NLP 第一周报告-梁继越

1. 知识点总结

1） TF-IDF

TF-IDF 是 Term Frequency - Inverse Document Frequency 的缩写，即“词频-逆文本频率”。其计算方法为：

$$IDF(x) = log\frac{N}{N(x)}$$

$$TF\_IDF(x) = TF(x) * IDF(x)$$

词频（TF）的值越高，说明该词出现的次数越多，其重要性应该提高。但对于一些几乎在所有文本都会高频出现的词，其重要性应当适当降低，这个值就体现在 IDF 值上。

2） 关于 Logistic Regression 和 SVM

Rule	Ex.	Method
n>>m	n=10000 m=10-1000	LR SVM with linear kernel
n小m适中	n=1-1000 m=10-10000	SVM
n<<m	通常效果不好，需要增加feature	LR SVM with linear kernel

n: feature 数目 m: sample 数目

3） word2vec 和 doc2vec

word2vec 就是将词表征为实数值向量的一种算法模型，其利用深度学习的思想，可以通过训练，把对文本内容的处理简化为 K 维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似。

Word2vec 忽略了词与词之间的顺序排列带来的语义变化，而 doc2vec 解决了这个问题。

2. 任务报告

任务一：

利用好未来提供的题库数据，能够将每一道题目用一个词袋向量表示。具体向量表示方法请分别采用 tf 和 tf-idf。同时进行自我检测。自我检测：自己随机选取 5 个题目，针对

自己选取的每一个题目，计算该题目向量和其他所有题目向量的相似度（cosine 相似度），然后选出和该题目最接近的 3 个题目，并将结果记录在本周最后要提交的报告中。

结果展示：

随机抽选的第 1/5 题：

题号：30726
题目内容：设 集合 由 所有 不 超过 并且 其 二进制 表示 式 中 恰有 两个 数字 为 的 正整数 组成 从中 随机 取出 一个 数则 这个 数能 被 整除 的 概率 为

相似度最高的三个题目为：[72327, 9664, 18124]

72327 在 和 两个 集合 中 各取 一个 数 组成 一个 两位数 则 这个 数能 被 整除 的 概率 是
9664 在 所有 的 两位数 中任取 一个 数则 这个 数能 被 或 整除 的 概率 是
18124 设 为 所有 介于 与 之间 且 二进制 表示 式 中 恰有 两个 其余 为 的 整数 组成 的 集合 从中 随机 取出 一个 数设 这个 数 被 整除 的 概率 为 其中 是 互素 的 正整数 试求 的 值

随机抽选的第 2/5 题：

题号：87995
题目内容：设 双曲线 的 虚轴 长为 焦距 为 则 它 的 渐近线 方程 为

相似度最高的三个题目为：[78914, 29293, 46910]

78914 双曲线 的 焦距 为 则 渐近线 方程 为
29293 双曲线 的 虚轴 长为
46910 设 双曲线 的 虚轴 长为 焦距 为 则 双曲线 的 渐近线 方程 为

随机抽选的第 3/5 题：

题号：20923
题目内容：对于 函数 部分 与 的 对应 关系 如下 表 数列 满足 且 对 任意 点 都 在 函数 的 图象 上 则 的 值 为

相似度最高的三个题目为：[67199, 63501, 88627]

67199 已知 函数 的 对应 关系 如下 表 所示 数列 满足 则
63501 对于 函数 部分 与 的 对应 关系 如下 表若 数列 满足 且 对 任意 点 都 在 函数 的 图像 上 则
88627 对于 函数 部分 与 的 对应 关系 如下 表 数列 满足 且 对 任意 点 都 在 函数 的 图象 上 则 的 值 为

随机抽选的第 4/5 题：

题号：47613
题目内容：已知 则

相似度最高的三个题目为：[42648, 51656, 11406]

42648 已知
51656 已知 则
11406 已知 且

随机抽选的第 5/5 题：

题号：5537
题目内容：求证 平面

相似度最高的三个题目为：[41818, 86831, 3788]

41818 求证 平面
86831 求证 平面
3788 求证 平面

总结：从随机选取的 5 个句子的相似题目来看，三道题目的相似度都很高，说明采用 tf-idf 已经可以很好的刻划句子特征。

任务二：

利用好未来提供的题库数据，抽取所有一级知识点为“三角函数与解三角形”和“函数与导数”的题目。能够将每一道题目用一个向量表示。具体向量表示方法请采用 tf-idf。用自己学习到的分类器（svm 或者 logistic regression），对“三角函数与解三角形”和“函数与导数”的题目进行分类，并汇报最终的 precision, recall, accuracy, F1 score, ROC and AUC，并将结果记录在本周最后要提交的报告中。

思路分析：

首先通过 knowledge_hierarchy.csv 的“name”标签获得“三角函数与解三角形”和“函数与导数”分别所对应的 id，再通过该 id 找到 question_knowledge_hierarchy.csv 中对应的题目，最后通过将 tiku_question_sx.csv 的“que_id”和 question_knowledge_hierarchy.csv 的“question_id”相同的内容进行合并，就找到了全部“三角函数与解三角形”和“函数与导数”的题目内容。

结果展示：

1）满足条件的一级知识点 id：

	id	name	degree
8850	hcwf4avcmp8l53s5iq010pelwlce000d	函数与导数	1
8890	hcwf4avcmp8l53s5iq010pelwlce0035	三角函数与解三角形	1

2）满足条件的题目内容：

	kh_id	question_id		content
0	hcwf4avcmp8l53s5iq010pelwlce000d	a65e49dd26a0476cb69e32e5f5e511e5	$f(x)$ $(0,+\infty)$	已知 为定义在 上的连续可导函数，且$$f...
1	hcwf4avcmp8l53s5iq010pelwlce000d	ff8080814db3e529014df75290f11e7e		已知函数$$f\left(x \right)=\left| \ln x \right...
2	hcwf4avcmp8l53s5iq010pelwlce000d	2965364a3fa14cbf9e3e953d27fb9254	R^2 f	将所有平面向量组成的集合记作 ， 是从$${{R}^{...
3	hcwf4avcmp8l53s5iq010pelwlce0035	d568513c37fa4621b1e516ebcaf5bd48	$f(x)$	请用“五点法”画出函数 在长度为一个周期的闭区间...
4	hcwf4avcmp8l53s5iq010pelwlce000d	a81eb192610c45c98b4480afb6400ee3	$f(x)$	求函数 的极值．

3）效果评价：

```
score:          0.8522794262134015
precision:      [0.87369304 0.78403949]
recall:         [0.9280181 0.6607731]
f_score:        [0.90003657 0.71714796]
```

总结：在这里学习到了通过 id 匹配来寻找需要的内容的方法，非常实用。但对于 Logistic 回归掌握的并不是很透彻，这里只是暂时会使用，需要进一步学习和练习。另外在计算中遇到了“Data is not binary and pos_label is not specified”的问题，目前尚未解决。

任务三：

利用好未来提供的题库数据，对所有知识点进行分词，记为 W。利用“网上已经训练好”的 word2vec 模型，计算 W 中每个词的词向量。同时进行自我检测。自我检测：自己随机选取 5 个词，针对自己选取的每一个词，计算该词向量和其他所有词向量的相似度（cosine 相似度），然后选出和该词最接近的 10 个词，并将结果记录在本周最后要提交的报告中。

结果展示：

随机抽选的第 1/5 个数字： 个体

相似度最高的 10 个词为：

群体 0.7204607725143433
生物体 0.697733461856842
性状 0.6883518695831299
隐性 0.6749706268310547
种群 0.6630771160125732
哺乳动物 0.6545930504798889
神经元 0.6477792263031006
变异 0.6394734382629395
知觉 0.6304793357849121
人类 0.6194202899932861

随机抽选的第 2/5 个数字： 葡萄糖

相似度最高的 10 个词为：

乙酰 0.7864810228347778
磷酸 0.7829124331474304
甘油 0.7749233841896057
脂肪酸 0.7702838182449341
谷氨酸 0.7680040001869202
酪氨酸 0.7672659158706665
丙酮酸 0.7665181756019592
氨基酸 0.7570896148681641
乳酸 0.7563743591308594

水解　　0.7553744316101074

随机抽选的第 3/5 个数字： 统一

相似度最高的 10 个词为：

规范　　0.5694348812103271
独立　　0.5470969080924988
中央集权　　0.5459444522857666
标准化　　0.5416475534439087
拟定　　0.5138547420501709
团结　　0.5058429837226868
规范化　　0.5009782314300537
确立　　0.49835145473480225
实现　　0.4970797896385193
政体　　0.49621057510375977

随机抽选的第 4/5 个数字： 东北

相似度最高的 10 个词为：

东南　　0.718100905418396
华北　　0.7001115679740906
东北地区　　0.6743338704109192
大兴安岭　　0.6342449188232422
西北地区　　0.6196337938308716
辽东半岛　　0.616443395614624
西北　　0.6151131987571716
华中　　0.6074798107147217
内蒙古　　0.6058622598648071
西南地区　　0.6057032942771912

随机抽选的第 5/5 个数字： 交通线

相似度最高的 10 个词为：

铁路干线　　0.6722004413604736
东线　　0.6307903528213501
南线　　0.6275345087051392
横贯　　0.62709641456604
铁路线　　0.6263600587844849
交通枢纽　　0.6193798184394836
纵贯　　0.6126803159713745
内河　　0.6051727533340454
山东半岛　　0.603766918182373
铁路沿线　　0.6025924682617188

总结：从以上结果可以看出网上下载的模型对相似词语预测的较好，但由于该模型中存在大量繁体字，可能使其准确度略有下降。

需要注意去除模型中不存在的字词，否则会出现找不到词语的情况。

任务四：

利用好未来提供的题库数据，自己训练针对教育场景中的 word2vec 模型。利用“自己训练好”的 word2vec 模型，计算 W 中每个词的词向量。同时进行自我检测。自我检测：自己随机选取 5 个词，针对自己选取的每一个词，计算该词向量和其他所有词向量的相似度（cosine 相似度），然后选出和该词最接近的 10 个词，并将结果记录在本周最后要提交的报告中。最后对比不同 word2vec 模型找出的相似词语的差别，并将结果记录在本周最后要提交的报告中。

结果展示：

随机抽选的第 1/5 个数字：　差型

相似度最高的 10 个词为：

化学性质　　0.41599100828170776
频率　　0.4097214341163635
遗传病　　0.40654873847961426
整千　　0.4059681296348572
其他　　0.4037994146347046
四边形　　0.40273135900497437

碳 0.40147125720977783
形式 0.3992994427680969
坐标 0.3976958096027374
平面 0.39723241329193115

随机抽选的第 2/5 个数字： 比解

相似度最高的 10 个词为：

光合作用 0.3777366578578949
交变 0.36065375804901123
体液 0.35255420207977295
环形 0.35213813185691833
不植 0.34862396121025085
生物进化 0.3473646640777588
功能 0.3470805287361145
公因式 0.34166157245635986
弹力 0.34001481533050537
电场 0.3383066654205322

随机抽选的第 3/5 个数字： 合成纤维

相似度最高的 10 个词为：

分析 0.5949720740318298
社会 0.5833403468132019
弦 0.5794479846954346
意义 0.5782973766326904
自然地理 0.576545000076294
定理 0.5761302709579468
有 0.5751723051071167
基本 0.5749151110649109
思维 0.5746855735778809
生物 0.5745463371276855

随机抽选的第 4/5 个数字： 指向

相似度最高的 10 个词为：

交叉　　0.42817533016204834
交变　　0.42758601903915405
国家　　0.4269413948059082
溶解　　0.42278918623924255
电路　　0.4203123152256012
计数问题　　0.41620123386383057
求和　　0.4141634404659271
计数　　0.41161829233169556
社会　　0.4103546142578125
方程组　　0.41027241945266724

随机抽选的第 5/5 个数字：　升值

相似度最高的 10 个词为：

资本主义　　0.4703207015991211
有机物　　0.44928082823753357
乘除　　0.4329111576080322
找　　0.4293348789215088
作用　　0.4273974895477295
人民代表大会　　0.42484137415885925
平衡　　0.42330265045166016
金属　　0.4230784475803375
水解　　0.4213751554489136
聚落　　0.4200463593006134

总结：

该模型采用题库中的题目数据进行训练，由于样本量太小，与任务 3 相比，自己训练的模型准确度较差。

任务五：

利用好未来提供的题库数据，抽取所有一级知识点为“三角函数与解三角形”和“函数与导数”的题目。利用 word2vec 或 doc2vec，将每一道题目用一个向量表示。再次用自己

学习到的分类器（svm 或者 logistic regression），对“三角函数与解三角形”和“函数与导数”这两类题目进行分类，并汇报最终的 precision, recall, accuracy, F1 score, ROC and AUC，并将结果记录在本周最后要提交的报告中。

分析：

该题目可继续利用任务二中抽取的数据，只是处理模型稍有不同。

结果：

```
score:  0.7211141678129298
 precision:  [0.72495455 0.58333333]
 recall:  [0.98423254 0.05581557]
 f_score:  [0.83492774 0.10188261]
```

总结：与任务二类似，由于对 Logistic 和 SVM 掌握的太肤浅，此处的模型使用可能有问题，需要进一步学习。

一周总结

这一周的学习让我从一个 NLP 的小白进化到了一个入门者，这一周接触和学习了词袋、词频、词性、TF-IDF、分词、Logistic Regression、SVM 等等一系列概念，也学习使用了词频计算、word2vec、doc2vec 等工具，让我初步具备了分词、向量化和相似性判断的能力。但这些能力都尚且停留在会使用工具的层面上，需要在今后的学习中进一步加深对其原理的理解和认识。