بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال دوم ۰۳-۰۲

مدرس: احسان‌الدین عسگری

تمرین دوم — **استخراج اطلاعات قاعده‌محور با عبارات منظم** — مهلت ارسال: ۲۱ اردیبهشت

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.
- در تمرین‌هایی که چند چالش دارند، فقط یک نفر از هر گروه در گوگل فرم باید چالش مورد نظر گروه را انتخاب کند. امکان تغییر چالش تا قبل از زمان ددلاین انتخاب چالش وجود دارد. البته ذکر این نکته ضروری است که هر چالش محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب چالش اقدام کنید.
- در طول ترم امکان ارسال با تاخیر برای هر تمرین ۵ روز و مجموع زمان مجاز تاخیر ۱۲ روز است. محل بارگزاری جواب تمرین‌ها مطابق زمان مشخص شده در تقویم، بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد. لازم به ذکر است که به دلیل تداخل زمان مجاز تاخیرها بین اعضای گروه در تمارین گروهی تمرین اول شامل تاخیر مجاز **نمی‌باشد**.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و مواردی از این دست) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- دقت داشته باشید، موارد امتیازی که در این تمرین آمده است، صرفا بر روی امتیاز همین تمرین اثر دارد و بر روی نمرات تمارین و یا بخش‌های دیگر درس، تاثیر ندارد.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به تیم تدریس خودداری کنید.

توضیحات کلی

در این تمرین، کدهای شما در قالب یک چت بات روی یک پیام‌رسان بر پایه‌ی پیام‌رسان Matrix پیاده‌سازی می‌شوند. نحوه‌ی تعامل با کد از طریق ارسال و دریافت پیام می‌باشد.

در این تمرین شما به حل مسائلی تازه در پردازش زبان فارسی خواهید پرداخت. مسائلی کاربردی، که عموما ابزاری برای آنها تولید نشده است. در این تمرین در بسیاری از بخش‌ها می‌توانید از حاصل کار عزیزان ترم‌های گذشته که با زحمات تدریسیاران درس در قالب کتابخانه parsi.io ایجاد شده بهره ببرید. به امید خدا در ترم‌های آینده حاصل تجمیع زحمات شما عزیزان در قالب محصولات متن‌باز (البته با ذکر نام خودتان) در اختیار دیگر دانشجویان و بلکه جامعه ایرانی قرار می‌گیرد تا در اثر این تلاش‌ها محصولاتی ارزشمند برای پردازش متن‌های فارسی و بلکه زبان‌های ایرانی و فراتر از آن داشته باشیم. می‌توانید به این کتابخانه از طریق این لینک دسترسی داشته باشید.

لطفا علاوه بر قوانین درس که در cw قرار گرفته اند، به توضیحات زیر در مورد تمرین ۲ توجه داشته باشید:

۱. در این تمرین شما قرار است که با روش‌های تشخیص به وسیله قواعد با تمرکز بر عبارات منظم و آنچه در ماژول ابتدایی درس آموخته اید، مساله‌های پردازش متن مختلفی را حل کنید. ملاک ارزیابی شما ،به ترتیب این موارد است: صحت، زمان اجرا، نتایج قابل بازتولید، مستندات.

۲. در زمینه صحت هم به شکل نسبی مقایسه انجام می‌شود. یعنی ممکن است در یک ترک خاص صحت ۴۰ درصد صحت بالایی محسوب شود.

۳. در زمان اجرا این موضوع مهم هست که زمان اجرای برنامه نسبت به ترک داده شده طولانی نباشد. اگر برنامه شما به شکل غیر بهینه پیاده‌سازی شده باشد بر روی نمره شما اثر منفی دارد.

۴. برنامه‌تان باید به گونه‌ای پیاده سازی شده باشد که دارای یک تابع

```
run(input: str)
```

باشد که این تابع با گرفتن ورودی خروجی مورد نظر را تولید می‌کند.

۵. فرمت خروجی باید رعایت شود. می‌توانید برای بازه هاspan از تایپ توپل پایتون نیز استفاده کنید. یعنی هر دوی حالات زیر مجاز هستند.

```
>>> span = (3,8)
>>> span = [3,8]
```

بازه شما باید به گونه‌ای باشد که اگر در پایتون به عنوان بازه‌ی substr استفاده شد، دقیقا متن مورد نظر بدون فاصله‌های ابتدا و انتها باشد. در مثال زیر بازه درست کلمه apple به شکل زیر است:

```
>>> input = "my apple is red"
>>> span = (3,8)
>>> input[span[0]: span[1]]
'apple'
```

۶. لازم است برای مسئله خود حداقل پنج نمونه آزمون بنویسید و کد شما روی این نمونه‌ها باید درست عمل کند.

۷. برای این تمرین شما مجاز هستید (حتی توصیه می‌شود) از مدل‌های زبانی بزرگ برای ساخت عبارات منظم مناسب براساس نمونه‌ها استفاده کنید.

توضیحات پیاده‌سازی بات‌ها

برای پیاده‌سازی بات‌ها از فریمورک OPSDroid استفاده خواهید کرد.آموزش نحوه‌ی ساخت اکانت و پیاده‌سازی بات در ویدیو آموزشی آورده شده، ولی پیشنهاد می‌شود حتما مستندات فریمورک را نیز مطالعه بفرمایید.

مشخصات سرور:

- آدرس پیام‌رسان: bot.quranic.network
- نام بات (نام نمایشی): NLP_HW۲_{GroupID}
- نام کاربری بات: bot_{functionality_of_the_bot}

هنگام پیاده‌سازی بات، می‌توانید آن را روی کامپیوتر شخصی خود اجرا کنید. نسخه‌ی پایانی بات‌ها روی سرورهای پیام‌رسان نصب می‌شوند و مورد استفاده قرار خواهند گرفت. اطلاعات تکمیلی مربوط به دیپلوی روی سرورهای اصلی به زودی در اختیارتان قرار خواهد گرفت ولی برای انجام تمرین از کامپیوتر شخصی خود استفاده نمایید.

نکته: می‌توانید برای شروع، از سرور matrix.org نیز به عنوان سرور استفاده کنید.

نکته: برای دسته‌بندی کلی پیام‌ها، می‌توانید از matcher regex فریمورک OPSDroid استفاده کنید، ولی برای بررسی دقیق‌تر هر دسته از پیام‌ها، به این matcher ها اکتفا نکنید.

نکته: برای بات خود، یک پیام خوش‌آمدگویی و یک دستور «راهنما» نیز پیاده‌سازی کنید تا نحوه‌ی کار و قابلیت‌های خود را به کاربر معرفی کند.

تجزیه و تحلیل اسناد حقوقی

در دنیای پرچالش حقوق، مراکز حقوقی به شدت به دستیارانی وابسته‌اند که می‌توانند فرآیندهای بررسی و تحلیل اسناد را با دقت بالا و در زمان کمتر انجام دهند. این امر، امکان مرور و بررسی اسناد حقوقی را آسان‌تر سازد و به بهبود کیفیت خدمات حقوقی کمک می‌کند. اسناد حقوقی، از جمله قراردادها، احکام دادگاه‌ها و قوانین، حاوی اطلاعات حیاتی هستند که شامل اصطلاحات تخصصی، ارجاعات به مقررات مختلف، و مشخصات مالی می‌شوند. این اطلاعات برای فهم عمیق‌تر تعهدات، حقوق و شرایط مندرج در هر سند بسیار مهم هستند. بنابراین، چالش اصلی این است که چگونه این موارد را به طور مؤثر استخراج کرده و آن‌ها را برای تحلیل‌های بیشتر، خلاصه‌سازی یا ورود به سیستم‌های فناوری حقوقی آماده کنیم .

شرح وظیفه:

با توجه به مجموعه ای از اسناد قانونی (میتوانید از لینک های ۱و۲و۳ , نمونه ای از اسناد مورد را نظر ملاحظه کنید)، باید عبارات regex بنویسید تا موجودیت های لازم را شناسایی و استخراج کند.

نمونه از اسناد مورد نظر:

"طبق ماده ۱۵ (۱) دستورالعمل‌های انطباق با مقررات، «کمیته» باید حداکثر تا ۳۱ اردیبهشت هر سال تشکیل جلسه دهد."

نمونه از خروجی مد نظر :

```
[
    {
        "Statute reference": "ماده ۱۵(۱)",
        "Date": " ۳۱ اردیبهشت هر سال ",
        "Defined terms": "کمیته",

    }
]

```

** **توجه:** بدیهی است که هرچه عناصر بیشتری از اسناد را در نظر بگیرید و در کد خود بگنجانید و آن را جامع تر کنید، امتیاز بالاتری دریافت خواهید کرد. خروجی نمونه ارائه شده صرفاً یک نمونه ساده است و انتظار می‌رود که خروجی گسترده‌تر و دقیق‌تری مشاهده شود.

** **توجه:** خروجی بالا صرفا خروجی به دست آمده از کد شما خواهد بود؛در نتیجه لازم است که خروجی بدست آمده از پیام رسان به صورت کاربر پسندتری نمایش داده شود.

نمونه های دیگر از اسناد مورد نظر :

" آیین نامه چگونگی بازرسی کار (نامه شماره ۷۵۸۶۹ مورخ ۱۴۰۱/۴/۴ وزارت تعاون، کار و رفاه اجتماعی) به پیوست "آیین‌نامه چگونگی بازرسی کار"، تدوین شده در شورای عالی حفاظت فنی موضوع تبصره ماده (۹۹) قانون کار، منضم به لوح فشرده آن که در تاریخ ۱۴۰۱/۱/۲۰ به توضیح و تصویب وزیر محترم تعاون، کار و رفاه اجتماعی رسیده است، برای درج در روزنامه رسمی کشور ارسال می‌گردد. معاون روابط کار- علی حسین رعیتی‌فرد آیین نامه چگونگی بازرسی کار به استناد تبصره ماده (۹۹) و تبصره (۱) ماده (۸۶) قانون کار جمهوری اسلامی ایران ، «آیین نامه چگونگی بازرسی کار» که در جلسه مورخ ۱۴۰۰/۱۲/۰۹ «شورای عالی حفاظت فنی» بازنگری و توسط آن شورا پیشنهاد شده است، به شرح زیر تصویب می‌گردد."

طبقه‌بندی کننده‌ی رجکسی

شما مسئول طراحی یک بات چت هستید که قادر است پیام‌های دریافتی را بر اساس محتوایشان طبقه‌بندی کند. این بات باید بتواند انواع مختلفی از پیام‌ها را تشخیص دهد، از جمله ایمیل‌ها، شماره‌های تلفن، آدرس‌ها، پیام‌های کوتاه، و پیام‌های بلند. علاوه بر این، کاربر باید قادر باشد با ارائه یک رجکس، بررسی کند که آیا یک پیام خاص مطابق با آن الگو است یا خیر. سیستم باید همچنین امکان اضافه کردن الگوهای جدید توسط کاربر را داشته باشد تا قابلیت‌های آن گسترش یابد. **وظایف**

- **طراحی Regex برای تشخیص**:
  - ایمیل‌ها
  - شماره‌های تلفن (با فرمت ایران)
  - آدرس‌ها (با استفاده از کلمات کلیدی مانند "محله، خیابان، کوچه، پلاک")
  - پیام‌های کوتاه و بلند (با تعریف تعداد کاراکترها یا کلمات برای هر یک)

- **توسعه یک تابع برای اضافه کردن رجکس‌های جدید**: این تابع باید امکان پذیرش یک نام و یک رجکس را داشته باشد و آن را به مجموعه قوانین موجود اضافه کند. مطلوب است که این مجموعه قواعد در یک پایگاه داده قرار بگیرد و به صورت دستی داخل کد پیاده‌سازی نشود.
- **توسعه یک تابع برای بررسی مطابقت پیام‌ها با الگوهای موجود**: این تابع باید قادر باشد یک پیام را دریافت کند و تمام مطابقت‌های یافت شده با الگوهای تعریف شده را برگرداند
- **توسعه یک تابع برای بررسی مطابقت یک پیام با یک رجکس خاص ارائه شده توسط کاربر** : این تابع باید بتواند تعیین کند آیا پیام داده شده مطابق با الگوی رجکس ارائه شده است یا خیر
- **تشخیص خودکار الگو از روی تعدادی مثال** قابلیت مهمی که این بات علاوه بر موارد ذکر شده باید داشته باشد، قابلیت تشخیص الگو رجکس بر اساس تعدادی مثال است. این کار باید کاملا به صورت قاعده محور و بدون استفاده از ابزارهای هوش مصنوعی مانند ابزارهای دسته‌بندی و غیره انجام پذیرد. هرچقدر الگوی به‌دست آمده دقیق‌تر و خاص‌تر باشد، امتیاز بالاتری به الگوی شما اختصاص داده خواهد شد. (قطعا جوابی مانند .* مورد نظر نیست.

توجه داشته باشید سیستم شما باید قادر باشد پیام‌هایی با چندین پترن (مثلا هم آدرس و هم تلفن) را تشخیص دهد و همه آنها را گزارش کند همچنین سیستم را به گونه‌ای طراحی کنید که قابلیت اضافه شدن الگوهای جدید به آسانی و بدون نیاز به تغییر کد اصلی را داشته باشد.
در کد باکس زیر یک مثال ساده آمده است که باید بر اساس عبارات منظم در بات پیاده‌سازی شود..

```
[

        example_text   =   "این یک تست است با ایمیل test@example.com و تلفن ۰۹۱۲۳۴۵۶۷۸۹.",
        matches_found = check_patterns(example_text)

]

```

که خروجی باید به صورت زیر شود.

```
[
    {
        "ایمیل": "test@example.com",
        "تلفن": "۰۹۱۲۳۴۵۶۷۸۹"
    }
]

```

البته، با توجه به این که خروجی در قالب یک پیام به کاربر نمایش داده می‌شود، سعی کنید فرمت خروجی مناسب برای خوانده شدن توسط کاربران باشد. می‌توانید از قابلیت‌های فرمت‌دهی پیام‌رسان برای بهینه‌سازی خروجی استفاده کنید. همچنین برای تشخیص ورودی و دستور کاربر می‌توانید از هر فرمت دلخواه (مانند استفاده از فرمت **/دستور ورودی های دستور** استفاده کنید.

بررسی نحو افعال

همان‌طور که می‌دانید در زبان فارسی عمده بار معنایی جملات بر روی فعل قرار دارد. فعل، شخص نهاد جمله، زمان انجام، و گاها توالی انجام اتفاقات را مشخص می‌کند. در زبان فارسی افعال به اشکال مختلفی ساخته می‌شوند.

افعال گذشته	افعال حال	افعال آینده
گذشته ساده	حال اخباری	آینده ساده
گذشته پیوسته	حال التزامی	
گذشته درخواستی	حال ملموس	
گذشته دور		
گذشته زنده		
گذشته ملموس		

این افعال در جملات می‌آیند و گاها اجزای مختلف آن از یک دیگر فاصله می‌گیرند. به عنوان مثال جمله

دارم با امید می‌آیم

را در نظر بگیرید که فعل دارم می‌آیم که از جنس حال ملموس است، به صورت جدا شده در جمله قرار گرفته است.
در این تمرین از شما خواسته می‌شود که در یک متن، افعال را تشخیص دهید. تشخیص افعال به این معنی است که

۱. تمام بازه‌هایی که افعال در آن‌ها قرار دارد را بیابید.

۲. بن فعل، زمان فعل و شخص فعل را مشخص کنید.

توجه کنید که ممکن است برخی از قواعد نگارشی از جمله نیم‌فاصله و جدا یا سرهم‌نویسی بخش‌های افعال رعایت نشود و روش شما باید نسبت به این موارد مقاوم باشد. همچنین در افعال محال التزامی گاهی اوقات الف ابتدای بن مضارع حذف می‌شود.
همچنین از شما خواسته می‌شود که نهاد و مفعول مرتبط به هر فعل را در صورت وجود پیدا کنید و آن را مشخص نمایید.
توجه نمایید که باید تطابق نهاد با شناسه فعل را بررسی نموده و در صورت عدم تطابق خروجی خود را اصلاح بفرمایید.
توجه بفرمایید که در زبان فارسی علاوه بر افعال عادی، افعال پیشوندی و افعال مرکب و حتی افعال پیشوندی‌ـ‌مرکب نیز وجود دارد. روش شما باید این مورد را نیز تشخیص دهد.
امتیازی: معمولا فرم اجزای جمله در اشعار به هم می‌ریزد. در این بخش از شما خواسته می‌شود که روشی ارائه بفرمایید که در اشعار نیز دقت مناسبی داشته باشد.
این تمرین را باید در قالب کتاب‌خانه parsi.io پیاده‌سازی کنید. جزئیات واسط برنامه شما بر عهده خودتان است. در ادامه یک مثال از یک واسط پیشنهادی برای این تمرین تقدیم می‌گردد. اما رعایت استانداردهای کتاب‌خانه parsi.io در این تمرین الزامی است. به عنوان مثال حتما span باید در ساختار درختی شما باشد و تمام خواسته‌های سوال نیز باید برآورده گردند.
به عنوان مثال برای ورودی جمله «من داشتم شیشه مربا را برمیداشتم» یک خروجی قابل قبول به شکل زیر خواهد بود.

```
[
    {
        "verb": {
            "span": [
                [3, 8],
                [23, 30]
            ],
            "root": "برداشتن",
```

```
            "structure": "simple",
            "person": "singular first",
            "tense": "reminding past"
        },
        "subject phrase": "من",
        "object phrase": "شیشه مربا"
    }
]

```

با به عنوان مثال در مورد فعل پیش‌وندی **پس‌افتاد** به صورت زیر خواهد.

```
[
    {
        "verb": {
            "span": [
                [0, 8],
            ],
            "root": "افتادن",
            "structure": "prefixed",
            "prefix": "پس",
            "person": "singular third",
            "tense": "simple past"
        }
    }
]

```

توجه بفرمایید که برای دسترسی به بن افعال، می‌توانید از مجموعه داده پیکره‌گان استفاده نمایید.

تشخیص کلمات شکسته، مختصر نویسی و غلط‌های املایی در متن فارسی

در این تمرین هدف شما ویرایش متن توییت‌ها و در آوردن یک متن استاندارد از آن‌ها است. متن توییت‌ها شامل غلط‌های املایی، نحوی، نگارشی، شکسته نویسی و غیره است. برای به دست آوردن یک متن استاندارد لازم است این موارد اصلاح شوند. این غلط‌ها را به سه دسته کلی تقسیم می‌کنیم که در هر دسته تعدادی از غلط‌های مربوط به آن دسته ذکر شده‌اند.

- غلط‌های املایی و تایپی:
  - یک یا تعدادی حرف اشتباه نوشته شده‌اند: حاظر ← حاضر، سلام ← شلام
  - یک یا تعدادی حرف کم نوشته شده است: پادسازی ← پیاده‌سازی
  - یک یا تعدادی حرف اضافه نوشته شده است: سلاام ← سلام
  - دو حرف جا به جا نوشته شده‌اند: خناه ← خانه
- غلط‌های دستور زبانی:
  - علائم نگارشی اشتباه به کار رفته یا به کار برده نشده است: سلام خوبی. ← سلام خوبی؟
  - اشتباه در استفاده از هکسره: حالت خوبِ؟ ← حالت خوبه؟، کتابه من ← کتابِ من
  - به کار بردن اشتباه حروف اضافه: کتابی که می‌خواستم را خریدم. ← کتابی را که می‌خواستم خریدم.
  - به کار بردن پشت سرهم دو قید پرسشی در یک جمله: آیا چرا من به دنیا آمده‌ام؟ ← چرا من به دنیا آمده‌ام؟
- شکسته نویسی:
  - عامیانه نویسی: معلمای من خیلی گلن ← معلم‌های من خیلی گل‌اند
  - اختصار نویسی: ج.ا. ← جمهوری اسلامی، نمت ← نمی‌توانم

نمونه ورودی	نمونه خروجی
امروز تعداد حاظرین به جلسه از آن چیزی که منه مدیر جلسه تضور می کردم بیشتر بود.	امروز، تعداد حاضرین در جلسه از آن چیزی که من مدیر جلسه تصور می‌کردم، بیش‌تر بود.

در ترم‌های گذشته روی بخشی از این غلط‌ها کار شده است. مانند اصلاح نیم‌فاصله و فاصله در متن و علائم نگارشی و اصلاح غلط‌های مربوط به هکسره. شما برای دریافت نمره کامل باید تمام قسمت‌های غلط‌های املایی و تایپی، یک قسمت از غلط‌های دستور زبانی به انتخاب خودتان و یک قسمت از شکسته‌نویسی به انتخاب خودتان را انجام دهید. در صورتی که قسمت انتخابی از بخش غلط‌های دستور زبانی در ترم‌های پیش کار شده بود در این قسمت شما باید با بررسی کارهای ترم‌های پیش ضعف‌های آن‌ها را پیدا کنید و این موارد را بهبود دهید.

* برای انجام این تمرین می‌توانید از پیکره کلمات فارسی کمک بگیرید.

مدیریت رویدادها و یادآوری‌ها

هدف از این تکلیف توسعه‌ی یک تجزیه‌گر مبتنی بر رجکس است که ورودی‌های زبان طبیعی مرتبط با برنامه‌ریزی کارهای روزانه را تفسیر کند. و باید قادر به انجام موارد زیر باشد:

- برنامه‌ریزی کارها: تجزیه عبارات برای برنامه‌ریزی یادآوری‌های جدید.
- لغو کارها: شناسایی دستورات برای لغو یادآوری‌های برنامه‌ریزی‌شده.
- تغییر کارها: تشخیص دستورالعمل‌ها برای تغییر زمان یادآوری‌های موجود.
- انجام شدن کار: شناسایی دستورات برای انجام شدن یادآوری‌های برنامه‌ریزی‌شده.
- بازگرداندن برنامه: بازیابی و بازگرداندن برنامه‌های روزانه و هفتگی.

۱ توضیح تکلیف

وظیفه شما پیاده‌سازی یک تجزیه‌گر زبان طبیعی با استفاده از رجکس است. این تجزیه‌گر باید اطلاعات ضروری را از پیام‌های کاربر برای برنامه‌ریزی، لغو یا تغییر یادآوری‌های کارهای "دوره‌ای" و "یک‌باره" از طریق بات استخراج کند. از طریق این لینک می‌توانید به فایل تکلیف "ایجاد و بروز رسانی وظایف با عبارات منظم" دسترسی داشته باشید و از آن برای این تمرین استفاده کنید تا زمان بیشتری برای رسیدگی به تنوع‌های زبانی مرتبط و جزییات دیگر داشته باشید.

۱.۱ بخش ۱: برنامه‌ریزی کارها

هدف: استخراج نام کارها و زمان‌های برنامه‌ریزی‌شده از پیام‌ها.
مثال ورودی: "یادم باشه هر روز ساعت ۸ صبح به جلسه اسکرام برم."
راهنمای رجکس: به دنبال کلمات کلیدی مانند "یادم باشه به" دنبال شده توسط یک عمل، و سپس مشخصات زمانی باشید.

```
[
    {
        "name": "رفتن به جلسه اسکرام",
        "time": "۸ صبح هر روز",
        "period": "یک روز",
        "done": "False",
        "cancel": "False",
    }
]

```

۲.۱ بخش ۲: لغو کارها

هدف: شناسایی کاری که باید لغو یا حذف شود.
مثال ورودی: "جلسه اسکرام روزانه‌ام را لغو کن."
راهنمای رجکس: عبارات کلیدی مانند "لغو کن" را شناسایی کنید.

```
[
    {
        "name": "جلسه اسکرام روزانه لغو شد",
        "time": "۸ صبح هر روز",
        "period": "یک روز",
        "done": "False",
        "cancel": "True",
    }
]

```

۳.۱ بخش ۳: تغییر برنامه‌های کار

هدف: تجزیه پیام‌ها برای شناسایی کاری که برنامه آن نیاز به تغییر دارد.
مثال ورودی: "زمان تماس با دوستم در ۱۲ فروردین را به ۹:۳۰ شب تغییر بده."
راهنمای رجکس: الگوهایی را که تغییر را توصیف می‌کنند، شامل تغییر کار و زمان را شناسایی کنید.

```
[
    {
        "name": "تماس با دوست",
        "time": " ساعت ۹:۳۰ شب، ۱۲ فروردین ",
        "period": "None",
        "done": "False",
        "cancel": "False",
    }
]

```

۴.۱ بخش ۴: انجام شدن کار

هدف: حذف کار انجام شده از برنامه هفتگی و روزانه
مثال ورودی: "کار تماس با دوستم انجام شد."
راهنمای رجکس: عبارات کلیدی مانند "انجام شد" یا "تمام شد" را شناسایی کنید.

```
[
    {
        "name": "کار تماس با دوست حذف شد.",
        "time": "ساعت ۹:۳۰ شب، ۱۲ فروردین ",
        "period": "None",
        "done": "True",
        "cancel": "False",
    }
]

```

۵.۱ بخش ۵: بازگرداندن برنامه‌های روزانه و هفتگی

هدف: بازیابی و بازگرداندن برنامه‌های روزانه و هفتگی بر اساس تاریخ‌های مشخص شده توسط کاربر.

مثال ورودی: "برنامه‌ام را برای ۲۵ اردیبهشت بگو" یا "برنامه هفتگی‌ام را نشان بده."
راهنمای رجکس: الگوهایی را برای شناسایی درخواست‌های بازگرداندن برنامه با توجه به تاریخ‌های مشخص یا بازه‌های زمانی (مانند "روزانه" یا "هفتگی") شناسایی کنید.
خروجی مورد انتظار: برای درخواست‌های مربوط به تاریخ مشخص، لیستی از کارهای برنامه‌ریزی‌شده برای آن تاریخ. برای درخواست‌های "هفتگی"، برنامه‌ای شامل تمام کارهای برنامه‌ریزی‌شده برای هفته مورد نظر برگرداند.

۲ نکات تکمیلی

- رسیدگی به درخواست‌های برنامه‌ریزی پیچیده‌تر (مثلاً هر دو روز یکبار، روزهای هفته، آخر هفته‌ها).
- تجزیه و رسیدگی به درخواست‌ها برای حذف چندین کار به طور همزمان.
- پیچیدگی را با تقسیم به قسمت‌های کوچکتر، مدیریت کنید. به عنوان مثال، تجزیه زمان می‌تواند یک رجکس باشد، در حالی که شناسایی عمل می‌تواند دیگری باشد.
- به تنوع‌های زبان طبیعی (مثلاً "یادم باشه به..." در مقابل "نیاز به یادآوری دارم به...") رسیدگی کنید. سعی کنید رجکس خود را تا حد امکان قوی و در عین حال بدون ایجاد پیچیدگی بیش از حد، بسازید.

۳ معیارهای ارزیابی

- **دقت:** توانایی تجزیه‌گر در استخراج دقیق نام‌های کار، زمان‌ها، و تغییرات از ورودی‌های مختلف.
- **انعطاف پذیری:** توانایی تجزیه‌گر در رسیدگی به یک دامنه وسیع از تنوع ورودی.
- **کارایی:** استفاده از ویژگی‌های رجکس برای دستیابی به هدف بدون الگوهای بیش از حد پیچیده.

گردآوری اطلاعات زمینه‌ای

در این تمرین می خواهیم که اطلاعات زمینه ای و خارجی را از متن استخراج کنیم. استخراج این اطلاعات پنهانی از متن ارزش بسیار بالایی دارد و می تواند در تجزیه و تحلیل احساسات، پیدا کردن متغیرهای آماری و ... کمک کند. دقت کنید که مثال های زده شده صرفا برای ایده دادن به شما هستند و هرگونه ایده و خلاقیت دیگری می توان برای پیدا کردن این اطلاعات زمینه ای استفاده کنید و اطلاعات زمینه ای بیشتری هم می توانید به دست آورید. این اطلاعات شامل موارد زیر می شود:

- استخراج اطلاعات شخصی:

  نام

  نام خانوادگی

  جنسیت

  گروه سنی:

  - (۱۴-۱۹ ساله) گروه سنی نوجوان
  - (۲۰-۳۹ ساله) گروه سنی بزرگسال
  - (۴۰-۶۴ ساله) گروه سنی میان‌سال
  - (۶۵-۹۰ ساله) گروه سنی پیر

  شغل در صورت وجود (برای شغل های معلمی، پزشکی، نویسندگی، آشپزی و نقاشی اطلاعات به صورت ضمنی داده می شود ولی برای بقیه شغل ها نیاز به استخراج شغل از روی اطلاعات ضمنی نیست و در آن حالت شغل به طور مستقیم داده می شود.)

- استخراج اطلاعات مکانی:

  شهر

  کشور (به صورت ضمنی از روی شهر قابل پیدا کردن است. از روی شهرهای مهم و پایتخت ها میتوان به این اطلاعات رسید. اگر شهری ذکر نشده بود، و اسم خود کشور هم وجود نداشت این قسمت را "نامعلوم" بگذارید)

- استخراج اطلاعات سلامتی:

  موارد بهداشتی و سلامتی در صورت وجود (یک سری بیماری های شایع و موارد این چنینی مانند کمردرد، سردرد، دندان درد، دیابت و ...)

- استخراج اطلاعات رویدادی:

  رویدادهای ذکر شده در صورت وجود (جشن ها و مراسم های معروف مثل عید نوروز، جشنواره غذا، شب یلدا، کریسمس و ...)

- تجزیه و تحلیل احساسات

  خلق و خوی کاربر که می تواند یکی از موارد خوشحال، ناراحت، خشمگین و یا بی احساس را داشته باشد.

- استخراج اطلاعات زمانی قرار

  تاریخ قرار قبلی و تاریخ قرار بعدی در صورت وجود

به عنوان مثال اگر در ورودی متن "من زهرا اسدی هستم و وقتی نوه‌ام را به مدرسه می‌بردم که قبلاً در آن درس می‌دادم، کمردرد گرفتم. بعد از بازنشستگی حوصله ام سر می رود. باید از رفتن به جشنواره برج میلاد تهران صرف نظر کنم، چقدر حیف! در بهار برج میلاد واقعا زیبا به نظر می‌رسد. برای عید نوروز برنامه داشتم): زنگ زدم که بهتون بگم نوبت ۱۰ اردیبهشت رو کنسل کنید و نوزدهم اردیبهشت ساعت جدید تعیین کنید. ممنونم" داده شود انتظار داریم خروجی به شکل زیر باشد. (خروجی شما می تواند به هر فرمت و شکل دیگری نیز باشد و تاجایی که نیازهای مساله را برطرف کند مشکلی ایجاد نمی شود)

```
[
    {
        "Name": "زهرا",
        "Surname": "اسدی",
        "Sex": "زن",
        "Age": "پیر",
        "Job": "معلم",
        "City": "تهران",
        "Country": "ایران",
        "Sickness": "کمردرد",
        "Event": [
        "عید نوروز",
        " جشنواره برج میلاد "
        ],
        "Sentiment": "ناراحت",
        "Add Appoinment":True,
        "New Date" : "نوزدهم اردیبهشت",
        "Cancel Appoinment":True,
        "Cancel Date" : "۱۰ اردیبهشت",
    }
]

```

در اینجا شغل معلمی از روی اطلاعات زمینه ای درس دادن استخراج شده است و سن هم از روی اطلاعات زمینه ای نوه و بازنشستگی می تواند استخراج شود. از روی جشنواره برج میلاد می توان به شهر تهران رسید و از تهران به کشور ایران. به این شکل اطلاعات پنهانی متن را استخراج کردیم.

مورد استخراج خلق و خوی کاربر امتیازی است

تبدیل فینگیلیش به فارسی

هدف از این تمرین توسعه یک برنامه برای تبدیل متن‌های نوشته شده به صورت فینگیلیش (استفاده از کاراکترهای انگلیسی به جای حروف فارسی) به فارسی با استفاده از عبارات منظم و سایر ابزارهای پیش پردازش متنی مورد نیاز، است.
برای انجام این تسک شما باید الگوهای مورد استفاده در متن فینگلیش را پیدا کرده و معادل‌های فارسی آن را بسازید.
موردهایی که باید پوشش داده شوند:

- حروفی که دارای نگاشت یک به یک یعنی به ازای هر کاراکتر انگلیسی فقط یک کاراکتر فارسی برای آن وجود دارد.
- ۲ حرفی‌ها یعنی آن دسته از الگوهایی که به ازای ۲ کاراکتر انگلیسی یک حرف فارسی برای آن وجود دارد و برعکس. مانند (chera –> چرا - ax —> عکس)
- حروفی که به ازای یک کاراکتر انگلیسی چندین حرف فارسی برای آن وجود دارد مانند (t –> ت‌ـط) برای این قسمت می‌توانید از لیست تکرار کلمات فارسی ویکیپدیا در این لینک استفاده کنید.
- کلمات پر کاربرد فارسی که دارای الگوی نوشتاری متفاوتی نسبت به آوای آنها وجود دارد مانند (خواهر)

به عنوان مثال داریم:

```
>>>input_text : "Ba zohoore modelhaye zabani bozorg mitavan dar zamane kam khorooji monaseb daryaft kard."
>>>output : "با ظهور مدل‌های زبانی بزرگ می‌توان در زمان کم خروجی مناسب دریافت کرد."

```

استخراج ویژگی تلفن‌همراه از نظرات

همواره در فروشگاه‌های اینترنتی تلفن‌های همراه یکی از پرفروش ترین کالاها به شمار میروند. معمولا اطلاعاتی از ویژگی‌های هر محصول در صفحه آن وجود دارد اما بدون شک مفید ترین بخش در این صفحه، نظرات کاربران در مورد آن محصول است. در این تمرین هدف استخراج ویژگی‌های متفاوت تلفن‌های همراه از نظرات خریداران است و این موضوع میتواند فرایند انتخاب محصول را برای مشتری تسریع کند.
به طور مشابه در ترم‌های گذشته روی استخراج ویژگی‌های عمومی انواع کالا با توجه به نظرات مشتریان کار شده است. این موضوع به شما برای توسعه این برنامه کمک خواهد کرد و طبیعتا انتظار میرود که نتیجه کار شما یک بهبود نسبت تلاش‌های قبلی باشد. اما به طور خاص در این تمرین تلفن همراه را از جوانب دیگری می‌توان بررسی کرد که در اینجا باید در نظر گرفته شود. در نتیجه علاوه بر مسائل کلی مثل قیمت یا گارانتی باید مواردی مثل:

- کیفیت تصویر/فیلم برداری
- عمر باطری
- عملکرد در پردازش سنگین
- و ...

نیز استخراج شوند. پیشنهاد می‌شود که با مطالعه بر روی نظرات خریداران در وبسایتی مثل دیجیکالا موارد پراهمیت را شناسایی کرده تا در اولویت پیاده‌سازی قراردهید. خروجی کد شما یک دیکشنری است که هر کلید آن یک ویژگی و مقدار متناظر با کلید مقدار ویژگی است.
برای نمونه بخشی از خروجی می‌تواند به شکل زیر باشد:

- ورودی:

  گوشی واقعا راضی کننده است . باتری گوشی یک روز پر کار رو کامل همراهی میکنه و تو نیم ساعت ۵۰ درصدش پر میشه دوربین فوق العاده ای داره و راضیتون میکنه پردازنده اش اصلا هنگ نمیکنه و داغ نمیشه و برنامه های سنگین رو راحت اجرا میکنه صفحه نمایشش با ۹۰۰ نیت روشنایی کاملا تصاویر رو واضح و با کیفیت نشون میده . ظاهر گوشی هم خیلی شیک و خوشگله .رابط کاربری روان و جذابی داره. در کل گوشی قابل قبولی هست و قیمتش نسبت به مشخصاتش خیلی ارزونه .

- خروجی:

  نظر کلی: واقعا راضی کننده است.

  قیمت و ارزش خرید: قیمتش نسبت به مشخصاتش خیلی ارزونه.

  باتری: باتری گوشی یک روز پر کار رو کامل همراهی میکنه و تو نیم ساعت ۵۰ درصدش پر میشه.

  عمکرد پردازشی: پردازنده اش اصلا هنگ نمیکنه و داغ نمیشه و برنامه های سنگین رو راحت اجرا میکنه

  ظاهر: خیلی شیک و خوشگله

  رابط کاربری: رابط کاربری روان و جذابی داره

درخواست تعریف چالش جدید برای تمرین ۲

در صورتی که پیشنهاد جدیدی دارید می‌توانید بات کاربردی جدیدی را در پروپوزال یک صفحه‌ای مشابه توضیحات و تعریف چالش‌های گفته شده، تعریف کرده و در صورت تصویب روی آن کار کنید.

فایل پروپوزال گروه خودتان را، به آدرس semad.zol۷۴@sharif.edu ایمیل کنید.