به نام یکتای هستی بخش

دانشگاه تهران، دانشکده‌ی مهندسی برق و کامپیوتر

پردازش زبان طبیعی، نیم‌سال دوم، سال تحصیلی ۱۴۰۱-۱۴۰۲

زمان آزمون ۱۵۰ دقیقه | امتحان میان‌ترم | پنجشنبه، ۱۴۰۱/۲/۱۴

solution

لطفاً به نکات زیر توجه فرمایید:

- پاسخ سؤالات را در همین برگه بنویسید.
- بارم سؤالات در جدول زیر آمده است. (لطفاً در جدول زیر چیزی ننویسید.)
- بارم از 111 است که 11 نمره اضافی در نظر گرفته شده است.
- استفاده از موبایل، تبلت و ... در هر حالتی تقلب به شمار می‌آید. لطفا موبایل خود را همین الان خاموش کنید.
- در انتهای امتحان تصویر پاسخها را در سامانه elearn اپلود کنید ولی حتما برگه ها را تحویل دهید.

شماره‌ی سؤال	بارم	نمره
۱	15	
۲	44	
۳	16	
۴	4	
۵	4	
۶	4	
۷	4	
۸	4	
۹	2	
۱۰	4	
۱۱	2	
۱۲	2	
۱۳	4	
۱۴	2	
جمع	111	

۱- بله خیر بگذارید (هر پاسخ درست یک نمره مثبت و هر پاسخ نادرست یک نمره منفی دارد)- سمت راست سوال بله یا خیر بگویید

a. زمانیکه تعداد داده های ما کم است عملکرد طبقه بند Naive Bayes احتمالا از طبقه بند Logistic Regression بهتر است. بله

b. میتوان استقلال دو کلمه از یکدیگر را توسط معیار PPMI سنجید. خیر (به دلیل آنکه ماکزیمم ۰ و PMI سنجیده میشود پس تنها با هم بودن دو کلمه را میتوان سنجید نه مستقل بودن آن ها را)

c. در مدل word2vec به دنبال آن هستیم که وزن های طبقه بند را در انتها استخراج کنیم.خیر (هدف تولید بردار های جانمایی برای کلمات است)

d. برای مسئله spam detection استفاده از مدل Naive Bayes احتمالا بهتر از مدل logistic regression خواهد بود.خیر (زیرا اهمیت برخی از کلمات در تشخیص spam بودن بیشتر از باقی کلمات است)

e. قوانین rule-based معمولا دارای precision بالا و recall پایین هستند. بله

f. استفاده از micro averaging به علت میانگین گیری بدون وزن بین کلاس ها، معمولا نسبت به macro averaging معیار ارزیابی ضعیف تری است.خیر (معیار macroageragin میانگیری گیری بدون وزن دارد نه معیار microaveraging)

g. هر چه تعداد ابعاد بردار بازنمایی حاصل از word2vec بیشتر باشد، خطای مدلسازی بازنمایی کلمات کمتر است. غلط - لزوما اینطوری نیست.

h. هر چه مقدار PPMI دو عدد کمتر باشد دو کلمه از یکدیگر وابستگی بیشتری دارند. خیر

i. بیشینه کردن احتمال وقوع یک جمله معادل بیشینه کردن perplexity می باشد. نادرست – معادل کمینه کردن perplexity می باشد.

j. اگر یک مدل unigram احتمال یکسانی به تمام کلمات بدهد، perplexity برای تمام جملات با طول های متفاوت یکسان خواهد بود. درست - $perplexity = p(w1w2w3 \dots wn)^{-\frac{1}{n}} = (p'^n)^{-\frac{1}{n}} = \frac{1}{p'}$

k. در صورتی که در الگوریتم edit distance تمام اعمال هزینه یکسانی داشته باشند، بیشترین هزینه ممکن برای تبدیل یک رشته به رشته دیگر برابر با طول رشته بزرگتر است. درست

l. از روش smoothing برای افزایش سرعت آموزش مدل استفاده می شود. نادرست – این روش برای جلوگیری از صفر شدن احتمال یک جمله به هنگام مشاهده کلمه دیده نشده استفاده می شود.

m. دو الگوی عبارت منظم /fire|ings?/ و /fir(e|ings)?/ یکسان هستند. نادرست – عبارت منظم اول ings را می پذیرد در حالی که الگوی دوم این عبارت را نمی پذیرد.

n. عمل Classification سخت تر از Sequence classification است . بله

o. برای اموزش HMM بایستی $n^2 m$ عدد آموزش داده که در آن n تعداد حالات و m تعداد انواع observation است.

۲- کلمه یا عبارت مناسب بگذارید: (هر یک ۲ نمره – جمعا ۱۶ نمره)

a. برای حذف تاثیر کلمات بسیار پرکاربرد در ساختن بردار های sparse بر اساس فرکانس کلمات میتوان از ـــ استفاده کرد tf-idf

b. برای حل مشکل PMI بزرگ کلمات نادر میتواند احتمال کلمات نادر را کمی ـــ در نظر گرفت. بیشتر (یا افزایش) (یا بتوان عددی کمتر از یک رساند)

c. مل Naive Bayes به سبب — میتواند زمانیکه طول بردار های ویژگی کم است عملکرد خوبی داشته باشد. فرض استقلال ویژگی ها

d. روش .....word2vec..... یک روش بازنمایی کلمات مستقل از بافتار (context) است

e. روش .....bert..... یک روش بازنمایی کلمات وابسته به بافتار است.

f. شاخص PPMI می‌تواند مقادیری در بازه عددی .....صفر..... و .....بینهایت..... داشته باشد.

g. هزینه زمانی الگوریتم MED برای دو رشته به طول های m ,n برابر .....O(mn)..... می باشد.

h. به فرایند تبدیل کلمات به ریشه و شکل استاندارد آنها ....lemmatization.... گفته می شود

i. فرایند حذف پسوند های مختلف برای رسیدن به یک ریشه مشترک.....stemming.... نام دارد.

j. معیار ارزیابی ذاتی (intrinsic evaluation) مدل زبانی ..... perplexity .... است.

k. معیار ارزیابی برونی (extrinsic evaluation) مدل زبانی ..... Spell Checker .... است

l. به مشکل آموزش یک مدل در حالتی که دیتای اموزشی کم باشد، .....Sparness..... مینامند.

m. اگر به تمام داده های آموزش یک واحد اضافه کنیم، به این روش هموار سازی .....Laplace(addone)..... مینامند.

n. به دانش مطالعه نحوه ساخت یک کلمه از اجزای کوچکتر آن ...Morphology..... گفته میشود.

o. یک نمونه از Multi token word ها در زبان انگلیسی .....New york.... است.

p. یک نمونه از Multi word token در زبان انگلیسی .....hi-light..... است.

q. این ایده که کلماتی که در bigram های بیشتری شرکت کنند، احتمال bigram بیشتری با کلمات دیده نشده خواهند داشت، در روش هموار سازی .....Kneser ney..... کاربرد دارد.

r. فرق classificationبا regression در آن است که کلاسها در اولی گسسته ولی در دومی پیوست است.....

s. تابع خطایی که در مساله محاسبات وزنهای بهینه در Linear Regression کمینه میشود .....MeanSquareError..... نام دارد.

t. در الگوریتم برنامه نویسی پویا Viterbi ، از یک ماتریس با نام viterbi[s,t] استفاده میشود که درایه viterbi[s,t] یعنی حداکثر احتمال تا t حرف اول رشته ورودی را ببینیم و در وضعیت s قرار بگیریم چقدر است؟

u. برای محاسبه sentiment یک متن، نمی توان از word2vec استفاده کرد زیرا ....word2vec برمبنای توزیع کلمات در یک متن تهیه شده است ونه برحسب Polarity و قطبیت کلمات ...مثلا کلمات good و bad دارای w2v نزدیک هم هستند درحالیکه قطبیت دوری از هم دارند

۳- به سوالات زیر بطور مختصر پاسخ دهید:

a. در چه صورت مدل Naive Bayes همان Language Model خواهد بود؟ در صورتی که ویژگی ها از جنس کلمات باشند و تمام کلمات را در اختیار داشته باشیم.

b. چه زمان استفاده از logistic regression بهتر از Naive Bayes است؟

زمانیکه تعداد داده ها زیاد باشد و فرض استقلال بین ویژگی ها قوی نباشد یا اهمیت تعداد از ویژگی ها بیشتر از باقی ویژگی ها باشد عملکرد مدل logistic regression احتمالا بهتر خواهد بود.

c. تفاوت مدل های skip-gram و CBOW را توضیح دهید.

مدل skip gram تلاش میکند که با در نظر گرفتن یک کلمه target، به کلماتی که در همسایگی آن کلمه ظاهر شده اند (context) احتمال بیشتری بدهد در حالیکه مدل CBOW با در نظر گرفتن کلمات همسایه (context) یک کلمه target سعی میکند که به کلمه target در مقایسه با دیگر کلمات احتمال بیشتری بدهد.

d. چرا در طبقه بندی مدل های generative غالبا اطلاعات بیشتر و قابل تفسیرتری از مدل های discriminative در اختیار ما قرار میدهد؟ از هر مدل یک مثال بیاورید.

زیرا یک مدل generative مانند Naive Bayes تلاش میکند که از یک کلاس یه نمونه تولید کند پس بررسی میکند که بر اساس ویژگی ها، یک نمونه به کدام کلاس شبیه تر است در نتیجه اطلاعات بیشتری دارد. ولی مدل های discriminative مانند logistic regression درکی از توصیف یک کلاس بر اساس ویژگی های آن ندارند و تنها سعی میکند تفاوت کلاس ها را ملاک تشخیص قرار دهند ها را ببیند.

e. آیا هر چه هم رخدادی بیشتر باشد لزوما بین دو کلمه شباهت بیشتری وجود دارد؟

خیر، هم رخدادی کلمات پر تکرار عموما معنی خاصی را به کلمه نخواهند داد.

f. توضیح دهید چطور می‌توان کیفیت یک مجموعه داده که توسط یک برچسب زن ، با برچسهای Part of speech برچسب خورده را ارزیابی کرد.

می‌توان بخشی از داده‌هایی که نفر اول برچسب زده است را به نفر دیگری داد و سپس با استفاده از آنها و معیار kappa کیفیت را ارزیابی کرد.

g. در مسئله pos tagging یک روش به عنوان سقف و یک روش به عنوان کف معرفی کنید.

روش‌های most frequent tags یک کف و معیار kappa در برچسب زنی چند برچسب زن یک سقف برای مسئله باشد.

h. Feature template هایی که در CRF میتوان تعریف کرد، بایستی چه محدودیتی را رعایت کنند؟ چرا؟ تنها محدودیتی که دارند آن است که برای برچسب زدن به کلمه i ام که به ان برچسب yi بگوییم، صرفا به y_{i-1} میتواند ارجاع داشته باشد و به برچسبهای کلمات بعدی و خیلی قبلتر نمیتواند ارجاع داشته باشد. دلیلش آن است که بتواند الگوریتم خطی Viterbi را بصورت DP از چپ به راست بتواند اعمال کند.

۴- یکی از مشکلات بردارهای تعبیه کلمات، مدیریت کلمات خارج از دیکشنری است. برای حل آن دو راهکار پیشنهاد می‌دهید.

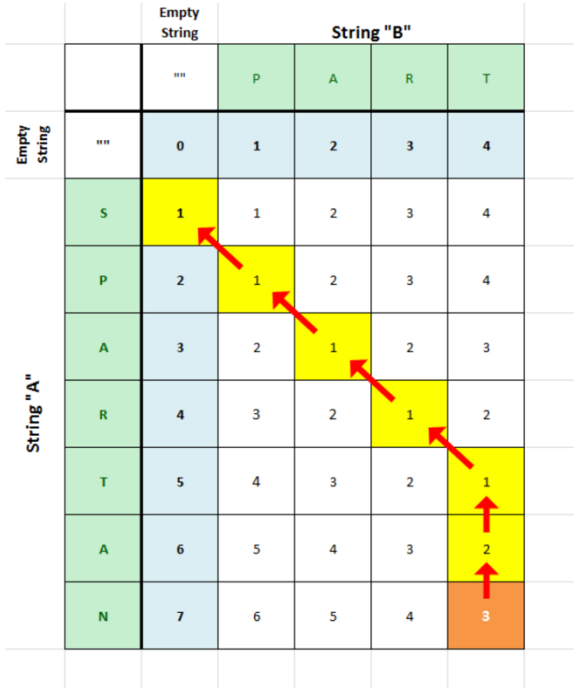
راهکار اول: برای حل آن می‌توان از یک کلمه به عنوان oov استفاده کرد و به طور مشترک از آن استفاده کرد. راهکار دوم : روش byte pair encoding هم خود می‌تواند گزینه مناسبی برای حل این چالش باشد.

۵- برای محاسبه مقادیر جداول transposition, substitution, deletion, insertion در الگوریتم MED چه روشی به ذهنتان میرسد؟ در دو حالت که دادگان آموزشی داشته باشیم یا نداشته باشیم، پیشنهاد دهید؟ دادگان آموزشی مورد نیاز چه فرمتی بایستی داشته باشند؟

اگر دادگان آموزشی باشد که موضوع خیلی ساده است. کافی است که از MLE مقادیر جداول را ترین کنیم. اگر جایی هم صفر شد میتوانیم از smoothing استفاده کنیم. دادگان آموزشی بایستی بصورت متون با خطاهای لغوی و متن اصلاح شده آن باشد. با فرض

انکه دادگان اموزشی نباشد، بهترین راه ارائه یک طبقه بند احتمالی چهارکلاسه یا چهار طبقه بند باینری است که از روی فیچرهای مختلف بتواند $P(insertion|x,y)$ را محاسبه کند. فیچرها میتواند نزدیکی حروف در کی برد یا در proximity Phonotics باشد.

۶- الگوریتم MED را برای تبدیل رشته spartan به part اجرا کنید(از سه عمل insert,delete,substitution با هزینه یک می توانید استفاده کنید) و مراحل لازم برای تبدیل این رشته را به دست آورید. کمترین هزینه برای تبدیل این رشته چه مقداری است؟



Step	Comparison	Edit necessary
1.	"S" to ""	delete: +1 edit
2.	"P" to "P"	no edits necessary!
3.	"A" to "A"	no edits necessary!
4.	"R" to "R"	no edits necessary!
5.	"T" to "T"	no edits necessary!
6.	"A" to "T"	delete: +1 edit

7. "N" to "T" delete: +1 edit

۷- فرض کنید سند زیر داده شده است. در صورت استفاده از طبقه بند Naïve Bayes به همراه هموار سازی add-one به نظر شما سند زیر چه برچسبی خواهد گرفت؟

“I loved the poor play” -

“I hated the play movie”

Document	Text	Class
1	I loved the movie	+
2	I hated the movie	-
3	a great movie. good movie	+
4	poor acting	-
5	great acting. a good movie	+

پاسخ:

مجموعه لغات:

```
< I, loved, the, movie, hated, a, great, poor, acting, good >
```

$$P(w_k \mid +) = \frac{n_k + 1}{n+ \mid \text{vocabulary}\mid}$$

$$P(+) = \frac{3}{5} = 0.6$$

$$P(I \mid +) = \frac{1+1}{14+10} = 0.0833$$

$$P(loved \mid +) = \frac{1+1}{14+10} = 0.0833$$

$$P(the \mid +) = \frac{1+1}{14+10} = 0.0833$$

$$P(movie \mid +) = \frac{4+1}{14+10} = 0.20833$$

$$P(a \mid +) = \frac{2+1}{14+10} = 0.125$$

$$P(great \mid +) = \frac{2+1}{14+10} = 0.125$$

$$P(acting \mid +) = \frac{1+1}{14+10} = 0.0833$$

$$P(good \mid +) = \frac{2+1}{14+10} = 0.125$$

$$P(poor \mid +) = \frac{0+1}{14+10} = 0.0417$$

$$P(hated \mid +) = \frac{0+1}{14+10} = 0.0417$$

P(UNKNOWNWORD| +) = 1/24 = 0.0417

$$P(-) = \frac{2}{5} = 0.4$$

$$P(I \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(loved \mid -) = \frac{0+1}{6+10} = 0.0625$$

$$P(the \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(movie \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(a \mid -) = \frac{0+1}{6+10} = 0.0625$$

$$P(great \mid -) = \frac{0+1}{6+10} = 0.0625$$

$$P(acting \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(good \mid -) = \frac{0+1}{6+10} = 0.0625$$

$$P(poor \mid -) = \frac{1+1}{6+10} = 0.125$$

$$P(hated \mid -) = \frac{1+1}{6+10} = 0.125$$

P(unkNOWNWORD|-) = 1/16 = 0.0625

$$V_{NB} = \arg\max_{v_j \in V} P(v_j) \sum_{w \in word} P(w \mid v_j)$$

“I loved the poor play” -

برای کلاس + احتمال انکه این جمله + باشد برابر است با

P(+) . p(I | +). P(LOVED|+) .p(the|+) . P(poor | +) . P(play|+) =

(3/5) * 0.0833 * 0.0833 * 0.0833 * 0.0417 * 0.0417 = $6.03 * 10^{-7}$

برای کلاس - احتمال انکه این جمله -باشد برابر است با

P(-) . p(I | -). P(LOVED|-) .p(the|-) . P(poor | -) . P(play|-) =

(2/5) * 0.125 * 0.0625 * 0.125 * 0.125 * 0.0625 = $3 * 10^{-6}$

یعنی جمله اول منفی است

“I hated the play movie”

مشابه فوق است

احتمال مثبت بودن

0.6 * 0.0833 * 0.0417 * 0.0833 * 0.0417 * 0.20833 = 1.508 * 10-6

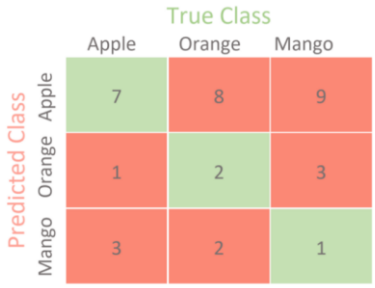
احتمال منفی بودن

0.4 * 0.125 * 0.125* 0.125* 0.0625 * 0.125 = 6 * 10-6

جمله دوم هم منفی است

از آن‌جایی که احتمال این جمله به شرط کلاس منفی بیشتر است پس به کلاس منفی تعلق دارد.

۸- به ازای confusion matrix زیر ابتدا معیار های precison,recall,F1 به ازای هر کلاس را محاسبه کنید و سپس معیار های Macro F1 و Micro F1 را محاسبه نمایید.



پاسخ:

$$TP = 7$$

$$TN = (2+3+2+1) = 8$$

$$FP = (8+9) = 17$$

$$FN = (1+3) = 4$$

$$Precision = 7/(7+17) = 0.29$$

$$Recall = 7/(7+4) = 0.64$$

$$F1\text{-}score = 0.40$$

$$Total\ TP = (7+2+1) = 10$$

$$Total\ FP = (8+9)+(1+3)+(3+2) = 26$$

$$Total\ FN = (1+3)+(8+2)+(9+3) = 26$$

$$Precision = 10/(10+26) = 0.28$$

$$Micro\ F1 = 0.28 \qquad Recall = 10/(10+26) = 0.28$$

برای مایکرو

Precision = Recall = Micro F1 = Accuracy

برای میکرو:

- *Class Apple F1-score = 0.40*
- *Class Orange F1-score = 0.22*
- *Class Mango F1-score = 0.11*

Hence,

$$Macro\ F1 = (0.40+0.22+0.11)/3 = 0.24$$

۹- مقدار perplexity برای مدل bigram آموزش یافته یک corpusبرابر با ۲۱۳ است. این عدد چه مفهومی دارد؟ مقدار perplexity برای مدل trigram آموزش یافته یک corpusبرابر با ۲۱۳ است. این عدد چه مفهومی دارد؟

پاسخ: در bigram اگر سرگشتگی (perplexity) برابر با ۲۱۳ باشد یعنی اگر یک کلمه خاص در زبان را در نظر بگیریم، بطور متوسط ۲۱۳ کلمه بعداز آن ظاهر میشود و میزان تنوع کلمات بعدی ان بطور میانگین ۲۱۳ است. برای trigram یعنی آنکه اگر دو کلمه متوالی را در نظر بگیرید بطور متوسط ۲۱۳ کلمه بعد از این دو کلمه ظاهر میشود

۱۰- اگر اندازه پنجره در محاسبه word2vec افزایش یابد، آنگاه میزان شباهت جفت کلمات زیر افزایش می یابد یا کاهش می یابد؟

a. نخ – سوزن
b. نخ - خیاط

اگر پنجره کوچک باشد، کلماتی مشابه هستند که syntactic similarity بالا داشته باشند مانند بند a ولی اگر پنجره بزرگ شود، topical similarity دارند مانند "نخ - خیاط"

۱۱- در دادگان آموزشی تعداد حضور کلمه edinbrugh از کلمه jahensen بیشتر است. در یک سامانه پیش بینی به عبارت "provision quality X freedom act" رسیده ایم که در آن X یکی از دو کلمه فوق الذکر است. بنظر شما کدام کلمه بهتر است انتخاب گردد؟ لازم بذکر است که هر دو کلمه فوق الذکر با هیچ یک از کلمات عبارت در دادگان اموزشی ظاهر نشده اند؟ از چه مدل احتمالی برای این انتخاب استفاده میکنید؟ نحوه محاسبه مدل احتمالی چیست؟

- از *number of different contexts word w has appeared in.*
- $$P_{\text{CONTINUATION}}(w) \propto |\{v : C(vw) > 0\}|$$

۱۲- یک مدل زبانی مبتنی بر شبکه عصبی feed forward را آموزش داده ایم برای محاسبه P(w1|w2) کندترین قسمت این شبکه چیست؟ بنظر شما محاسبه LM از روی شبکه عصبی کندتر است یا ngram ؟

مخرج کسر برای تابع softmax کندترین بخش شبکه است. شبکه عصبی کندتر است.

۱۳- در مورد LSTM و GRU به سوالات زیر پاسخ دهید:

a. LSTM چه بخشهایی دارد؟ سه گیت forget/input/output

b. GRU چه بخشهایی دارد و چه تفاوتهایی با lstm دارد؟ دو گیت reset/update – input , forget با هم ادغام شدند و update را تشکیل دادند. Cell و hidden نیز با هم ادغام شدند.

c. آیا اموزش LSTM نیاز به دادگان اموزشی بیشتری دارد یا GRU؟ LSTM شبکه بزرگتری با پارامترهای بیشتری است که نیاز به دادگان آموزشی بیشتری دارد.

d. بنظر شما کجاها بهتراست که LSTM استفاده کنیم کجاها GRU؟ هر جایی که شبکه کوچکتری نیاز داشته باشد یا دیتاهای اموزشی کمتری وجود داشته باشد، gru مناسب تر است.

۱۴- چون NLM مدل bengio از سه کلمه اخر به عنوان ورودی شبکه استفاده میکند و کلمه بعدی را پیش بینی میکند، آیا میتوان گفت که آن شبکه معادل یک 4 gram است؟ برای پاسخ خود دلیل بیاورید؟

خیر، چون از embedding کلمات استفاده میشود که در انصورت خیلی از کلماتی که oov هستند ولی embedding مشابه دارند، میتواند پیش بینی کند.