联网(121)	物联网(164)	虚拟化(62)
	google api(1)	虚拟化(11)	云安全(24)	图书馆(57)	图书馆(81)	大数据(109)	物联网(59)
	安全(1)	cloud computing(10)	物联网(24)	云存储(53)	应用(77)	hadoop(108)	hadoop(54)

数据说明

爬取知网科技文献10大类168小类，共计168000篇文章。



1542246811548CpzMj8.lab	2018/11/15 9:53	LAB 文件	1 KB
1542246811548CpzMj8.txt	2018/11/15 9:53	TXT 文件	1 KB
15422468115546v3lyG.lab	2018/11/15 9:53	LAB 文件	1 KB
15422468115546v3lyG.txt	2018/11/15 9:53	TXT 文件	1 KB

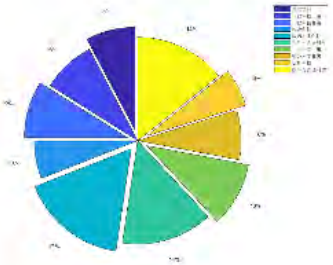

8%
14%
8%
6%
9%
8%
6%
10%
17%
14%
基础学科
工程科技Ⅰ级
工程科技Ⅱ级
农业科技
医药卫生科技
哲学与人文科学
社会科学Ⅰ辑
社会科学Ⅱ辑
信息科技
经济与管理科学

68000篇文章。

ID	标题	摘要	类别1	类别2
1	1971-2010年中国干湿区降水资源变化特征分析	针对降水在农业生产领域、科学技术领域以及预报洪涝灾害、分析水资源等方面意义重大,本文利用全国1971-2010年743个气象站点的逐日降水数据,基于干旱区、半干旱区、半湿润区和湿润区的区域划分,对降水量和降水日数按等级进行划分,运用线性倾向率、累积距平、MK检验和滑动t检验等方法系统分析了四大区域内不同等级降水量和降水日数资源变化特征以及各区域年降水量时间演变特征。结果表明:(1)过去40a各干湿区年降水日数分布存在差异,其中湿润区年降水日数呈中部高、周边偏低分布趋势;(2)干旱区北部小雨的降水贡献率和频率较高,半干旱区大到暴雨的降水贡献率和频率低下;(3)在年代际变化中,1991-2000年湿润区年降水量增加幅度最大,2001-2010年半湿润区年降水量变化不明显;(4)各干湿分区的降水量近40年突变趋势不一致,湿润区在2002年发生从多到少突变,而半干旱区未发生突变。	天文学、地球科学	大气科学（气象学）
2	1973年以来黄河三角洲形态与入海水沙通量关系研究	黄河是世界上著名的多沙河流,每年携带数亿吨泥沙经黄河三角洲入海。为了分析黄河三角洲面积与入海水沙通量关系,选取1973-2016年41景影像资料,借助ARCGIS 10.2得到研究区域面积,并把三角洲分割成河口地区与北部河滩地区进行研究。最后结合利津站水沙资料,通过SPSS数据分析软件,分析黄河入海水沙、黄河三角洲及其各区域随时间变化情况,并进一步探讨入海水沙与河口年造陆面积之间的关系。研究结果表明:河口地区年造陆面积与入海年输沙量之间具有显著正相关性。由于黄河中游一系列生态环境工程建设,使黄河水沙通量总体呈减少趋势,其中输沙量减少尤为显著,2015年利津站年输沙量比1973年减少97.5%。随着泥沙锐减,黄河三角洲及河口面积已由增大转为减小,转折点在1996年;人工固岸工程对控制三角洲北部河滩面积的稳定具有重要意义。此外,调水调沙依旧不能改变年造陆面积减小的趋势。	工业技术	水利工程
3	1976年龙陵地震诱发滑坡的影响因子敏感性分析	基于前人的研究和龙陵地震滑坡的调查资料,选取了地层岩性、断裂、地震烈度、震中距、地形坡度、坡向、高程、水系等8个因子作为1976年龙陵地震诱发滑坡的影响因子。利用GIS强大的空间分析能力,结合滑坡确定性系数(CF)的方法,对1976年龙陵地震诱发滑坡的诸影响因子进行敏感性分析,确定了该区域内各因子最利于地震滑坡发育的数值区间,为进一步区域地震滑坡稳定性评价奠定基础。	天文学、地球科学 天文学、地球科学	地球物理学 水文地质学与工程地质学
4	1979—2012年莱州湾南岸海水入侵与区域海岸线变动时空耦合分析	在整合区域海水入侵历史观测数据的基础上,结合现代遥感综合观测手段,开展莱州湾南岸海水入侵时空动态变化过程和海岸线历史变迁耦合机制研究。完成了1979—2012年区域海岸线的多期遥感监测和区域咸水入侵锋线演化的时间过程与空间特征数字重建,并引入端点速率法(end point ratio,EPR)分析模型对二者的时空耦合关系进行了探索研究。研究结果表明:(1)区域海水入侵经历了由快到慢的变化过程,1990年以后入侵速率明显减缓,入侵锋线基本稳定在1995年锋线附近,且2008—2012年入侵锋线局部发生后退;(2)区域海岸线除局部人工造陆导致岸线向海扩张外,区域海岸以蚀退型为主;(3)海水入侵锋线变化与海岸线进退二者之间在时空上存在强耦合关系,相关系数达到0. 407,显著性水平P <0. 01(双侧)。研究结果可为区域海水入侵的预防和治理提供数据支撑和科学依据。	天文学、地球科学	海洋学

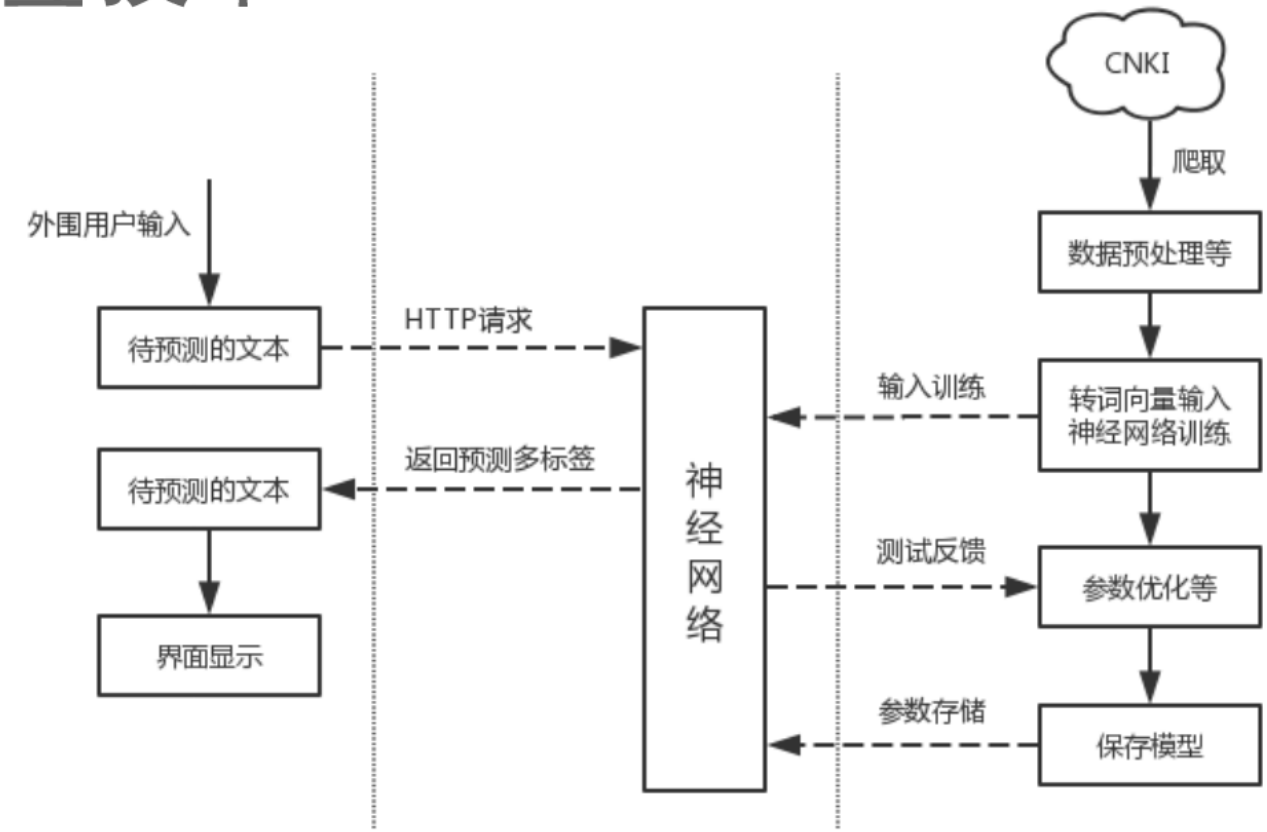
数据说明

爬取知网科技文献10大类168小类，共计168000篇文章。



1542246811548CpzMj8.lab	2018/11/15 9:53	LAB 文件	1 KB
1542246811548CpzMj8.txt	2018/11/15 9:53	TXT 文件	1 KB
15422468115546v3lyG.lab	2018/11/15 9:53	LAB 文件	1 KB
15422468115546v3lyG.txt	2018/11/15 9:53	TXT 文件	1 KB

多标签技术


CNKI
爬取
数据预处理等
转词向量输入
神经网络训练
参数优化等
保存模型
输入训练
测试反馈
参数存储
神经网络
外围用户输入
待预测的文本
HTTP请求
待预测的文本
返回预测多标签
界面显示

多标签技术

基于word2vec的词向量训练

- One-hot

例如：科学 技术 研究 情报 所

编码：科学 (1, 0, 0, 0, 0)
技术 (0, 1, 0, 0, 0)
研究 (0, 0, 1, 0, 0)
情报 (0, 0, 0, 1, 0)
所 (0, 0, 0, 0, 1)

科学 研究
=>[(1, 0, 0, 0, 0), (0, 0, 1, 0, 0)]

国王 - 王后 ≈ 男 - 女

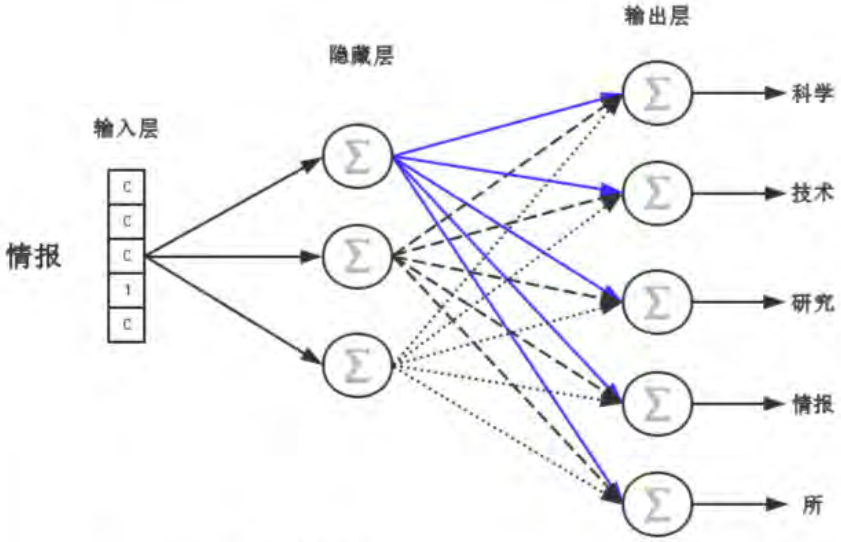
弊端：
1) 矩阵过于稀疏
2) 词语之间关联性

多标签技术

基于word2vec的词向量训练

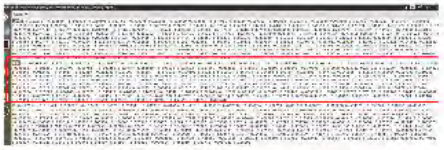
- word embedding

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

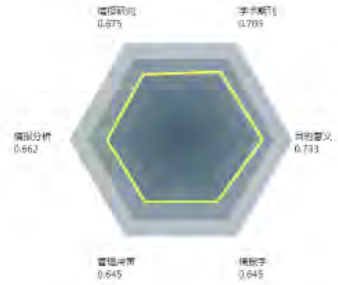


$$y = f(x)$$

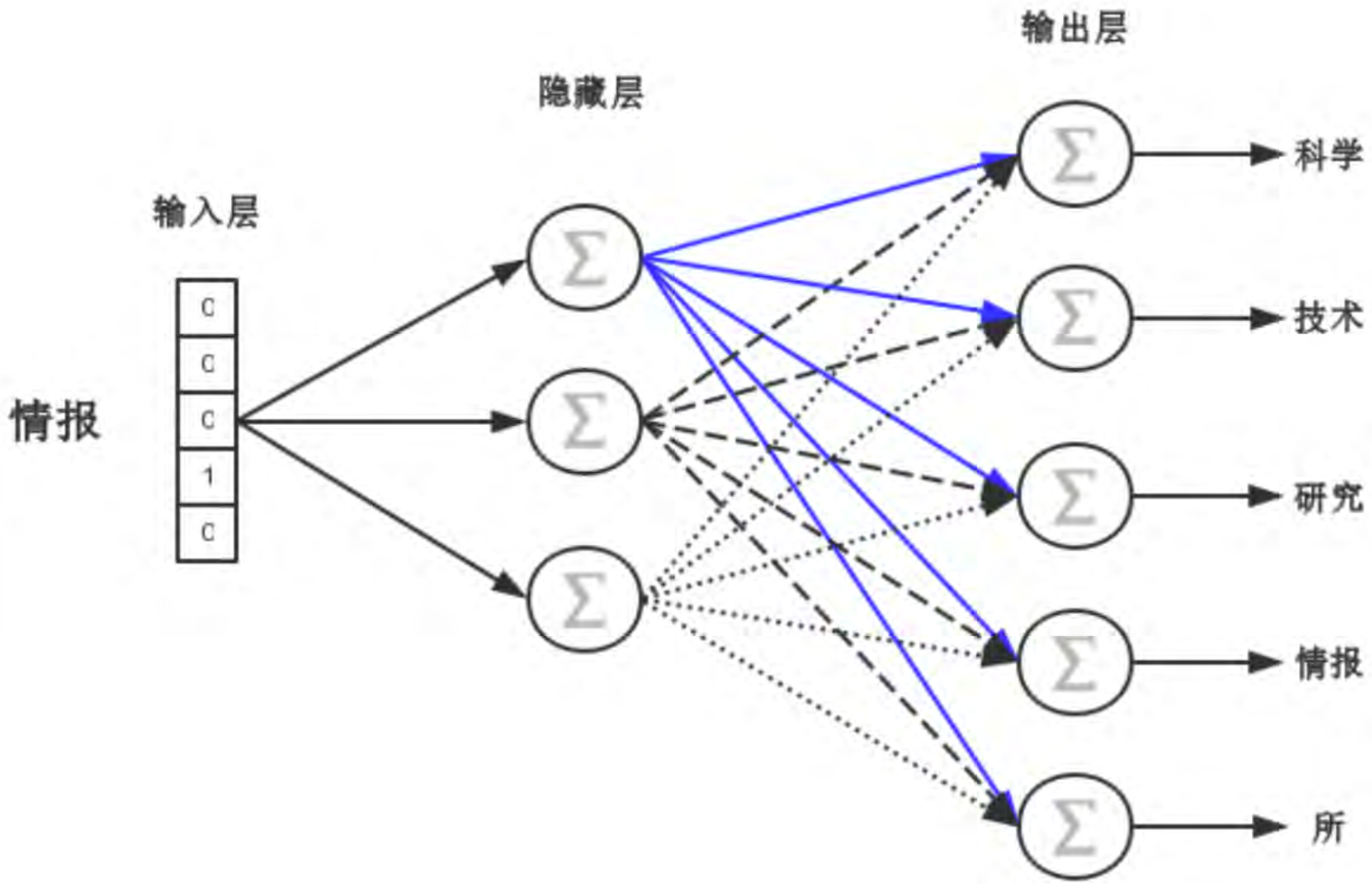
- 静态



- 动态



- word embedding



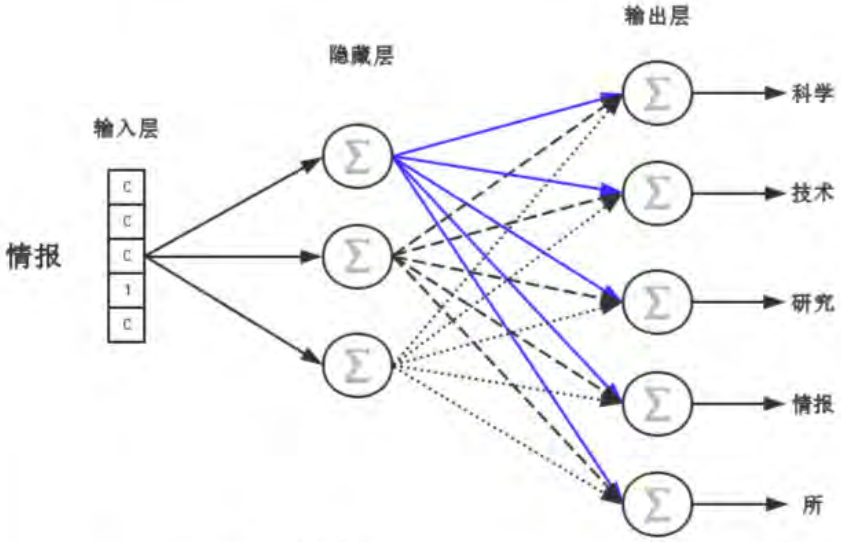
$$y = f(x)$$

多标签技术

基于word2vec的词向量训练

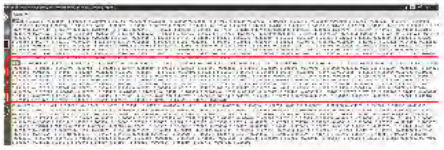
- word embedding

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

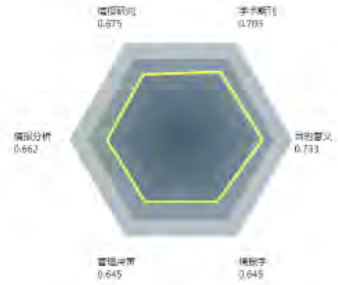


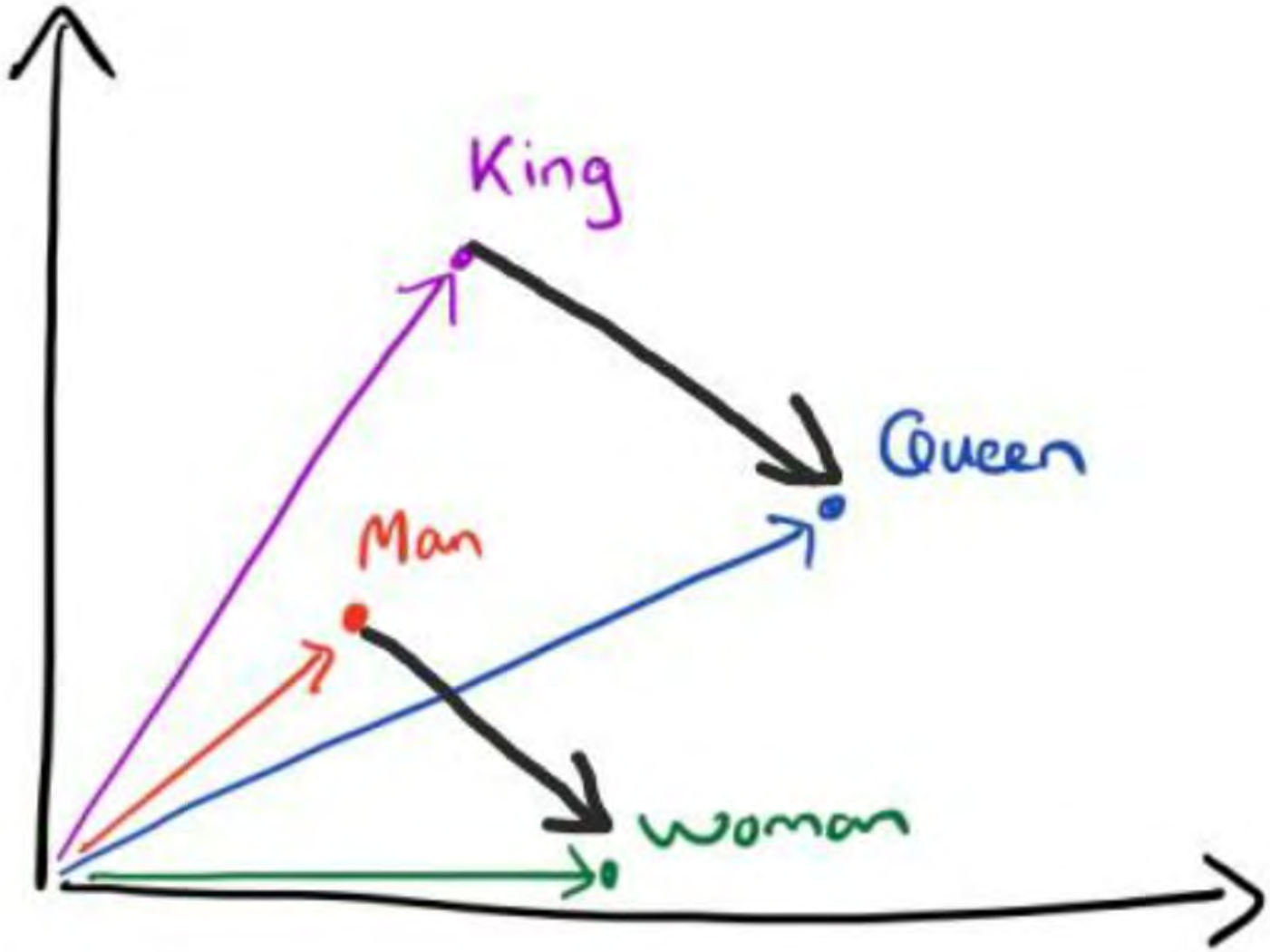
$$y = f(x)$$

- 静态



- 动态





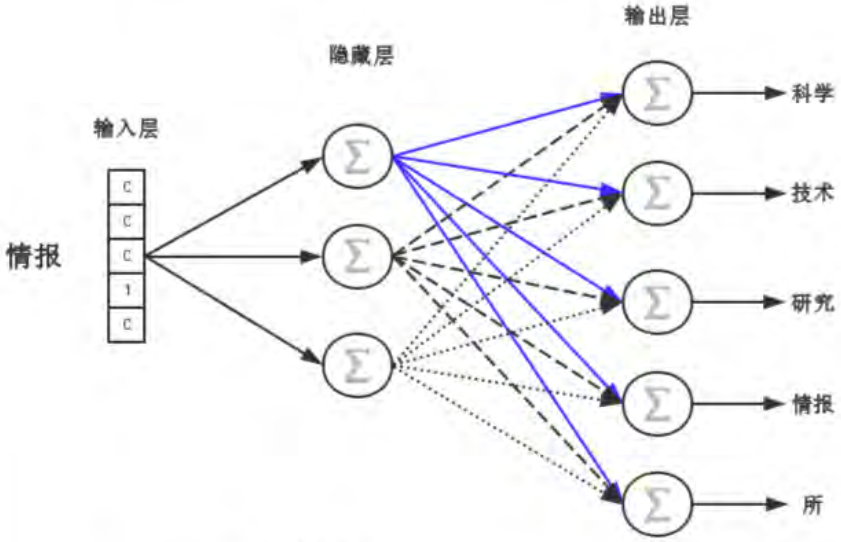
Word Vectors

多标签技术

基于word2vec的词向量训练

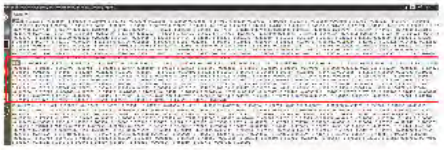
- word embedding

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

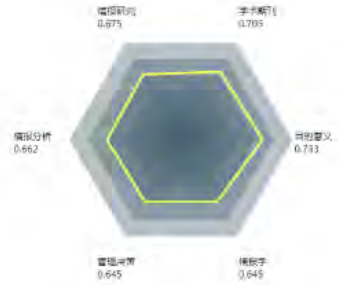


$$y = f(x)$$

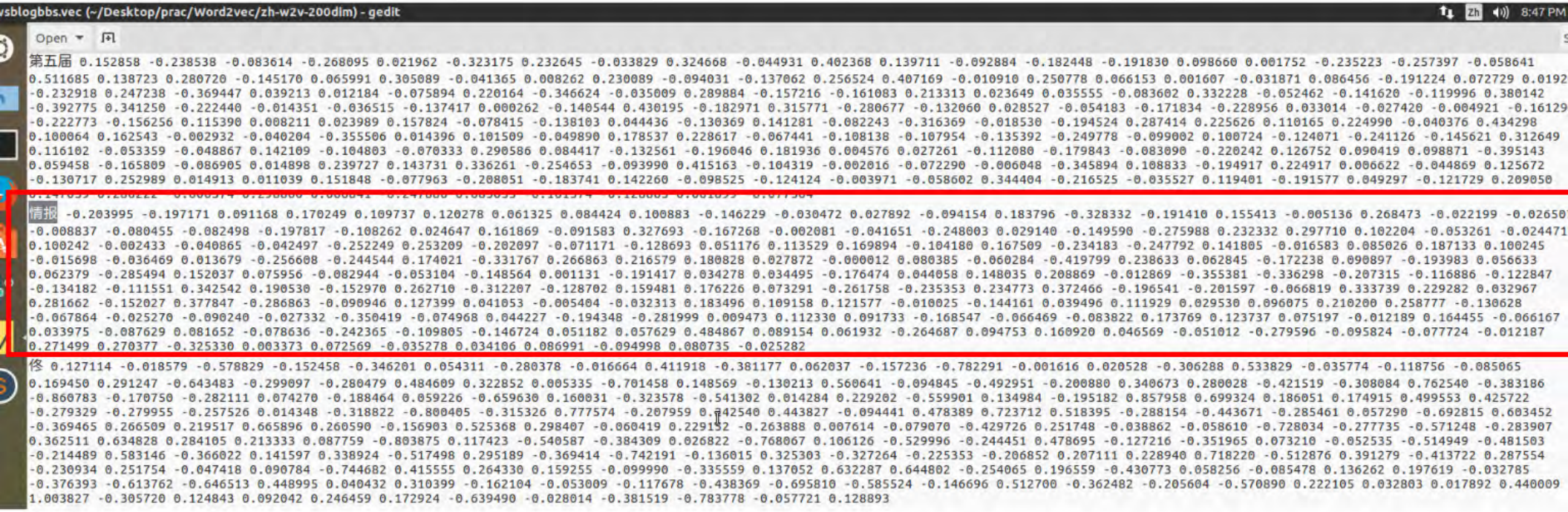
- 静态



- 动态



静态

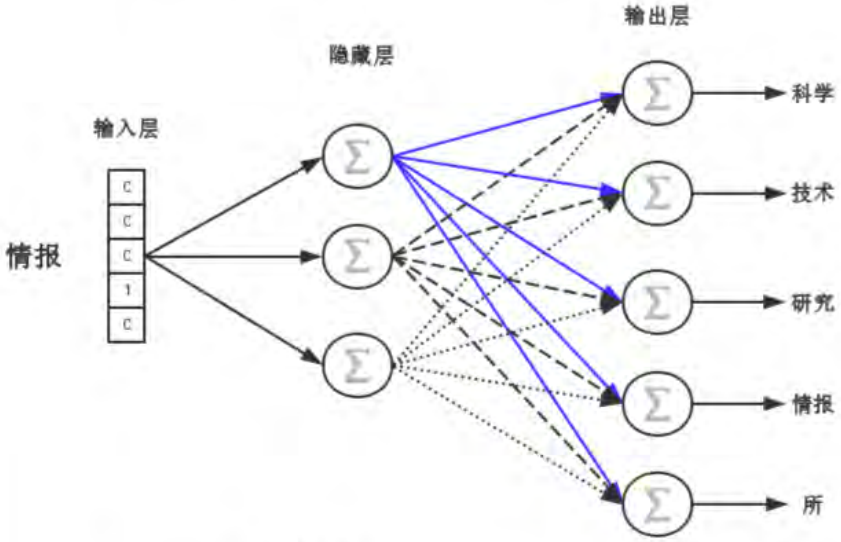

vsblogbbs.vec (~/Desktop/prac/Word2vec/zh-w2v-200dim) - gedit
Open
第五届 0.152858 -0.238538 -0.083614 -0.268095 0.021962 -0.323175 0.232645 -0.033829 0.324668 -0.044931 0.402368 0.139711 -0.092884 -0.182448 -0.191830 0.098660 0.001752 -0.235223 -0.257397 -0.058641
0.511685 0.138723 0.280720 -0.145170 0.065991 0.305089 -0.041365 0.008262 0.230089 -0.094031 -0.137062 0.256524 0.407169 -0.010910 0.250778 0.066153 0.001607 -0.031871 0.086456 -0.191224 0.072729 0.0192
-0.232918 0.247238 -0.369447 0.039213 0.012184 -0.075894 0.220164 -0.346624 -0.035009 0.289884 -0.157216 -0.161083 0.213313 0.023649 0.035555 -0.083602 0.332228 -0.052462 -0.141620 -0.119996 0.380142
-0.392775 0.341250 -0.222440 -0.014351 -0.036515 -0.137417 0.000262 -0.140544 0.430195 -0.182971 0.315771 -0.280677 -0.132060 0.028527 -0.054183 -0.171834 -0.228956 0.033014 -0.027420 -0.004921 -0.16129
-0.222773 -0.156256 0.115390 0.008211 0.023989 0.157824 -0.078415 -0.138103 0.044436 -0.130369 0.141281 -0.082243 -0.316369 -0.018530 -0.194524 0.287414 0.225626 0.110165 0.224990 -0.040376 0.434298
0.100064 0.162543 -0.002932 -0.040204 -0.355506 0.014396 0.101509 -0.049890 0.178537 0.228617 -0.067441 -0.108138 -0.107954 -0.135392 -0.249778 -0.099002 0.100724 -0.124071 -0.241126 -0.145621 0.312649
0.116102 -0.053359 -0.048867 0.142109 -0.104803 -0.070333 0.290586 0.084417 -0.132561 -0.196046 0.181936 0.004576 0.027261 -0.112080 -0.179843 -0.083090 -0.220242 0.126752 0.090419 0.098871 -0.395143
0.059458 -0.165809 -0.086905 0.014898 0.239727 0.143731 0.336261 -0.254653 -0.093990 0.415163 -0.104319 -0.002016 -0.072290 -0.006048 -0.345894 0.108833 -0.194917 0.224917 0.006622 -0.044869 0.125672
-0.130717 0.252989 0.014913 0.011039 0.151848 -0.077963 -0.208051 -0.183741 0.142260 -0.098525 -0.124124 -0.003971 -0.058602 0.344404 -0.216525 -0.035527 0.119401 -0.191577 0.049297 -0.121729 0.209050
情报 -0.203995 -0.197171 0.091168 0.170249 0.109737 0.120278 0.061325 0.084424 0.100883 -0.146229 -0.030472 0.027892 -0.094154 0.183796 -0.328332 -0.191410 0.155413 -0.005136 0.268473 -0.022199 -0.02650
-0.008837 -0.080455 -0.082498 -0.197817 -0.108262 0.024647 0.161869 -0.091583 0.327693 -0.167268 -0.002081 -0.041651 -0.248003 0.029140 -0.149590 -0.275988 0.232332 0.297710 0.102204 -0.053261 -0.024471
0.100242 -0.002433 -0.040865 -0.042497 -0.252249 0.253209 -0.202097 -0.071171 -0.128693 0.051176 0.113529 0.169894 -0.104180 0.167509 -0.234183 -0.247792 0.141805 -0.016583 0.085026 0.187133 0.100245
-0.015698 -0.036469 0.013679 -0.256608 -0.244544 0.174021 -0.331767 0.266863 0.216579 0.180828 0.027872 -0.000012 0.080385 -0.060284 -0.419799 0.238633 0.062845 -0.172238 0.090897 -0.193983 0.056633
0.062379 -0.285494 0.152037 0.075956 -0.082944 -0.053104 -0.148564 0.001131 -0.191417 0.034278 0.034495 -0.176474 0.044058 0.148035 0.208869 -0.012869 -0.355381 -0.336298 -0.207315 -0.116886 -0.122847
-0.134182 -0.111551 0.342542 0.190530 -0.152970 0.262710 -0.312207 -0.128702 0.159481 0.176226 0.073291 -0.261758 -0.235353 0.234773 0.372466 -0.196541 -0.201597 -0.066819 0.333739 0.229282 0.032967
0.281662 -0.152027 0.377847 -0.286863 -0.090946 0.127399 0.041053 -0.005404 -0.032313 0.183496 0.109158 0.121577 -0.010025 -0.144161 0.039496 0.111929 0.029530 0.096075 0.210200 0.258777 -0.130628
-0.067864 -0.025270 -0.090240 -0.027332 -0.350419 -0.074968 0.044227 -0.194348 -0.281999 0.009473 0.112330 0.091733 -0.168547 -0.066469 -0.083822 0.173769 0.123737 0.075197 -0.012189 0.164455 -0.066167
0.033975 -0.087629 0.081652 -0.078636 -0.242365 -0.109805 -0.146724 0.051182 0.057629 0.484867 0.089154 0.061932 -0.264687 0.094753 0.160920 0.046569 -0.051012 -0.279596 -0.095824 -0.077724 -0.012187
0.271499 0.270377 -0.325330 0.003373 0.072569 -0.035278 0.034106 0.086991 -0.094998 0.080735 -0.025282
佟 0.127114 -0.018579 -0.578829 -0.152458 -0.346201 0.054311 -0.280378 -0.016664 0.411918 -0.381177 0.062037 -0.157236 -0.782291 -0.001616 0.020528 -0.306288 0.533829 -0.035774 -0.118756 -0.085065
0.169450 0.291247 -0.643483 -0.299097 -0.280479 0.484609 0.322852 0.005335 -0.701458 0.148569 -0.130213 0.560641 -0.094845 -0.492951 -0.200880 0.340673 0.280028 -0.421519 -0.308084 0.762540 -0.383186
-0.860783 -0.170750 -0.282111 0.074270 -0.188464 0.059226 -0.659630 0.160031 -0.323578 -0.541302 0.014284 0.229202 -0.559901 0.134984 -0.195182 0.857958 0.699324 0.186051 0.174915 0.499553 0.425722
-0.279329 -0.279955 -0.257526 0.014348 -0.318822 -0.800405 -0.315326 0.777574 -0.207959 0.342540 0.443827 -0.094441 0.478389 0.723712 0.518395 -0.288154 -0.443671 -0.285461 0.057290 -0.692815 0.603452
-0.369465 0.266509 0.219517 0.665896 0.260590 -0.156903 0.525368 0.298407 -0.060419 0.229152 -0.263888 0.007614 -0.079070 -0.429726 0.251748 -0.038862 -0.058610 -0.728034 -0.277735 -0.571248 -0.283907
0.362511 0.634828 0.284105 0.213333 0.087759 -0.803875 0.117423 -0.540587 -0.384309 0.026822 -0.768067 0.106126 -0.529996 -0.244451 0.478695 -0.127216 -0.351965 0.073210 -0.052535 -0.514949 -0.481503
-0.214489 0.583146 -0.366022 0.141597 0.338924 -0.517498 0.295189 -0.369414 -0.742191 -0.136015 0.325303 -0.327264 -0.225353 -0.206852 0.207111 0.228940 0.718220 -0.512876 0.391279 -0.413722 0.287554
-0.230934 0.251754 -0.047418 0.090784 -0.744682 0.415555 0.264330 0.159255 -0.099990 -0.335559 0.137052 0.632287 0.644802 -0.254065 0.196559 -0.430773 0.058256 -0.085478 0.136262 0.197619 -0.032785
-0.376393 -0.613762 -0.646513 0.448995 0.040432 0.310399 -0.162104 -0.053009 -0.117678 -0.438369 -0.695810 -0.585524 -0.146696 0.512700 -0.362482 -0.205604 -0.570890 0.222105 0.032803 0.017892 0.440009
1.003827 -0.305720 0.124843 0.092042 0.246459 0.172924 -0.639490 -0.028014 -0.381519 -0.783778 -0.057721 0.128893

多标签技术

基于word2vec的词向量训练

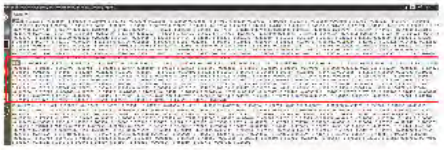
- word embedding

$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

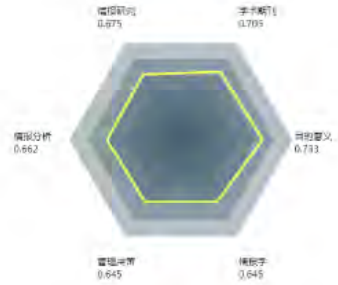


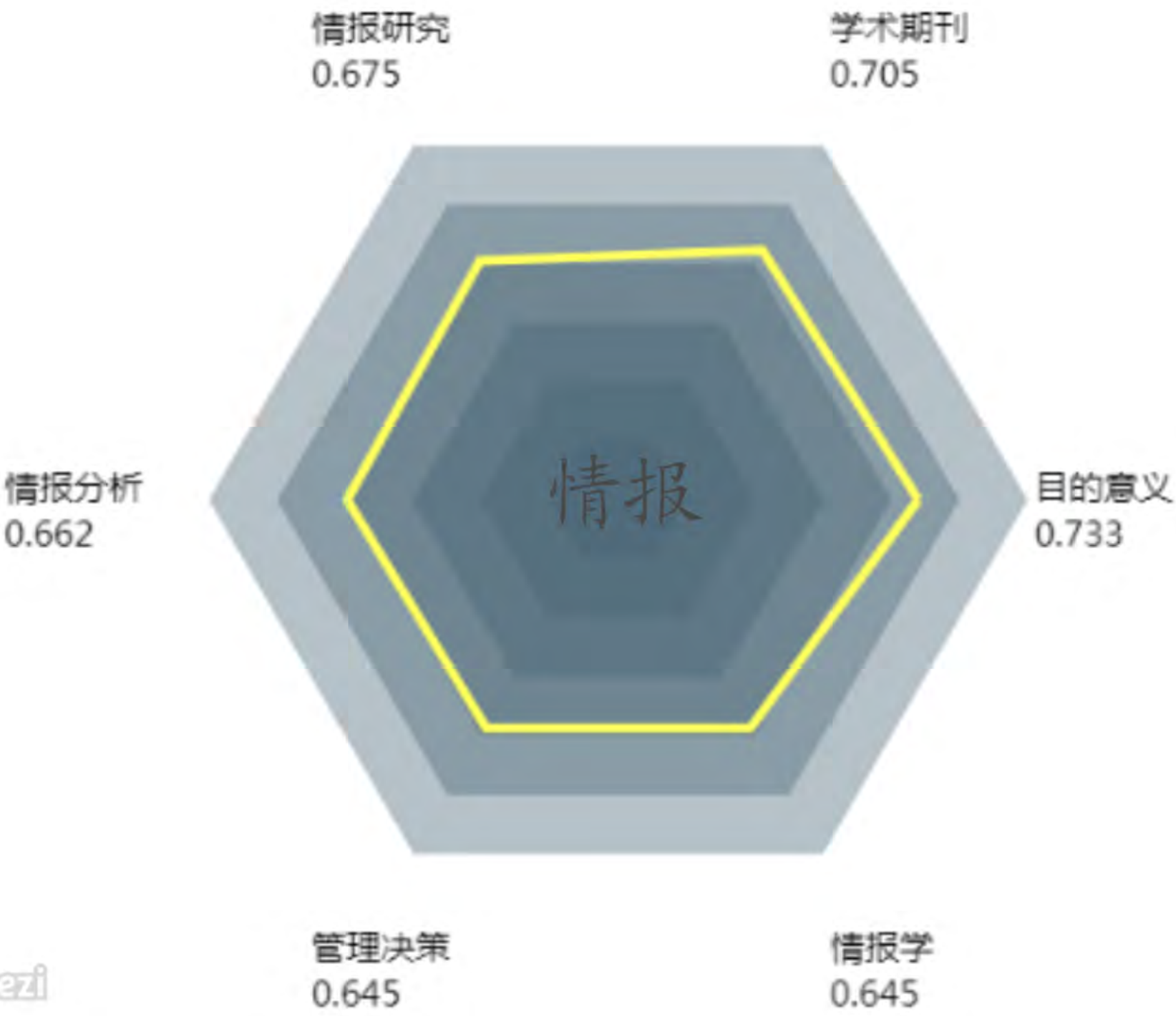
$$y = f(x)$$

- 静态



- 动态




情报研究
0.675
学术期刊
0.705
情报分析
0.662
情报
目的意义
0.733
管理决策
0.645
情报学
0.645

例如：科学 技术 研究 情报 所

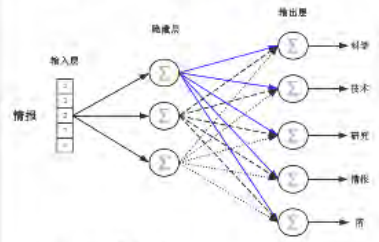
编码：科学 (1, 0, 0, 0, 0)
技术 (0, 1, 0, 0, 0)
研究 (0, 0, 1, 0, 0)
情报 (0, 0, 0, 1, 0)
所 (0, 0, 0, 0, 1)

科学 研究
=>[(1, 0, 0, 0, 0), (0, 0, 1, 0, 0)]

国王 − 王后 ≈ 男 − 女

弊端：
1) 矩阵过于稀疏
2) 词语之间关联性

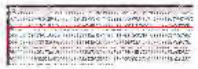
• word embedding



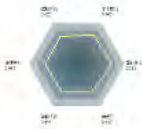
y = f(x)

[0 0 0 1 0] × [4 6 13; 10 12 19; 11 18 25] = [10 12 19]

• 静态

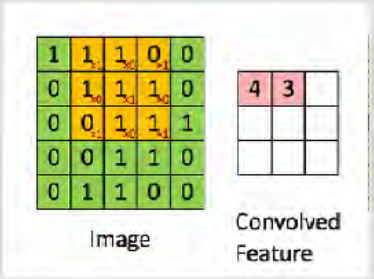


• 动态

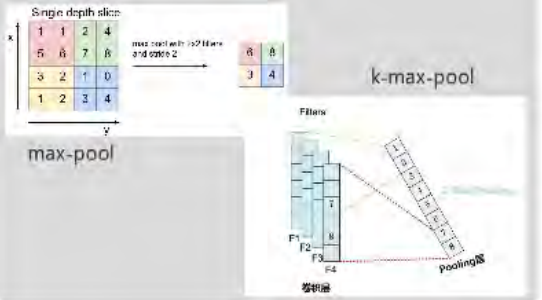


多标签技术 TextCNN网络

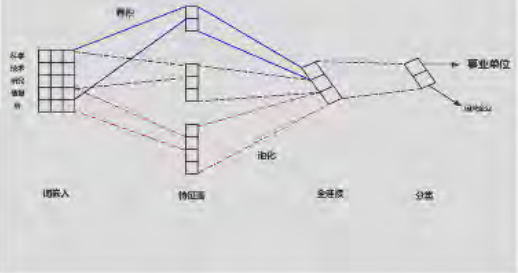
. 卷积



. 池化



. TextCNN分类

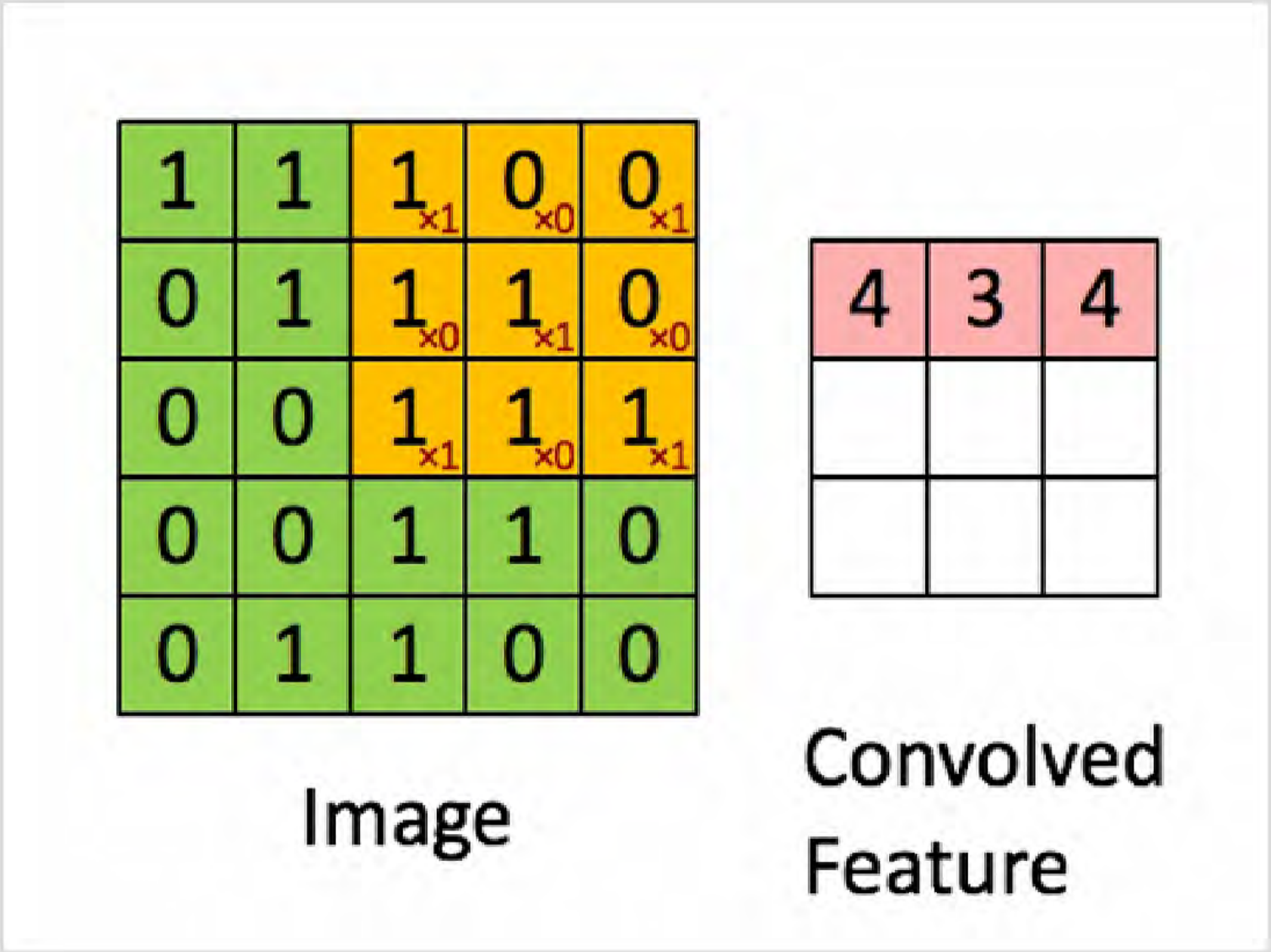


新知识发现 案例分析

文献摘要	多标签	新分类

文献摘要	多标签	新分类

- 卷积


1 1 1 0 0
0 1 1 1 0
0 0 1 1 1
0 0 1 1 0
0 1 1 0 0
1 0 1
0 1 0
1 0 1
Image
4 3 4
Convolved Feature

例如：科学 技术 研究 情报 所

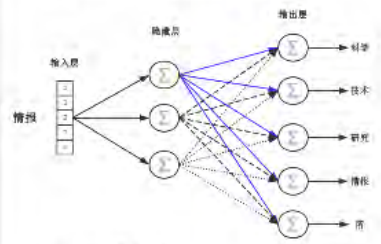
编码：科学 (1, 0, 0, 0, 0)
技术 (0, 1, 0, 0, 0)
研究 (0, 0, 1, 0, 0)
情报 (0, 0, 0, 1, 0)
所 (0, 0, 0, 0, 1)

科学 研究
=>[(1, 0, 0, 0, 0), (0, 0, 1, 0, 0)]

国王 − 王后 ≈ 男 − 女

弊端：
1) 矩阵过于稀疏
2) 词语之间关联性

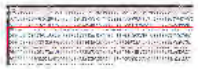
• word embedding



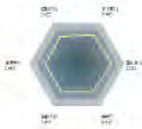
y = f(x)

[0 0 0 1 0] × [4 6 13; 10 12 19; 11 18 25] = [10 12 19]

• 静态

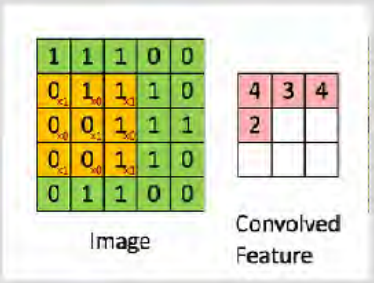


• 动态

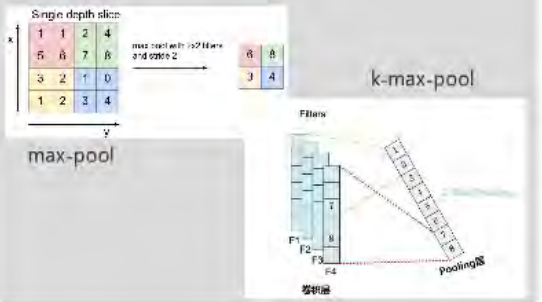


多标签技术 TextCNN网络

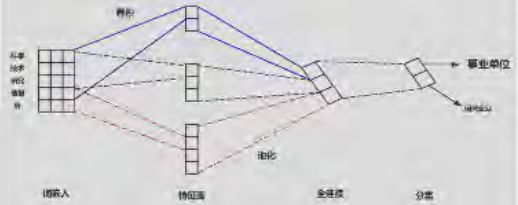
. 卷积



. 池化



. TextCNN分类

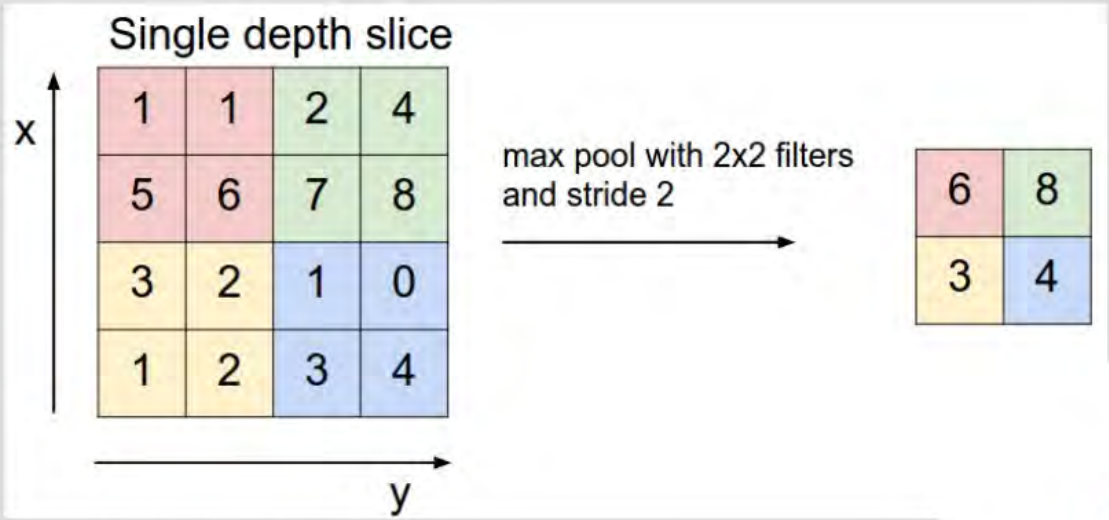


新知识发现 案例分析

文献摘要	多标签	新分类

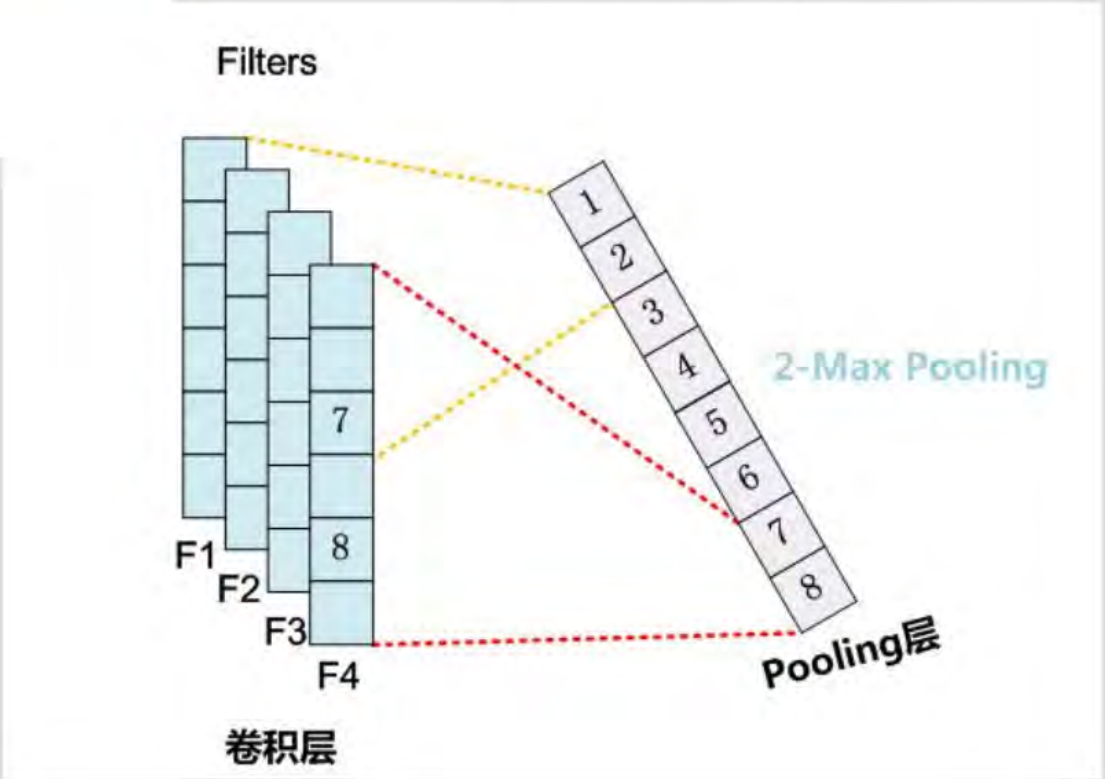
文献摘要	多标签	新分类

池化

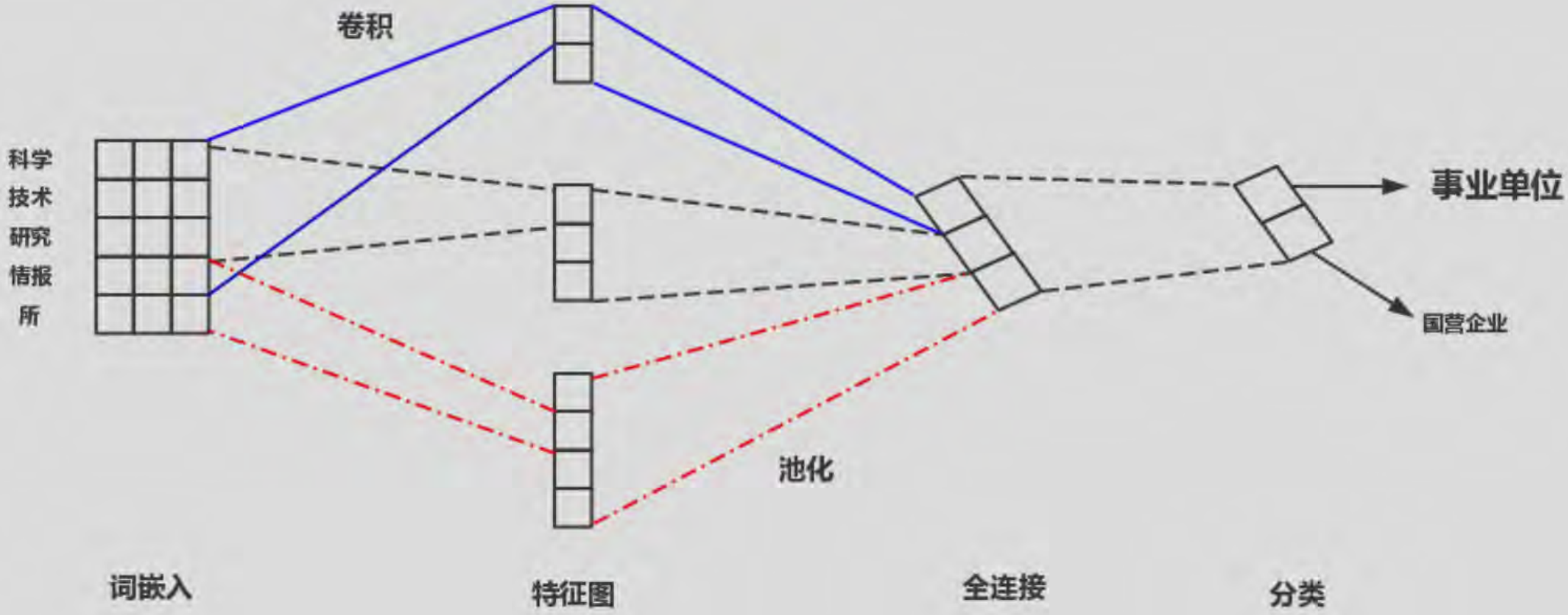


max-pool

k-max-pool



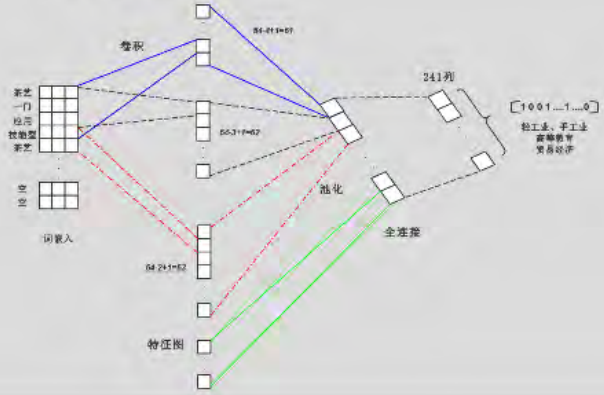
. TextCNN分类



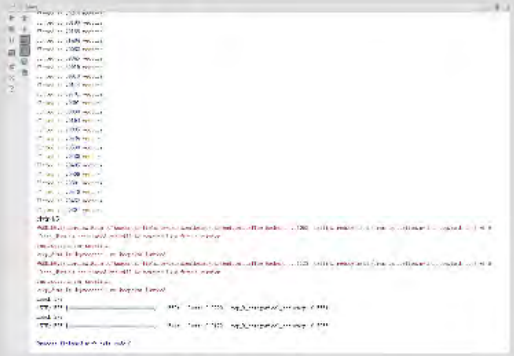
多标签技术　　网络的训练与优化

. 数据增强

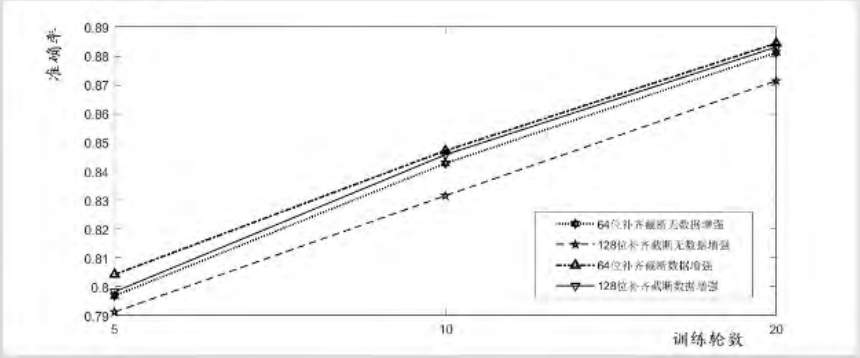
. 网络简图



. 训练过程



. 参数优化



. 预测结果

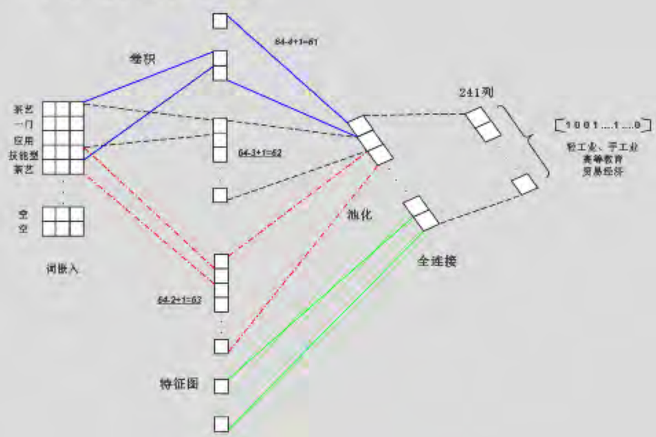
文章摘要：
一带 一路 倡议 实施 高级 国际 商务 人才 提出 新的 要求 茶叶 一带 一路 商务活动 中 地位 独特 具有 一定 教育 功能 茶文化 国际交流 重要 符号 茶文化 融入 国际 商务 专硕 培养 很有 必要 目前 国际 商务 硕士 培养 存在 文化 缺失 跨 文化教育 较少等 问题 一带 一路 背景 探索 茶文化 融入 国际 商务 专硕 培养 路径 一种 有益 尝试 空 空 空

预测结果：
('轻工业、手工业', 0.8076695)
('高等教育', 0.39927575)
('贸易经济', 0.22038086)
('法律', 0.04906398)
('职业技术教育', 0.048415482)

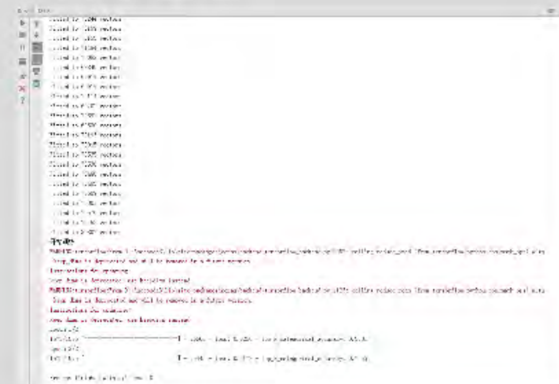
多标签技术　网络的训练与优化

数据增强

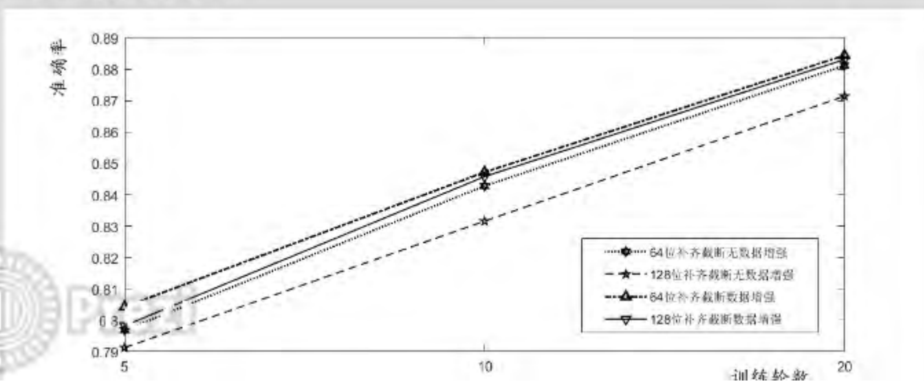
网络简图



训练过程



参数优化



预测结果

文章摘要：
一带 一路 倡议 实施 高级 国际 商务 人才 提出 新的 要求 茶叶
一带 一路 商务活动 中 地位 独特 具有 一定 教育 功能 茶文化
国际交流 重要 符号 茶文化 融入 国际 商务 专硕 培养 很有 必要
目前 国际 商务 硕士 培养 存在 文化 缺失 跨 文化教育 较少等
问题 一带 一路 背景 探索 茶文化 融入 国际 商务 专硕 培养
路径 一种 有益 尝试 空 空 空

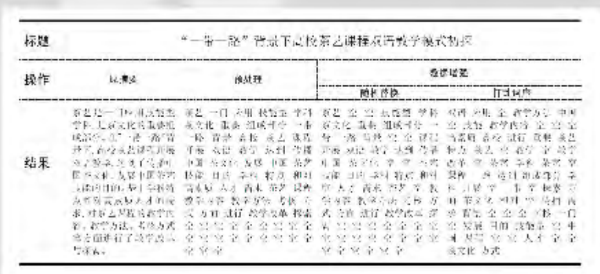
预测结果：
('轻工业、手工业', 0.8076695)
('高等教育', 0.39927575)
('贸易经济', 0.22038086)
('法律', 0.04906398)
('职业技术教育', 0.048415482)

• 数据增强

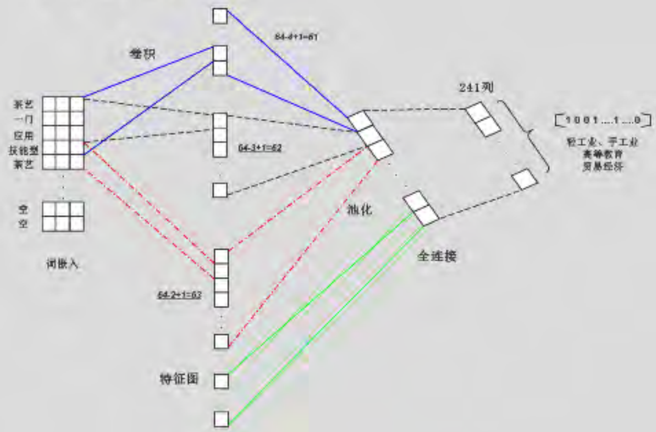
标题	“一带一路”背景下高校茶艺课程双语教学模式初探			
操作	原摘要	预处理	数据增强	
			随机替换	打乱词序
结果	茶艺是一门应用技能型学科，是茶文化的重要组成部分。在"一带一路"背景下，高校茶艺课程开展双语教学，达到了传播中国茶文化，发展中国茶艺技能的目的。基于学科特点和对高素质人才的需求，对茶艺课程的教学内容、教学方法、考核方式等方面进行了教学改革与探索。	茶艺 一门 应用 技能型 学科 茶文化 重要 组成部分 一带 一路 背景 高校 茶艺 课程 开展 双语 教学 达到 传播 中国 茶文化 发展 中国 茶艺 技能 目的 学科 特点 和对 高素质 人才 需求 茶艺 课程 教学内容 教学方法 考核 方式 方面 进行 教学改革 探索 空	茶艺 空 空 技能型 学科 茶文化 重要 组成部分 一带 一路 背景 空 空 课程 开展 双语 教学 达到 传播 中国 茶文化 空 空 茶艺 技能 目的 学科 特点 和对 空 人才 需求 茶艺 空 教学内容 教学方法 考核 方式 方面 进行 教学改革 探索 空	双语 应用 空 教学方法 中国 空 技能 教学内容 空 空 空 高素质 高校 进行 重要 茶艺 特点 茶艺 空 教学 空 教学 改革 空 茶艺 学科 茶艺 空 课程 一路 达到 组成部分 学科 开展 空 一带 空 探索 方面 茶文化 和对 空 传播 需求 背景 空 空 空 考核 一门 空 发展 目的 技能型 空 中国 课程 空 空 人才 空 空 茶文化 方式

多标签技术　网络的训练与优化

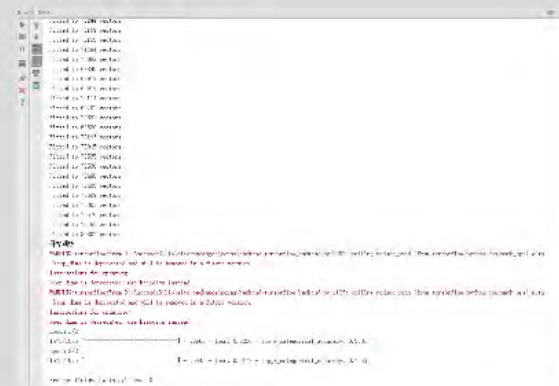
• 数据增强



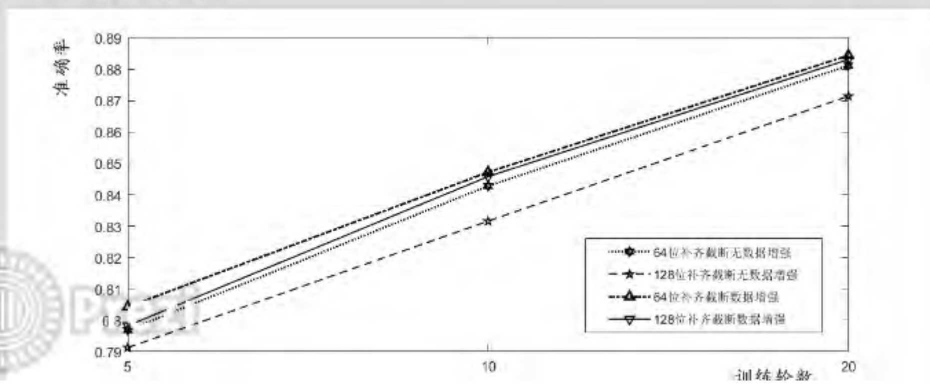
• 网络简图



• 训练过程



• 参数优化

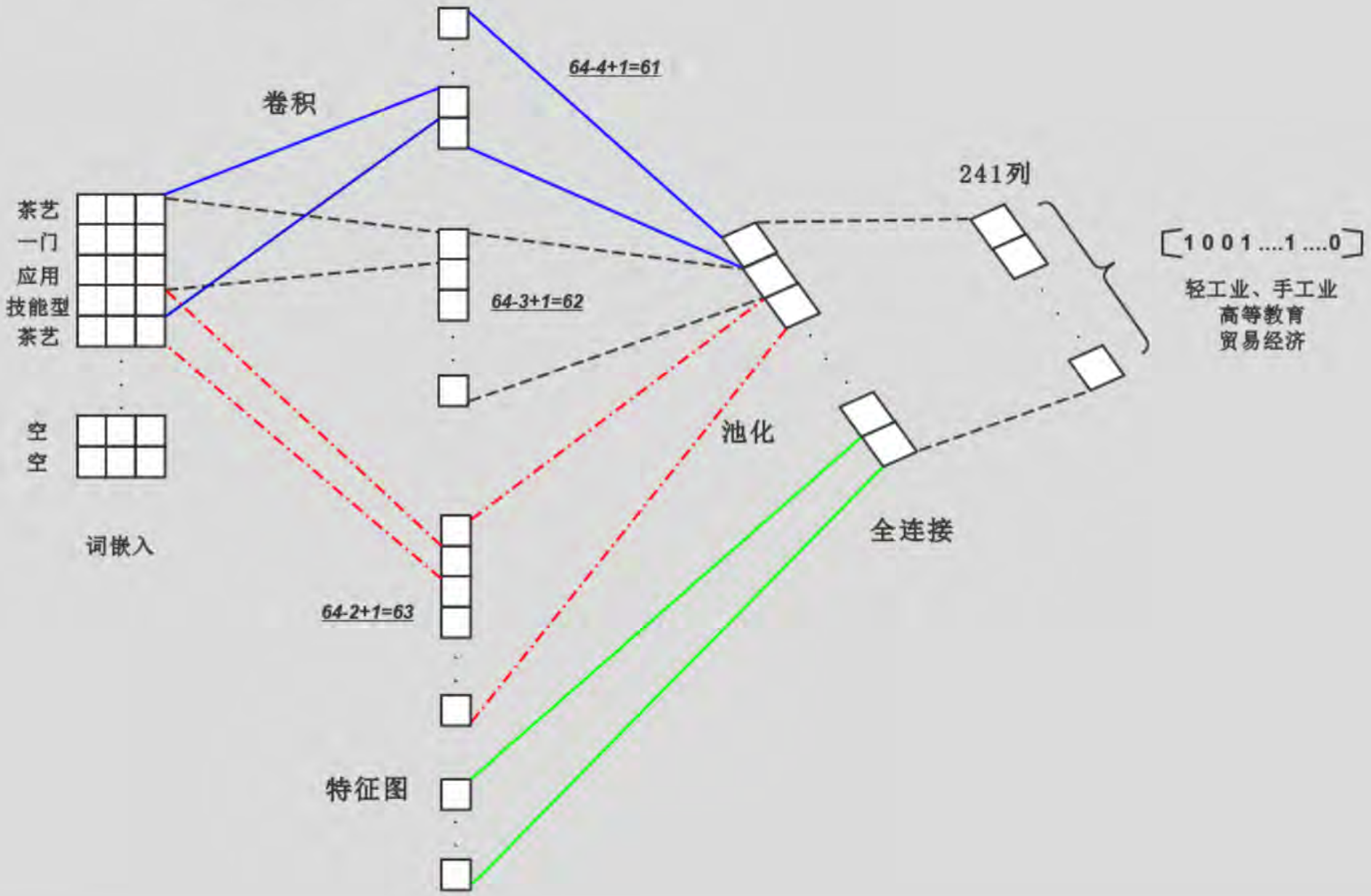


• 预测结果

文章摘要：
一带 一路 倡议 实施 高级 国际 商务 人才 提出 新的 要求 茶叶
一带 一路 商务活动 中 地位 独特 具有 一定 教育 功能 茶文化
国际交流 重要 符号 茶文化 融入 国际 商务 专硕 培养 很有 必要
目前 国际 商务 硕士 培养 存在 文化 缺失 跨 文化教育 较少等
问题 一带 一路 背景 探索 茶文化 融入 国际 商务 专硕 培养
路径 一种 有益 尝试 空 空 空

预测结果：
('轻工业、手工业', 0.8076695)
('高等教育', 0.39927575)
('贸易经济', 0.22038086)
('法律', 0.04906398)
('职业技术教育', 0.048415482)

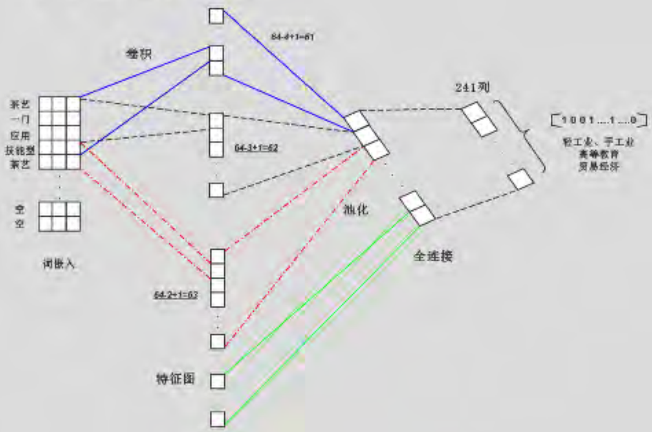
• 网络简图


64-4+1=61
卷积
241列
茶艺
一门
应用
技能型
茶艺
64-3+1=62
[1 0 0 1 ....1 ....0]
轻工业、手工业
高等教育
贸易经济
池化
空
空
词嵌入
全连接
64-2+1=63
特征图

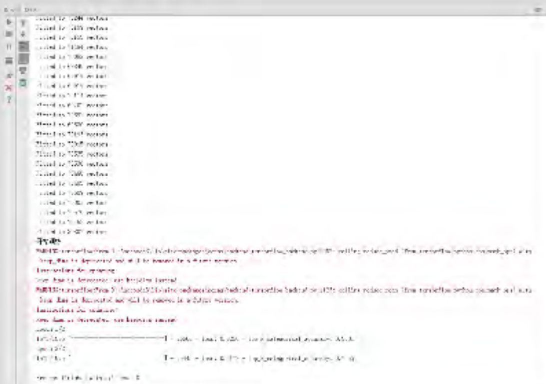
多标签技术　网络的训练与优化

数据增强

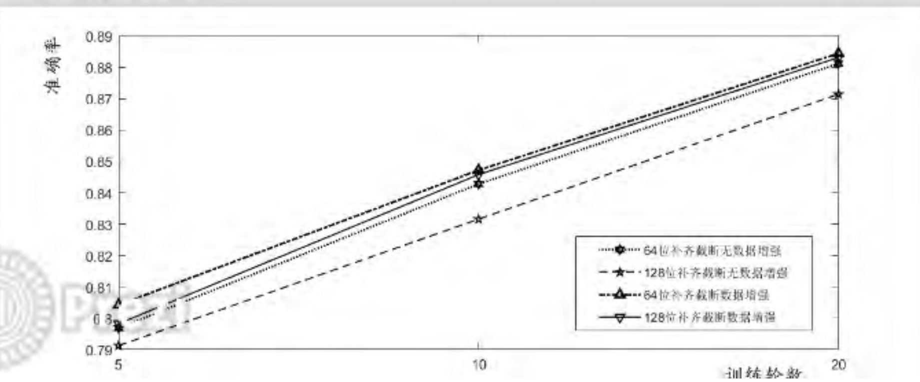
网络简图



训练过程



参数优化



预测结果

文章摘要：
一带 一路 倡议 实施 高级 国际 商务 人才 提出 新的 要求 茶叶
一带 一路 商务活动 中 地位 独特 具有 一定 教育 功能 茶文化
国际交流 重要 符号 茶文化 融入 国际 商务 专硕 培养 很有 必要
目前 国际 商务 硕士 培养 存在 文化 缺失 跨 文化教育 较少等
问题 一带 一路 背景 探索 茶文化 融入 国际 商务 专硕 培养
路径 一种 有益 尝试 空 空 空

预测结果：
('轻工业、手工业', 0.8076695)
('高等教育', 0.39927575)
('贸易经济', 0.22038086)
('法律', 0.04906398)
('职业技术教育', 0.048415482)

Run Main

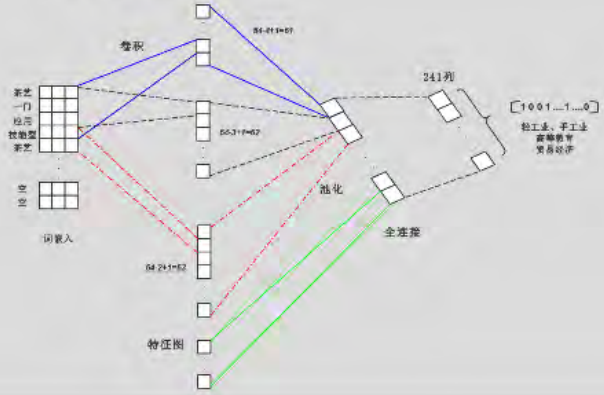
```
Fitted to 70244 vectors
Fitted to 70199 vectors
Fitted to 70155 vectors
Fitted to 71484 vectors
Fitted to 70082 vectors
Fitted to 69845 vectors
Fitted to 69919 vectors
Fitted to 69919 vectors
Fitted to 70174 vectors
Fitted to 69701 vectors
Fitted to 70391 vectors
Fitted to 69830 vectors
Fitted to 70193 vectors
Fitted to 70095 vectors
Fitted to 70535 vectors
Fitted to 70530 vectors
Fitted to 70480 vectors
Fitted to 70405 vectors
Fitted to 70509 vectors
Fitted to 70301 vectors
Fitted to 70570 vectors
Fitted to 70452 vectors
Fitted to 30807 vectors
开始训练
WARNING:tensorflow:From D:\Anaconda3\lib\site-packages\keras\backend\tensorflow_backend.py:1062: calling reduce_prod (from tensorflow.python.ops.math_ops) with
 keep_dims is deprecated and will be removed in a future version.
Instructions for updating:
keep_dims is deprecated, use keepdims instead
WARNING:tensorflow:From D:\Anaconda3\lib\site-packages\keras\backend\tensorflow_backend.py:1123: calling reduce_mean (from tensorflow.python.ops.math_ops) with
 keep_dims is deprecated and will be removed in a future version.
Instructions for updating:
keep_dims is deprecated, use keepdims instead
Epoch 1/2
1575/1575 [==============================] - 1560s - loss: 0.0220 - top_k_categorical_accuracy: 0.5631
Epoch 2/2
1575/1575 [==============================] - 1544s - loss: 0.0172 - top_k_categorical_accuracy: 0.7053

Process finished with exit code 0
```

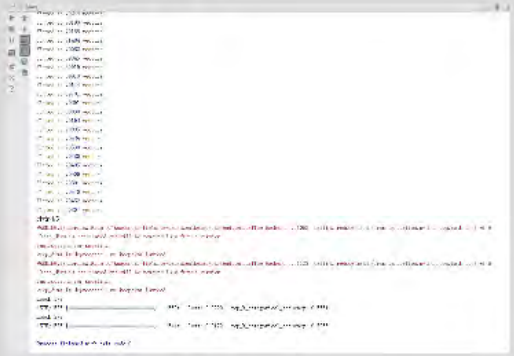
多标签技术　　网络的训练与优化

. 数据增强

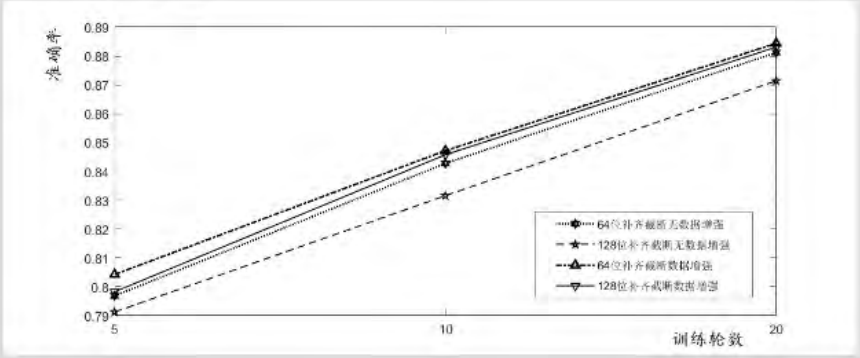
. 网络简图



. 训练过程



. 参数优化



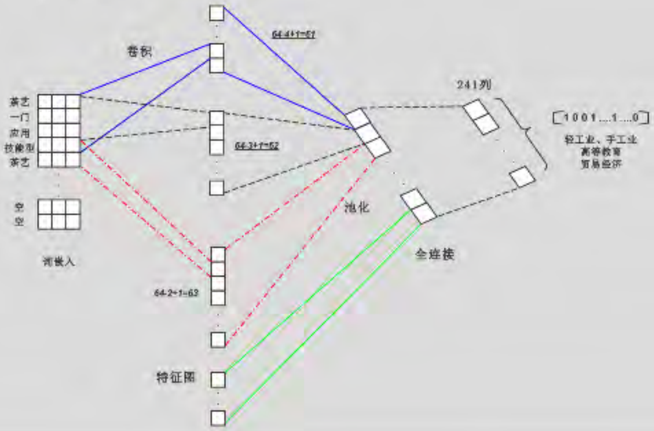
. 预测结果

文章摘要：
一带 一路 倡议 实施 高级 国际 商务 人才 提出 新的 要求 茶叶 一带 一路 商务活动 中 地位 独特 具有 一定 教育 功能 茶文化 国际交流 重要 符号 茶文化 融入 国际 商务 专硕 培养 很有 必要 目前 国际 商务 硕士 培养 存在 文化 缺失 跨 文化教育 较少等 问题 一带 一路 背景 探索 茶文化 融入 国际 商务 专硕 培养 路径 一种 有益 尝试 空 空 空

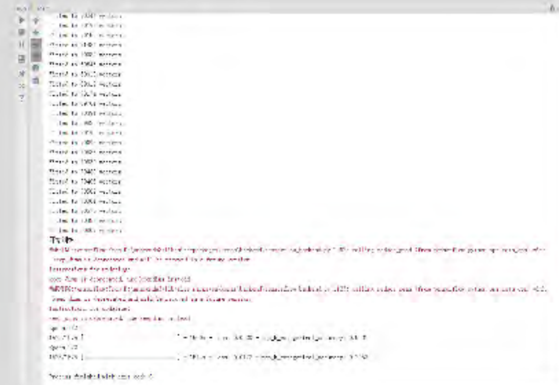
预测结果：
('轻工业、手工业', 0.8076695)
('高等教育', 0.39927575)
('贸易经济', 0.22038086)
('法律', 0.04906398)
('职业技术教育', 0.048415482)

- 数据增强

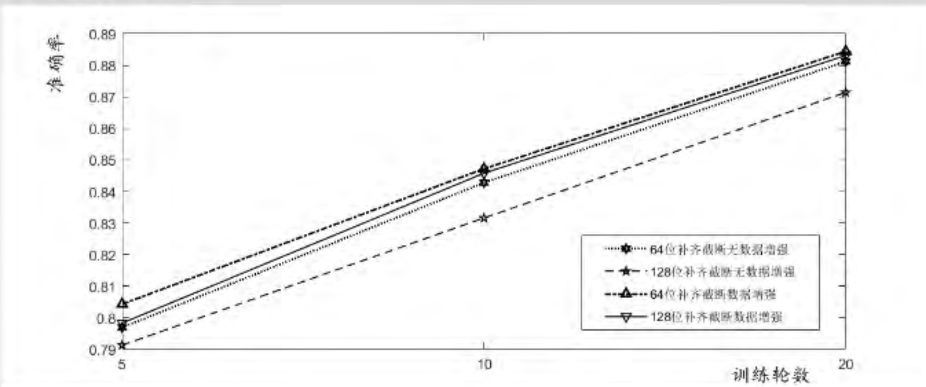
- 网络简图



- 训练过程



- 参数优化



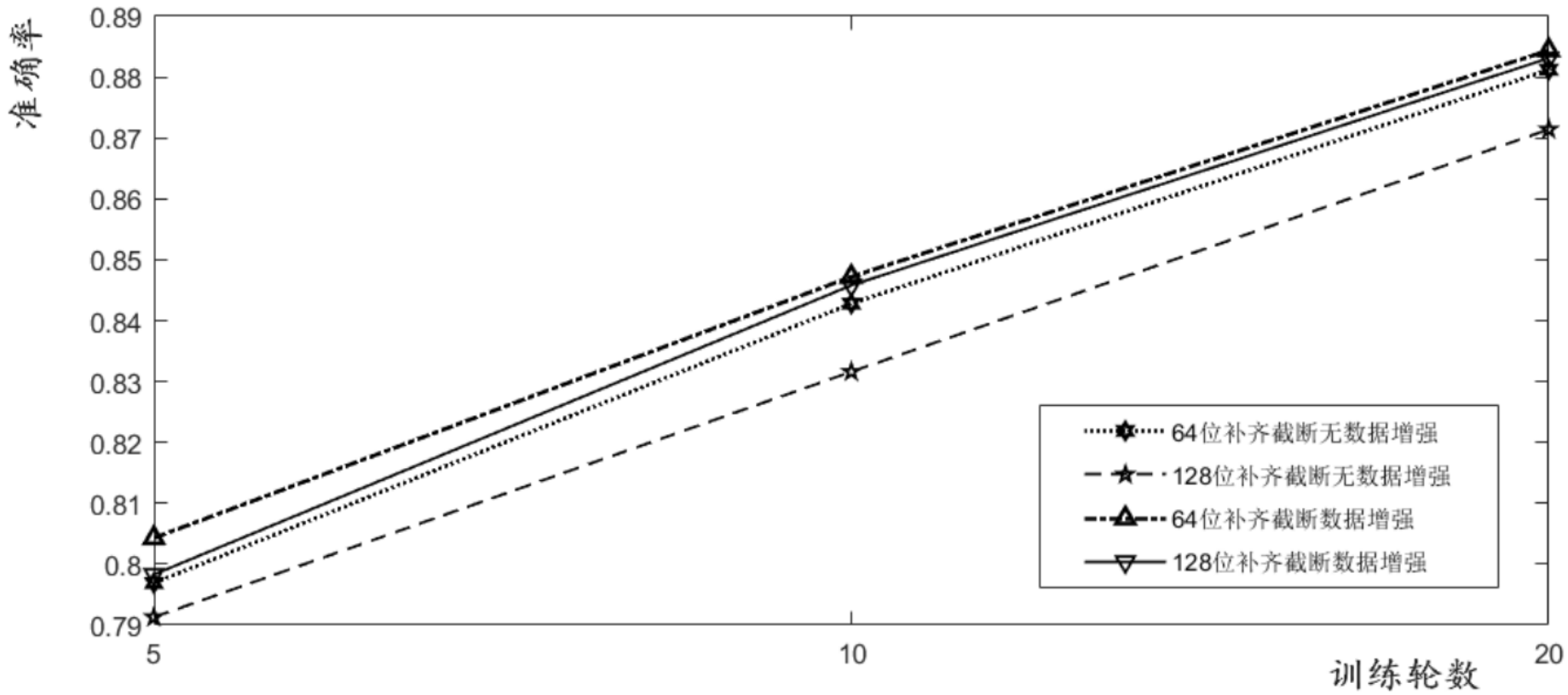
- 预测结果

文章摘要：
一带 一路 倡议 实施 高级 国际 商务 人才 提出 新的 要求 茶叶
一带 一路 商务活动 中 地位 独特 具有 一定 教育 功能 茶文化
国际交流 重要 符号 茶文化 融入 国际 商务 专硕 培养 很有 必要
目前 国际 商务 硕士 培养 存在 文化 缺失 跨 文化教育 较少等
问题 一带 一路 背景 探索 茶文化 融入 国际 商务 专硕 培养
路径 一种 有益 尝试 空 空 空

预测结果：
('轻工业、手工业', 0.8076695)
('高等教育', 0.39927575)
('贸易经济', 0.22038086)
('法律', 0.04906398)
('职业技术教育', 0.048415482)

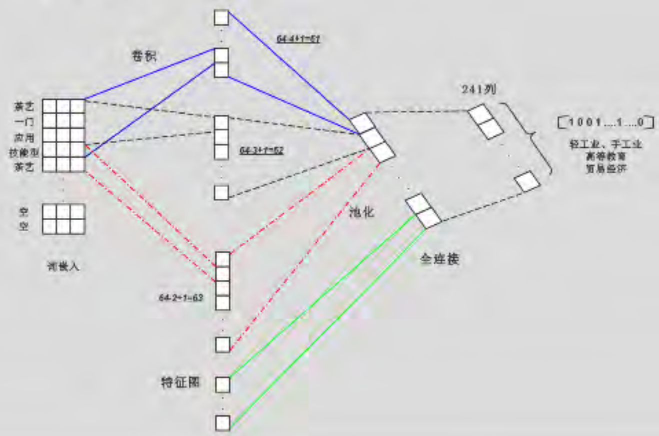
Prezi

、参数优化

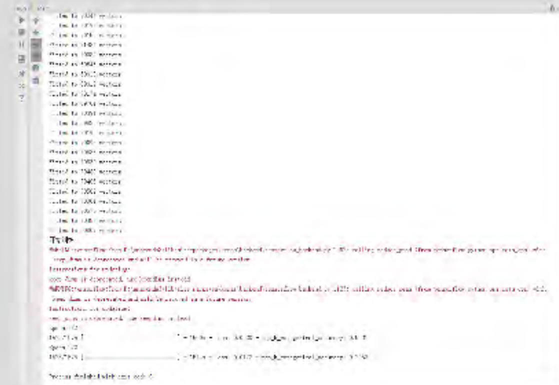

准确率
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0.82
0.81
0.8
0.79
5
10
20
训练轮数
64位补齐截断无数据增强
128位补齐截断无数据增强
64位补齐截断数据增强
128位补齐截断数据增强

• 数据增强

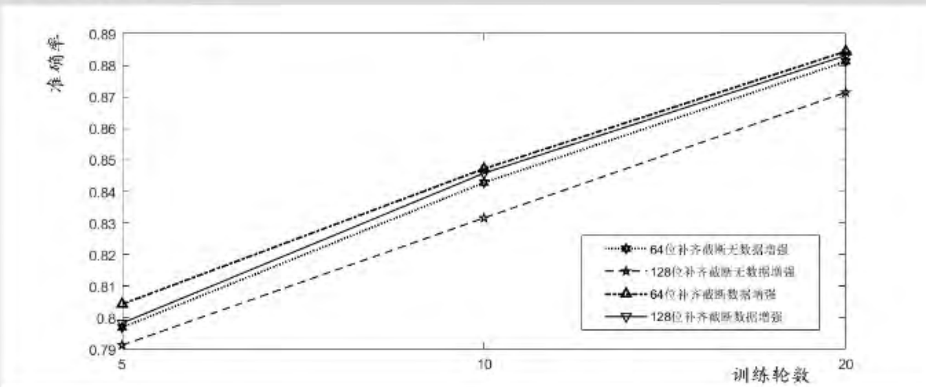
• 网络简图



• 训练过程



• 参数优化



• 预测结果

文章摘要：
一带 一路 倡议 实施 高级 国际 商务 人才 提出 新的 要求 茶叶
一带 一路 商务活动 中 地位 独特 具有 一定 教育 功能 茶文化
国际交流 重要 符号 茶文化 融入 国际 商务 专硕 培养 很有 必要
目前 国际 商务 硕士 培养 存在 文化 缺失 跨 文化教育 较少等
问题 一带 一路 背景 探索 茶文化 融入 国际 商务 专硕 培养
路径 一种 有益 尝试 空 空 空

预测结果：
('轻工业、手工业', 0.8076695)
('高等教育', 0.39927575)
('贸易经济', 0.22038086)
('法律', 0.04906398)
('职业技术教育', 0.048415482)

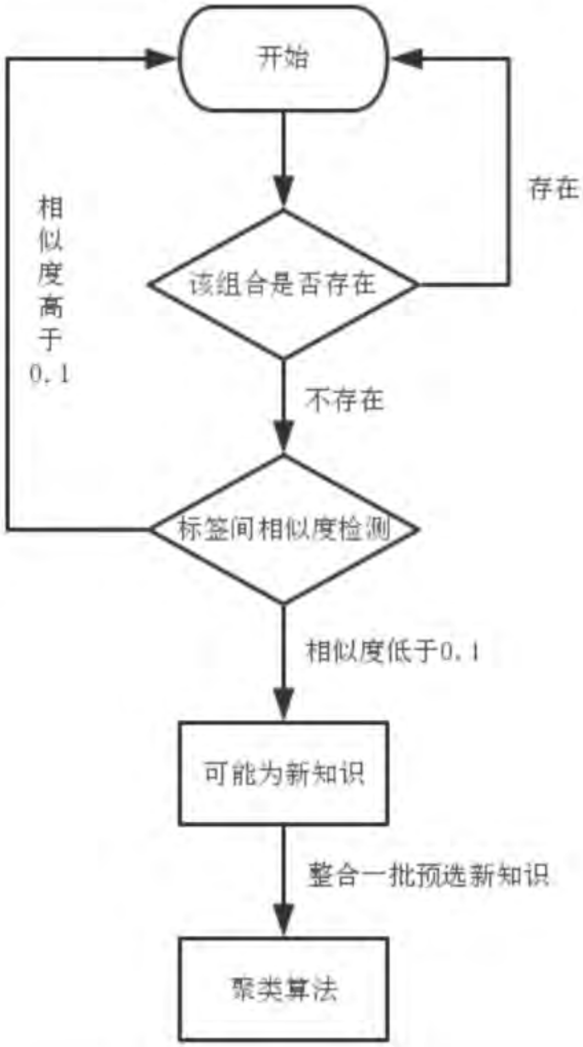
Prezi

预测结果

文章摘要：
一带 一路 倡议 实施 高级 国际 商务 人才 提出 新的 要求 茶叶
一带 一路 商务活动 中 地位 独特 具有 一定 教育 功能 茶文化
国际交流 重要 符号 茶文化 融入 国际 商务 专硕 培养 很有 必要
目前 国际 商务 硕士 培养 存在 文化 缺失 跨 文化教育 较少等
问题 一带 一路 背景 探索 茶文化 融入 国际 商务 专硕 培养
路径 一种 有益 尝试 空 空 空

预测结果：
('轻工业、手工业', 0.8076695)
('高等教育', 0.39927575)
('贸易经济', 0.22038086)
('法律', 0.04906398)
('职业技术教育', 0.048415482)

新知识发现



自动化技术、计算机技术*物理学 1543299199649Agtdfy
工业经济*自然科学现状及发展 1543299199657wMhF1z
自然科学现状及发展*工业经济 1543299199661jgwbJU
自动化技术、计算机技术*数学 1543299204959qKBtGe
自动化技术、计算机技术*计算数学 1543299204993Ig50Ah
公路运输*水路运输 1543299204999D5stZM
物理学*无线电电子学、电信技术 1543299205005ad4kqF
物理学*无线电电子学、电信技术 1543299205008qMhSwG
农业经济*系统科学 1543299205015v6jtMd

0.0006745048075 82931 ['财政', '金融'] ['大文学'] 1543299243444XSNJOb
0.084015963677 62992 ['测绘', '学'] ['农业', '经济'] 1543299300317JcGINi
0.0796606787797 9058 ['轻工业', '手工业'] ['高等教育'] 1543299452762ia4e0t
-0.071047375001 06474 ['中国', '政治'] ['水', '路', '运输'] 1543299536999cYLRzp
0.084015963677 62992 ['测绘', '学'] ['农业', '经济'] 1543299590963 4eELjn
0.07439861715 1357 ['化学工业'] ['药学'] 1543299614168F2MzcY
0.028550832889 086195 ['预防', '医学', '卫生学'] ['中国', '政治'] 1543299676100 5pqXIC
0.0850909057627 8043 ['中国', '史'] ['农业', '经济'] 1543300351891XTAgu
-0.198795237612 29465 ['中国', '史'] ['水', '路', '运输'] 1543300074113THF8V2
0.032514532094 01979 ['交通', '运输', '经济'] ['中国', '史'] 1543300104648MfI29D
0.027485811018 6326 ['职业', '技术', '教育'] ['中国共产党'] 1543300182500DANegV
-0.076254217483 04136 ['法律'] ['水', '路', '运输'] 1543300296365XxfhbZ
0.0796606787797 9058 ['轻工业', '手工业'] ['高等教育'] 1543300373611FcIY8f
0.088391560623 28918 ['外', '科学'] ['教育'] 1543300368474Fsl1r6
0.041902804861 64409 ['自动化', '技术', '计算机', '技术'] ['中国', '政治'] 1543300434198enAtRZ
0.028550832889 086195 ['中国', '政治'] ['预防', '医学', '卫生学'] 1543300484722b0NdIC
0.024079827000 535382 ['中国', '人物', '传记'] ['工业', '经济'] 1543300524496eNjdpC

['系统科学', '农业工程', '自动化技术、计算机技术'] 1543299212687UPmzZh
['轻工业、手工业', '化学工业', '基础医学'] 1543299235518MxiPBO
['财政', '金融'] ['天文学'] 1543299243444XSNJOb
['矿床学', '地球物理勘探', '石油、天然气工业'] 1543299280078 3mcL6M
['测绘', '学'] ['农业', '经济'] 1543299300317JcGINi
['工业经济', '经济计划与管理', '贸易经济'] 1543299389125 0Clgzt
['金属学与金属工艺', '化学工业', '一般工业技术'] 1543299410557ymFvLe
['一般工业技术', '金属学与金属工艺', '化学工业'] 1543299427910JLCbVO
['工业经济', '废物处理与综合利用', '金属学与金属工艺'] 1543299427946wVxi10
['农作物', '化学', '轻工业、手工业'] 1543299444860 3fwMel
['轻工业、手工业', '高等教育', '信息与知识传播'] 1543299452754Mg4m9q
['轻工业', '手工业'] ['高等教育'] 1543299452762ia4e0t

自动化技术、计算机技术*物理学 1543299199649Agtdfy

工业经济*自然科学现状及发展 1543299199657wMhF1z

自然科学现状及发展*工业经济 1543299199661jgwbJU

自动化技术、计算机技术*数学 1543299204959qKBtGe

自动化技术、计算机技术*计算数学 1543299204993Ig5OAh

公路运输*水路运输 1543299204999D5stZM

物理学*无线电电子学、电信技术 1543299205005ad4kqF

物理学*无线电电子学、电信技术 1543299205008qMhSwG

农业经济*系统科学 1543299205015v6jtMd

0.000874504807582931 ['财政', '金融'] ['天文学']

0.000874504807582931 ['财政', '金融'] ['天文学'] 1543299243444XSNJOb
0.08401596367762992 ['测绘', '学'] ['农业', '经济'] 1543299300317JcGINi
0.07966067877979058 ['轻工业', '手工业'] ['高等教育'] 1543299452762ia4e0t
-0.07104737500106474 ['中国', '政治'] ['水', '路', '运输'] 1543299536999cYlRop
0.08401596367762992 ['测绘', '学'] ['农业', '经济'] 15432995909634eELjn
0.074398617151357 ['化学工业'] ['药学'] 1543299614168F2WzeY
0.028550832889086195 ['预防', '医学', '卫生学'] ['中国', '政治'] 15432996761005pqXIC
0.06509090576278043 ['中国', '史'] ['农业', '经济'] 1543300035189lXTAgu
-0.19879523761229465 ['中国', '史'] ['水', '路', '运输'] 1543300074113THF8V2
0.03251453209401979 ['交通', '运输', '经济'] ['中国', '史'] 1543300104648MfIZ9D
0.0274858110186926 ['职业', '技术', '教育'] ['中国共产党'] 15433001825O0DANsgV
-0.07625421748304136 ['法律'] ['水', '路', '运输'] 1543300296365XxfhbZ
0.07966067877979058 ['轻工业', '手工业'] ['高等教育'] 1543300373611FcIY8f
0.08839156063328918 ['外', '科学'] ['教育'] 1543300388474FsI1r6
0.04190280486164409 ['自动化', '技术', '计算机', '技术'] ['中国', '政治'] 1543300434188enAtRZ
0.028550832889086195 ['中国', '政治'] ['预防', '医学', '卫生学'] 1543300484722bONdIC
0.024079827000535382 ['中国', '人物', '传记'] ['工业', '经济'] 1543300524496eNjdpC

['人物', '传记'] ['工业', '经济'] 1543300524496eNjdpC

['系统科学', '农业工程', '自动化技术、计算机技术'] 1543299212687UPmzZh
['轻工业、手工业', '化学工业', '基础医学'] 1543299235518MxiPBO
['财政', '金融'] ['天文学'] 1543299243444XSNJOb
['矿床学', '地球物理勘探', '石油、天然气工业'] 1543299280078 3mcL6M
['测绘', '学'] ['农业', '经济'] 1543299300317JcGINi
['工业经济', '经济计划与管理', '贸易经济'] 1543299389125OClgzt
['金属学与金属工艺', '化学工业', '一般工业技术'] 1543299410557ymFvLe
['一般工业技术', '金属学与金属工艺', '化学工业'] 1543299427910JLCbVO
['工业经济', '废物处理与综合利用', '金属学与金属工艺'] 1543299427946wVxi10
['农作物', '化学', '轻工业、手工业'] 1543299444860 3fwMel
['轻工业、手工业', '高等教育', '信息与知识传播'] 1543299452754Mg4m9q
['轻工业', '手工业'] ['高等教育'] 1543299452762ia4e0t

新知识发现　案例分析

文献摘要	多标签	新分类
在分析农业决策支持系统存在问题的基础上，以Jay W. Forrester教授创立的系统动力学为基础，建立了玉米种业仿真动态模型：用面向对象的C++进行程序设计，完成了玉米种业系统的仿真；科学的建模方法与先进计算机编程技术的结合，为建立更加科学与符合实际的玉米种业仿真系统提供关键技术。	系统科学 农业工程 自动化技术 计算机技术	互联网“+”
随着我国科学技术的快速发展，现代化生产与生活中能够应用的各项技术不断增多，近些年3D技术应用较为广泛，3D技术在机械零部件逆向工程中的有效应用具有较高价值。过去传统机械零部件逆向工程都是通过CAD设计机械零部件的基本模型，但是CAD是通过平面方式构建立体化模型，此类形式便于直接进行外部观察。在各类技术发展背景下，机械零部件逆向工程发展速度加快，机械零部件工程在逐步完善的基础上推动了我国工业化发展。	机械仪表工业 自动化技术 计算机技术 金属学与金属工艺	互联网“+”
铁路机务段设计具有涉及专业多、专业接口复杂的特点，传统二维设计容易出现管线碰撞及设计缺陷。为提高设计质量，以迁建西安机务段项目为依托，对BIM技术在大型铁路工程中的应用进行探索。本项目运用Inventor、Revit、Navisworks等设计软件作为BIM设计工具，将所涉及的14个专业工程内容全部进行BIM设计。通过建立族库、各专业BIM设计、模型整合的流程，构建一个覆盖全专业的数字化仿真模型，并总结出一套机务段BIM设计流程。实现了模型可视化展示、二维出图、属性信息添加、工程量统计、碰撞检测等应用。机务段工程设计复杂，专业接口众多，运用BIM技术可直观展示设计意图，并能轻易发现各专业设计中的错、漏、碰、缺问题，为提高设计质量提供有力保障。	铁路运输 公路运输 自动化技术 计算机技术	互联网“+”
本文从湖南省县级耕地质量等别年度更新评价工作实际出发，阐述如何利用ArcGIS选择工具快速完成耕地质量等别年度更新评价的底图制作，利用ArcGIS中选择工具来实现耕地质量等别年度更新评价工作中变化数据的快速提取，从而使这项繁琐的工作变得简单快捷，为高效开展耕地质量等别年度更新评价工作提供经验。	测绘学 农业 经济 计算机技术	互联网“+”

文献摘要	多标签	新分类
4G网络时代手机自媒体迅速发展，传统茶文化不断发扬光大。本文探究手机自媒体与传统茶文化在思政教育上创新结合的价值与路径，推动思想政治教育模式不断创新。	轻工业 手工业 高等教育 信息与知识传播	轻工业、传统文化和高等教育
茶艺是一门应用技能型学科，是茶文化的重要组成部分。在“一带一路”背景下，高校茶艺课程开展双语教学，达到了传播中国茶文化，发展中国茶艺技能的目的。基于学科特点和对高素质人才的需求，对茶艺课程的教学内容、教学方法、考核方式等方面进行了教学改革与探索。	轻工业 手工业 高等教育	轻工业、传统文化和高等教育
“一带一路”倡议的实施，对高级国际商务人才提出新的要求。由于茶叶在“一带一路”商务活动中地位独特、具有一定教育功能以及茶文化是国际交流的重要符号，茶文化融入国际商务专硕培养很有必要。但是目前国际商务硕士培养存在文化缺失、跨文化教育较少等问题，因此，在“一带一路”背景下，探索茶文化融入国际商务专硕培养路径是一种有益地尝试。	轻工业 手工业 高等教育	轻工业、传统文化和高等教育


BELT ROAD SUMMIT
一帶一路高峰論壇
28/6/2018・香港會議展覽中心
www.beltandroadsummit.hk
Belt and Road Summit
28/6/2018 · HKCEC
www.beltandroadsummit.hk

21世纪海上丝绸之路的重要枢纽

Strategically Located Sri Lanka

应用场景讨论

预测目标：**问题的话题(topic)**

问题的标题(title)

知乎　人工智能　数据挖掘　机器学习

如何评价 2017 知乎看山杯机器学习比赛？

寻找算法界最强大脑，2017 知乎 - 看山杯机器学习挑战赛开始了！
比赛链接：Introduction

和其他机器学习比赛相比，看山杯有哪些特点和难点？

问题的描述(Description)

为博取眼球乱贴、多贴标签

网站标签的自动生成

预测目标：**问题的话题(topic)**

问题的标题(title)

如何评价 2017 知乎看山杯机器学习比赛？

寻找算法界最强大脑， 2017 知乎 · 看山杯机器学习挑战赛开始了！

比赛链接：Introduction

和其他机器学习比赛相比，看山杯有哪些特点和难点？

问题的描述(Description)

网站标签的自动生成

iption)

CSDN 博客 学院 下载 图文课 论坛 APP 问答 商城 VIP会员 活动 招聘 ITeye GitChat 搜CSDN

下载首页 精品专辑 排行榜 我的资源 上传资源赚积分 已下载 我的收藏

下载 > 大数据 > 算法与数据结构 > word2vec情感分析实例

ZIP

word2vec情感分析实例

python平台情感分析实例，使用gensim中的doc2vec实现，可用于新版gensim。

★ 3 情感分析, gensim, word2vec

2017-08-01 上传 大小：43.14MB

所需: 50积分/C币 开通VIP 立即下载

分享 收藏(10) 举报

相关视频课程： 【VIP免费】 <4>数据结构与算法（C/C++实现）视频... 【VIP免费】 数据结构与算法在实战项目中的应用

评论

共7条

qq_21578125：不是我想要的，浪费了三十积分 ★★☆☆☆ 2018-09-14

hiheiheicdn：非常有用，谢谢 ★★★★★ 2017-11-29

weixin_41012593：imdb.d2v没有 ★☆☆☆☆ 2017-11-15

上一页 1 下一页

应用场景讨论

预测目标：**问题的话题(topic)**

问题的标题(title)

知乎　人工智能　数据挖掘　机器学习

如何评价 2017 知乎看山杯机器学习比赛？

寻找算法界最强大脑，2017 知乎 - 看山杯机器学习挑战赛开始了！
比赛链接：Introduction

和其他机器学习比赛相比，看山杯有哪些特点和难点？

问题的描述(Description)

为博取眼球乱贴、多贴标签

网站标签的自动生成

应用场景讨论

某企业想邀请一位大数据专家来进行项目评审。

该如何去判断一位专家是不是大数据专家呢?

打听，询问，或查看homepage。

如何判断专家的标签是否准确?

由论文、会议和授课等信息生成其专长领域多标签

专家领域自动生成

[Teaching] [DataGroup@Utah][The InitialD Lab][InitialD GitHub][LinkedIn] [Contact]

I am a Professor in the School of Computing at University of Utah. I obtained my Ph.D in computer science from the Computer Science Department at Boston University in summer 2007, and was an Assitant Professor in the Computer Science Department at Florida State University from Aug 2007 to Aug 2011. A more detailed version about myself is here [PDF]. I am a member of the Data Group at Utah.

Data Driven Science

- Dong Xie is one of the 10 Microsoft Research PhD Fellows for Class of 2018-2019! See the news release for more details.
- We are excited to open source the first release of the XDB (approXimate DB) system. The first release contains the PosgreSQL (9.4.2) version that has fully integrated online aggregation (with the support for SPJ and Group By queries, including joins over multiple tables) into the kernel of PostgreSQL.
- We have released both the kernel-version and the stand-alone version of Simba system, which is a Spark/Spark-SQL extention to in-memory cluster-based large scale spatial data analytical engine.
- Our study on using deep learning techniques to perform sentiment analysis over geo-tagged tweets for analyzing and predicting this year's Presidential election is covered by Salt Lake Tribune and Deseret News. See our project website at http://www.cstorm.org
- The joint work on using social media data to build healthy/unhealthy geo-indicators with quynh nguyen from department of health promotion and education and others is featured on time.com. the project hashtaghealth is open sourced on github and an interactive map layer is also available.
- Our STORM project is highlighted in NSF's news release on ``Data Driven Science'': College of Engineering News Report., NSF News Release.

Research Interests

Database systems and large-scale data management, systems, and analytics. Security issues in data management and systems. Mining and machine learning driven methods for system performance tuning, monitoring, and data analytics. I am grateful to NSF for supporting my past and ongoing research on:

- database security issues (projet completed)
- ranking and aggregate query processing in probabilistic data (projet completed)
- distributed data management system with potentially fuzzy and uncertain data
- new architectures and systems for big data workloads
- building trustworthy cloud platforms

该如何去判断一位专家是不是大数据专家呢？

打听，询问，或查看homepage。

如何判断专家的

由论文、会议和授课等信息生成其专

专家领域自

[Teaching] [DataGroup@Utah][The InitialD Lab][InitialD GitHub][Contact]

I am a Professor in the School of Computing at University of Utah. I obtained my Ph.D in computer science from the Computer Science Department at Boston University in summer 2007, and was an Assitant Professor in the Computer Science Department at Florida State University from Aug 2007 to Aug 2011. A more detailed version about myself is here [PDF]. I am a member of the Data Group at Utah.

Data Driven Science

- Dong Xie is one of the 10 Microsoft Research PhD Fellows for Class of 2018-2019! See the news release for more details.
- We are excited to open source the first release of the XDB (approXimate DB) system. The first release contains the PosgreSQL (9.4.2) version that has fully integrated online aggregation (with the support for SPJ and Group By queries, including joins over multiple tables) into the kernel of PostgreSQL.
- We have released both the kernel-version and the stand-alone version of Simba system, which is a Spark/Spark-SQL extention to in-memory cluster-based large scale spatial data analytical engine
- Our study on using deep learning techniques to perform sentiment analysis over geo-tagged tweets for analyzing and predicting this year's Presidential election is covered by Salt Lake Tribune and Deseret News. See our project website at http://www.estorm.org
- The joint work on using social media data to build healthy/unhealthy geo-indicators with quynh nguyen from department of health promotion and education and others is featured on time.com. the project hashtaghealth is open-sourced on github and an interactive map layer is also available.
- Our STORM project is highlighted in NSF's news release on ``Data Driven Science'': College of Engineering News Report., NSF News Release.

Research Interests

Database systems and large-scale data management, systems, and analytics. Security issues in data management and systems. Mining and machine learning driven methods for system performance tuning, monitoring, and data analytics. I am grateful to NSF for supporting my past and ongoing research on:

- database security issues (projet completed)
- ranking and aggregate query processing in probabilistic data (projet completed)
- distributed data management system with potentially fuzzy and uncertain data
- new architectures and systems for big data workloads
- building trustworthy cloud platforms

应用场景讨论

某企业想邀请一位大数据专家来进行项目评审。

该如何去判断一位专家是不是大数据专家呢?

打听，询问，或查看homepage。

如何判断专家的标签是否准确?

由论文、会议和授课等信息生成其专长领域多标签

专家领域自动生成

[Teaching] [DataGroup@Utah][The InitialD Lab][InitialD GitHub][LinkedIn] [Contact]

I am a Professor in the School of Computing at University of Utah. I obtained my Ph.D in computer science from the Computer Science Department at Boston University in summer 2007, and was an Assitant Professor in the Computer Science Department at Florida State University from Aug 2007 to Aug 2011. A more detailed version about myself is here [PDF]. I am a member of the Data Group at Utah.

Data Driven Science

- Dong Xie is one of the 10 Microsoft Research PhD Fellows for Class of 2018-2019! See the news release for more details.
- We are excited to open source the first release of the XDB (approXimate DB) system. The first release contains the PosgreSQL (9.4.2) version that has fully integrated online aggregation (with the support for SPJ and Group By queries, including joins over multiple tables) into the kernel of PostgreSQL.
- We have released both the kernel-version and the stand-alone version of Simba system, which is a Spark/Spark-SQL extention to in-memory cluster-based large scale spatial data analytical engine.
- Our study on using deep learning techniques to perform sentiment analysis over geo-tagged tweets for analyzing and predicting this year's Presidential election is covered by Salt Lake Tribune and Deseret News. See our project website at http://www.cstorm.org
- The joint work on using social media data to build healthy/unhealthy geo-indicators with quynh nguyen from department of health promotion and education and others is featured on time.com. the project hashtaghealth is open sourced on github and an interactive map layer is also available.
- Our STORM project is highlighted in NSF's news release on ``Data Driven Science'': College of Engineering News Report., NSF News Release.

Research Interests

Database systems and large-scale data management, systems, and analytics. Security issues in data management and systems. Mining and machine learning driven methods for system performance tuning, monitoring, and data analytics. I am grateful to NSF for supporting my past and ongoing research on:

- database security issues (projet completed)
- ranking and aggregate query processing in probabilistic data (projet completed)
- distributed data management system with potentially fuzzy and uncertain data
- new architectures and systems for big data workloads
- building trustworthy cloud platforms

感谢聆听！

扫一扫上面的二维码图案，加我微信