



## پردازش زبان طبیعی

نیم سال دوم ۱۴۰۰-۰۱

استاد: احسان الدین عسگری

### تمرین اول

### عبارات منظم

مهلت ارسال: ۲۴ فروردین

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص شده است.
- در تمرین هایی 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 می کنند، دارد. بنابراین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین ها بعد از ۳ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ های ارسال شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- توجه داشته باشید که نوت بوک های 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 نوت بوک وجود داشته باشد.
- تمامی فایل های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW آپلود 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 نوت بوک و مستندات قرار دهید.
- در پروژه های 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 آن ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده اید توضیح دهید. بلکه باید به شکل کلی ایده تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی های مساله را در گزارش بیاورید و براساس آن رفتار برنامه تان 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 کوثرای 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 عموماً ابزاری برای آنها تولید نشده است. در این تمرین در بسیاری از بخشها می توانید از حاصل کار عزیزان ترم گذشته که با زحمات تدریساران درس در قالب کتابخانه [parsio.io](http://parsio.io) ایجاد شده بهره ببرید. به امید خدا در ترم های آینده حاصل جمع زحمات شما عزیزان در قالب محصولات متن باز (البته با ذکر نام خودتان) در اختیار دیگر دانشجویان و بلکه جامعه ایرانی قرار می گیرد تا در اثر این تلاشها محصولاتی ارزشمند برای پردازش متن های فارسی و بلکه زبانهای ایرانی و فراتر از آن داشته باشیم. می توانید به این کتابخانه از طریق [این لینک](#) دسترسی داشته باشید

لطفاً 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 روش های تشخیص به وسیله قواعد با تمرکز بر عبارات منظم و آنچه در مازول ابتدایی درس آموخته اید، مساله های 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 پیاده‌سازی 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 بازه‌ها span 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، دقیقاً 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 آسیب‌پذیر جامعه می‌شود.	در هنگام وقوع بلایای طبیعی، بیشترین خسارت متوجه قشر آسیب‌پذیر جامعه می‌شود.
آن‌ها چهاردهم مارس (مارچ) را بعنوان روز جهانی ی عدد پی نام‌گذاری کرده‌اند.	آن‌ها چهاردهم مارس (مارچ) را به‌عنوان روز جهانی عدد پی نام‌گذاری کرده‌اند.
لاک‌پشت با هوش از همهٔ آنها زیرکانه‌تر 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 خروجی 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 مواردی 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.)	<pre>[   {     "type": "گفت‌وگو و مذاکرات و توافق",     "text": "گفتگوهای صلح اوکراین",     "span": [0,20],     "place": "",     "time": "فردا"   } ]</pre>
دو هفته از استعفای نخست‌وزیر بریتانیا می‌گذرد.	<pre>[   {     "type": "قراردادهای رسمی",     "text": "امضای قرارداد ۲۲۳۱",     "span": [22,51],     "place": "",     "time": ""   } ]</pre>
اکثر کشورهای دنیا با امضای قرارداد ۲۲۳۱ سازمان ملل موافق هستند.	<pre>[   {     "type": "قراردادهای رسمی",     "text": "امضای قرارداد ۲۲۳۱",     "span": [22,51],     "place": "",     "time": ""   } ]</pre>

<pre>[ {   "type": "جنگ",   "text": "حمله سنگین ارتش عراق به ایران",   "span": [21,50],   "place": "",   "time": "سال ۱۳۵۹" } ]</pre>	<p>پس از انقلاب اسلامی، حمله سنگین ارتش عراق به ایران در سال ۱۳۵۹ آغاز گردید.</p>
<pre>[ {   "type": "تغییر قیمت",   "text": "افزایش بهای طلا",   "span": [11,26],   "place": "صرافی‌های تهران",   "time": "امروز" } ]</pre>	<p>امروز شاهد افزایش بهای طلا در صرافی‌های تهران بودیم.</p>
<pre>[ {   "type": "مرگ",   "text": "درگذشت امیرکبیر",   "span": [27,42],   "place": "",   "time": "بیست دی ماه" } ]</pre>	<p>امروز بیست دی ماه مصادف با درگذشت امیرکبیر است.</p>
<pre>[ {   "type": "بیماری",   "text": "شیوع کوید ۱۹",   "span": [6,18],   "place": "جهان",   "time": "" } ]</pre>	<p>پس از شیوع کوید ۱۹ بسیاری در جهان جان خود را از دست دادند.</p>
<pre>[ {   "type": "واردات و صادرات",   "text": "افزایش واردات میوه و تره‌بار",   "span": [33,61],   "place": "",   "time": "سال آینده" } ]</pre>	<p>به علت خشکسالی در سال آینده شاهد افزایش واردات میوه و تره‌بار خواهیم بود.</p>
<pre>[ {   "type": "قرار ملاقات",   "text": "دیدار پوتین و مکرون",   "span": [0,19],   "place": "کاخ کرملین",   "time": "امروز عصر" } ]</pre>	<p>دیدار پوتین و مکرون امروز عصر در کاخ کرملین برگزار خواهد شد.</p>

<pre>[   {     "type": "تحریم و رفع تحریم",     "text": "اجرای تحریم‌های جدیدی علیه ایران",     "span": [27,59],     "place": "",     "time": ""   } ]</pre>	<p>آزمایش موشک کروز جدید باعث اجرای تحریم‌های جدیدی علیه ایران شد.</p>
<pre>[   {     "type": "صعود و سقوط و حذف",     "text": "تیم اتلتیکو مادرید به مرحله پایانی مسابقات صعود کرد",     "span": [29,80],     "place": "مادرید",     "time": "شب گذشته"   } ]</pre>	<p>در دیدار شب گذشته در مادرید تیم اتلتیکو مادرید به مرحله پایانی مسابقات صعود کرد.</p>
<pre>[   {     "type": "قهرمانی و نایب قهرمانی",     "text": "بازیکن هلندی تبار عنوان نایب قهرمانی را کسب کرد",     "span": [0,47],     "place": "",     "time": ""   } ]</pre>	<p>بازیکن هلندی تبار عنوان نایب قهرمانی را کسب کرد.</p>
<pre>[   {     "type": "کسب مدال",     "text": "کسب مدال نقره",     "span": [51,64],     "place": "قزاقستان",     "time": ""   } ]</pre>	<p>در جام جهانی والیبال در قزاقستان تیم ایران موفق به کسب مدال نقره شد.</p>

---

## 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 بنویسید 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).



خروجی	ورودی
<pre>[   {     "type": "وزن",     "amount": 3.5,     "unit": "کیلوگرم",     "item": "آرد",     "marker": "کیلوگرم آرد ۳.۵",     "span": [4,19]   },   {     "type": "توان",     "amount": 0.85,     "unit": "وات",     "item": "",     "marker": "هشتاد و پنج صدم وات",     "span": [40,59]   } ]</pre>	<p>علی ۳.۵ کیلوگرم آرد خرید و باتری خود را هشتاد و پنج صدم وات شارژ کرد.</p>
<pre>[   {     "type": "سرعت",     "amount": 15,     "unit": "کیلومتر بر ثانیه",     "item": "",     "marker": "۱۵ km/s تندی",     "span": [13,25]   } ]</pre>	<p>شهاب سنگی به تندی ۱۵ km/s وارد جو زمین شد.</p>
<pre>[   {     "type": "سرعت",     "amount": "",     "unit": "",     "item": "",     "marker": "سرعت زیاد",     "span": [12,21]   } ]</pre>	<p>یک خودرو با سرعت زیاد از ما سبقت گرفت.</p>

---

## 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 موجودیت‌های استخراج شده است. به منظور جمع‌آوری عنوان مارک‌های بورسی و کلمات کلیدی می‌توانید از سایت‌هایی نظیر سهام‌یاب علی‌الخصوص [بخش توئیتهای و بخش گفت‌وگوی ره‌آورد ۳۶۵](#) و [سایت اصلی تالار بورس](#) استفاده کنید. در گزارش خود دسته‌بندی‌های استخراج شده (type) و مثال‌هایی 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 آن‌ها توجه کنید موارد زیر هستند:

- نام شرکت‌های 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 رانتهی
- ...

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

خروجی	ورودی
<pre>[   {     "type": "نماد",     "marker": "برکت",     "span": [1,5]   },   {     "type": "واقعه",     "marker": "افشای ب",     "span": [11,18]   },   {     "type": "واقعه",     "marker": "سه درصد مثبت",     "span": [31,43]   } ]</pre>	<p>برکت همین افشای ب باعث شد سهم سه درصد مثبت شود. بخاطر همین میگم پیگیر باشید.</p>
<pre>[   {     "type": "نماد",     "marker": "فزر",     "span": [19,22]   },   {     "type": "واقعه",     "marker": "واحد تاثیر مثبت 2.15",     "span": [26,46]   } ]</pre>	<p>جریان آغاز معاملات فزر با ۱۵.۲ واحد تاثیر مثبت بر روند صعودی بازار فرابورس اثر گذار بود.</p>
<pre>[   {     "type": "نماد",     "marker": "دیران",     "span": [58,63]   },   {     "type": "گزارش",     "marker": "فعالیت ماهانه",     "span": [6,19]   } ]</pre>	<p>گزارش فعالیت ماهانه دوره ۱ ماهه منتهی به ۱۴۰۰/۰۹/۳۰ برای دیران منتشر شد.</p>

## 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",
  "recipe": "دستور غذا",
  "span_ingredients": "لیستی از مکان مواد اولیه در متن بر اساس کاراکتر",
  "span_recipe": "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
<p>ورودی نمونه:</p> <p>مواد لازم:</p> <p>اردک درسته ۱ عدد</p> <p>پیاز ۱ عدد</p> <p>انار ۳ قاشق غذاخوری</p> <p>رب گوجه ۱ قاشق غذاخوری</p> <p>رب نارنج ۲ ق غ</p> <p>نمک و فلفل و ادویه</p> <p>سبزی معطر</p> <p>طرز تهیه:</p> <p>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 را با همراه نصف لیوان آب می‌گذاریم تا ملایم بپزد.</p>	<pre>{   "Ingredients": [     "اردک درسته",     "پیاز",     "انار",     "رب گوجه",     "رب نارنج",     "نمک و فلفل و ادویه",     "سبزی معطر"   ],   "quantity": [     "۱ عدد",     "۱ عدد",     "۳ قاشق غذاخوری",     "۱ قاشق غذاخوری",     "۲ ق غ",     "",     ""   ],   "recipe": "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 را با همراه نصف لیوان آب می‌گذاریم تا ملایم بپزد.",   "span_ingredients": [     [11,27],[28,38],[39,58],[59,81],[82,96],[97,115],[116,125]   ],   "span_recipe": [136,567],   "time": 1 }</pre>

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

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.

خروجی	ورودی
<p>نمایش شکل های مختلف هر کلمه:</p> <p>اهْدُنَا (ظاهر کلمه) - هَدَيْ (لمای کلمه) - هدى (ریشه کلمه)</p> <p>الصِّرَاطَ (ظاهر کلمه) - صِرْطَ (لمای کلمه) - صرط (ریشه کلمه)</p>	<p>کلمات : صراط، هَدْي</p> <p>فاصله: سه کلمه</p>
<p>آیات مورد نظر:</p> <p>142##2 سَيَقُولُ السُّفَهَاءُ مِنَ النَّاسِ مَا وَلَّاهُمْ عَنْ قِبْلَتِهِمُ الَّتِي كَانُوا عَلَيْهَا قُلْ لِلَّهِ الْمَشْرِقُ وَالْمَغْرِبُ يَهْدِي مَنْ يَشَاءُ إِلَى صِرَاطٍ مُسْتَقِيمٍ</p> <p>213##2 كَانَ النَّاسُ أُمَّةً وَاحِدَةً فَبَعَثَ اللَّهُ النَّبِيِّينَ مُبَشِّرِينَ وَمُنْذِرِينَ وَأَنْزَلَ مَعَهُمُ الْكِتَابَ بِالْحَقِّ لِيَحْكُمَ بَيْنَ النَّاسِ فِي مَا اخْتَلَفُوا فِيهِ وَمَا اخْتَلَفَ فِيهِ إِلَّا الَّذِينَ أُوتُوهُ مِنْ بَعْدِ مَا جَاءَتْهُمْ الْبَيِّنَاتُ بَغْيًا بَيْنَهُمْ فَهَدَى اللَّهُ الَّذِينَ آمَنُوا لِمَا اخْتَلَفُوا فِيهِ مِنَ الْحَقِّ بِإِذْنِهِ وَاللَّهُ يَهْدِي مَنْ يَشَاءُ إِلَى صِرَاطٍ مُسْتَقِيمٍ</p> <p>54##22 وَلِيَعْلَمَ الَّذِينَ أُوتُوا الْعِلْمَ أَنَّهُ الْحَقُّ مِنْ رَبِّكَ فَيُؤْمِنُوا بِهِ فَتُخْبِتَ لَهُ قُلُوبُهُمْ وَإِنَّ اللَّهَ لَهَادٍ الَّذِينَ آمَنُوا إِلَى صِرَاطٍ مُسْتَقِيمٍ</p> <p>46##24 لَقَدْ أَنْزَلْنَا آيَاتٍ مُبَيِّنَاتٍ وَاللَّهُ يَهْدِي مَنْ يَشَاءُ إِلَى صِرَاطٍ مُسْتَقِيمٍ</p>	<p>انتخاب یک فرم از هر کلمه:</p> <p>هدى (لمای کلمه) - صرط (ریشه کلمه)</p>

## 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 و Answer است. دقت کنید که این ترک یک سوال باز (open problem) است و از هر ایده خوبی می توانید استفاده کنید. پرسش هایی که از متن در می آورید نیاز نیست محدود به حالات خاص باشند. واضحا در این تمرین هدف استخراج پرسش هایی با منطق پیچیده نیست. چند مثال در زیر آورده شده که می توانید از آن ها ایده بگیرید.

ورودی	خروجی
از این موضوع می توان نتیجه گرفت که اتهام مربوطه به او وارد است.	<pre>[   {     "Question": "چه نتیجه ای از این موضوع می توان گرفت؟",     "Answer": "اتهام مربوطه به او وارد است"   },   {     "Question": "از این موضوع چه نتیجه ای می توان گرفت؟",     "Answer": "اتهام مربوطه به او وارد است"   } ]</pre>
حرکت بار الکتریکی باعث ایجاد میدان الکترومغناطیسی در فضا می شود.	<pre>[   {     "Question": "حرکت بار الکتریکی باعث چه چیزی می شود؟",     "Answer": "ایجاد میدان الکترومغناطیسی"   },   {     "Question": "چه چیزی باعث ایجاد میدان الکترومغناطیسی در فضا می شود؟",     "Answer": "حرکت بار الکتریکی"   } ]</pre>
سرور دانشگاه پایین است چرا که دیروز مشکل سخت افزاری پیش آمد.	<pre>[   {     "Question": "به چه علت سرورهای دانشگاه پایین است؟",     "Answer": "چرا که (چون) دیروز مشکل سخت افزاری پیش آمد"   } ]</pre>
جاذبه باعث افتادن سیب از درخت می شود.	<pre>[   {     "Question": "چه چیزی باعث افتادن سیب از درخت می شود؟",     "Answer": "جاذبه"   } ]</pre>
در امتحان موفق شدم زیرا تمرین های زیادی حل کردم.	<pre>[   {     "Question": "چرا در امتحان موفق شدم؟",     "Answer": "زیرا تمرین ها زیادی حل کردم"   } ]</pre>

## 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 می‌آورد. یکی از چالش‌های امروز در که در پیش پردازش متن‌ها وجود دارد گمنام‌سازی اطلاعات شخصی است. برای مثال ابزار پرسیدو یکی از تلاش‌های صورت گرفته در این رابطه در زبان انگلیسی است. همچنین گمنام‌سازی وجهه‌های 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 [parsio.io](http://parsio.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، speech 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 پرسیدو رو در این لینک می‌توانید مشاهده کنید.

ورودی	خروجی
شماره تلفن من ۰۲۱۳۳۴۷۸ است و من در خیابان آبخار 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.

• 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:

• 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.

• ۱۶۴۷۷۲۱۸۰۰ - ۱۶۴۷۸۹۴۶۰۰

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

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