# به نام خدا

# گزارش تمرین پنجم درس پردازش زبان طبیعی

#### استاد درس:

### جناب دکتر برادران

#### نام و نام خانوادگی دانشجو:

### امیررضا صدیقین

#### شماره دانشجویی:

## 993614024



تمام کدها و خروجی‌های هر بخش در فایل جوپیتر مربوط به هر بخش در دسترس است. (هم خروجی ipynb و html)

# بخش 1

در این بخش فایل‌های sport.txt , tech.txt خوانده شده است.

# بخش 2 و 3

در این بخش برای کلمات داده‌شده synset آن‌ها به دست آمد و name و definition(تعریف) و مثال‌هایی و hypernyms و part\_of\_speech آن مشخص شده است.

هر synset دارای فرمت name.pos.index است. که name خود کلمه و pos نوع کلمه که آیا اسم است یا فعل یا ... (part of speech) و index آن برای unique کردن آن است و صرفا برای تمایز بین synsetهایی است که name و pos یکسانی دارند.

# بخش 4

در این بخش تابعی به نام get\_lemmas نوشته شده است که کلمه را در ورودی می‌گیرد و تمام lemma های تمام synset های کلمه را در قالب یک مجموعه (تکراری‌ها حذف می‌شود) برمی‌گرداند.

این تابع پارامتری به به نام print\_lemmas است که در صورت true بود در هر مرحله به صورت جداگانه برای هر synset ، lemmaهای آن چاپ شده است.

سپس برای کلمات داده شده در صورت سوال این تابع اجرا شده است.

# بخش 5

در این بخش تابعی نوشته شده است که در آن هم‌معنا ها و متضاد‌های یک کلمه چاپ می‌شود. به این صورت که برای هم‌معنا ها lemma ها چاپ شده و برای متضاد‌ها از تابع antonyms روی هر lemma به دست می‌آید.

برای کلمات داده شده تابع اجرا شده است.

# بخش 6

در این بخش تابعی نوشته شده است که hypernyms و Hyponyms کلمه به دست می‌آید که در آن hypernyms و hyponyms برای تمام synset ها بدست می‌آید.

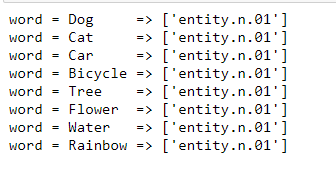
سپس برای کلمات داده شده این تابع اعمال شده است.

# بخش 7

## بخش a:

در این بخش برای کلمات داده‌ شده root\_hypernyms برای synset اول (synsetهای گفته شده در صورت سوال) به دست آمده است.

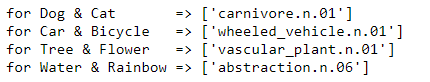
محتویات جدول به صورت زیر پر می‌شود.



## بخش b:

در این بخش برای زوج کلمات داده‌ شده lowest\_common\_hypernyms به دست آمده است. (در این بخش از synsetهای گفته شده استفاده شده است.)

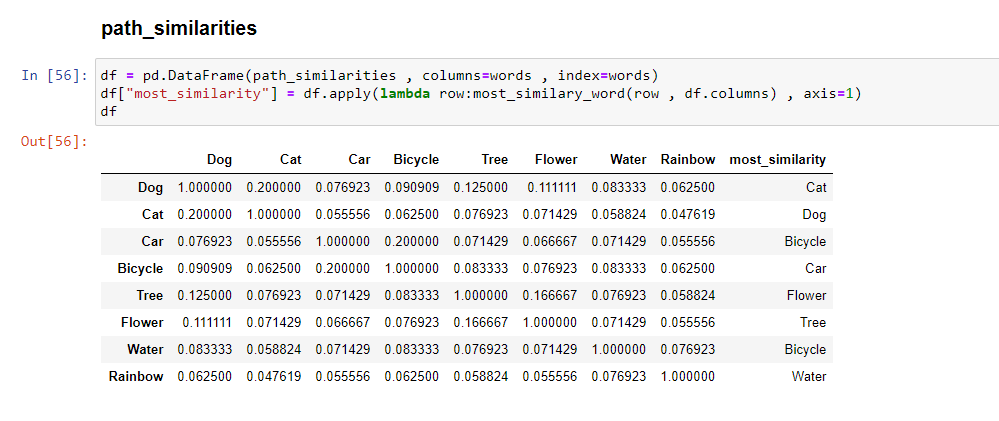
محتویات جدول به صورت زیر پر می‌شود.

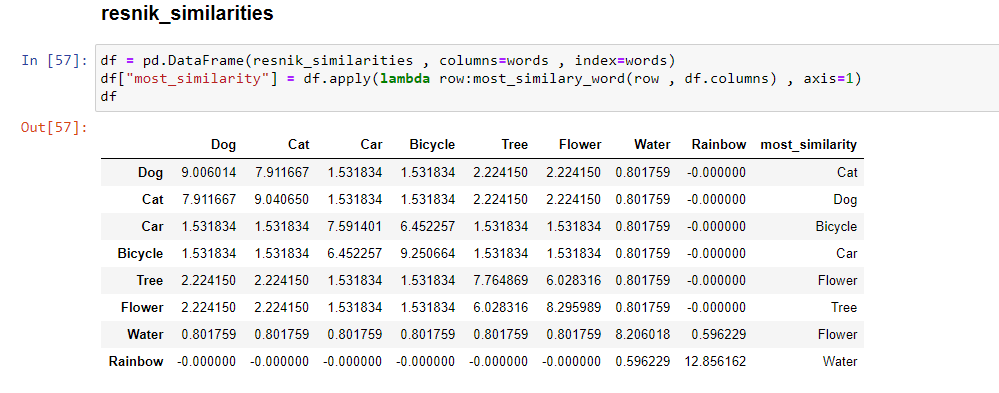


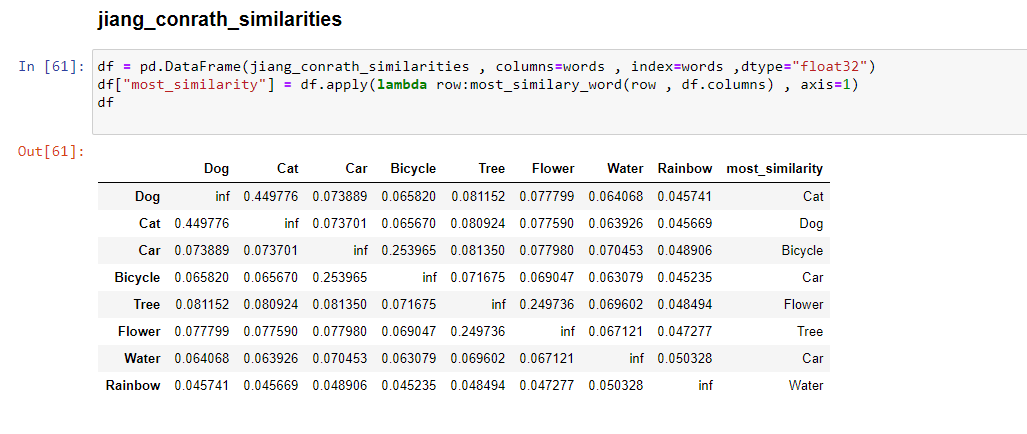
# بخش 8

در این بخش path similarity و resnik similarity و jiang-conrath similarity برای دو به دو کلمات داده شده به دست می‌آید و در ماتریسی ذخیره می‌شود و به صورت جدولی نمایش داده شده است. همچنین تایع نوشته شده است که اتربیوتی که بیشترین شباهت داشته باشه در هر سطر را پیدا می‌کند و در قالب ستون most\_similarity به دست آمده است.

که به صورت زیر جداول به دست آمده است.







# بخش 9

## بخش a:

در این بخش توکن‌های هر متن استخراج شده اند.

## بخش b:

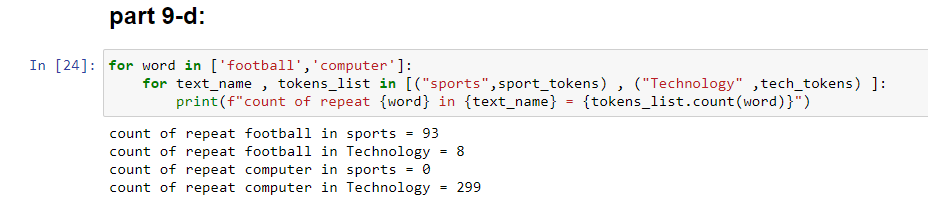
در این بخش کلمات stopwords و تک حرفی‌ها از توکن‌های استخراج شده حذف می‌شوند.

## بخش c :

در این بخش typeهای توکن‌های هر متن به‌دست آمده است که در متن‌های sport 11010 تا type وجود دارد و در متن‌های tech 13404 تا type بدست آمده است.

## بخش d:

در این بخش تعداد تکرار کلمات football و computer در هر متن بدست آمده است.



## بخش e:

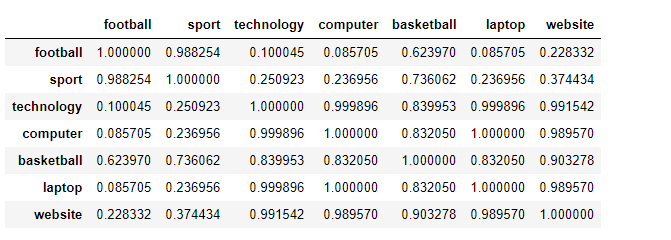
در این بخش ماتریس term\_document به دست آمده است که به ازای هر کلمه‌ی موجود در دو متن تعداد تکرار هر کدام در هر متن مشخص می‌شود.



## بخش f:

در این بخش شباهت کسینوسی برای کلمات زیر به دست آمده است. (دو به دو) ["football","sport","technology","computer","basketball","laptop","website"]

که شباهت کسینوسی به معنای 1- cos(alpha) است که منظور از alpha زاویه‌ی بین دو بردار است. (ورودی تابع cosine دو تا بردار است.)

‌

|  |  |  |  |
| --- | --- | --- | --- |
| بیشترین شباهت | شباهت کسینوسی | کلمات | کلمه‌ی مرجع |
| Sport | 0.988 | sport | Football |
| 0.1000 | technology |
| 0.085 | computer |
| basketball | 0.236 | computer | Sport |
| 0.250 | technology |
| 0.736 | basketball |
| laptop | 0.832 | basketball | computer |
| 0.999 | technology |
| 1.000 | laptop |
| technology | 0.989 | laptop | website |
| 0.991 | technology |
| 0.228 | football |