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 - ۹۷۱۲۳۴۵۶

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:

EX

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

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:

EX

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 به 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:

OUTPUT

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:

OUTPUT

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. مثال:

OUTPUT

یا اگر یک فعل مرکب تشخیص دهیم، سعی می‌کنیم فاعل جمله را نیز به دست آوریم تا مفوم واقعه کامل باشد. مثال:
requiremets