复查测验提交: Homework 1

用户	生物医学工程学院 吉泓光
课程	自然语言处理
测试	Homework 1
已开始	24-3-19 下午7:15
已提交	24-3-19 下午9:21
截止日期	24-3-19 下午11:59
状态	已完成
尝试分数	得 102 分，满分 112 分
已用时间	2 小时 6 分钟

说明

我们将在本学期作业中允许多次尝试，不限制提交次数。请注意：

- 作业将使用最后一次尝试的成绩作为最终成绩；
- 未提交的尝试将被记为0分；
- 当开始新的尝试时，所填入的答案将被完全清除。

因此，当决定提交作业时，请在其他设备上妥善保存已经完成的答案；否则，请保存答案但不要提交。在截止日期之前，请确保作业的最后一次尝试已经提交。在截止日期之后，如果发现作业成绩有任何问题，可以随时联系助教处理。

FAQ

1. 作业有grace day吗?

本学期所有作业没有grace day。针对特殊情况（如身体不适、需要参加比赛、其他课程作业太多等等）需要延期提交作业的，请在**作业截止日期前**向任课老师发送邮件（tukw@），经过批准后由助教处理延期提交的作业。由于答案会在截止日期自动发布，任何作业截止日期后提出的申请将不被批准。

2. 我忘记提交作业了，可以请助教帮忙提交吗?

在同时满足以下条件时，你可以联系助教在ddl之后为你提交作业：
a. 你的当前作业没有成绩，没有提交记录；
b. 你的作业完成记录显示你的所有操作在ddl之前完成。

注意，BB会记录助教的所有操作，这些操作也都将需要归档。

显示的结果 所有答案, 已提交的答案, 正确答案

问题 1

得 10 分，满分 10 分

Choose all of the following in which the regular expression accepts the string. “Accept” means all characters are matched. For example, for a RE “abc”, string “abc123” is rejected because “123” is not matchable.

所选答案：
✓ RE: "`10.5`", String: "`10.5`"
✓ RE: "`(\w+).w?`", String: "`www`"
✓ RE: "`(\..)+...?`", String: "`.+_+.`"

答案：
✓ RE: "`10.5`", String: "`10.5`"
RE: "`10+5`", String: "`10+5`"
✓ RE: "`(\w+).w?`", String: "`www`"
✓ RE: "`(\..)+...?`", String: "`.+_+.`"

问题 2

得 0 分，满分 10 分

Assume a BPE tokenizer with vocab `{_, t, o, g, e, h, r, he, the, er, r_, er_, to, ge, get}` in the learned order, `together` will be tokenized into

所选答案：
✗ `to get he r_`

答案：
`t o g e t h e r _`
`to get he r_`
✓ `to ge the r_`
`to ge th er_`

问题 3

得 10 分，满分 10 分

Consider the sentence: **Kitty’s cat was found behind the television.**
Choose the technique corresponding to the processed result:

Kitti cat wa found behind the televis - **[x1]**
Kitty 's cat be find behind the television - **[x2]**
kitty’s cat was found behind the television - **[x3]**

所选答案： Consider the sentence: **Kitty’s cat was found behind the television.**
Choose the technique corresponding to the processed result:

Kitti cat wa found behind the televis - ✓ **stemming**
Kitty 's cat be find behind the television - ✓ **lemmatization**
kitty’s cat was found behind the television - ✓ **case folding**

答案： Consider the sentence: **Kitty’s cat was found behind the television.**
Choose the technique corresponding to the processed result:

Kitti cat wa found behind the televis - ✓ **stemming**
Kitty 's cat be find behind the television - ✓ **lemmatization**
kitty’s cat was found behind the television - ✓ **case folding**

所有答案选项

- stemming
- lemmatization
- case folding

问题 4

得 10 分，满分 10 分

Choose all of the word vector models that produce **static** and **dense** word embeddings.

所选答案： ✓ Word2vec skip-grams

✓ Latent Semantic Analysis (LSA)

答案： ✓ Word2vec skip-grams

PPMI vectors (without SVD)

✓ Latent Semantic Analysis (LSA)

One-hot vectors

BERT (dynamic, will be introduced later. This option won't count for scores.)

问题 5

得 10 分，满分 10 分

Select all correct statements

所选答案：

✓ A. One-hot word vectors are unable to capture word similarity.

✓ B. A word-word co-occurrence matrix captures some word similarity.

✓ C.
PPMI matrix is sparse, but it can produce dense word embeddings after applying SVD decomposition.

✓ D.
Word-word PPMI matrix, one-hot vectors, and Word2vec assign a fixed embedding for each word independent of contexts.

答案：

✓ A. One-hot word vectors are unable to capture word similarity.

✓ B. A word-word co-occurrence matrix captures some word similarity.

✓ C.
PPMI matrix is sparse, but it can produce dense word embeddings after applying SVD decomposition.

✓ D.
Word-word PPMI matrix, one-hot vectors, and Word2vec assign a fixed embedding for each word independent of contexts.

问题 6

得 12 分，满分 12 分

Given a term-context (word-word) co-occurrence matrix:

puppy and panda are words, (pet, cute, china) are contexts

co-occurrence	pet	cute	china
puppy	5	10	0
panda	0	15	20

Q1: Calculate the number of all occurrences: **[Q1]**.

Q2: Calculate p(word=panda, context=china): **[Q2]**.

Q3: Calculate the PMI(panda, china). use $\log_2$, not weighted, and use two decimal places, (e.g., 0.37). **[Q3]**.

Q4: Calculate the PPMI(panda, cute). use $\log_2$, not weighted, and use two decimal places, (e.g., 0.37). **[Q4]**

Q5: Calculate the PPMI(panda, china). use $\log_2$, not weighted, and use two decimal places, (e.g., 0.37). **[Q5]**

Q6: Calculate the add-2 smoothed PPMI(panda, china). use $\log_2$, not weighted, and use two decimal places. **[Q6]**

Q1 的指定答案： ✓ 50

Q2 的指定答案： ✓ 0.4

Q3 的指定答案： ✓ 0.51

Q4 的指定答案： ✓ 0

Q5 的指定答案： ✓ 0.51

Q6 的指定答案： ✓ 0.47

Q1 的正确答案：

评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	50	
✓ *完全匹配*	50.00	

Q2 的正确答案：

评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	0.4	
✓ *完全匹配*	0.40	

Q3 的正确答案：

评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	0.51	

Q4 的正确答案：

评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	0	
✓ *完全匹配*	0.00	

Q5 的正确答案：

评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	0.51	

Q6 的正确答案：

评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	0.47	

问题 7

得 10 分，满分 10 分

Train a multinomial naive Bayes with add-1 smoothing, on the following document counts for key sentiment words, with positive or negative class assigned as noted.

Note: the vocabulary consists of three words: good, poor and great. All other words are ignored.

doc	"good"	"poor"	"great"	(class)
d1.	3	0	3	pos
d2.	0	1	2	pos
d3.	1	3	1	neg
d4.	1	5	2	neg
d5.	1	3	0	neg

Use both naive Bayes models to analysis sentence:

> A good, good plot and great characters, but poor acting.

Recall that with naive Bayes text classification, we simply ignore (throw out) any word that never occurred in the training document. (We don't throw out words that appear in some classes but not others; that's what add-1 smoothing is for.)

The score(prior*likelihood) that **the multinomial naive Bayes** assign it to class **positive** is **[Q1]**.

The score that **the multinomial naive Bayes** assign it to class **negative** is **[Q2]**.

All answers should be irreducible fractions, e.g., 1/2.

Q1 的指定答案： ✓ 1/270

Q2 的指定答案： ✓ 9/3125

Q1 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	1/270	
Q2 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *完全匹配*	9/3125	

问题 8

得 10 分，满分 10 分

Consider the following confusion matrix:

	gold positive	**gold negative**
system positive	102	9
system negative	54	140

In this question, please use decimals to represent the final result and keep 3 decimal points (e.g. 0.500, 0.167).

What is the precision?

[x1]

What is the recall?

[x2]

What is the accuracy?

[x3]

What is the f1-score?

[x4]

x1 的指定答案： ✓ 0.919

x2 的指定答案： ✓ 0.654

x3 的指定答案： ✓ 0.793

x4 的指定答案： ✓ 0.764

x1 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *模式匹配*	0\.91\d*	
✓ *完全匹配*	0.919	

x2 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *模式匹配*	0\.65\d*	
✓ *完全匹配*	0.654	

x3 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *模式匹配*	0\.79\d*	
✓ *完全匹配*	0.793	

x4 的正确答案：		
评估方式	**正确答案**	**区分大小写**
✓ *模式匹配*	0\.76\d*	
✓ *完全匹配*	0.764	

问题 9

得 10 分，满分 10 分

Consider the following confusion matrix:

	gold tag 1	**gold tag 2**	**gold tag 3**
system tag 1	42	7	11
system tag 2	24	140	3
system tag 3	9	16	132

In this question, please use decimals to represent the final result and keep 3 decimal points (e.g. 0.500, 0.167).

What is the macro-average f1-score?

[x1]

What is the micro-average f1-score?

[x2]

x1 的指定答案： ✓ 0.781

x2 的指定答案： ✓ 0.818

x1 的正确答案：

评估方式	**正确答案**	**区分大小写**
✓ *模式匹配*	0\.78\d*	
✓ *完全匹配*	0.781	

x2 的正确答案：

评估方式	**正确答案**	**区分大小写**
✓ *模式匹配*	0\.81\d*	
✓ *完全匹配*	0.818	

问题 10

得 10 分，满分 10 分

Which of the following clustering algorithms require(s) a distance measure?

所选答案：
✓ K-means
✓ Hierarchical Agglomerative Clustering (HAC)

答案：
✓ K-means
✓ Hierarchical Agglomerative Clustering (HAC)
Expectation-maximization with Mixture of Gaussian (MoG)

问题 11

得 10 分，满分 10 分

For unsupervised Naive Bayes, we

所选答案：
✓
revise each cluster based on its proportionately assigned words in M step.

答案：
assign sentences proportionately to different clusters in M step.

✓
revise each cluster based on its proportionately assigned words in M step.

minimize the log likelihood of each word given its sentence in M step.

take a cup of coffee and do nothing in M step.

2024年5月29日 星期三 下午07时30分39秒 CST

← 确定