文本内容对基于doc2vec的标准文本与答案文本相似性影响的研究

- **问题描述：**

探讨相似性反映的本质问题，以及那些因素是影响相似性的主要因素。为后期基于文档向量的语义特征和开放题评分性能提升作铺垫。

目标：确认相似性是否可以体现语义信息

名词解释：

标准文本与答案文本相似性，指标准文本的文档向量均值与答案文本向量之间的cos夹角的大小，取值范围[-1,1]，值越大相似性越高，1为最大值

数据集：eng_hyt140326 （两套卷）与AES数据的混合

- **研究中主要涉及的问题：**

问题1：基于标准答案间的平均向量是否可以作为标准答案的文档向量
问题2 ：不同主题之间的相似性？
问题3：同一个主题内部，中心句与标准答案文档向量的相似性，以及中心句中关键词（实体词）的影响，或者中心句中词的排续对相似性的影响
问题4：接问题3，若中心句改为反义的或与主题表达相反的信息，相似性的影响情况
问题5：多个中心句的纯重复，对相似的影响
问题6：题型的宽泛性差异对相似性特征的影响

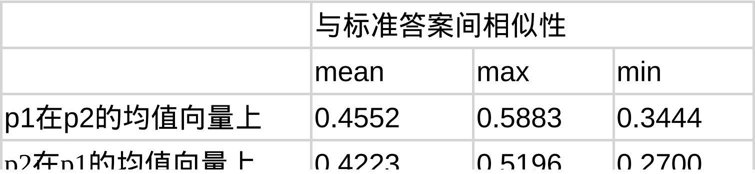
问题1：基于标准答案间的平均向量是否可以作为标准答案的文档向量
方案：不同标准答案与生成的标准答案文档向量间的相似性情况，与整个训练集该paper下的相似性均值差异；
PAPER-000001-QT-000002 与 PAPER-000002-QT-000002题型差异：
PAPER-000001-QT-000002 描述的是car给人们带来的好的方面与不好的方面，比较具体，文章结构比较统一；而PAPER-000002-QT-000002描述的让你happiness的行为（如读书，sports等），主题偏宽泛，文章结果相对不太统一
表1.1
配置 Doc2Vec(dm=1, dm_concat=1, size=50, window=4, negative=15, hs=0, min_count=1, workers=cores, sample=0)

	与标准答案间相似性			总体相似性	
	mean	max	min	train	test
PAPER-000001-QT-000002	0.9335	0.9729	0.8700	0.7357	0.7167
PAPER-000002-QT-000002	0.8544	0.9189	0.7459	0.6882	0.6789
mean				0.7119	0.6977

表1.2
配置 Doc2Vec(dm=1, dm_concat=1, size=50, window=4, negative=20, hs=0, min_count=1, workers=cores, sample=0)
向量 infer alpha=0.001, min_alpha=0.0001, steps=40

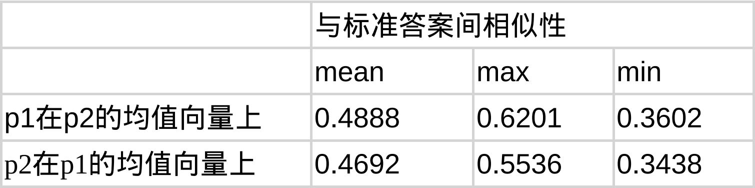
	与标准答案间相似性			总体相似性	
	mean	max	min	train	test
PAPER-000001-QT-000002	0.9404	0.9757	0.8844	0.8617	0.8611
PAPER-000002-QT-000002	0.8746	0.9226	0.7990	0.8204	0.8280
mean				0.8411	0.84448

问题2 ：不同主题之间的相似性？
方案：观察数据集在某个不相关主题上相似性变化，均值，适当分析相似高的文本
在eng_hyt140326的两套题都做以上分析
p1 :PAPER-000001-QT-000002　 p2:PAPER-000002-QT-000002
PAPER-000001-QT-000002_02与PAPER-000002-QT-000002_02间的相似性：0.36(model1)0.39（model2）
表2.1



	与标准答案间相似性		
	mean	max	min
p1在p2的均值向量上	0.4552	0.5883	0.3444
p2在p1的均值向量上	0.4223	0.5196	0.2700

表2.2



	与标准答案间相似性		
	mean	max	min
p1在p2的均值向量上	0.4888	0.6201	0.3602
p2在p1的均值向量上	0.4692	0.5536	0.3438

问题3：同一个主题内部，中心句与标准答案文档向量的相似性，以及中心句中关键词（实体词）的影响，或者中心句中词的排续对相似性的影响

中心句定义：能够概括文章主要内容的句子就是“中心句”。一般在文章开头或结尾。

注：若去掉个别关键词对相似性影响不大，或中心句的排序对相似的影响较大，可以说明相似体现的是语义上的信息

方案：选取两个题型中与标准答案向量相似性最大的文章作为研究对象，挑选出这些文章中的中心句。再作进一步分析

PAPER-000001-QT-000002_02(0.9689) :

car bring many convenience we can go wherever we want and quickly get to our destination and it bring many problems either it cause environment pollution and make our living around wrong and even bring many disease that do harm to our health because too many many cars in the city it cause traffic jam in my opinion we can do something to deal with this for example we can go out for a walk instead of driving cars we can use public traffic transportation instead of cars and plant many trees to improve our quality i think the car bring our convenience and we should care more about the environment so lets live in green life

分句

1 car bring many convenience we can go wherever we want and quickly get to our destination and it bring many problems either

2 it cause environment pollution and make our living around wrong and even bring many disease that do harm to our health

3 because too many many cars in the city it cause traffic jam

4 in my opinion we can do something to deal with this

5 for example we can go out for a walk instead of driving cars

6 we can use public traffic transportation instead of cars and plant many trees to improve our quality

7 i think the car bring our convenience and we should care more about the environment so lets live in green life

中心句：

1、car bring many convenience we can go wherever we want and quickly get to our destination and it bring many problems either

random ： quickly go our and want destination to either convenience car we and many problems we bring get can wherever bring many it

2、i think the car bring our convenience and we should care more about the environment so lets live in green life

random： the car lets and so bring care we think should convenience our in about the i green more life live environment

PAPER-000002-QT-000002_02 (0.9168):

my topic is about happiness different people have different opinions what makes me happy is reading reading is meaningful it also tells me something that i couldn't learn from my life for example my favorite boob is romeo and juliet it tells us the story about young lovers fight against for their freedom and love what i learned from the book is that we should be encouragement and i also like to share my feeling with others when i finished a book when i shared my feeling with other people we can exchange different feelings and this makes me happier that's why i like to share my happiness with each other that's all thank you

分句：

1 my topic is about happiness

2 different people have different opinions what makes me happy is reading

3 reading is meaningful

4 it also tells me something that i couldn't learn from my life

5 for example my favorite boob is romeo and juliet

6 it tells us the story about young lovers fight against for their freedom and love
7 what i learned from the book is that we should be encouragement
8 and i also like to share my feeling with others when i finished a book
9 when i shared my feeling with other people we can exchange different feelings and this makes me happier that's why i like to share my happiness with each other
10 that's all thank you

中心句：
1 different people have different opinions what makes me happy is reading
random ：is people what makes different me happy different have opinions reading
2 when i shared my feeling with other people we can exchange different feelings and this makes me happier that's why i like to share my happiness with each other
random ：can my happier different feelings and with with shared other i when i makes that's my each this like exchange to me share we feeling happiness why other people

表3.1

	整段与标准答案	答案平均相似性	答案的最低相似性	中心句	去除关键词	no key 相似性	乱序 随机
PAPER-000001-QT-000002_02	0.9689	0.9335	0.8844	0.8991/ 0.9175	car	1 0.8865/ 1 0.8748	0.6208/ 0.5652
PAPER-000002-QT-000002_02	0.9168	0.8544	0.7990	0.7569/ 0.8417	happy，happiness	1 0.4180/ 2 0.8646	0.6384/ 0.5660

小结：文档向量中包含了文本的篇章结构顺序；文章的最后一句中心句与文档向量相关性较高；中心句中不止是关键词包含了语义信息。

*补充：将段落中各句作为文档向量，的文本相似性情况，及中心句的相似高是否有体现。
PAPER-000001-QT-000002_02
id mean   doc vec
1 0.8805 0.91 car bring many convenience we can go wherever we want and quickly get to our destination and it bring many problems either
2 0.7337 0.82 it cause environment pollution and make our living around wrong and even bring many disease that do harm to our health
3 0.8188 0.80 because too many many cars in the city it cause traffic jam
4 0.6109 0.66 in my opinion we can do something to deal with this
5 0.7973 0.75 for example we can go out for a walk instead of driving cars
6 0.8866 0.83 we can use public traffic transportation instead of cars and plant many trees to improve our quality
7 0.9298 0.92 i think the car bring our convenience and we should care more about the environment so lets live in green life

PAPER-000002-QT-000002_02
id mean   doc vec
1 0.6488 0.54 my topic is about happiness
2 0.7583 0.58 different people have different opinions what makes me happy is reading
3 0.2915 0.28 reading is meaningful
4 0.1842 0.04 it also tells me something that i couldn't learn from my life
5 0.4725 0.36 for example my favorite boob is romeo and juliet
6 0.3990 0.45 it tells us the story about young lovers fight against for their freedom and love
7 0.6199 0.41 what i learned from the book is that we should be encouragement
8 0.6852 0.55 and i also like to share my feeling with others when i finished a book
9 0.8675 0.66 when i shared my feeling with other people we can exchange different feelings and this makes me happier that's why i like to share my happiness with each other

10 0.02 -0.07 that's all thank you

注：因文档向量初始值是随机的，故以上相似性会有波动。

小结：从以上结果中，可以看到，相似性高的句子就是中心句。

问题4：接问题3，若中心句改为反义的或与主题表达相反的信息，相似性的影响情况

PAPER-000001-QT-000002_02

i think the car bring our convenience and we should care more about the environment so lets live in green life

反义句：i think the car can not bring our convenience and we should not care more about the environment so lets live in bad life

PAPER-000002-QT-000002_02

when i shared my feeling with other people we can exchange different feelings and this makes me happier that's why i like to share my happiness with each other

反义句：when i shared my feeling with other people we can not exchange different feelings and this makes me no happier that's why not i like to share my happiness with each other

表4.1

	中心句	反义句
PAPER1	0.92	0.92
PAPER2	0.84	0.81

说明：反义句与原句的区分性几乎没有。

问题5：多个中心句的纯重复，对相似的影响

PAPER-000001-QT-000002_02

i think the car bring our convenience and we should care more about the environment so lets live in green life

PAPER-000002-QT-000002_02

when i shared my feeling with other people we can exchange different feelings and this makes me happier that's why i like to share my happiness with each other

表5.1

	1sent	2sent	3sent	5sent
PAPER1	0.93	0.92	0.94	0.94
PAPER2	0.86	0.87	0.87	0.84

问题6：题型的宽泛性差异对相似性特征的影响

paper描述：

PAPER-000001-QT-000002 与 PAPER-000002-QT-000002题型差异：

PAPER-000001-QT-000002 描述的是car给人们带来的好的方面与不好的方面，比较具体，文章结构比较统一；而PAPER-000002-QT-000002描述的让你happiness的行为（如读书，sports等），主题偏宽泛，文章结果相对不太统一

与分数间的皮尔森相关性

表6.1 test

PAPER-000001-QT-000002	0.550
PAPER-000002-QT-000002	0.401

表6.2 test

PAPER-000001-QT-000002	0.566
PAPER-000002-QT-000002	0.432

总结：

1、从问题1可以看到总体上基于标准答案的文档向量均值可以作为标准答案的向量；同时，结合问题6也注意到，如果paper的答案过于宽泛，是不利于均值向量的，直接影响到特征的表达。后期可以考虑用多个向量来表达标准答案的向量。

2、从问题2可以看到不同主题间的文档向量差异明显。

3、关于中心句，从文档向量中有明显体现。但如果表达意思虽与主题相关但相反是无法从文档向量中体现

4、文档向量包含了文章的结构信息和字符顺序的信息。

后期实验思考：

1、解决标准答案的文档向量表示；多个向量表示

2、文档向量间差异的增大；