به نام خدا

پاسخ بخش تئوری تمرین اول درس پردازش زبان طبیعی

استاد درس: دکتر یاسمن برشبان

علی افشار دگرسی ۱۴۰۴۱۲۲۶۱۵۱۰۱

سوال اول: معیار ارزیابی

برای یک دسته بندی دو کلاسه نتایج زیر به دست آمده است.

الف) FP,FN ,TP,TN را مشخص کنید.

پاسخ:

TN = 30, TP = 45, FN = 15, FP = 10

ب) معیارهای precision, recall, F1 و accuracy را محاسبه کنید.

پاسخ:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{75}{100}$$

$$Precision = \frac{TP}{TP + FP} = \frac{45}{55}$$

$$Recall = \frac{TP}{TP + FN} = \frac{45}{60}$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2}{\frac{55}{45} + \frac{60}{45}} = \frac{2}{\frac{115}{45}} = \frac{90}{115}$$

سوال دوم: نایوبیز

با توجه به جدول زیر مشخص کنید سند شماره 6 مربوط به چه کلاسی است. از روش نایو بیز با استفاده از هموارسازی add-1 استفاده کنید.

Documents	Words	Class
1	Technology, Stemming, Lemmatization, Learning	AI
2	Recognition, Learning, Data	AI
3	Neural, Network, Lemmatization	AI
4	Router, Switch, Network	Network
5	Network, Bandwidth, Router	Network
6	Recognition, Learning, Router	?

پاسخ:

$$\hat{P}(w|c) = \frac{count(w,c)+1}{count(c)+|V|} \qquad |V| = 11$$

$$P(AI) = \frac{3}{5} \qquad P(Network) = \frac{2}{5}$$

$$count(AI) = 10 \qquad count(Network) = 6$$

$$\hat{P}(Recognition|AI) = \frac{1+1}{10+11} = \frac{2}{21} \qquad \hat{P}(Learning|AI) = \frac{2+1}{10+11} = \frac{3}{21}$$

$$\hat{P}(Router|AI) = \frac{0+1}{10+11} = \frac{1}{21} \qquad P(AI|d6) \propto \frac{3}{5} * \frac{2}{21} * \frac{3}{21} * \frac{1}{21} \simeq 0.0004$$

$$\hat{P}(Recognition|Network) = \frac{0+1}{6+11} = \frac{1}{17} \qquad \hat{P}(Learning|AI) = \frac{0+1}{6+11} = \frac{1}{17}$$

$$\hat{P}(Router|AI) = \frac{2+1}{6+11} = \frac{3}{17} \qquad P(Network|d6) \propto \frac{2}{5} * (\frac{1}{21})^2 * \frac{3}{21} \simeq 0.0001$$

$$\boxed{P(AI|d6) > P(Network|d6) \Rightarrow so\ answer\ is\ AI}$$

سوال سوم: پیش پردازش + نایو بیز

ج) مقایسه داده خام و داده پیش پردازش شده

با توجه به نتایج به دست آمده از مرحله قبل به سوالات زیر پاسخ دهید.

1. کدام مراحل پیش پردازش بیشترین تاثیر را بر بهبود مدل داشتهاند؟

پاسخ: مرحله فقط نرمال سازی بدون کارهای دیگر، بیشترین بهبود را داشته است.

2. آیا حذف stopwords باعث کاهش یا افزایش دقت شده؟ چرا؟

پاسخ: کاهش، به این علت که بعضی از توکن هایی که به عنوان stop word حذف شده اند در تفاوت معنای جملات تاثیر داشته و باعث تمایز بهتر می شدند.

3. پیش پردازش چگونه روی کلاسهای مختلف (مثبت/منفی/خنثی) تأثیر گذاشته؟

پاسخ: بیشترین تاثیر را روی کلاس های مثبت و منفی داشته و کمترین تاثیر را روی کلاس خنثی داشته است. نرمال سازی باعث بهبود و lemmatization و حذف stop word نیز باعث افت عملکرد شده اند.

سوال چهارم: پیاده سازی سیستم بازیابی اطلاعات (IR) فارسی با TF-IDF بر رویPQuAD

د) مقایسه عملکرد با و بدون پیش پردازش

تمام مراحل بخش ب و ج را یک بار دیگر انجام دهید، اما این بار بدون هیچ پیش پردازشی. نتایج دو حالت را مقایسه کنید:

- دقت و MRR چقدر تغییر کرده اند؟
- کدام مقادیر بهبود یافته یا بدتر شده اند؟

پاسخ: precision@5 و recall@5 بدتر شده اند ولی MRR بهبود یافته است.

- چرا پیش پردازش در برخی پرسش ها اثر بیشتری گذاشته است؟

پاسخ: با حذف stopword ها یا علامت های نگارشی، سایز vocab کمتر شده و فضای مسئله ما دارای ابعاد کمتری نسبت به حالت عادی می باشد، پس هم معیار فاصله ما بهتر متمایز می کند هم توجه به توکن هایی که خاص تر می باشند بیشتر می شود. از طرفی ممکن است توکن هایی که در واقع اهمیت بیشتری دارند به اشتباه حذف شوند.

ه) تحلیل خطا و ارائه پیشنهادهای بهبود

برای 3 تا ۵ سوال که سیستم در آنها عملکرد ضعیفی داشته (مثلًا سند درست در رتبه های پایین قرار گرفته)

1. متن سؤال را نشان دهید.
2. سند درست را نمایش دهید.
3. سه سند اول رتبه بندی شده را مقایسه کنید.

4. توضیح دهید چرا IDF-TF اشتباه کرده است .

5. حداقل سه راهکار برای بهبود سیستم پیشنهاد کنید .

پاسخ:

سوال اول: "رندر چیست؟"

سند درست:

"بالا بردن میزان نرخ زمانی کلاک و که سبب تولید هش‌ریت (میزان محاسبه اطلاعات توسط پردازنده در واحد زمان است)، بیشتر در واحد زمان و انجام سریع تری محاسبه در بازه زمانی می‌شود. پردازنده‌هایی که قفلشان بازگشایی شده باشد را می‌توان اورکلاک کرد و مزیت اورکلاک، انجام سریع تر پردازش و رندرهای سنگین توسط پردازنده است. مضرات اورکلاک بالا رفتن دمای پردازنده برای محاسبه و در نتیجه استفاده مداوم سبب پایین آمدن عمر پردازش گر می‌شود. عمل Overclocking نیازمند دانش کافی در زمینه سخت افزار می‌باشد و هرگونه اقدام نادرست، آسیب‌های جبران ناپذیری به پردازنده ها وارد می کند. تنها پردازنده‌هایی اورکلاک می‌شوند که قابلیت اورکلاک شدن را در پسوند خود داشته باشند...."

سند اول بازیابی شده:

"در سه سالگی، او شروع به شرکت در آموزش‌های رقص در زادگاهش، کنتوود، لوئیزیانا کرد و برای اجرا به عنوان یک هنرمند انفرادی در مراسم سالانه موسیقی انتخاب شد. اسپیرز در سن پنج سالگی، اولین حضور خود را در صحنه محلی با اجرای آواز «این کودک چیست؟» در کودکستان فارغ‌التحصیلی خود آغاز کرد. او همچنین در دوران کودکی خود درس‌های ژیمناستیک و آواز را آموخته بود و در بسیاری از مسابقات در سطح ایالت و نمایش‌های استعدادیابی کودکان پیروز شد. او دربارهٔ اهداف خود در دوران کودکی گفت، «در دنیای خودم بودم، [...] من در سنین پایین متوجه شدم که قرار است چه کاری انجام دهم»"

سند دوم بازیابی شده:

"نخستین استفاده از واژه جنگ سرد برای اشاره به رقابت ژئوپلیتیکی پساجنگ بین ایالات متحده آمریکا و اتحاد جماهیر شوروی در سخنرانی برنارد باروک از مشاوران مهم رئیس جمهوران دموکراتیک آمریکا و در ۱۶ آوریل ۱۹۴۶ آمده است. در این سخنرانی که توسط روزنامه‌نگار هربرت بایارد اسووپ[واژه‌نامه 3] نوشته شده است چنین آمد: «نباید فریب بخوریم، ما امروز در میانه یک جنگ سرد به سر می‌بریم». والتر لیپمن روزنامه‌نگار با کتابش جنگ سرد به این اصطلاح رواج بیشتری داد. در ۱۹۴۷ وقتی از لیپمن پرسیده شد خاستگاه این اصطلاح چیست وی گفت که برگرفته از یک اصطلاح فرانسوی دهه ۱۹۳۰ است: la guerre froide"

سند سوم بازیابی شده:

"به تعریفی دیگر تاریخ، دانش واکاوی پیشامدهای گذشته‌است، علمی که بشر بر مبنای آن این امکان را می‌یابد که بگوید حوادث تاریخی در یک سلسله زمانی رخ داده‌اند. بحث تعریف تاریخ و علم تاریخ در نزد گذشتگان و معاصران ابداعی نیست و در نگاه اجمالی به مفهوم زمان تأکید ویژه‌ای دارد. هر تعریفی که ارائه می‌شود بر منظری است که از آن زاویه مطرح می‌شود. مفهوم تاریخ نیز از مفاهیمی مانند گذشته، طبیعت و هستی متفاوت است. تاریخ مربوط به گذشته‌است ولی فایده تاریخ برای گذشته نیست بلکه برای زمان حال و آینده تأثیر بسیاری دارد. تاریخ بازسازی گذشته بر اساس مدارک و اسناد است و اینکه مدرک و سند چیست و چه چیزی را سند و مدرک تلقی کنیم بحث مهمی در علم تاریخ به‌شمار می‌آید. نظریه‌ای که یک مورخ دارد بایستی به تأیید مورخان برسد و مورخان دیگر نیز با آن نظریه موافقت کنند، اگر مورخان بر نظریه‌ای که مبتنی بر واقعیت‌هایی که از اسناد و مدارک ذکر می‌شود، توافق کنند آن نظریه باعث می‌شود حقایق به واقعیت‌های تاریخی بدل شوند."

علت اشتباه: به این علت که ما lemmatization یا stemming انجام نداده ایم، توکن "رندر" با اینکه در دیتاست نادر است و idf بالایی دارد ولی با توکن "رندرهای" متفاوت در نظر گرفته شده و معیار مقایسه بیشتر روی توکن "چیست" تمرکز کرده و سند هایی که دارای این توکن هستند را در ابتدا می آورد.

راهکارها:

1. با استفاده از lemmatization یا stemming میتوانیم توکن "رندرهای" و "رندر" را به یک ریشه مشترک تبدیل کنیم تا یک توکن در نظر گرفته شوند.
2. وقتی توکن "رندر" را می بینیم میتوانیم با گسترش متن سوال و اضافه کردن مترادف های آن مثل "پردازش" یا ... نتیجه را بهبود دهیم.
3. حتی می توانیم در هنگامی که vocab خود را میسازیم و توکن ها را استخراج میکنیم، علاوه بر ریشه یابی، تحلیل معنایی نیز بکنیم و از لحاظ معنایی به یک توکن واحد مرتبطشان کنیم. که هم سایز vocab را ممکن است کمتر کند هم ارتباط را بهتر در نظر بگیرد

سوال دوم: "الگوریتم به چه معناست؟"

سند درست:

"کارایی پردازش کامپیوتر ها، با استفاده از پردازنده های چند هسته ای که اساساً اتصال دو یا بیش از دو پردازنده ی مجزا (با نام هسته ها، در این زمینه) در یک مدار مجتمع است، افزایش می یابد. به‌طور ایده آل، یک پردازنده ی دو هسته ای تقریباً قدرت معادل دو برابر یک پردازنده تک هسته‌ای دارد، اما در واقعیت، افزایش کارایی بسیار کمتر از این است و حدود تقریبا ۵۰% است که دلیل آن، الگوریتم های نرم افزاری و پیاده سازی های ناکامل است. افزایش تعداد هسته ها در یک پردازنده (یعنی دو هسته ای، چهار هسته ای و ...)

موجب افزایش میزان بار کاری قابل انجام توسط پردازنده می شود. این بدان معنی است که این پردازنده ها می توانند وقایع ناهمگام، وقفه ها و ... را که در حجم زیاد می توانند اثرات مخربی روی عملکرد پردازنده داشته باشند، مدیریت کنند. این هسته ها را می توان به شکل طبقات مختلف در یک ساختمان پردازش در نظر گرفت که هر طبقه یک وظیفه ی متفاوت را انجام میدهد. گاهی این هسته ها، در زمانی که یک هسته به تنهایی برای مدیریت اطلاعات کافی نباشد، می توانند کاری مشابه با هسته های مجاور خود انجام دهند."

سند اول بازیابی شده:

"مناطق جغرافیایی خاص می‌توانند مورد مطالعهٔ تاریخی قرار گیرند، برای نمونه، قاره‌ها، کشورها و شهرها. درک این که چرا مکان وقایع تاریخی صورت گرفته مهم است، برای انجام این کار، مورخان اغلب به جغرافیا روی می‌آورند. الگوهای آب و هوایی، منابع آب موجود و چشم‌انداز از مکان زندگی، همه، بر روی مردمی که در آنجا زندگی می کنند، اثر می گذارد. برای نمونه، برای توضیح اینکه چرا مصریان باستان تمدنی موفق را گسترش دادند، مطالعهٔ جغرافیای مصر ضروری است. تمدن مصری در ساحل رود نیل ساخته شده بود که طغیان‌های سالانهٔ نیل، خاک حاصل‌خیزی را به ساحل می‌آورد. خاک غنی می‌تواند به کشاورزان کمک کند تا محصولات کافی را رشد دهند و به تأمین خوراک مردم شهر بپردازند. این بدان معناست که هرکس مزرعهٔ کشاورزی نداشت، بنابراین دیگر مردم می‌توانستند کارهای دیگری انجام دهند که به رشد و توسعهٔ تمدن کمک می کرد."

سند دوم بازیابی شده:

"باستان‌شناسی رشته‌ای علمی است که به‌طور ویژه به بررسی و کاوش محوطه‌های باستانی و اشیاء قدیمی که کشف شده‌اند می‌پردازد و به مطالعهٔ تاریخ کمک می کند. اما باستان‌شناسی به‌ندرت می‌تواند به‌تنهایی مورد استفاده قرار گیرد، به این معنا که در این دانش، از منابع روایی و مکتوب تاریخی برای تکمیل آنچه کشف‌شده‌است استفاده می‌شود. البته، باستان‌شناسی

متشکل از بازهٔ گوناگونی از روش‌شناسی‌ها و رویکردهاست که هر کدام آن‌ها مستقل از تاریخند؛ این بدان معناست که باستان‌شناسی خلأها و شکاف‌های ناشی از منابع مکتوب را پُر نمی‌کند. در واقع، باستان‌شناسی تاریخی شاخه‌ای ویژه از باستان‌شناسی است که بیشتر، نتیجه‌گیری‌های آن با آنچه منابع متنی می‌گویند، متضاد است."

سند سوم بازیابی شده:

"سالِ ۱۳۴۹ کانون پرورش فکری کودکان و نوجوانان سبب شد تا بیضایی دو کارِ نکرده را نخستین بار بیازماید: داستان نوشتن برای کودکان و فیلم‌سازی. حقیقت و مرد دانا را به خواهشِ فیروز شیروانلو نوشت؛ و این نخستین و واپسین آزمونش در این کار بود؛ ولی فیلم‌سازی همان بویه‌ای بود که از نوجوانی در پی‌اش بود و یک بار هم به‌طوری محدود با یک دوربین ساده دست داده بود، ولی این بار می‌توانست فیلمی برای نمایش روی پردهٔ بزرگ بسازد. حاصلِ کار عمو سیبیلو شد، که راهی بود برای آن که بیضایی سرانجام بتواند کارِ سینمایی کند. فیلم‌سازی جدّی‌ترش از دههٔ ۱۳۵۰ مقدور شد، ولی همچنان نشد که فیلمنامهٔ بزرگی چون عیّار تنها را فیلم کند."

علت اشتباه: مثل سوال قبلی ولی اینبار برای توکن های "الگوریتم" و "الگوریتم های"

راهکارها: مثل سوال قبل

سوال سوم: "ناصرالدین‌شاه چه اقدامات مهمی را در زمان پادشاهی خود انجام داد؟"

سند درست:

"در آستانهٔ مراسم پنجاهمین سال تاجگذاری در سال ۱۲۷۵ هجری خورشیدی (۱۷ ذی‌القعده۱۳۱۳ هجری قمری) به دست میرزا رضای کرمانی یکی از پیروان سید جمال الدین اسدآبادی و به تحریک او در حرم شاه عبدالعظیم در شهر ری ترور شد. او در هنگام ترور

پنجاهمین سالگرد سلطنت خویش را جشن می گرفت. وی در زیارتگاه شاه عبدالعظیم در شهر ری در نزدیکی تهران دفن است. سنگ قبر یک پارچهٔ مرمری وی که تمثال کامل وی بر آن حکاکی شده هم‌اکنون در موزهٔ کاخ گلستان در تهران نگهداری می‌شود و به یکی از شاهکارهای کنده‌کاری دوره قاجار معروف است."

سند اول بازیابی شده:

"ناصرالدین‌شاه (۲۵ تیر ۱۲۱۰ - ۱۲ اردیبهشت ۱۲۷۵) که پیش از دوران پادشاهی ناصرالدین میرزا خوانده می‌شد، معروف به «قبلهٔ عالم»، «سلطان صاحبقران» و بعد از کشته شدن توسط میرزا رضای کرمانی «شاه شهید»، چهارمین شاه از دودمان قاجار ایران بود. او با نزدیک به ۵۰ سال پادشاهی، پس از شاپور دوم ساسانی و تهماسب اول صفوی طولانی‌ترین دورهٔ پادشاهی در میان تمامی شاه‌های تاریخ ایران را داراست. او به افتخار نیم قرن سلطنت بر ایران، خود را صاحبقران نامید. او همچنین نخستین پادشاه ایران بود که خاطرات خود را نوشت. همچنین، ناصرالدین‌شاه اولین پادشاه ایرانی بود که در رأس هیئت حاکمه برای بازدید از تمدن و تکنولوژی غرب عازم اروپای نوین شد. او اولین صدراعظم خود، امیرکبیر را کشت و سید محمدعلی، مؤسس آئین بابی را نیز توسط امیرکبیر اعدام کرد. ناصرالدین‌شاه در سال ۱۲۷۵ خورشیدی و در آستانه مراسم پنجاهمین سال پادشاهی‌اش به ضرب گلوله در سن ۶۵ سالگی ترور شد. او را در حرم عبدالعظیم حسنی در شهر ری در جنوب شهر تهران به خاک سپردند."

سند دوم بازیابی شده:

"تأسیس مدرسه‌های خارجی در ایران از زمان محمدشاه و با فعالیت گروه‌های میسیونر آغاز شده بود. در اواخر دوره ناصرالدین‌شاه مدرسه‌های خارجی گسترش بیشتری در ایران یافتند. ناصرالدین شاه که در نخستین سفر اروپا با آدولف کرمیو رئیس اتحاد جهانی آلیانس ملاقات کرده بود، به او قول داد که اجازه فعالیت این مؤسسه در ایران و تأسیس مدارس را بدهد. در سال ۱۳۰۶ قمری دو شعبه از مدرسه آلیانس در تهران و شیراز تأسیس شد. ناصرالدین‌شاه

پس از چند سال به سبب شایعاتی که پیرامون فعالیت‌های ضدسلطنتی و ضدمذهبی این مدرسه به وجود آمده بود، مجوز فعالیت آلیانس را لغو کرد. برای جلب اعتماد شاه، مدیر مدرسه، ریاست عالی مدرسه را به شاه اهداء کرد و شاه بازگشایی دوباره آلیانس را پذیرفت. شمار مدارس میسیونری آمریکایی در دوران ناصرالدین شاه افزایش یافت و در تهران و سایر شهرهای ایران مدارس بیشتری تأسیس شد به‌طوری که در سال ۱۲۷۴ خورشیدی، تنها ۱۱۷ مدرسه در شمال غربی ایران (آذربایجان) فعالیت می کردند."

سند سوم بازیابی شده:

"در شهریور ۱۳۲۰، ایران به دست نیروهای دو کشور انگلستان و شوروی (و بعدها ایالات متحده) به بهانه همراهی رضاشاه با قوای آلمان مورد هجوم گسترده قرار گرفت که از تاب تحمل ارتش نوپای ایران خارج بود و رضاشاه پهلوی پیش از رسیدن قوای متفقین به تهران از پادشاهی استعفا داد و فرزندش را به عنوان جایگزین به مجلس شورای ملی معرفی کرد که مورد تصویب قرار گرفت. این اقدامات به پیشنهاد محمد علی فروغی صورت گرفت تا راه سوء استفاده متفقین را ببندد. بدین ترتیب سالهای آغازین پادشاهی وی با اشغال ایران و پایان جنگ دوم جهانی مصادف شد.[نیازمند منبع]"

علت اشتباه: اینطور به نظر میرسد که سند درست ارتباط زیادی به سوال ندارد، و اشتباه از ارتباط گذاری اولیه است و مدل به درستی کار کرده است.

راهکارها: -

ممنون از توجهتون