بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

بازیابی پیشرفته اطلاعات

نیم‌سال دوم ۰۱-۰۰

استاد: احسان‌الدین عسگری

تمرین اول — **عبارات منظم** — مهلت ارسال: ۲۸ فروردین

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در تمرین‌هایی که چند ترک دارند، فقط یک نفر از هر گروه در سامانه CW باید ترک مورد نظر گروه را انتخاب کند. امکان تغییر ترک تا قبل از زمان ددلاین انتخاب ترک وجود دارد. البته ذکر این نکته ضروری است که هر ترک محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها بعد از ۳ روز 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.
- 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 ...) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

توضیحات کلی

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 ترم 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 ۱ 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'
```

تشخیص و اصلاح نیم‌فاصله‌ها و فاصله‌ها در متن و علائم نگارشی

در این تمرین هدف شما اصلاح کردن فرم نویسه‌های فارسی و یکسان‌سازی آن‌هاست. به طور ویژه، انتظار می‌رود که برنامه شما قادر به اصلاح اشتباهات در استفاده از نیم‌فاصله و فاصله، جدانویسی و سرهم‌نویسی کلمات و افعال مختلف، و فاصله‌گذاری‌های مربوط به علائم نگارشی باشد. برای آشنایی با شکل استاندارد نوشتار، می‌توانید از مواردی که در این نوشته به آن‌ها اشاره شده، این مطلب و همچنین این مطلب استفاده کنید.
در این تمرین استفاده از فهرستی از افعال فارسی یا لیست‌های مشابه می‌تواند کمک‌دهنده باشد. در گزارش خود به الگوهایی که کد شما می‌تواند آن‌ها را اصلاح کند، اشاره کنید.

ورودی	خروجی
در هنگام وقوع بلایای طبیعی ،بیش‌ترین خسارت متوجه قشر آسیب پذیر جامعه می شود.	در هنگام وقوع بلایای طبیعی، بیشترین خسارت متوجه قشر آسیب‌پذیر جامعه می‌شود.
آن ها چهاردهم مارس(مارچ) را بعنوان روز جهانی ی عدد پی نام گذاری کرده اند.	آن‌ها چهاردهم مارس (مارچ) را به‌عنوان روز جهانی عدد پی نام‌گذاری کرده‌اند.
لاک پشت با هوش از همهٔ آنها زیرکانه تر عمل کرد!	لاک‌پشت باهوش از همه‌ی آن‌ها زیرکانه‌تر عمل کرد!

استخراج واقعه از متن (استخراج زمان و مکان و محتوا)

طی این تمرین باید با استفاده از regex وقایع موجود در متن را استخراج کنید. وقایع می‌توانند انواع گوناگونی داشته باشند که برخی از آن‌ها در ادامه ذکر شده‌اند. ورودی مسئله یک متن و خروجی مسئله بازه‌های شروع و پایان یک واقعه به همراه نوع واقعه است و زمان و مکان واقعه است. این خرجی باید به شکل یک لیست از چند دیکشنری پایتون برگردانده شود که کلیدهای دیکشنری مانند خروجی زیر هستند. به عنوان مثال :
ورودی : "تابستان ۱۴۰۱ پایان مذاکرات ایران با آمریکا در شهر ژنو بود."
خروجی:

- نوع واقعه (type) : گفت‌وگو و مذاکرات و توافق
- متن واقعه (text) : پایان مذاکرات ایران با آمریکا
- بازه واقعه : (۱۳،۴۲)
- زمان (time) : تابستان ۱۴۰۱
- مکان (place) : شهر ژنو

در انجام این ترک حتما به نکات زیر دقت کنید:

۱. توجه داشته باشید که هدف تمرین پوشش کامل تمامی وقایع نیست. لازم است به انتخاب خود دسته یا و دسته‌هایی از وقایع را انتخاب کنید و سعی کنید تا جای ممکن آن‌ها را پوشش دهید. به منظور تشخیص وقایع می‌تونید از ابزارهای خروجی ترم قبل نظیر تشخیص زمان و یا مکان استفاده کنید.

۲. دقت داشته باشد اگر زمان یا مکان واقعه نامعلوم باشد باید برای زمان یا مکان مقدار null بازگردانده شود.

۳. در مورد اینکه واقعه چه بخشی از جمله است، فکر کنید و استدلال خود را در مستند تمرین بیان کنید. سعی کنید وقایع مطابق با تعریفی که از واقعه ارائه کرده اید، استخراج شوند. در پایان به تلاش شما به منظور استخراج وقایع نمره داده خواهد شد.

۴. در صورتی که علاوه بر توکن زمان در مواردی بتوانید مقدار آن با فرمت استاندارد را استخراج کنید، نمره اضافه خواهد داشت.

۵. در صورتی که علاوه بر توکن مکان در موارد بتوانید نوع مکان نظیر کشور، شهر، اتاق، اتاق مجازی و ... را استخراج کنید، نمره اضافه خواهد داشت.

۶. **فقط کافی است از ۱۵ واقعه موجود ۱۰ واقعه را به انتخاب خود بزنید. وقایع انتخابی را در گزارش خود بیاورید.**

لیستی از انواع وقایع در زیر آمده است.

- وقایع مهم
  - قرار ملاقات
  - بیماری (بیماری، همه‌گیری، ...)
  - مرگ (مرگ، درگذشت، شهادت، ...)
- وقایع سیاسی

- قراردادهای رسمی (قطع‌نامه، عهدنامه، ...)
- جنگ (جنگ، حمله، نزاع ، ...)
- عزل و نصب و استعفا و انتخاب
- گفت‌وگو و مذاکرات و توافق
- تحریم و رفع تحریم
- بیانیه سیاسی

- وقایع ورزشی

- برد و باخت و تساوی
- صعود و سقوط و حذف
- قهرمانی و نایب قهرمانی
- کسب مدال

- وقایع اقتصادی

- تغییر قیمت (افزایش، کاهش، صعود، ...)
- واردات و صادرات

ورودی	خروجی

گفتگوهای صلح اوکراین، فردا از سر گرفته خواهد شد.
(تشخیص این واقعه دشوار است زیرا اوکراین ممکن است به عنوان مکان در نظر گرفته شود.)
	[
	{
	"type": "گفت‌وگو و مذاکرات و توافق",
	"text": "گفتگوهای صلح اوکراین",
	"span": [0,20],
	"place": "",
	"time": "فردا"
	}
	]

دو هفته از استعفای نخست‌وزیر بریتانیا می‌گذرد.	[
	{
	"type": "قراردادهای رسمی",
	"text": "امضای قرارداد ۲۲۳۱",
	"span": [22,51],
	"place": "",
	"time": ""
	}
	]

اکثر کشورهای دنیا با امضای قرارداد ۲۲۳۱ سازمان ملل موافق هستند.	[
	{
	"type": "قراردادهای رسمی",
	"text": "امضای قرارداد ۲۲۳۱",
	"span": [22,51],
	"place": "",
	"time": ""
	}
	]

[
  {
  "type": "جنگ",
  "text": "حمله سنگین ارتش عراق به ایران",
  "span": [21,50],
  "place": "",
  "time": "سال ۱۳۵۹"
  }
]
	پس از انقلاب اسلامی، حمله سنگین ارتش عراق به ایران در سال ۱۳۵۹ آغاز گردید.

[
  {
  "type": "تغییر قیمت",
  "text": "افزایش بهای طلا",
  "span": [11,26],
  "place": "صرافی‌های تهران",
  "time": "امروز"
  }
]
	امروز شاهد افزایش بهای طلا در صرافی‌های تهران بودیم.

[
  {
  "type": "مرگ",
  "text": "درگذشت امیرکبیر",
  "span": [27,42],
  "place": "",
  "time": "بیست دی ماه"
  }
]
	امروز بیست دی ماه مصادف با درگذشت امیرکبیر است.

[
  {
  "type": "بیماری",
  "text": "شیوع کوید ۱۹",
  "span": [6,18],
  "place": "جهان",
  "time": ""
  }
]
	پس از شیوع کوید ۱۹ بسیاری در جهان جان خود را از دست دادند.

[
  {
  "type": "واردات و صادرات",
  "text": "افزایش واردات میوه و تره‌بار",
  "span": [33,61],
  "place": "",
  "time": "سال آینده"
  }
]
	به علت خشکسالی در سال آینده شاهد افزایش واردات میوه و تره‌بار خواهیم بود.

[
  {
  "type": "قرار ملاقات",
  "text": "دیدار پوتین و مکرون",
  "span": [0,19],
  "place": "کاخ کرملین",
  "time": "امروز عصر"
  }
]
	دیدار پوتین و مکرون امروز عصر در کاخ کرملین برگزار خواهد شد.

آزمایش موشک کروز جدید باعث اجرای تحریم‌های جدیدی علیه ایران شد.	[
	{
	"type": "تحریم و رفع تحریم",
	"text": "اجرای تحریم‌های جدیدی علیه ایران",
	"span": [27,59],
	"place": "",
	"time": ""
	}
	]

در دیدار شب گذشته در مادرید تیم اتلتیکو مادرید به مرحله پایانی مسابقات صعود کرد.	[
	{
	"type": "صعود و سقوط و حذف",
	"text": "تیم اتلتیکو مادرید به مرحله پایانی مسابقات صعود کرد",
	"span": [29,80],
	"place": "مادرید",
	"time": "شب گذشته"
	}
	]

بازیکن هلندی تبار عنوان نایب قهرمانی را کسب کرد.	[
	{
	"type": "قهرمانی و نایب قهرمانی",
	"text": "بازیکن هلندی تبار عنوان نایب قهرمانی را کسب کرد",
	"span": [0,47],
	"place": "",
	"time": ""
	}
	]

در جام‌جهانی والیبال در قزاقستان تیم ایران موفق به کسب مدال نقره شد.	[
	{
	"type": "کسب مدال",
	"text": "کسب مدال نقره",
	"span": [51,64],
	"place": "قزاقستان",
	"time": ""
	}
	]

واحدهای اندازه‌گیری

در این تمرین هدف پیدا کردن واحدهای اندازه‌گیری است. بدین ترتیب از برنامه شما انتظار می‌رود برای هر واحد اندازه‌گیری پیدا شده در صورت وجود مقدار اندازه گرفته شده و مورد اندازه‌گیری را نیز خروجی دهد. برای آشنایی با کمیت و واحدهای اندازه‌گیری به سایت باحساب می‌توانید مراجعه کنید.
در این تمرین لازم است ابتدا تمامی واحدهای اندازه‌گیری و کمیت‌های متناسب آن‌ها را جمع‌آوری کنید، سپس عبارت‌های منظمی بنویسسید که بتواند واحدهای اندازه‌گیری، مقدار اندازه گرفته شده و مورد اندازه‌گیری (هر کدام از این موارد ممکن است غایب باشد) را گزارش کند.
به این منظور در کتابخانه parsi.io ابزاری جهت تشخیص اعداد (به خصوص اعداد اعشاری) به شما داده شده است، از این ابزار استفاده کنید تا بتوانید اعداد را تشخیص دهید و بدین ترتیب مقدار اندازه‌گیری شده را گزارش دهید. در این تمرین انتظار می‌رود ابتدا به معماری مناسب و مقیاس پذیر فکر شود به گونه‌ای که برای افزایش و بروزرسانی الگوها مشارکت‌کننده بعدی در این پروژه دچار مشکل نشود.
قسمت امتیازی: در صورتی که قسمت قبل به درستی کامل کار کند در نظر گرفته می‌شود: در این قسمت تبدیل‌ها را نیز در نظر می‌گیرید و از آنجا که هر واحد متعلق به یک کمیت است؛ مقادیر تبدیل شده این واحد نسبت به واحدهای دیگر در آن کمیت را نیز گزارش می‌کنید.
در جدول زیر خروجی و ورودی انواع حالت‌های مختلف نمایش داده شده است (برای span لازم است تا index شروع و پایان در یک لیست مشخص شوند).

ورودی	خروجی

علی ۳.۵ کیلوگرم آرد خرید و باتری خود را هشتاد و پنج صدم وات شارژ کرد.	[
	{
	"type": "وزن",
	"amount": 3.5,
	"unit": "کیلوگرم",
	"item": "آرد",
	"marker": "۳.۵ آرد کیلوگرم",
	"span": [4,19]
	},
	{
	"type": "توان",
	"amount": 0.85,
	"unit": "وات",
	"item": "",
	"marker": "هشتاد و پنج صدم وات",
	"span": [40,59]
	}
	]

شهاب سنگی به تندی ۱۵ km/s وارد جو زمین شد.	[
	{
	"type": "سرعت",
	"amount": 15,
	"unit": "کیلومتر بر ثانیه",
	"item": "",
	"marker": "۱۵ تندی km/s",
	"span": [13,25]
	}
	]

یک خودرو با سرعت زیاد از ما سبقت گرفت.	[
	{
	"type": "سرعت",
	"amount": "",
	"unit": "",
	"item": "",
	"marker": "سرعت زیاد",
	"span": [12,21]
	}
	]

استخراج وقایع و علائم بورس ایران

در این تمرین هدف استخراج موجودیت‌های نام‌دار بورسی و وقایع مربوط به بازار بورس با استفاده از regex از متن است. مساله یک متن را به عنوان ورودی دریافت می‌کند و خروجی آن آرایه ای از موجودیت های استخراج شده است.
به منظور جمع آوری عنوان مارکرهای بورسی و کلمات کلیدی می توانید از سایت‌هایی نظیر سهام‌یاب علی الخصوص بخش توئیت‌ها و بخش گفت‌وگوی ره‌آورد ۳۶۵ و سایت اصلی تالار بورس استفاده کنید. در گزارش خود دسته بندی های استخراج شده (type) و مثال هایی از هر مورد را شرح دهید.
در این تمرین باید سعی شود تا هر با استفاده از regex ها کلمات و عبارات کلیدی مرتبط با بورس از داخل متن استخراج شوند. به صلاح دید خود می‌توانید دسته‌بندی وقایع را دقیق‌تر کنید. از جمله مواردی که باید به آن ها توجه کنید موارد زیر هستند:

- نام شرکت های بورس ایران
- نام نمادهای معاملاتی در بورس ایران
- کلمات و اصطلاحات خاص:
  - ضرر
  - سود
  - اطلاعیه
  - حقیقی/حقوقی
  - افزایش سرمایه
  - تقسیم سود
  - دامنه نوسان
  - نوسان شدید
  - سهم رانتی
  - ...

در جدول زیر نمونه هایی از ورودی مسئله و خروجی مرتبط با هرکدام آورده شده است.

ورودی	خروجی

برکت همین افشای ب باعث شد سهم سه درصد مثبت شود. بخاطر همین میگم پیگیر باشید.	[
	{
	"type": "نماد",
	"marker": "برکت",
	"span": [1,5]
	},
	{
	"type": "واقعه",
	"marker": "افشای ب",
	"span": [11,18]
	},
	{
	"type": "واقعه",
	"marker": "سه درصد مثبت",
	"span": [31,43]
	}
	]

جریان آغاز معاملات فزر با ۱۵.۲ واحد تاثیر مثبت بر روند صعودی بازار فرابورس اثر گذار بود.	[
	{
	"type": "نماد",
	"marker": "فزر ",
	"span": [19,22]
	},
	{
	"type": "واقعه",
	"marker": "2.15 واحد تاثیر مثبت",
	"span": [26,46]
	}
	]

گزارش فعالیت ماهانه دوره ۱ ماهه منتهی به ۱۴۰۰/۰۹/۳۰ برای دیران منتشر شد.	[
	{
	"type": "نماد",
	"marker": "دیران ",
	"span": [58,63]
	},
	{
	"type": "گزارش",
	"marker": "فعالیت ماهانه",
	"span": [6,19]
	}
	]

استخراج دستور غذا از متن

در این تمرین به شما متن تعدادی دستور غذا داده می‌شود و با استفاده از آن باید سه مورد زیر را به کمک عبارات منظم استخراج کنید:

۱. پیدا کردن مواد اولیه بدون مقدار آن‌ها.

۲. پیدا کردن مقدار استفاده شده برای هر ماده‌ی اولیه.

۳. پیدا کردن متن دستور غذا.

خروجی شما به ازای هر ورودی باید به شکل یک دیکشنری به فرمت زیر باشد.

```
{
  "ingredients": "لیستی از مواد اولیه",
  "quantity": "لیستی از حجم مواد اولیه",
  "recipie": "دستور غذا",
  "span_ingredients": "لیستی از مکان مواد اولیه در متن بر اساس کاراکتر",
  "span_recipie": "مکان دستور غذا در متن بر اساس کاراکتر",
  "time": "زمان اجرای برنامه"
}
```

در زیر یک نمونه از مثال این تمرین آورده شده است. برای نمونه‌های بیشتر به این لینک مراجعه فرمایید.

ورودی	خروجی

ورودی نمونه:
مواد لازم:
اردک درسته ۱ عدد
پیاز ۱ عدد
انار ۳ قاشق غذاخوری
رب گوجه ۱ قاشق غذاخوری
رب نارنج ۲ ق غ
نمک و فلفل و ادویه
سبزی معطر
طرز تهیه:
اردک را بعد از شستشو و تمیز کردن با ادویه و نمک و فلفل مزه دار می‌کنیم؛ و پیاز را طلایی کرده و به آن زردچوبه اضافه می‌کنیم سپس انار و سبزی را اضافه کرده و خیلی کم تفت می‌دهیم و در آخر ادویه را اضافه می‌کنیم. بعد از خنک شدن مواد داخل شکم اردک را از مواد پر کرده و با خلال دندان شکم اردک را می‌بندیم و در تابه‌ای آن را سرخ می‌کنیم. سپس در قابلمه‌ای که اندازه اردک باشه رب انار و گوجه همراه نصف لیوان آب می گذاریم تا ملایم بپزد.
	{
	"Ingredients": [
	"اردک درسته",
	"پیاز",
	"انار",
	"رب گوجه",
	"رب نارنج",
	"نمک و فلفل و ادویه",
	"سبزی معطر"
	],
	"quantity": [
	"عدد ۱",
	"عدد ۱",
	"قاشق غذاخوری ۳",
	"قاشق غذاخوری ۱",
	"ق غ ۲",
	"",
	""
	],
	"recipie": "اردک را بعد از شستشو و تمیز کردن با ادویه و نمک و فلفل مزه دار
	می‌کنیم؛ و پیاز را طلایی کرده و به آن زردچوبه اضافه می‌کنیم
	سپس انار و سبزی را اضافه کرده و خیلی کم تفت می‌دهیم
	و در آخر ادویه را اضافه می‌کنیم. بعد از خنک شدن مواد داخل شکم اردک
	را از مواد پر کرده و با خلال دندان شکم اردک را می‌بندیم
	و در تابه‌ای آن را سرخ می‌کنیم
	سپس در قابلمه‌ای که اندازه اردک باشه رب انار و گوجه
	همراه نصف لیوان آب می گذاریم تا ملایم بپزد.",
	"span_ingredients": [
	[11,27],[28,38],[39,58],[59,81],[82,96],[97,115],[116,125]
	],
	"span_recipie": [136,567],
	"time": 1
	}

مچ کردن سریع ریشه‌ها و ریشه دربابها و کلمات قرآنی برای موتور جست وجو

در این تمرین، هدف انتخاب فرم های مختلف یک کلمه قرآنی، و جستجو با فاصله های متفاوتی از یکدیگر در آیات براساس فرم های انتخاب شده است.
کاربرد این تمرین به این صورت است که کاربر کلمه‌هایی را وارد میکند و سیستم به کاربر فرم‌های مختلف کلمه‌ها را پیشنهاد میدهد که کاربر فرم مورد نظر خود را انتخاب کند و سپس بتواند با تعیین فاصله بین کلمه‌های وارد شده آیات مورد نظر خود را ببیند.
شما باید دو عبارت منظم طراحی کنید، عبارت منظم اول شکل های مختلف یک کلمه را در قرآن پیدا میکند و عبارت منظم دوم آیه ها رو براساس شکل ها مختلف یک کلمه و فاصله بین آن ها برمی‌گرداند.
در ورودی کلمه‌هایی برای جستجو و فاصله‌ای بین کلمات تعریف شده است، تمامی فرم‌های مختلف هر کلمه را نشان می‌دهیم تا با انتخاب هر فرم جستجویی براساس آن صورت بگیرد.

ورودی	خروجی
کلمات : صراط، هَدَي فاصله: سه کلمه	نمایش شکل های مختلف هر کلمه: اهْدِنَا (ظاهر کلمه) ـ هَدَي (لمای کلمه) ـ هدی (ریشه کلمه) الصِّرَاطَ (ظاهر کلمه) ـ صِرْط (لمای کلمه) ـ صرط (ریشه کلمه)
انتخاب یک فرم از هر کلمه: هدی (لمای کلمه) ـ صرط (ریشه کلمه)	آیات مورد نظر: 2##142 سَيَقُولُ السُّفَهَاءُ مِنَ النَّاسِ مَا وَلَّاهُمْ عَنْ قِبْلَتِهِمُ الَّتِي كَانُوا عَلَيْهَا قُلْ لِلَّهِ الْمَشْرِقُ وَ الْمَغْرِبُ يَهْدِي مَنْ يَشَاءُ إِلَىٰ صِرَاطٍ مُسْتَقِيمٍ 2##213 كَانَ النَّاسُ أُمَّةً وَاحِدَةً فَبَعَثَ اللَّهُ النَّبِيِّينَ مُبَشِّرِينَ وَ مُنْذِرِينَ وَ أَنْزَلَ مَعَهُمُ الْكِتَابَ بِالْحَقِّ لِيَحْكُمَ بَيْنَ النَّاسِ فِيمَا اخْتَلَفُوا فِيهِ وَ مَا اخْتَلَفَ فِيهِ إِلَّا الَّذِينَ أُوتُوهُ مِنْ بَعْدِ مَا جَاءَتْهُمُ الْبَيِّنَاتُ بَغْيًا بَيْنَهُمْ فَهَدَى اللَّهُ الَّذِينَ آمَنُوا لِمَا اخْتَلَفُوا فِيهِ مِنَ الْحَقِّ بِإِذْنِهِ وَ اللَّهُ يَهْدِي مَنْ يَشَاءُ إِلَىٰ صِرَاطٍ مُسْتَقِيمٍ 22##54 وَ لِيَعْلَمَ الَّذِينَ أُوتُوا الْعِلْمَ أَنَّهُ الْحَقُّ مِنْ رَبِّكَ فَيُؤْمِنُوا بِهِ فَتُخْبِتَ لَهُ قُلُوبُهُمْ وَ إِنَّ اللَّهَ لَهَادِ الَّذِينَ آمَنُوا إِلَىٰ صِرَاطٍ مُسْتَقِيمٍ 24##46 لَقَدْ أَنْزَلْنَا آيَاتٍ مُبَيِّنَاتٍ وَ اللَّهُ يَهْدِي مَنْ يَشَاءُ إِلَىٰ صِرَاطٍ مُسْتَقِيمٍ

استخراج پرسش و پاسخ

در این تمرین هدف استخراج تعدادی پرسش و پاسخ از متن با استفاده از عبارات منظم است.
هدف استخراج زوج پرسش و پاسخ است. خروجی یک لیست از پرسش و پاسخ های متن است که هر عنصر آن لیست یک دیکشنری پایتون با کلیدهای Question وAnswer است.
دقت کنید که این ترک یک سوال باز (open problem) است و از هر ایده خوبی می‌توانید استفاده کنید. پرسش‌هایی که از متن در می‌آورید نیاز نیست محدود به حالات خاص باشند.
واضحا در این تمرین هدف استخراج پرسش‌هایی با منطق پیچیده نیست.
چند مثال در زیر آورده شده که می‌توانید از آن ها ایده بگیرید.

ورودی	خروجی

از این موضوع می‌توان نتیجه گرفت که اتهام مربوطه به او وارد است.	[
	{
	"Question": "چه نتیجه‌ای از این موضوع می‌توان گرفت؟",
	"Answer": "اتهام مربوطه به او وارد است"
	},
	{
	"Question": "از این موضوع چه نتیجه‌ای می‌توان گرفت؟",
	"Answer": "اتهام مربوطه به او وارد است"
	}
	]

حرکت بار الکتریکی باعث ایجاد میدان الکترومغناطیسی در فضا می‌شود.	[
	{
	"Question": "حرکت بار الکتریکی باعث چه چیزی می‌شود؟",
	"Answer": "ایجاد میدان الکترومغناطیسی"
	},
	{
	"Question": "چه چیزی باعث ایجاد میدان الکترومغناطیسی در فضا می‌شود؟",
	"Answer": "حرکت بار الکتریکی"
	}
	]

سرور دانشگاه پایین است چرا که دیروز مشکل سخت‌افزاری پیش آمد.	[
	{
	"Question": "به چه علت سرورهای دانشگاه پایین است؟",
	"Answer": "چرا که (چون) دیروز مشکل سخت‌افزاری پیش آمد"
	}
	]

جاذبه باعث افتادن سیب از درخت می‌شود.	[
	{
	"Question": "چه چیزی باعث افتادن سیب از درخت می‌شود؟",
	"Answer": "جاذبه"
	}
	]

در امتحان موفق شدم زیرا تمرین‌های زیادی حل کردم.	[
	{
	"Question": "چرا در امتحان موفق شدم؟",
	"Answer": "زیرا تمرین‌ها زیادی حل کردم"
	}
	]

ناشناس کردن اطلاعات شخصی متن (نام افراد، نام شرکت، آدرس، تلفن‌ها، ...)

افزایش روزافزون داده‌های دیجیتال نگرانی‌هایی همچون حفاظت از حریم شخصی افراد را نیز با خود به همراه می‌اورد. یکی از چالش‌های امروز در که در پیش پردازش متن‌ها وجود دارد گمنام‌سازی اطلاعات شخصی است. برای مثال ابزار پرسیدیو یکی از تلاش‌های صورت گرفته در این رابطه در زبان انگلیسی است. همچنین گمنام‌سازی وجه‌های دیگری نیز دارد و موجب ایجاد ساختار داده‌ای در متن می‌شود که کارهای بعدی می‌توانند از این ساختار استفاده کنند و مدل‌های عمومی‌تری بسازند.
در این تمرین قصد بر آن است که نام افراد، شرکت‌ها، شهرها، آدرس‌ها، شماره حساب و مواردی نظیر آن که به اطلاعات شخصی مربوط است تشخیص دهید و با توکن مناسب جایگزین کنید. به این نمونه نگاه کنید.
به این منظور در کتابخانه parsi.io ابزاری جهت شناسایی آدرس‌ها، شماره تلفن و ایمیل و زمان به شما داده می‌شود. بقیه موارد نظیر نام‌های افراد، شرکت‌ها و ... توسط شما باید پیاده سازی شود. به این منظور نیاز است جمع‌آوری داده در ابعاد مناسبی انجام شود تا بیشترین حالت را شامل شود.
در این تمرین می‌توانید از کراول دادگان مناسب، speach of part و recognition entity named استفاده کنید و rule هایی بنویسید که خروجی مدنظر محقق شود. برای مثال به این نمونه توجه کنید، البته دقت کنید که در این تمرین هدف استفاده بیشتر از الگوها و در کل نوشتن rule هاست. اطلاعات بیشتر از ابزار پرسیدیو رو در این لینک می‌توانید مشاهده کنید.

ورودی	خروجی
شماره تلفن من ۰۲۱۳۳۴۷۸ است و من در خیابان آبشار تهران زندگی می‌کنم.	شماره تلفن من <#phone#> است و من در <#address#> زندگی می‌کنم.
علی به شرکت گوگل رفت.	<#person_name#> به شرکت <#compny_name#> رفت.

استخراج مقدار زمان استاندارد از یک توکن زمانی

در این تمرین هدف استخراج زمان و تاریخ به فرمت استاندارد از یک توکن زمان و تاریخ است. در این تمرین باید تابع extractor value ابزار پارس‌ت‌دکس(Parstdex) را گسترش دهید تا زمان‌های معادل توکن‌های زمانی و تاریخی را به فرمت استاندارد تبدیل کند. به عنوان مثال فرض کنید جمله‌ی "امروز درست در ساعت پنج و چهل و یک دقیقه صدای گوش خراشی از چهارراه نزدیک خانه به گوش می‌رسید" را به عنوان ورودی به پارس‌ت‌دکس داده شده است. خروجی پارس‌ت‌دکس به شکل زیر خواهد بود.
ورودی:

- دیروز درست در ساعت پنج و چهل و یک دقیقه صدای گوش خراشی از چهارراه نزدیک خانه به گوش می‌رسید.

خروجی:

- دیروز (۰، ۶)
- ساعت پنج و چهل و یک دقیقه (۱۴، ۴۰)

حال فرض کنید امروز ۱۴۰۱/۰۱/۰۱ است. لازم است شما خروجی‌ای به صورت زیر تولید کنید.

- ۱۶۴۷۷۸۵۷۰۰

عدد فوق timestamp ساعت ۱۷:۴۵ روز ۱۴۰۰/۱۲/۲۹ است که می‌توانید از این سایت نیز آن را مشاهده کنید. گاهی ممکن است یک ساعت دقیق مشخص نشده باشد. به عنوان مثال:
ورودی:

- اتفاقات امروز تا دو روز دیگر می‌تواند برای مردم ایران بسیار حساس باشد.

خروجی:

- امروز تا دو روز دیگر (۸، ۲۹)

در چنین حالتی شما timestamp حال را در نظر بگیرید و یک بازه timestamp نظیر خروجی زیر را اعلام کنید.

- ۱۶۴۷۸۹۴۶۰۰_۱۶۴۷۷۲۱۸۰۰

و در نهایت گاهی بازه‌های زمانی تکرار پذیری در توکن‌های خروجی وجود دارند. به عنوان مثال:
ورودی:

- کامران هر روز هفته به مدرسه می‌رود.

خروجی:

- هر روز هفته (۷، ۱۹)

در چنین حالت از شما انتظار می‌رود خروجی‌ای با فرمت کرون‌تایم داشته باشید.

- * * * , ,

لذا خروجی شما باید در یکی از سه دسته زیر قرار گیرد.

کرون‌تایم	بازه‌ی زمانی	زمان دقیق

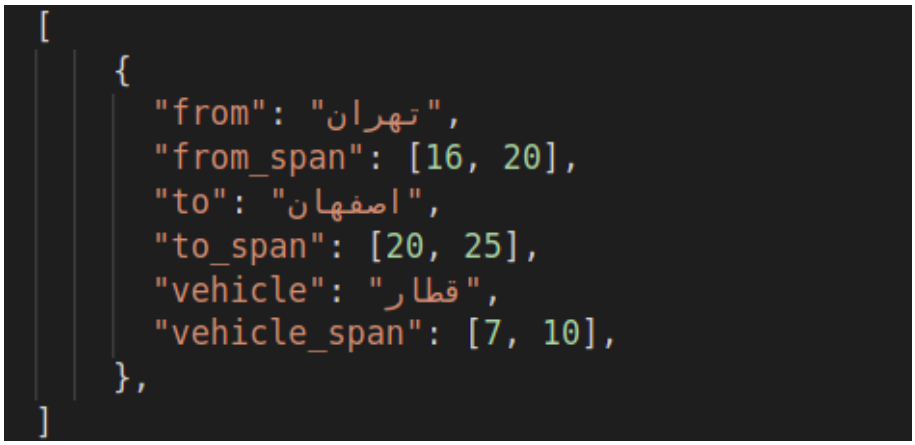
{
"type": "crontime",
"text": "token",
"span" : "(0,20)",
"value": "* * * 0 0"
}
	{
	"type": "duration",
	"text": "token",
	"span": "(0,20)",
	"value": [
	1647721800,
	1647894600
	]
	}
		{
		"type": "exact",
		"text": "token",
		"span": "(0,20)",
		"value": 1647721800
		}

تشخیص انواع جابه‌جایی وسایل نقلیه

در این بخش هدف آن است که بتوان در هر متن داده شده، وسایل نقلیه‌ی استفاده شده را استخراج کرد. همچنین در صورتی که مبدا و مقصد حرکت نیز مشخص است باید این دو مورد نیز مشخص شوند. به یاد داشته باشید که باید محدوده‌ی مبدا، مقصد و همچنین خود وسیله‌ی نقلیه مشخص شود. برای نمونه:
ورودی:

- من با قطار از تهران به اصفهان می‌روم

خروجی:



ورودی:

- چون بلیت قطار پر شده بود مجبور شدم با پرایدم به تهران بروم.

خروجی:

ورودی:

- من و خواهرم تیبا معمولا با هواپیما مسافرت می‌کنیم.

خروجی:

```
[
    {
      "from": "",
      "from_span": [-1, -1],
      "to": "",
      "to_span": [-1, -1],
      "vehicle": "هواپیما",
      "vehicle_span": [29, 35] ,
    },
]
```


ورودی:

- در حال رانندگی با خودروی تیبا هستم.

خروجی:


```
[
    {
      "from": "",
      "from_span": [-1, -1],
      "to": "",
      "to_span": [-1, -1],
      "vehicle": "تیبا",
      "vehicle_span": [27, 30] ,
    },
]
```

تشخیص جمله‌های امری از پرسشی

در این بخش هدف آن است که بتوان اولا تشخیص داد که آیا یک جمله‌ی داده‌ی شده، به صورت امری است و یا خیر؟ و در گام دوم نیز در صورتی که جمله امری است مشخص شود جمله، امری مثبت است و یا منفی؟ همچنین خود فعل امری نیز مشخص شود. چند نمونه
ورودی:

- از جلوی این خانه برو.

خروجی:

- type: امری مثبت
- verb: برو

ورودی:

- چرا گفتی که این کار را انجام نمی‌دهی؟

خروجی:

- type: پرسشی
- verb: نمی‌دهی

ورودی:

- در این گونه موارد به هیچ وجه نظر نده.

خروجی:

- type: امری منفی
- verb: نظر نده

ورودی:

- بیش از این اینجا نمان.

خروجی:

- type: امری منفی
- verb: نمان

توسعه‌ی هضم برای نرمالایز کردن و ریشه‌یابی متن‌های تاریخی

در این بخش هدف آن است که بتوانید ابزار هضم را برای متن‌های تاریخی توسعه دهید. برای نمونه، نرمال‌سازی و ریشه‌یابی که در ابزار هضم بر روی متن‌های رسمی انجام می‌شود باید بر روی متن‌های تاریخی نیز بتواند اعمال شود. در نظر گرفتن فعل‌های تاریخی، حروف اضافه‌ی مورد استفاده در گذشته و ... این توسعه می‌تواند انجام بشود. حوزه‌ی تمرکز شما می‌تواند در مورد مثلا یک قرن خاص باشد و یا دفتر شعر یک شاعر خاص را برای توسعه‌ی هضم استفاده کنید. برای نمونه می‌توانید غزلیات خافظ را برای این کار انتخاب کنید. در پایان کد شما باید یک Wrapper باشد که بتواند قابلیت‌های جدید را افزون‌بر قابلیت‌های کتابخانه‌ی هضم انجام دهد. ورودی این ماژول یک رشته‌ی متنی (جمله) است که باید بتوان دو قابلیت ریشه‌یابی و همچنین نرمال سازی را در صورت دل‌خواه فعال کرد و خروجی نیز به صورت رشته (string) بوده و با توجه به مواردی که کاربر فعال کرده است (برای نمونه کاربر ممکن است هم ریشه‌یابی و هم نرمال‌سازی را فعال کرده باشد) نمایش داده شود.

ورودی:

- درویشی را شنیدم که در آتش فاقه می سوخت و رقعه بر خرقه همی دوخت و تسکین خاطر مسکین را همی گفت

خروجی:

- ریشه‌یابی: درویشی را شنیدم که در آتش قاقه سوختن و رقعه بر خرقه دوختن و تسکین خاطر مسکین را گفتن (گلستان سعدی)

ورودی:

- حاتم طایی را گفتند: از تو بزرگ همت تر در جهان دیده ای یا شنیده ای؟ گفت: بلی! روزی چهل شتر قربان کرده بودم امرای عرب را، پس به گوشهٔ صحرایی به حاجتی برون رفته بودم، خارکنی را دیدم پشته فراهم آورده. گفتمش: به مهمانی حاتم چرا می نروی که خلقی بر سماط او گرد آمده اند؟ گفت:

  هر که نان از عمل خویش خورد منت حاتم طایی نبرد

  من او را به همت و جوانمردی از خود برتر دیدم. (گلستان سعدی)

خروجی:

- نرمال‌سازی: حاتم طایی را گفتند: از تو بزرگ همت تر در جهان دیده‌ای یا شنیده‌ای؟ گفت: بلی! روزی چهل شتر قربان کرده بودم امرای عرب را، پس به گوشهٔ صحرایی به حاجتی برون رفته بودم، خارکنی را دیدم پشته فراهم آورده. گفتمش: به مهمانی حاتم چرا می‌نروی که خلقی بر سماط او گرد آمده‌اند؟ گفت:

  هر که نان از عمل خویش خورد منت حاتم طایی نبرد

  من او را به همت و جوان‌مردی از خود برتر دیدم.

ورودی:

- ریشه‌یابی و نرمال‌سازی: مصطفی صلی اللهّ علیه و سلم باصحاب نشسته بود کافران اعتراض آغاز کردند فرمود که آخر شما همه متفّقید که در عالم یکی هست که صاحب وحی اوست وحی برو فرو میآید بر هر کسی فرو نمیآید و آنکس را علامتها و نشانها باشد. (فیه ما فیه)

خروجی:

- ریشه‌یابی و نرمال‌سازی: مصطفی صلی اللهّ علیه و سلم با اصحاب نشستن کافران اعتراض آغاز کردن فرمودن که آخر شما همه متفق بودن که در عالم یکی بودن که صاحب وحی اوست وحی برو فرود آمدن بر هر کسی فرود آمدن و آن‌کس را علامت‌ها و نشانه‌ها بودن.

استخراج نام و لقب‌های ائمه‌ی معصومین و عبارت‌هایی دعایی از متن‌های حدیث

در این بخش با در اختیار داشتن مجموعه‌ای از حدیث‌های مربوط به ائمه، شما باید بتوانید نام و لقب‌های مربوط به ائمه را از آن‌ها استخراج کنید. همچنین عبارت‌های دعایی که معمولا پس از نام آن‌ها می‌آید را نیز باید استخراج کنید. منظور از عبارت‌های دعایی،عبارت‌هایی مانند «علیه‌السلام»، متن صلی الله علیه و آله و سلم » و ... است که بیش‌تر وقت‌ها پس از نام ایشان آورده می‌شود. خروجی شما باید شامل آرایه‌ای باشد که از ابتدای جمله‌ی ورودی هر نامی از ائمه مشاهده می‌کند را در نظر گرفته و محدوده‌ی آن اسم و یا لقب و یا عبارت دعایی را به ترتیب در این لیست خروجی بنویسد. برای دریافت داده‌ها از این لینک استفاده کنید. دریافت داده‌ها در خروجی شما باید نوع خروجی نیز مشخص شود که آیا از جنس نام ائمه است و یا از جنس عبارت‌های دعایی برای نمونه:

ورودی:

- كَان رَسُولُ اَللَّهِ صَلَّي اَللَّهُ عَلَيْهِ وَ اٰلِهِ صُومُ شَعْبَانَ وَ رَمَضَانَ يَصِلُهُمَا وَ يَقُولُ هُمَا شَهْرُ

خروجی:

```
[
    {
      "type": "name",  #نام ائمه
      "text": "رَسُولُ اَللَّهِ",
      "span": [5, 20],

    },
    {
      "type": "salute",  #عبارت دعایی
      "text": "صَلَّي اَللَّهُ عَلَيْهِ وَ اٰلِهِ",
      "span": [22, 56] ,
    },
]
```

استخراج اطلاعات از فرهنگ لغت لری

هدف این تمرین تبدیل داده‌ی خام با فرمت این لینک به داده‌ی ساختاریافته است. این لینک تعدادی واژه از یک لغت‌نامه لری را در بر دارد که باید اجزای مختلف معنی هر کدام را استخراج و بصورت ساختارمندی ذخیره کنید. البته کد شما باید قابل تعمیم به باقی واژه‌های لغت‌نامه نیز باشد. یکی از کاربردهای داده‌ی حاصل در ابعاد بزرگ‌تر، می‌تواند ترجمه‌ی ماشینی از زبان لری به زبان فارسی باشد. هر واژه‌ی مجزایی که در لینک مشاهده می‌کنید، بصورت بولد نوشته شده و تلفظ و معانی آن به همراه اصطلاحات مرتبط (در صورت وجود) ذکر شده‌اند. لازم است برای هر کلمه موارد زیر استخراج شوند.

- خود کلمه
- تلفظ کلمه
- معنی ذکر شده در مقابل کلمه (در صورتی که چندین معنی برای کلمه ذکر شده باشد، باید یک لیست متشکل از معانی به طور مجزا برای مدخل معنی ذخیره شود).
- لیستی از اصطلاحات مرتبط با کلمه به همراه معنی و تلفظ

برخی واژه‌ها ممکن است توضیحات اضافی ذیل خود داشته باشند که باید از آن‌ها صرف نظر شود. برای مثال کلمه‌ی «هقیده» را ببینید. یک نمونه از خروجی قابل قبول برای این کلمه بصورت زیر است:

```
d = {
    'vocab': 'هقیده',
    'pronunciation': 'haqiδa',
    'meaning': ['عقیده، نظر یا اندیشه‌ای که انسان به آن باور دارد.',
                'اندیشه‌ای که مردم جامعه آن را پذیرفته‌اند.'],
    'phrases': [
                {
                    'هقیده داشتن': (
                        'haqiδa dāštan',
                        ['عقیده داشتن', 'باور داشتن', 'ایمان داشتن']
                    ),
                    'هقیده زدن': (
                        'haqiδa zaδan',
                        ['عقیده داشتن', 'باور داشتن']
                    ),
                    'هقیدم اِزنه': (
                        'haqiδam izane',
                        ['عقیده‌ام میزند', 'بر این باورم']
                    ),
                }
    ],
}
```

می‌توانید از هر ساختار مشابهی برای ارائه‌ی خروجی استفاده کنید؛ تنها شرط لازم ارائه‌ی اطلاعات مربوط به هر کلمه بصورت مجزا و ساختارمند است تا بتوان در پروژه‌های مختلف آن را بعنوان ورودی مسئله به کار گرفت. ارزیابی تمرین با استفاده از واژه‌های متفاوتی از همین لغت‌نامه انجام می‌شود و ورودی آدرس نسبی یک فایل docx شامل واژه‌های تست است. کد شما باید متن خام را از فایل docx استخراج کند و اطلاعات مورد انتظار را به ازای کل کلمات فایل خروجی دهد. هر چه اطلاعات (بخصوص بخش موازی) بیشتر و مفیدتری از کلمات استخراج شوند، نمره‌ی بیشتری به پاسخ تعلق می‌گیرد. می‌توانید از هر تکنیکی (اعم از عبارات منظم) و هر کتابخانه‌ی دلخواهی برای تولید خروجی موردنظر استفاده کنید.