# به نام خدا

# گزارش تمرین اول درس پردازش زبان طبیعی

#### استاد درس:

### جناب دکتر برادران

#### نام و نام خانوادگی دانشجو:

### امیررضا صدیقین

#### شماره دانشجویی:

## 993614024



# مرحله 1:

در این بخش متغیر‌های متنی مورد استفاده در این تمرین تعریف شده اند.

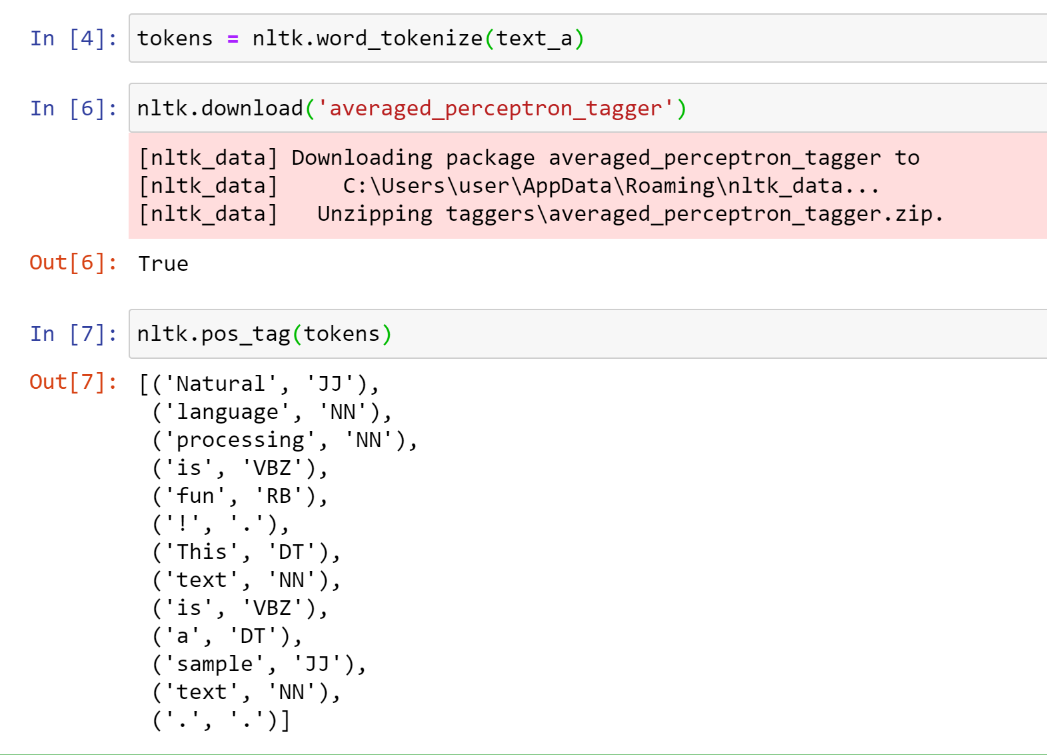
# مرحله ‌2:

## بخش a:

Chunking یک پروسه‌ی استخراج عبارات از یک متن ساختار نیافته است. و به هر عبارت یک chunk گویند. فرق chunk با توکن در آن است که chunk یه عبارت معنی دار را جدا می‌کند در حالی که token صرفا بر اساس یک سری قوانین این کار را انجام میدهد. برای مثال South Africa در پروسه‌ی chunking یک عبارت در نظر گرفته می‌شود به جای این که دو توکن South و Africa شود. همچنین noun-phrase chunking پروسه‌ی استخراج عبارات اسمی است.

## بخش b:

در این بخش متن اول توکنایز شده و سپس عملیات POS tagging روی آن اعمال شده است.

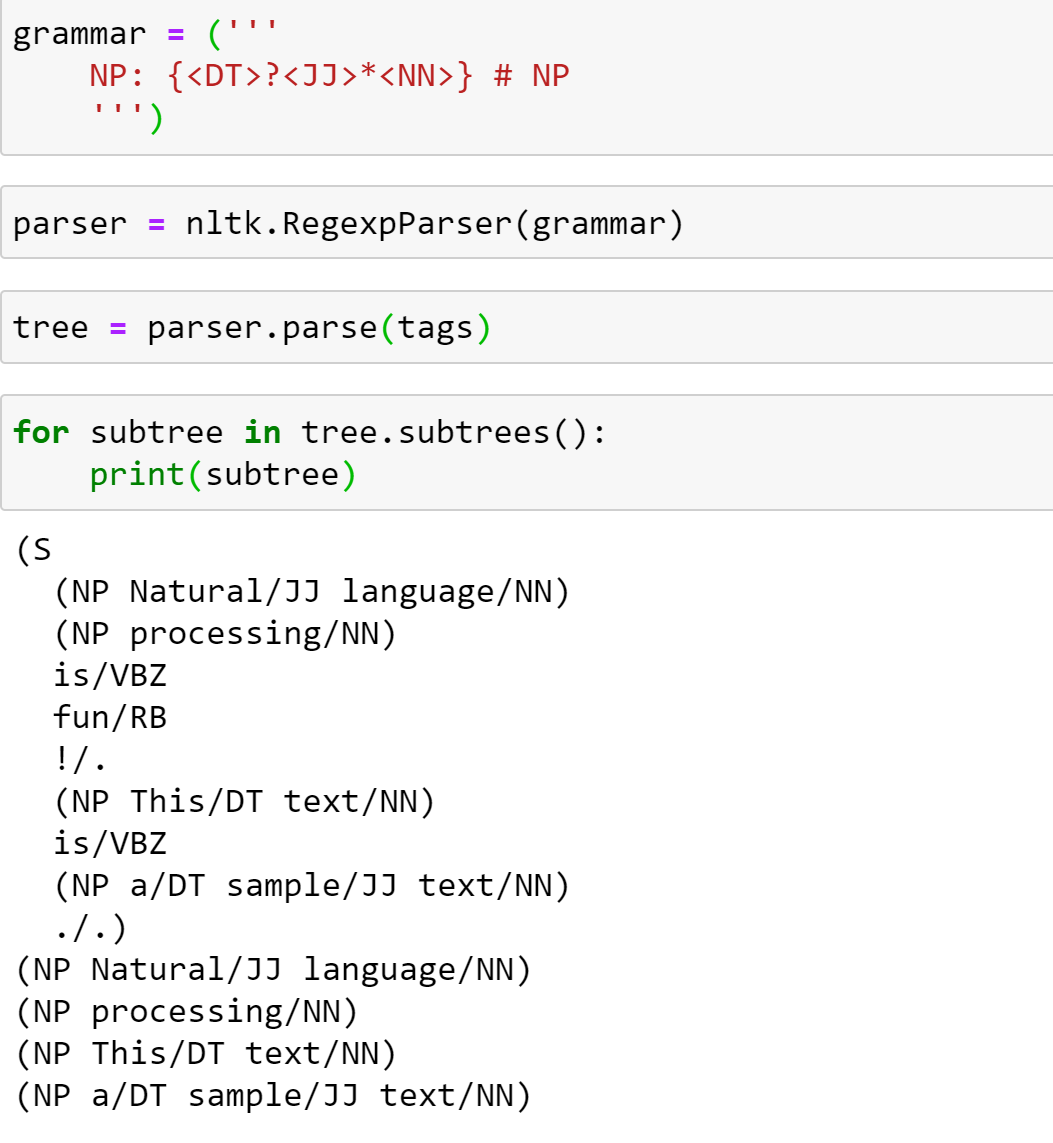


## بخش c:

می‌توان با استفاده از کلاس RegexpParser عملیات chunking را با استفاده از یک عبارت منظم روی یک متن انجام داد. ( در ورودی کلاس RegexpParser یک گرامر گرفته می‌شود که میتوان عبارت منظم را در قالب گرامر به این کلاس داد.). همچنین با استفاده از تابع parse در این کلاس میتوان درخت آن را به دست آورد.

## بخش d:

در این بخش گرامر گفته شده تعریف شد و parserی با استفاده از RegexpParser و گرامر گفته شده ساخته شد.



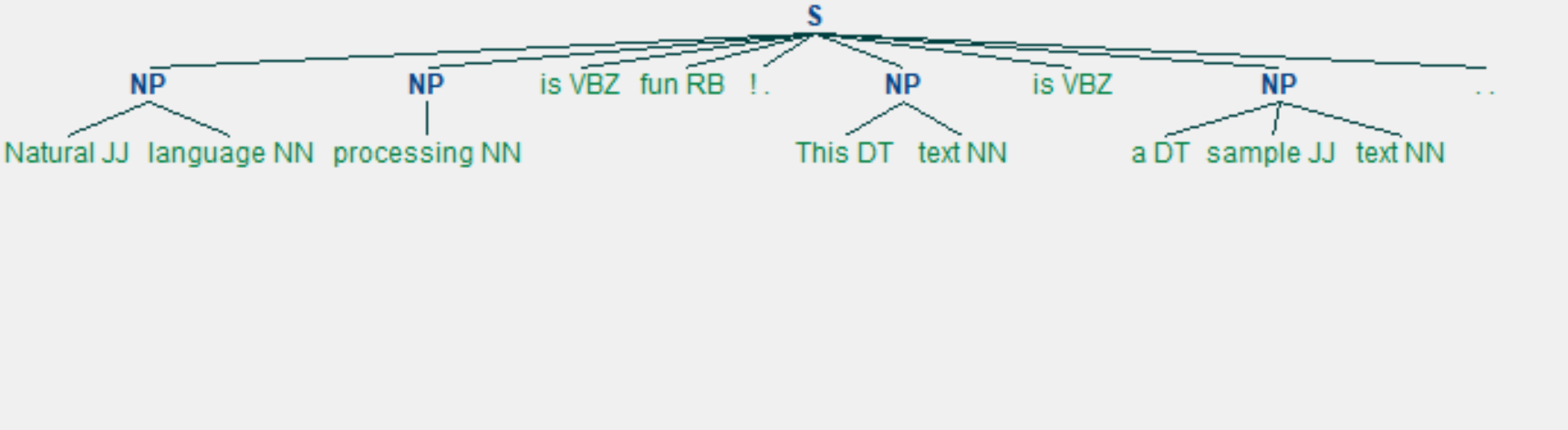
## بخش e:

با توجه به گرامر گفته شده، منظور آن است که اسم‌های مفرد را به عنوان یک Noun Phrase در نظر بگیر و همچنین اگر قبل از آن‌ها نیز صفت‌هایی آمده باشد، آن صفت‌ها به همراه اسم را نیز به عنوان Noun Phrase در نظر می‌گیریم و اگر قبل از این موارد کلمه‌ی Determinerی وجود داشت نیز آن هم به عنوان عبارت اسمی در نظر میگیریم. ( کلمات Determiner مثل the یا a و ... است)

## بخش f:

با استفاده از متد draw روی درخت ایجاد شده، درخت زیر بدست می‌آید.

درخت آن به صورت زیر است.



## بخش g:

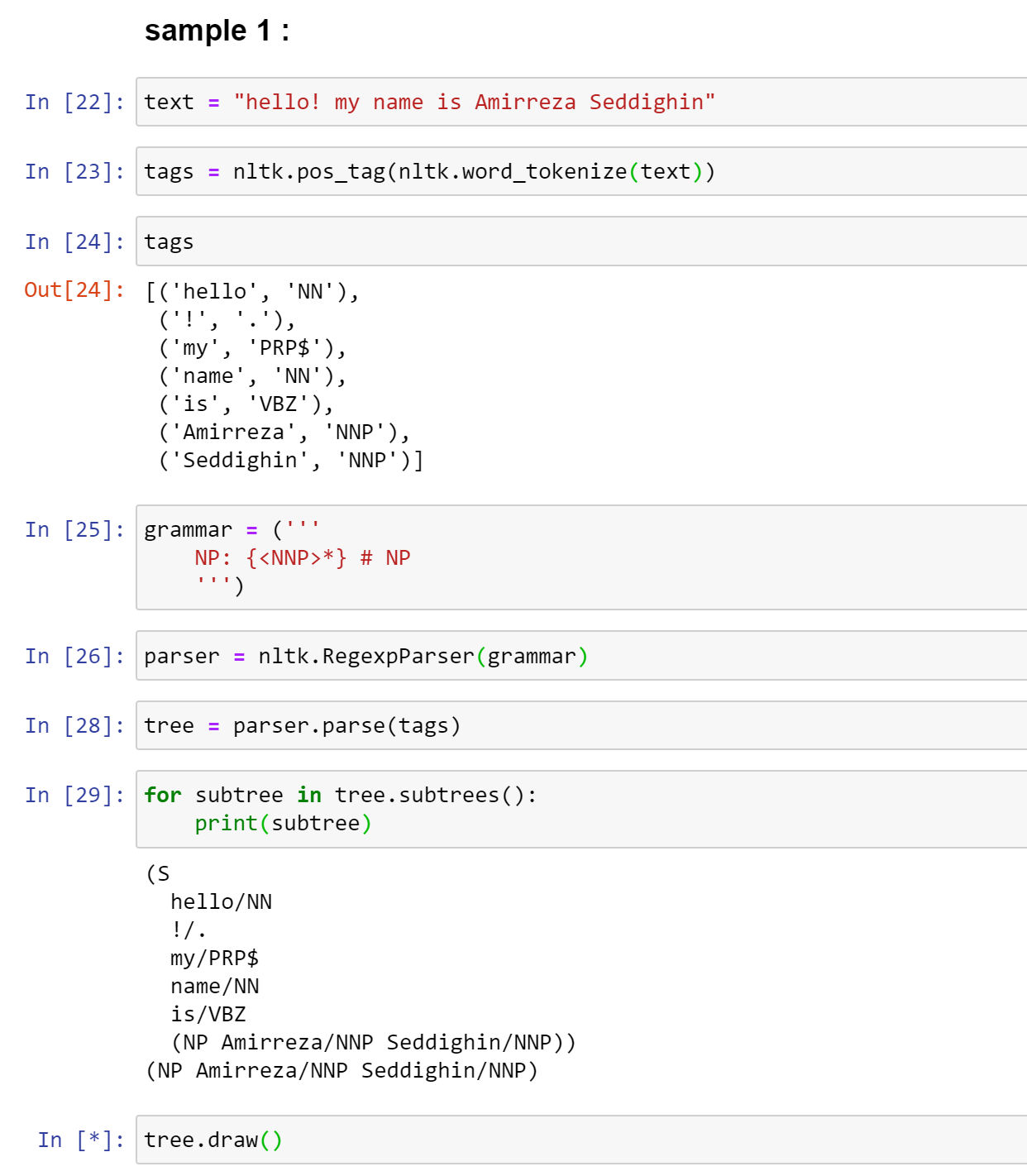
در این بخش دو گرامر و دو متن دیگر مثال زده شده است.

### مثال اول:

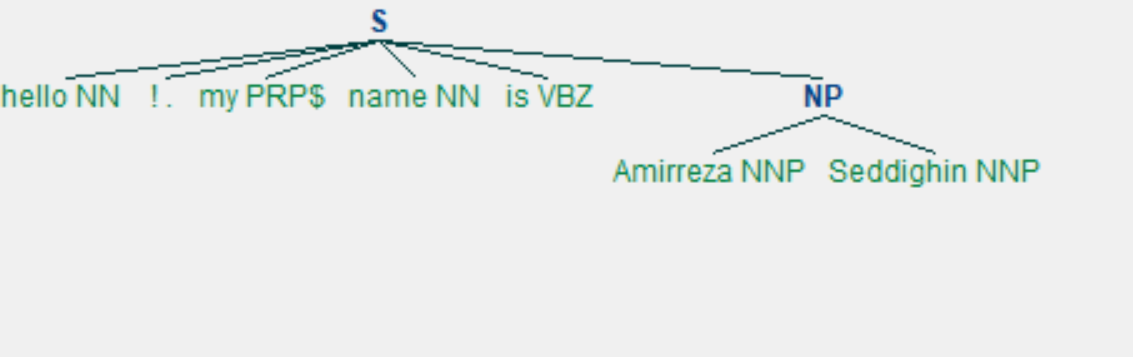
* متن نمونه : hello! My name is Amirreza Seddighin
* پترن : دنباله‌ی اسامی خاص پشت سرهم.

<NNP>\*

تکه کد آن نیز به صورت زیر است.



درخت آن نیز به صورت زیر است.

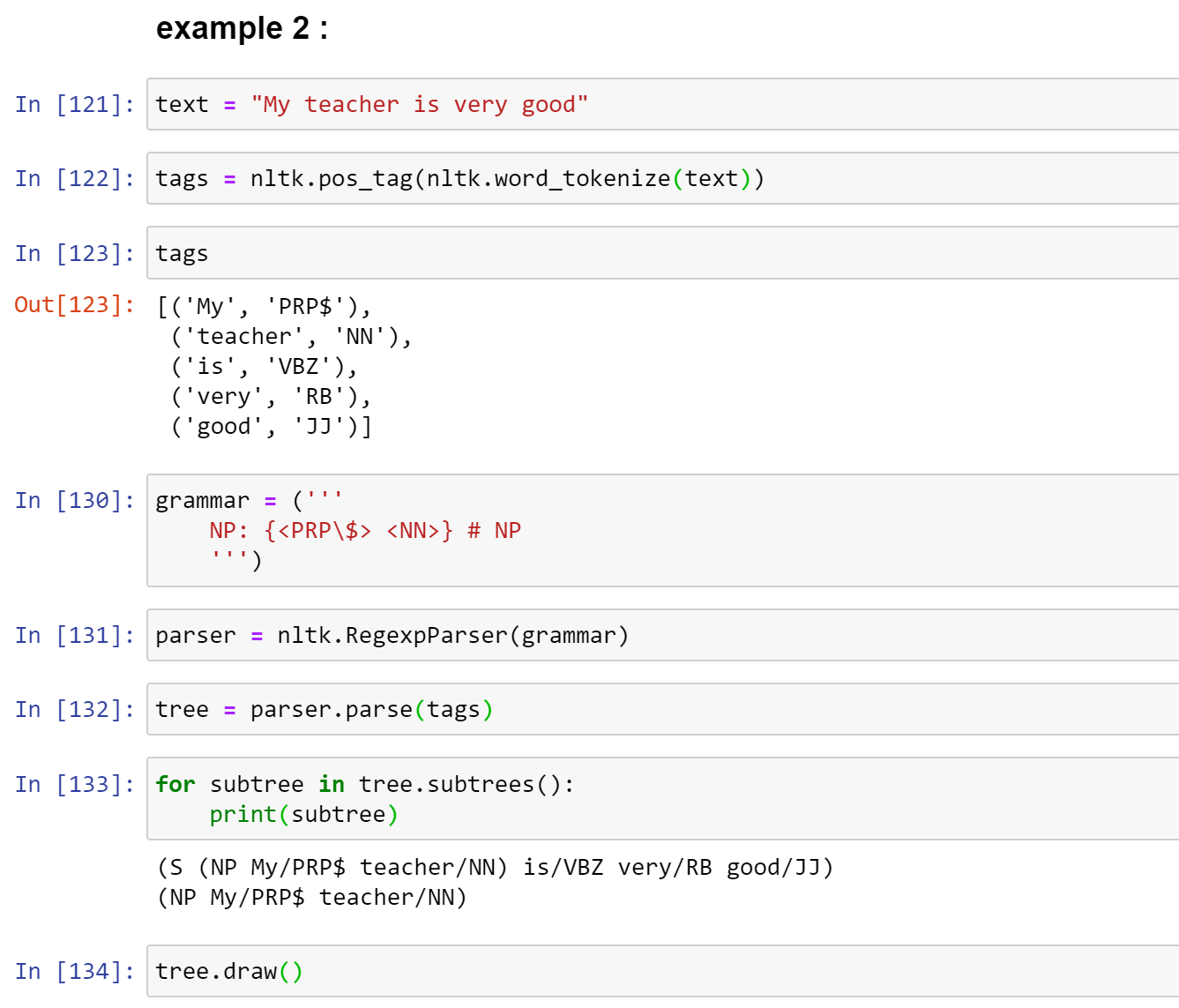


### مثال دوم:

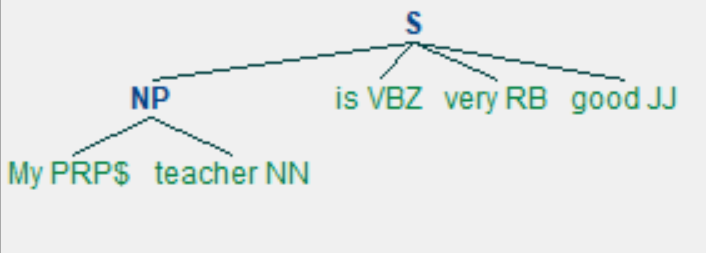
* متن نمونه : My teacher is good
* پترن : ضمیر به علاوه‌ی یک اسم مفرد

<PRP\$> <NN>

تکه کد آن نیز به صورت زیر است.



درخت آن نیز به صورت زیر است.



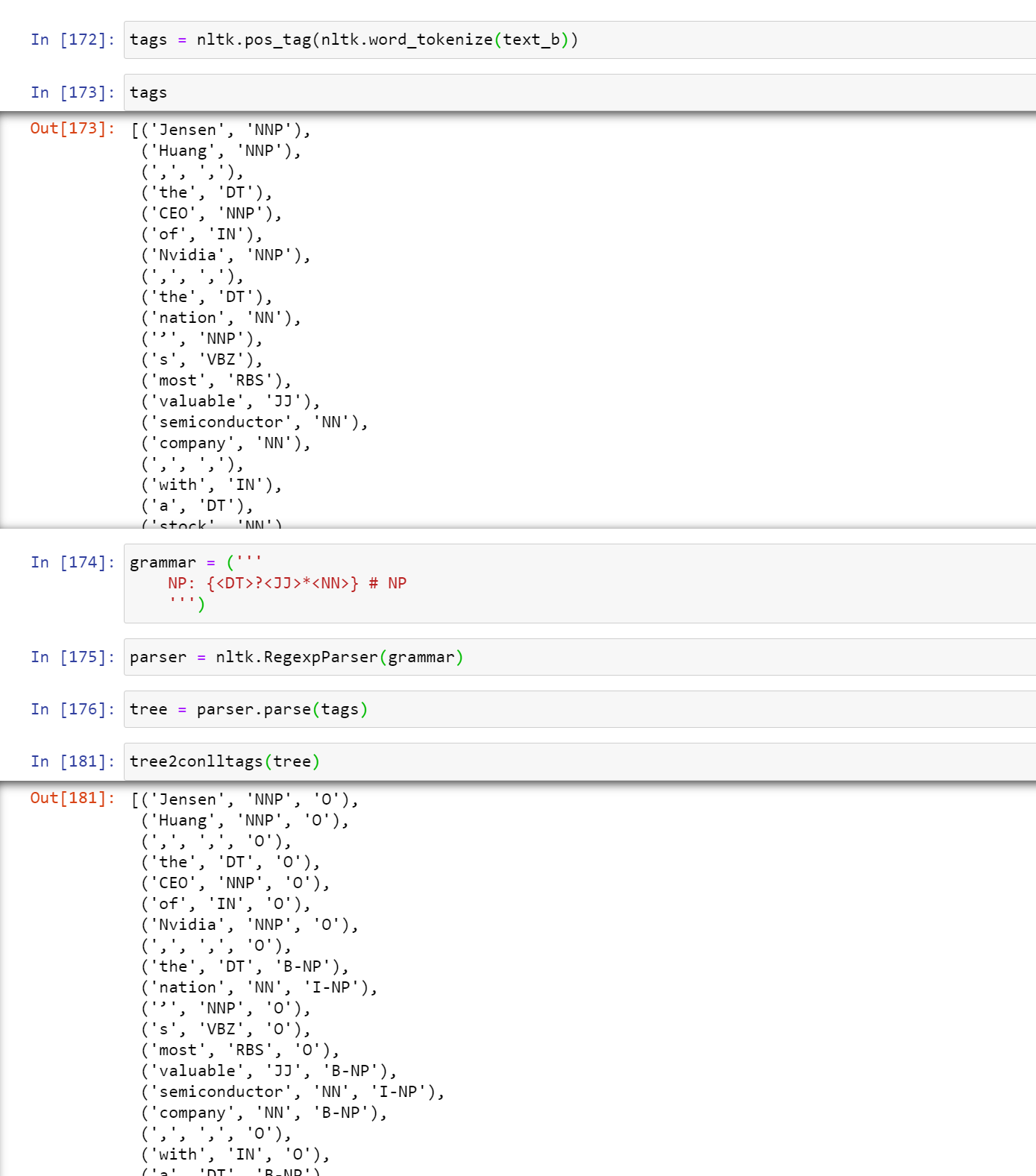
# مرحله‌ی 3:

## بخش a:

برای IOB encoding کافی است از تابع tree2conlltags استفاده کنیم و در ورودی آن درخت تجزیه‌ی خود را بدهیم.

## بخش b:

در این بخش عملیات IOB encoding بر روی متن دوم اعمال شده است که به صورت زیر است.(همچنین بخش‌هایی از جواب‌ها نیز به نمایش در آمده است.)



## بخش c:

در این بخش با توجه به گرامر گفته شده، عبارات به سه بخش O و B\_NP و I\_NP است. که منظور از O یعنی آن که خارج از محدوده است و NP نیست. منظور از B\_NP یعنی Begin NP به معنای توکنی است که نشان دهنده‌ی اول عبارت اسمی متناسب با regex تعریف شده‌ی ما است و I\_NP یعنی inside NP که به معنای توکنی است که به عنوان شروع کننده‌ی عبارت اسمی نیست و بعد از یک B\_NP آمده است و جزوی از آن عبارت اسمی است.

معمولا B\_NP یک Determiner یا در صورت نبود یک صفت است و اگر صفت هم نبود یک عبارت اسم مفرد است.

# مرحله 4 :

## بخش a:

برای استفاده از مدل از پیش آموزش داده شده‌ی Stanford باید از کلاس StanfordNERTagger در nltk استفاده کنیم و در ورودی آن مسیر فایل مدل و مسیر فایل jar را بدهیم و بعد با استفاده از تابع .tag موجود در کلاس آن مجموعه توکن‌ها را تگ گذاری کنیم.

## بخش b:

در این بخش عملیات NER بر روی متن b اعمال شد. قطعه کد آن و خروجی به صورت زیر است.



## بخش c:

در این بخش موجودیت‌های سازمانی و نام افراد استخراج شده از متن b به نمایش درامده است.



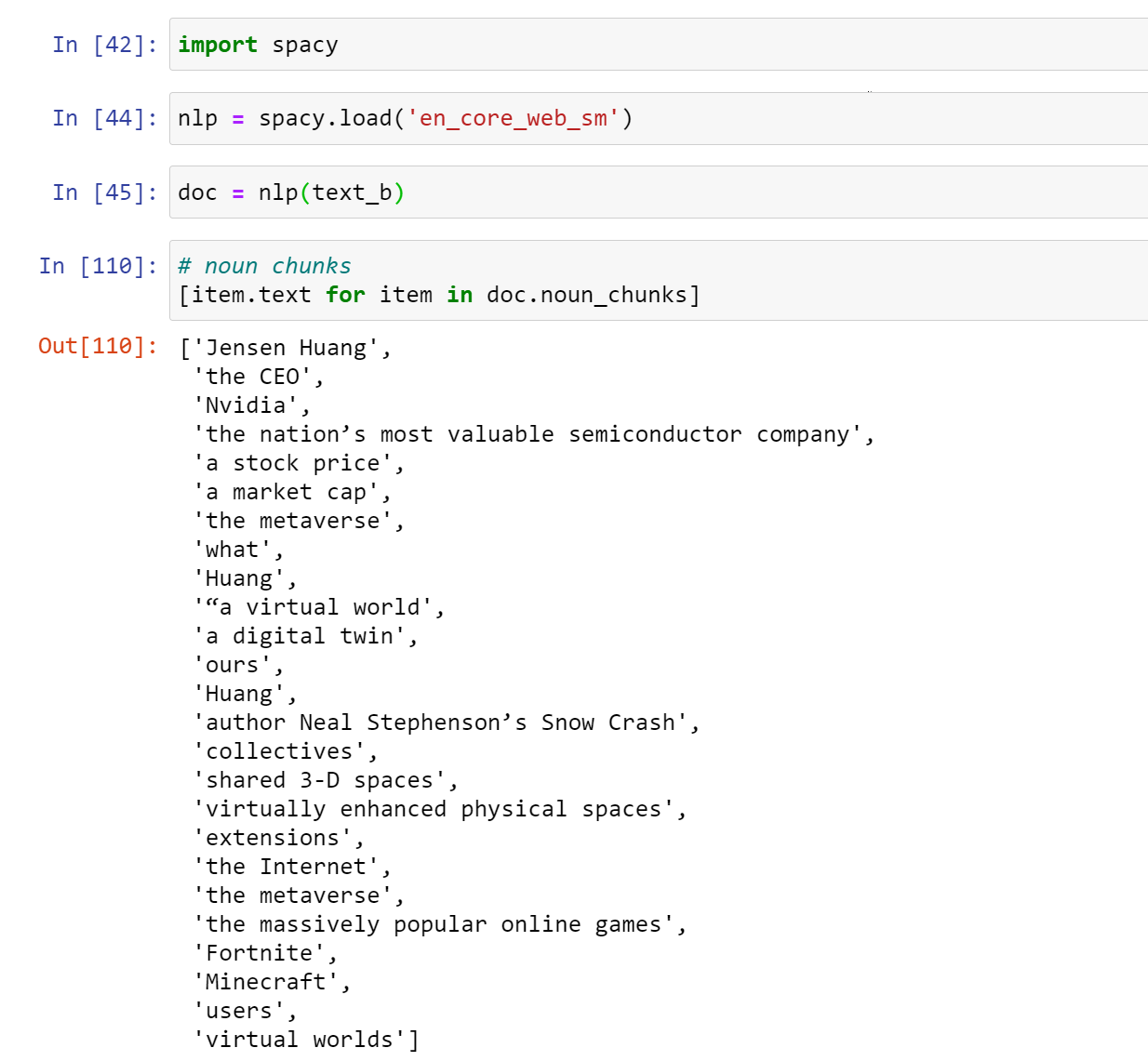
# مرحله‌ی 5:

## بخش a:

برای استفاده از کتابخانه‌ی Spcay باید آن را و مدل خواسته شده را نصب کرد و مدل en\_core\_web\_sm را لود کرد. سپس متن داده شده را به مدل لود شده داد.

## بخش b:

در این بخش عملیات NER را برروی متن b انجام شد. قطعه کد و noun\_chunk ها به صورت زیر آمده است.



## بخش c:

در این بخش موجودیت‌های نام افراد، نام سازمان، پول، اعداد استخراج شده اند.



# مرحله 6 :

در کتابخانه‌ی Spacy توابع متعدد برای دیدن entity ها خاص در یکجا وجود دارد و همچنین به نسبت nltk دقیق تر است و قابلیت آن را دارد که موجودیت‌های اسمی جزئی تری (مثل money) را پیدا کند و همچنین در تشخیص موجودیت‌های اسمی دو بخشی نسبت به nltk قوی تر کار کرده است مثل نام و نام خانوادگی.