# پردازش زبان‌های طبیعی

نیم‌سال اول ۱۴۰۰-۱۴۰۱

مدرس: احسان‌الدین عسگری

دانشکده مهندسی کامپیوتر

**تمرین سری اول**

نام و نام‌خانوادگی: امین کشیری، فاطمه توحیدیان، سید علیرضا موسوی

شماره دانشجویی: ۹۷۱۰۱۰۲۶ - ۹۷۰۰۳۵۴ - ۹۷۱۰۶۲۸۴

## ملاحضات

- کتابخانه‌های مورد نیاز در ابتدای کد نصب می‌شوند، اما در صورتی که بخواهید محیط اجرا را عینا شبیه‌سازی کنید می‌توانید از فایل requirements.txt استفاده کنید.
- لطفا فایل symbols_info.txt که همراه کد و داکیومنت آپلود شده است را کنار کد قرار دهید.

## روند اجرای کد

در این تمرین سعی شده است تا بتوانیم وقایع و علائم بورسی را از متن استخراج کنیم. منطق اصلی کد ما می‌تواند به سه قسمت اصلی تقسیم شود. در قسمت اول، به کمک کلمات کلیدی از پیش تعیین شده‌ای، قسمت‌هایی از متن را به دست می‌آوریم که قسمتی از یک واقعه باشند. در قسمت دوم، هر قسمت را بررسی می‌کنیم و سعی می‌کنیم متن کامل واقعه را به دست آوریم. در قسمت سوم نیز وقایع تکراری را حذف می‌کنیم و در صورتی که بتوانیم بعضی از وقایع را با هم ترکیب می‌کنیم تا واقعه بهتری را پیدا کنیم.

دسته‌بندی اتفاقات داخل متن، به صورت زیر صورت گرفته است:

۱. **نماد**

نمادهای معاملاتی در بورس ایران

۲. **شرکت**

نام شرکت‌های بورسی ایران

۳. **اعلان**

اصطلاحات اداری همانند گزارش، اطلاعیه و ...

۴. **تحلیل**

اصطلاحات خاص بورسی و تحلیلی مانند تحلیل تکنیکال، واگرایی و ...

۵. **شخصیت**

اتفاقات مربوط به شخصیت‌های حاضر در بازار مانند نوسان‌گیر، بازیگر و ...

۶. **واقعه**

سایر وقایع مهم بورسی که در دسته های بالا جای نگیرند و وقایع مختلف بورسی را نشان می‌دهند. مانند صف خرید، تقسیم سود، مجمع عمومی و ...

دقت کنید که این دسته‌ها لزوما کل اتفاقات را افراز نمی‌کنند. برای مثال ممکن است در یک واقعه، یک نماد بورسی نیز وجود داشته باشد، و در این حالت ما هردوی این اتفاقات را گزارش می‌کنیم. برای برخی از وقایع تنها قسمتی از متن را به عنوان واقعه گزارش می‌دهیم، اما برخی وقایع پیچیده ترند و سعی می‌کنیم اجزای دیگری از آن را نیز خروجی دهیم. برای مثال، برای ورودی زیر، وقایع را به همراه فاعل آن‌ها پیدا می‌کنیم:

```
> 12 مثال
ارزش سهام مخابرات ایران امروز کاهش زیادی یافت. [U+200E]
{
"type": "شرکت",
"marker": "مخابرات ایران",
"span": [10,23]
}
{
"type": "واقعه",
"marker": "کاهش زیادی یافت",
"span": [30,45],
"subject": "ارزش سهام مخابرات ایران",
"span_subject": [0,23]
}
```

برای آشنایی بیشتر با خروجی کد ما، می‌توانید مثال‌های زیر را ببینید.

```
> 1 مثال
برکت امروز اطلاعیه‌ای مهم منتشر میکند.
{
"type": "نماد",
"marker": "برکت",
"span": [0,4]
}
{
"type": "اعلان",
"marker": "اطلاعیه‌ای مهم",
"span": [11,25]
}
```

```
> 9 مثال
یک نکته‌ی تکنیکالی هم در صورت دستکاری نشدن اضافه کنم، کندلی که روز سه شنبه‌ی گذشته ثبت کرد کامل است
{
"type": "تحلیل",
"marker": "یک نکته‌ی تکنیکالی",
"span": [0,18]
}
{
"type": "تحلیل",
"marker": "کندلی که روز سه شنبه‌ی گذشته ثبت کرد",
"span": [54,90]
}
```

```
> 10 مثال
آ س پ امروز روند مثبتی داشت.
{
"type": "نماد",
"marker": "آ س پ",
"span": [0,5]
}
{
"type": "واقعه",
"marker": "روند مثبتی",
"span": [12,22]
}
```

```
> 11 مثال
آمریکا موجب ریزش بازار شد
{
"type": "واقعه",
"marker": "ریزش بازار",
"span": [12,22]
}
```

```
> 13 مثال
.قیمت سهام زیاد است. رشد قیمت‌ها باعث کاهش قیمت شد. ریزش بازار هم به همین دلیل بود
{
"type": "واقعه",
"marker": "رشد قیمت‌ها",
"span": [20,31]
}
{
"type": "واقعه",
"marker": "کاهش قیمت",
"span": [37,46]
}
{
"type": "واقعه",
"marker": "ریزش بازار",
"span": [51,61]
}
```

```
> 14 مثال
بورس سه شنبه افت بیش از ۳۳ هزار واحدی داشت
{
"type": "نماد",
"marker": "بورس",
"span": [0,4]
}
{
"type": "واقعه",
"marker": "افت بیش از ۳۳ هزار واحدی",
"span": [13,37]
}
```

```
> 15 مثال
شاخص معاملات ۱۱ درصد افت پیدا کرد، ۱۲٫۵ درصد رشد کرد و ۲٫۲ درصد افزایش یافت
NOT IMP - child dep is:  cc
{
"type": "واقعه",
"marker": "افت کرد",
"span": [21,33],
"subject": "شاخص معاملات",
"span_subject": [0,12]
}
{
"type": "واقعه",
"marker": "رشد کرد",
"span": [45,52],
"subject": "۱۲٫۵ درصد",
"span_subject": [35,44]
}
{
"type": "واقعه",
"marker": "افزایش یافت",
"span": [64,75],
"subject": "۲٫۲ درصد",
"span_subject": [55,63]
}
```

## به دست آوردن کلمات کلیدی

در این قسمت، به دنبال کلمات کلیدی از پیش تعیین شده‌ای در متن می‌گردیم. این کلمات در دسته‌های مختلفی قرار می‌گیرند و اتفاقات مختلفی را گزارش می‌دهند. در صورتی که اتفاقی از قلم افتاده باشد، کافی است که یک کلمه‌ی کلیدی مربوط به آن اتفاق را به دسته‌بندی‌های خود اضافه کنیم.

بعضی از کلمه‌های کلیدی دو بخشی‌اند. مانند «عرضه اولیه». در چنین شرایطی، تمام حالات ممکن این عبارت کلیدی را نیز پیدا می‌کنیم. برای مثال می‌توانید به خروجی‌های زیر نگاه کنید:

```
> مثال 2
نماد برکت امروز عرضه‌ی اولیه خیلی خوبی داره.
{
"type": "نماد",
"marker": "برکت",
"span": [5,9]
}
{
"type": "واقعه",
"marker": "عرضه‌ی اولیه خوبی",
"span": [16,38]
}
```

```
> مثال 3
نماد برکت امروز عرضه‌ی اولیه خیلی خوبی داره.
{
"type": "نماد",
"marker": "برکت",
"span": [5,9]
}
{
"type": "واقعه",
"marker": "عرضه‌ی اولیه خوبی",
"span": [16,38]
}
```

```
> مثال 4
نماد برکت امروز عرضه‌ی اولیه دارد.
{
"type": "نماد",
"marker": "برکت",
"span": [5,9]
}
{
"type": "واقعه",
"marker": "عرضه‌ی اولیه",
"span": [16,28]
}
```

```
> مثال 5
عرضه‌های اولیه امروز خوب هستند.
{
"type": "واقعه",
"marker": "عرضه‌های اولیه",
"span": [0,14]
}
```

```
> مثال 6
عرضه اولیه‌های امروز خوب هستند.
{
"type": "واقعه",
"marker": "عرضه اولیه‌های امروز",
"span": [0,20]
}
```

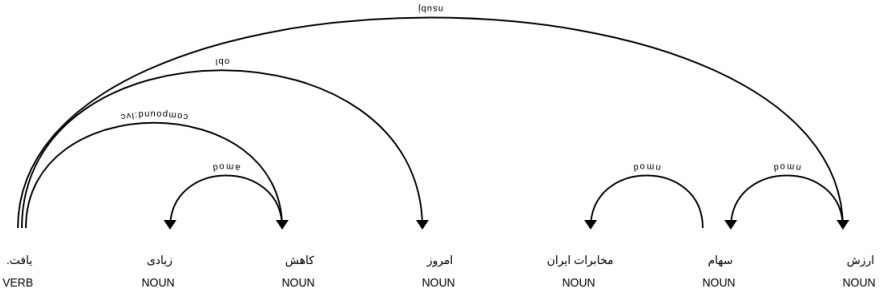
```
> مثال 7
این عرضه‌ی اولیه خیلی خوبه
{
"type": "واقعه",
"marker": "این عرضه‌ی اولیه",
"span": [0,16]
}
```

برای پیدا کردن کلمات کلیدی، از regex ها و کلاس Matcher در کتابخانه‌ی spacy کمک گرفته‌ایم. توکن‌های چند کلمه‌ای پس از این که پیدا شدند، به عنوان یک توکن واحد در نظر گرفته می‌شوند تا کنار یک دیگر معنی پیدا کنند.

برای پیدا کردن نمادها و اسم شرکت‌های بورسی، با crawl کردن توانستیم یک فایل csv تهیه کنیم. سپس به کمک کتابخانه‌ی pandas اسم نمادهای بورسی و شرکت‌ها را در متغیرهای جداگانه ذخیره کردیم، و با regex به دنبال آن‌ها گشتیم. این اسامی در کد ما به عنوان Named Entity شناخته می‌شوند، و در صورتی استفاده از خروجی displacy در کتابخانه spacy به صورت زیر نمایش داده می‌شوند:

## پیدا کردن متن کامل واقعه

بعد از این که کلمات کلیدی وقایع را به دست آوردیم، سعی می‌کنیم آن کلمات را گسترش دهیم تا شامل یک واقعه‌ی کامل شوند. مثلا، تاثیر یا مثبت به تنهایی یک واقعه تشکیل نمی‌دهند، اما ۵ واحد تاثیر مثبت یک واقعه تشکیل می‌دهد. برای این کار، از کتابخانه‌ی StanfordNLP استفاده می‌کنیم. استنباط ما دو کمک کننده‌ی اساسی دارد. اول POS tagها و دوم استفاده از Dependecy Treeها. درخت روابط، درختی است که روابط اجزای مختلف یک جمله با هم را نشان می‌دهد. یک مثال از این درخت‌ها به صورت زیر است:




حال وقتی به یک کلمه می‌رسیم، با استفاده از این دو، تشخیص می‌دهیم که واقعه‌ی اصلی چیست. برای مثال، وقتی کلمه‌ی کلیدی ما یک مضاف‌الیه است، سعی می‌کنیم هسته‌ی گروه اسمی را به دست آوریم و گروه اسمی را خروجی دهیم. مثلا در متن زیر، به جای این که تنها «صف خرید» را به عنوان واقعه تشخیص دهیم، کل واقعه یا به عبارتی «ایجاد صف خرید در سهم پرشیا» را تشخیص داده‌ایم.

```
> 8 مثال
رشد قیمت‌ها باعث ایجاد صف خرید در سهم پرشیا شد
{
"type": "نماد",
"marker": "پرشیا",
"span": [38,43]
}
{
"type": "واقعه",
"marker": "رشد قیمت‌ها",
"span": [0,11]
}
{
"type": "واقعه",
"marker": "ایجاد صف خرید در سهم پرشیا",
"span": [17,43]
}
```

یا اگر یک فعل مرکب تشخیص دهیم، سعی می‌کنیم فاعل جمله را نیز به دست آوریم تا مفوم واقعه کامل باشد. مثلا در ورودی زیر، نه تنها «کاهش زیادی یافتن» را پیدا کرده‌ایم، بلکه فاعل آن یعنی «ارزش سهام مخابرات ایران» را نیز تشخیص داده‌ایم

```
> 12 مثال
ارزش سهام مخابرات ایران امروز کاهش زیادی یافت. [U+200E]
{
"type": "شرکت",
"marker": "مخابرات ایران",
"span": [10,23]
}
{
"type": "واقعه",
"marker": "کاهش زیادی یافت",
"span": [30,45],
"subject": "ارزش سهام مخابرات ایران",
"span_subject": [0,23]
}
```

توجه کنید که در روند اجرای برنامه هر کجا منطقی برای پیدا کردن اتفاقات کامل‌تر پیاده‌سازی شده متناظرا کدهای لازم برای حفظ درستی span ها نیز نوشته شده.

## خروجی کد و ترکیب وقایع

در صورتی که دو اتفاق از یک نوع یکسان باشند و هر دو به اتفاقی مشترک اشاره کنند این چند اتفاق را با هم ترکیب کرده و در نهایت کامل ترین اتفاق را خروجی می‌دهیم. این کار با الگوریتمی با زمان اجرای $O(n \log n)$ این کار را انجام می‌دهد. در صورتی که اتفاقات تکراری برای ما مهم نباشد، نیازی به اجرای این قسمت از کد نیز نیست و برنامه سریع تر می‌شود. نحوه‌ی عملکرد الگوریتم به این صورت است که با بررسی span اتفاقات رویدادهای مشترک را اجتماع می‌گیرد و از متن اصلی آن‌ها را انتخاب می کند. یک مثال از اجرای این الگوریتم به صورت زیر است که با این که رشد و مثبت هر دو کلمه‌ای کلیدی محسوب می‌شوند، اما ما تنها یک واقعه که «رشد مثبت ...» را تشخیص داده‌ایم.

```
> 16 مثال
رشد مثبت نماد فملی قابل توجه بود.
{
"type": "نماد",
"marker": "فملی",
"span": [14,18]
}
{
"type": "واقعه",
"marker": "رشد مثبت نماد فملی",
"span": [0,18]
}
```