بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

پردازش زبان طبیعی

نیم‌سال اول ۰۲-۰۱

استاد: احسان‌الدین عسگری

تمرین دوم | عبارات منظم | مهلت ارسال: ۴ آذر

- 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 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'
```

استخراج روابط علت و معلولی

در این ترک هدف استخراج روابط علت و معلولی از جمله‌ی ورودی است. اینگونه روابط می‌تواند در استخراج پرسش و پاسخ از متن، بدست آوردن روابط علت و معلولی، استخراج اطلاعات و نیازمندی‌های دیگری استفاده بشوند. در این ماژول شما باید به کمک عبارت‌های منظم تشخیص دهید که آیا جمله‌ی ورودی شامل علت و معلول است یا خیر. این تمرین ادامه‌ی مسیر ماژول $cause-effect-extraction$ موجود در کتابخوانه‌ی $parsi.io$ است. از شما انتظار می‌رود که با گسترش دادن این ماژول سعی کنید روابط علت و معلولی پیچیده‌تری را از متن بدست بیاورید. ورودی شما یک رشته متن خواهد بود و شما باید در خروجی مشخص کنید که آیا در این متن رابطه‌ی علت و معلولی وجود دارد یا خیر، و در صورت وجود نشانگر مشخص‌کننده‌ی آن، علت و معلول را بدست بیاورید. ورودی شما یک رشته متن است و خروجی آن یک دیکشنری زبان پایتون است. در جدول زیرچند مثال برای درک بهتر آورده شده است.

ورودی	خروجی
چون نمی‌خواستم اون چیزی از ماجرا بفهمه، مجبور به تظاهر شدم.	`{"flag": true, "cause": "نمی‌خواستم اون چیزی از ماجرا بفهمه", "effect": "مجبور به تظاهر شدم", "marker": "چون"}`
حرکت بار الکتریکی باعث ایجاد میدان الکترومغناطیسی در فضا می‌شود.	`{"flag": true, "cause":"حرکت بار الکتریکی", "effect": "ایجاد میدان الکترومغناطیسی در فضا می‌شود", "marker": "باعث"}`
اسب پرواز کرد.	`{"flag": false, "cause":"", "effect": "", "marker": ""}`

استخراج اطلاعات رزومه

کارفرمایان برای هر فرصت شغلی صدها رزومه دریافت می‌کنند. به منظور ساده‌تر شدن فرآیند استخدام و سرعت بخشیدن به آن، نرم‌افزارهای کامپیوتری با عنوان Application Tracking Software توسعه یافته‌اند. نرم‌افزارهای ATS از چندین فایل رزومه ورودی، اطلاعات لازم را استخراج کرده و افراد را بر این اساس رتبه‌بندی می‌کنند. هدف از این تمرین، توسعه‌ی یک نرم‌افزار ATS به صورت rule-based است. ورودی برنامه‌ی شما چندین رزومه مختلف فارسی و خروجی مورد انتظار، اطلاعات این رزومه‌ها در قالب یک فایل csv است. به منظور تبدیل فایل PDF به متن، می‌توانید از کتابخانه‌ی PyPDF یا کتابخانه‌های نظیر آن استفاده نمائید.

```python
from PyPDF2 import PdfReader

reader = PdfReader("example.pdf")
page = reader.pages[0]
print(page.extract_text())
```

اجزای اصلی یک رزومه می‌تواند شامل موارد زیر باشد:

- اطلاعات فردی
  - نام
  - نام خانوادگی
  - عنوان شغلی
  - استان محل سکونت
  - شهر محل سکونت
  - تاریخ تولد
- اطلاعات شغلی
  - وضعیت اشتغال
  - نوع شغل موردنظر
  - حقوق موردانتظار
  - تکنولوژی‌های موردعلاقه برای کار
- سوابق تحصیلی
- سوابق شغلی
- مهارت‌ها
- پروژه‌ها
- دستاوردها

استخراج ویژگی کالا از نظرات

در این ترک هدف استخراج ویژگی‌های یک کالا از صفحه محصول یا نظرات مشتریان است. سامانه‌های فروش کالا معمولا از متن یا ویژگی‌هایی که توسط فروشنده کالا تهیه می‌شود به عنوان اطلاعات برای پیشنهاد دادن آن به کاربران استفاده می‌کند هرچند که اطلاعات کالاها ممکن است تا حدودی نادقیق و غیرواقعی باشند. در این ترک قصد داریم برای بهبود سامانه‌های پیشنهاددهنده[1] اطلاعات یک کالا را از نظرات مشتریان نیز استخراج کنیم. شما باید سامانه‌ای برای استخراج ویژگی‌های کالا از نظرات تهیه کنید.

اینکه چه ویژگی‌هایی را استخراج کنید کاملا بر عهده خودتان است و می‌توانید مواردی مانند قیمت، کارکرد، اندازه و غیره را در نظر بگیرید. البته واضح است که برای کارکرد بهتر کدتان بهتر است روی چند ویژگی (حداقل ۶ ویژگی) تمرکز کنید و تمام قواعد[2] مرتبط با آن‌ها را شناسایی کنید.

ورودی کد شما یک نظر و خروجی آن یک دیکشنری زبان پایتون است. هر کلید دیکشنری یک ویژگی و مقدار متناظر با کلید مقدار ویژگی است. در جدول زیر چند مثال برای درک بهتر آورده شده است.

ورودی	خروجی
برای ماکتی به این کوچیکی قیمت خیلی گرونه همچنین بعد از فقط ۲ ماه یکی از پایه‌هاش شکست.	`{ "قیمت": "زیاد", "اندازه": "کوچک", "زمان کارکرد": "کم" }`
زمان خرید قیمتش کمی زیاد بود ولی واقعا خوب کار می‌کنه و کارخونه هم برای خرابی به خوبی پشتیبانی می‌کنه فقط کاش چندتا رنگ ازش وجود داشت.	`{ "گارانتی": "خوب", "رنگ": "تک", "قیمت": "نسبتا زیاد", "کارکرد": "خوب" }`

[1] recommendation system
[2] pattern

استخراج زمان جمله و فعل

هدف از این ترک استخراج زمان فعل همراه با مشخصات آن است. غالب افعال دارای زمان هستند و می‌توان آن‌ها را با توجه به اینکه در چه زمانی رخ داده یا می‌دهند. تقسیم کرد. افعال به سه دسته کلی زیر تقسیم می‌شود.

- فعل‌هایی که به زمان **گذشته** اشاره می‌کنند.
- فعل‌هایی که به زمان **حال** اشاره می‌کنند.
- فعل‌هایی که به زمان **آینده** اشاره می‌کنند.

در این ترک انتظار می‌رود اطلاعات فعل‌های یک جمله نظیر **زمان** فعل، **نوع** فعل، **بن** فعل و همچنین **شخص** فعل استخراج و مشخص شود.

- انواع فعل‌های گذشته: گذشته ساده، گذشته بعید، گذشته نقلی، گذشته استمراری، گذشته استمراری، گذشته مستمر
- انواع فعل‌های حال: حال اخباری، حال التزامی، حال مستمر

به نمونه زیر توجه کنید.

ورودی	خروجی

کتاب را به کتابخانه می‌برم.	{
	"زمان": "حال",
	"نوع": "اخباری",
	"بن فعل": "بر",
	"شخص": "اول شخص مفرد"
	}

به عنوان نمره امتیازی می‌توانید استخراج زمان جمله و فعل را برای زبان فارسی کهن نیز پیاده‌سازی کنید. در صورت تمایل به این موضوع کتاب دستور تاریخی در اختیارتان قرار می‌گیرد.

شناسایی تهدید یا خشونت در متن

با گسترش شبکه‌های اجتماعی، تشخیص پست‌ها و نظرات تهدید آمیز مورد توجه قرار گرفته است. دادگان زیادی در زبان انگلیسی برای متون تهدید آمیز ساخته شده است در زبان فارسی نیز کارهایی در این حوزه انجام شده است که می‌توان به لینک اشاره کرد. هدف از این ترک گرفتن یک متن فارسی و پیدا کردن جملات تهدید آمیز با استفاده از عبارات منظم است تا بتوان با روش‌های قاعده‌مند مجموعه دادگانی جدید برای کارهای آینده در این حوزه درست کرد. برای بررسی خروجی‌های خود می‌توانید از مجموعه دادگان توئیتر فارسی در این لینک استفاده کنید. در این ترک تهدیدها به دو دسته فیزیکی (به معنای آسیب فیزیکی رساندن) و غیرفیزیکی (به معنای آسیب غیر فیزیکی مانند روانی، مالی و غیره) تقسیم می‌شوند که باید نوع آن‌ها را نیز برای هر تهدید شناسایی کنید. در جدول زیر چند نمونه از ورودی‌ها و خروجی‌ها آورده شده است.

ورودی	خروجی

بوریس جانسون وعده داده است تا بریتانیا را در روز ۳۱ اکتبر با توافق یا بدون توافق از اتحادیه اروپا خارج کند. این در شرایطی است که برخی از نمایندگان محافظه کار تهدید کرده‌اند که برای جلوگیری از برگزیت بدون توافق، علیه وی رای خواهند داد.	{
	"flag": true,
	"type": "not physical",
	"span": [109, 234]
	}

ساعت کاری شرکت ما از هفت صبح شروع می‌شود هرکسی که یک روز تاخیر داشته باشد اخراج می‌شود.	{
	"flag": true,
	"type": "not physical",
	"span": [42, 87]
	}

او مخالفان خود را تهدید به مرگ کرد	{
	"flag": true,
	"type": "physical",
	"span": [1, 35]
	}

آسمان امروز آبی می‌باشد.	{
	"flag": false,
	}

اگر جرئت داری فردا بیا میدون امام حسین تا بفهمی دنیا دست کیه	{
	"flag": true,
	"type": "physical",
	"span": [1, 61]
	}

استخراج مکان و پارس آن

هدف از این تمرین استخراج آدرس از متن، استانداردسازی و استخراج اطلاعات از آن است. به منظور انجام این تمرین مراحل زیر باید طی شوند.

۱. استخراج متن آدرس: ابزار AddressExtraction کتابخانه Parsio را بررسی کنید و در صورت نیاز به منظور رفع کاستی‌های استخراج متن آدرس از متن آن را بهبود دهید.

۲. نرمال سازی آدرس: آدرس‌ها معمولا به یک شیوه نوشته نمی‌شوند. به عنوان مثال گاهی بجای پلاک ۵ از پ ۵ استفاده می‌شود. در این بخش باید آدرس‌ها را به شیوه استاندارد، تبدیل کنید.

۳. تعیین نوع آدرس (type) : یک آدرس لزوما یک آدرس جغرافیایی مشخص نیست برای مثال لینک یک جلسه در گوگل میت و یا سامانه‌ی اسکای‌روم نیز یک آدرس مجازی حساب می‌شود. در این بخش باید آدرس‌ها را به نوع‌های متفاوت (به عنوان مثال آدرس ادارات، رستوران، مجازی و غیره) دسته‌بندی کنید.

۴. تبدیل به اطلاعات (coordinate-url) : در این بخش باید آدرس‌های فیزیکی را به coordianate جغرافیایی و آدرس‌های مجازی را به url تبدیل کنید.

۵. استخراج اطلاعات راهنما (notes) : یک آدرس ممکن است جزئیاتی نظیر رنگ در یک ساختمان و یا نام قدیم یک کوچه و غیره داشته باشد، که در مسیریابی کمک کند. این جزئیات به صورت استاندارد در آدرس ذکر نمی‌شوند و در این بخش باید این اطلاعات را استخراج و به خروجی اضافه کنید.

خروجی با فرض اینکه ممکن است در یک متن چندین آدرس وجود داشته باشد آرایه‌ای از دیکشنری‌ها خواهد بود.

پیش از شروع تمرین حتما لینک‌های زیر را بررسی کنید.

لینک map street open

لینک یک آموزش مفید

لینک کتابخانه‌ی پایتون geocoder برای تبدیل آدرس استاندارد به عرض و طول جغرافیایی

گسترش شناسایی وقایع از متن (زیر مجموعه)

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 : " تقدیر از برگزیدگان جشنواره ی فیلم فجر، دیشب در برج میلاد برگزار شد."
خروجی:

- نوع واقعه (type) : جشنواره ها و برگزیدگان
- متن واقعه (text) : تقدیر از برگزیدگان جشنواره ی فیلم فجر
- بازه واقعه : (۰،۳۷)
- زمان (time) : دیشب
- مکان (place) : برج میلاد

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.

لیستی از انواع وقایع در زیر آمده است.

- هماهنگی زمانی و مکانی
  - قرار ملاقات
  - ملاقات های رسمی
- وقایع فرهنگی - هنری
  - نمایشگاه ها و مکان ها ی فرهنگی
  - جشنواره ها و برگزیدگان

- صنایع دستی و هنر های بومی
- وقایع تاریخی
- جنگ و صلح
- قرارداد ها و عهد نامه ها
- ظهور و سقوط
- نصب و خلع

ورودی	خروجی
امروز که هم رو ندیدیم اما فردا ساعت ۷ صبح دم دانشگاه میبینمت.	[{"type":"قرار ملاقات","text":"ساعت ۷ صبح دم دانشگاه میبینمت","span":[26,60],"place":"دانشگاه","time":" فردا ساعت ۷ صبح"}]
دیدار پوتین و مکرون امروز عصر در کاخ کرملین برگزار خواهد شد.	[{"type":"ملاقات های رسمی","text":"دیدار پوتین و مکرون","span":[0,19],"place":"کاخ کرملین","time":" امروز عصر"}]
نمایشگاه بین المللی کتاب، از فردا به مدت یک ماه میزبان علاقه مندان به کتاب خواهد بود.	[{"type":"نمایشگاه ها و مکان ها ی فرهنگی","text":"نمایشگاه بین المللی کتاب","span":[0,25],"place":"null","time":"از فردا به مدت یک ماه"}]
تمدید دوباره ی تئاتر شکوفه های گیلاس در سالن اصلی تئاتر شهر.	[{"type":"نمایشگاه ها و مکان ها ی فرهنگی","text":"تمدید دوباره ی تئاتر شکوفه های گیلاس","span":[0,36],"place":"سالن اصلی تئاتر شهر","time":"null"}]
امروزه رشد تعداد کارگاه های محلی گلیم بافی را در کاشان شاهد هستیم.	[{"type":"صنایع دستی و هنر های بومی","text":"رشد تعداد کارگاه های محلی گلیم بافی","span":[7,42],"place":"کاشان","time":"null"}]

[
{
"type":"جنگ و صلح",
"text":"درگیری های بین روسیه و اوکراین",
"span":[28,59],
"place":["روسیه","اوکراین"],
"time":"null"
}
]
	تا کنون بیش از ۱۳۰ شهروند در درگیری های بین روسیه و اوکراین کشته شده اند.

[
{
"type":"قرارداد ها و عهد نامه ها",
"text":"عهدنامه ی گلستان",
"span":[0,16],
"place":"گلستان",
"time":"۳ آبان ۱۱۹۲"
}
]
	عهدنامه ی گلستان در تاریخ ۳ آبان ۱۱۹۲ خورشیدی در پی جنگ‌های ایران و روسیه در دوره قاجار بین این دو کشور در روستای گلستان امضا شد.

[
{
"type":"ظهور و سقوط",
"text":"سقوط امپراتوری بیزانس",
"span":[104,125],
"place":"قسطنطنیه",
"time":"۲۹ سال مه ۱۴۵۲"
}
]
	تصرف شهر قسطنطنیه توسط ترکان مسلمان عثمانی به رهبری سلطان محمد دوم در روز ۲۹ مه سال ۱۴۵۳ میلادی منجر به سقوط امپراتوری بیزانس شد.

استخراج اجزای سخن و ریشه‌ی کلمات برای زبان‌های ایرانی

در این ترک هدف استخراج اجزای سخن[۳] و ریشه‌ی کلمات زبان‌هایی نظیر ترکی آذربایجانی، لری، کردی و ... است. با توجه به نبود و یا کمبود ابزارهایی برای پیش پردازش زبان‌هایی با منابع محدود، ایجاد و توسعه‌ی چنین ابزارهایی حائز اهمیت است. در زیر توضیحاتی در رابطه با گسترش این ابزار برای ترکی آذربایجانی به تفصیل آورده شده است اما می‌توانید با هماهنگی تیم تدریس[۴]، این ابزار را برای سایر زبان‌های ایرانی نیز توسعه دهید.

ترکی آذربایجانی

برای انتخاب این ترک لزومی به تسلط بر زبان ترکی آذربایجانی نیست و در صورت وجود هر گونه ابهام می‌توانید از منابع معرفی شده و نیز دستیاران آموزشی مربوطه کمک بگیرید.

خروجی مورد انتظار شامل موارد زیر است:

- **Token**
- **Span**
- **POS** برای سهولت کار، تشخیص اسم، ضمیر، صفت، قید و فعل کافی است. در صورتیکه نتوان یکی از موارد فوق را بصورت قطعی به کلمه‌ای نسبت داد، تمامی احتمالات در قالب یک لیست برگردانده شود.
- **Lemma** ریشه‌ی کلمه برگردانده شود. (با توجه به پسوندی بودن زبان ترکی آذربایجانی، در بخش قابل توجهی از کلمات با حذف این پسوندها به ریشه‌ی کلمه می‌رسیم. برای اطلاع از پسوندهای کلمات از این لینک می‌توانید کمک بگیرید).
- **Tense** در صورتیکه نقش کلمه فعل باشد، زمان فعل برگردانده شود. لیست این زمان‌ها در صفحات ابتدایی این لینک قابل دسترسی است.
- **Pronoun Info** در صورتیکه نقش کلمه فعل باشد، شخص فعل برگردانده شود.

در صورت تمایل، می‌توانید روی جزییات یک یا چند نقش خاص متمرکز شده و خروجی مطلوب‌تری برای آن‌ها تولید کنید.

---

[۳] Part of Speech

[۴] reihane.zohrabi@gmail.com، marziehnouri۱۹۹۹@gmail.com، mahsa.ama۱۳۹۱@gmail.com

ورودی	خروجی

آخشام آنام بازاردان بیر ساری داراق آلدی.	[
	{
	"token": "آخشام",
	"span": [0,5],
	"POS": ["قید"],
	"lemma": "",
	"tense": "",
	"pronoun info": ""
	},
	{
	"token": "آنام",
	"span": [6,10],
	"POS": ["اسم"],
	"lemma": "آنا",
	"tense": "",
	"pronoun info": ""
	},
	{
	"token": "بازاردان",
	"span": [11,19],
	"POS": ["اسم"],
	"lemma": "بازار",
	"tense": "",
	"pronoun info": ""
	},
	{
	"token": "بیر",
	"span": [20,23],
	"POS": ["صفت"],
	"lemma": "بیر",
	"tense": "",
	"pronoun info": ""
	},
	{
	"token": "ساری",
	"span": [24,28],
	"POS": ["صفت"],
	"lemma": "ساری",
	"tense": "",
	"pronoun info": ""
	},
	{
	"token": "داراق",
	"span": [29,34],
	"POS": ["اسم"],
	"lemma": "دارا",
	"tense": "",
	"pronoun info": ""
	},
	{
	"token": "آلدی",
	"span": [35,39],
	"POS": ["فعل"],
	"lemma": "آل",
	"tense": "ماضی ساده خبری",
	"pronoun info": "سوم شخص مفرد"
	}
	]

ورودی	خروجی

بو آلمالاری آلاجاکلار.	[
	{
	"token": "بو",
	"span": [0,2],
	"POS": ["ضمیر"],
	"lemma": "",
	"tense": "",
	"pronoun info": ""
	},
	{
	"token": "آلمالاری",
	"span": [3,11],
	"POS": ["اسم"],
	"lemma": "آلما",
	"tense": "",
	"pronoun info": ""
	},
	{
	"token": "آلاجاکلار",
	"span": [12,21],
	"POS": ["فعل"],
	"lemma": "آل",
	"tense": "آینده دور خبری",
	"pronoun info": "سوم شخص جمع"
	}
	]

پیش پردازش از نوع کهن

برای زبان فارسی معیار که امروزه استفاده می‌شود کتاب‌خانه‌های متعددی برای پیش پردازش وجود دارد. معروف‌ترین کتابخانه برای این موضوع کتابخانه هضم است. زبان موجودی زنده است و به خاطر همین موضوع زبان فارسی معیار کنونی تفاوت‌هایی با زبان فارسی که در قرون گذشته استفاده می‌شده است دارد. هدف شما در این ترک این است که کتاب‌خانه‌ای مانند هضم برای زبان فارسی کهن طراحی کنید. این ترک بنچ‌مارک دقیق و مشخصی ندارد و براساس زحمات و خلاقیت شما نمره‌دهی می‌شود.
برای انجام این ترک pdf کتاب دستور تاریخی در اختیارتان قرار می‌گیرد که بتوانید با بررسی آن به تفاوت‌های فارسی کنونی با فارسی کهن پی ببرید. همچنین در انتخاب نوع متن (نثر یا نظم) یا زمان متن (متن در چه قرنی باشد) خودتان آزادی عمل برای انتخاب دارید. می‌توانید برای انجام این ترک از کد موجود هضم استفاده کنید و آن را متناسب با متون کهن تغییر دهید.
در دوره‌های کهن فارسی ساختمان فعلی متفاوت وجود داشته است و شما باید بتوانید ریشه‌یابی، نرمال‌سازی و حذف کلمات زائد را برای این دوره‌ها انجام دهید. در این ترک ساختار ورودی و خروجی مشخصی وجود ندارد چون توابع مختلفی باید پیاده سازی شوند اما در زیر مثال‌هایی از تفاوت فارسی کهن با فارسی کنونی آمده است.

مثال	توضیح
گفته بودی که: بیایم، چو به جان آیی تو من به جان آمدم، اینک تو چرا می‌نایی؟	می‌نایی در گذشته معادل نمی‌آیی بوده است. و اگر ابزاری برای این دوره فارسی درست شود باید آن را به ریشه: آمد## بیا ببرد.
حکیمی پسران را پند همی‌داد که جانان پدر هنر آموزید که ملک و دولت دنیا اعتماد را نشاید	در اینجا همی همان نشانه استمرار در زبان فارسی است و باید در نرمالسازی به می برده شود!

استخراج لیست قیمت محصولات و خدمات

هدف از این ترک، استخراج لیست قیمت تمامی محصولات و خدمات ذکر شده در یک متن با استفاده از عبارت‌های منظم است. این ابزار می‌تواند در جمع‌آوری خودکار قیمت کالاها و سرویس‌هایی که در متن اخبار یا وب‌سایت‌های فروش محصولات و ارائه خدمات منتشر می‌شود، کاربرد داشته باشد. ورودی کد شما یک متن و خروجی آن یک دیکشنری در زبان پایتون به ازای هر محصول یا خدمت است که باید شامل موارد زیر باشد:

۱) نام محصول یا خدمت (product_name)
۲) بازه نام محصول یا خدمت (product_name_span)
۳) مقدار محصول یا خدمت (product_quantity)
۴) واحد محصول یا خدمت (product_unit)
۵) مقدار قیمت محصول یا خدمت (price_amount)
۶) واحد قیمت محصول یا خدمت (price_unit)

در نهایت خروجی پیاده‌سازی شما، می‌بایست به صورت یک لیست از دیکشنری‌ها باشد. برای نمونه به مثال‌های زیر توجه کنید:

ورودی	خروجی

به گزارش خبرنگار اقتصادی خبرگزاری تسنیم، هم‌اکنون در بازار آزاد، قیمت طلای ۱۸ عیار هر گرم یک میلیون و ۳۵۹ هزار تومان و قیمت سکه تمام‌بهار آزادی طرح جدید ۱۵ میلیون و ۱۰۰ هزار تومان است.	<pre>[
	  {
	  "product_name": "طلای ۱۸ عیار",
	  "product_name_span": [70,81],
	  "product_amount": "۱" ,
	  "product_unit": "گرم",
	  "price_amount": "یک میلیون و 359 هزار",
	  "price_unit": "تومان",
	  },
	  {
	  "product_name": "سکه تمام‌بهار آزادی طرح جدید",
	  "product_name_span": [124, 152],
	  "product_amount": "۱" ,
	  "product_unit": "عدد",
	  "price_amount": "۱۵ هزار ۱۰۰ میلیون و",
	  "price_unit": "تومان",
	  },
	]</pre>

دبیر انجمن تصفیه کنندگان شکر گفت: در شرایط کنونی هر تن شکر ۵۳۰ دلار تحویل بنادر می شود.	<pre>[
	  {
	  "product_name": "شکر",
	  "product_name_span": [55, 58]
	  "product_amount": "۱" ,
	  "product_unit": "تن",
	  "price_amount": "۵۳۰",
	  "price_unit": "دلار",
	  }
	]</pre>

در شرکت اسپارد، قیمت یک سرویس نظافت ۵ ساعته ۲۴ هزار تومان می‌باشد.	<pre>[
	  {
	  "product_name": "نظافت",
	  "product_name_span": [30, 35]
	  "product_amount": "۵" ,
	  "product_unit": "ساعت",
	  "price_amount": "هزار ۲۴",
	  "price_unit": "تومان",
	  }
	]</pre>

برای آزمایش پیاده‌سازی خود می‌توانید از متن اخبار موجود در وب‌سایت‌های مختلف یا هر منبع دیگری استفاده کنید. همچنین در نهایت می‌بایست پیاده‌سازی خود را با استفاده از چند نمونه متنوع ارزیابی کرده و به همراه خروجی در گزارش تمرین ذکر کنید. در زمان تحویل تمرین به دستیاران، از چند نمونه داده متنوع برای ارزیابی پیاده‌سازی شما استفاده می‌شود.