هدف تمرین:

- آشنایی با wordnet و مفاهیم آن در NLTK
- آشنایی با انواع محاسبه‌ی شباهت کلمات مبتنی بر تزاروس در NLTK
- آشنایی با محاسبه شباهت کلمات بر مبنای ماتریس term-document

مراحل:

۱. در این تمرین شما با دو فایل sport.txt و tech.txt که شامل مجموعه‌ای از متون اخبار مربوط به ورزش و تکنولوژی است کار می‌کنید. توجه کنید در طول تمرین، سوالاتی که تکمیل جدول خواسته شده است، علاوه بر تکمیل جدول و ارائه آن در گزارش، کدهای مربوط به بدست آوردن اطلاعات مربوط به جدول حتما در فایل Jupyter ارسالی وجود داشته باشد.

۲. در ابتدا می‌خواهیم با wordnet در NLTK کار کنیم. ابتدا بررسی کنید چگونه می‌توان synsetهای یک کلمه انگلیسی را بدست آورد، سپس synsetهای کلمات pharmacy، flower و dog را بدست آورید. تحلیل کنید فرمت هر synset چگونه است و هر بخش آن مربوط به چیست؟

۳. بررسی کنید چگونه می‌توان توضیحات مربوط به synsetهای یک کلمه را بدست آورد و توضیحات مربوط به تمامی synsetهای کلمات pharmacy و flower را بدست آورید.

۴. تابعی تعریف کنید که یک کلمه را به عنوان ورودی دریافت کند و lemmaهای مربوط به تمامی synsetهای کلمه دریافت شده را به صورت لیست برگرداند. سپس با استفاده از این تابع، lemma مربوط به synsetهای کلمات banks و sung را بدست آورید.

۵. حال میخواهیم با استفاده از NLTK، هم‌معنی و متضاد کلمات را بدست بیاوریم. تابعی تعریف کنید که با گرفتن یک کلمه به عنوان ورودی، هم‌معنی‌ها و متضاد آن کلمه را (مربوط به همه‌ی synsetهای یک کلمه) چاپ کند سپس هم‌معنی‌ها و متضادهای کلمات good، better، dark، long و car را چاپ کنید.
(راهنمایی: از تابع تعریف شده در بخش ٤ استفاده کنید. از متد name بر روی هر شی lemma نیز برای چاپ استفاده کنید.)

۶. در ادامه می‌خواهم Hyponyms و Hypernyms مربوط به کلمات را بدست بیاوریم. تابعی تعریف کنید که یک کلمه را به عنوان ورودی بگیرد و Hyponyms و Hypernyms مربوط به آن را برگرداند. Hyponyms و Hypernyms مربوط به کلمات car، frog و tree را بدست آورید و چاپ کنید.

۷. بررسی کنید چگونه می‌توان Hypernyms ریشه (root hypernyms) یک کلمه و همچنین اولین hypernym مشترک بین دو کلمه (Lowest common hypernyms) را در NLTK بدست آورد. سپس جداول زیر را کامل کنید. توجه کنید که در جدول اول، synset مدنظر مربوط به هر کلمه را آورده‌ایم تا در ادامه با آن synset کار کنید.

کلمه	**Synset مدنظر**
Dog	dog.n.01
Cat	cat.n.01
Car	car.n.01
Bicycle	bicycle.n.01
Tree	tree.n.01
Flower	flower.n.01
Water	water.n.01
Rainbow	rainbow.n.01

در جدول زیر Hypernyms ریشه هر کدام از کلمات را بدست آورده و در مقابل آن بنویسید.

کلمه	**Hypernyms ریشه**
Dog	
Cat	
Car	
Bicycle	
Tree	
Flower	
Water	
Rainbow	

در جدول زیر اولین hypernyms مشترک بین دو کلمه داده شده را بدست آورید و در مقابل آن بنویسید.

کلمات	اولین hypernyms مشترک بین دو کلمه
Dog and Cat	
Car and Bicycle	
Tree and Flower	
Water and Rainbow	

۸. در ادامه می‌خواهیم شباهت بین کلمات را بدست بیاوریم. ابتدا با محاسبه شباهت مبتنی بر تزاروس آغاز می‌کنیم.

a. بررسی کنید چگونه می‌توان path similarity بین دو کلمه را بدست آورد.
b. بررسی کنید چگونه می‌توان resnik similarity بین دو کلمه را بدست آورد.
c. بررسی کنید چگونه می‌توان jiang-conrath similarity بین دو کلمه را بدست آورد.
d. برای هرکدام از شباهت‌های گفته شده در قسمت a، b و c، جدول زیر را کامل کنید. هر خانه از جدول، شباهت محاسبه شده از طریق شیوه گفته شده است. سپس به ازای هر کلمه، شبیه‌ترین کلمه به آن (غیر از شباهت آن کلمه با خودش) را مشخص کنید. (در نهایت باید سه جدول مانند جدول زیر در گزارش وجود داشته باشد.)

	Dog	Cat	Car	Bicycle	Tree	Flower	Water	Rainbow
Dog								
Cat								
Car								
Bicycle								
Tree								
Flower								
Water								
Rainbow								

۹. در ادامه میخواهیم با محاسبه ماتریس term-document بر روی کلمات دو فایل tech.txt و sport.txt به محاسبه شباهت بین کلمات بپردازیم.

a. پس از باز کردن این دو فایل، آنها را به کلمات آنها تجزیه‌کنید به گونه‌ای که علامت‌های اضافه حذف شوند.
b. Stopwordهای موجود و حروف تک کاراکتری را حذف کنید.
c. Typeهای موجود در هر متن و تعداد هرکدام را بدست آورید.
d. تعداد تکرار کلمه football و کلمه computer در هر متن را چاپ کنید.
e. ماتریس term-document کلمات را بدست آورید. **(راهنمایی: به ازای هر کلمه یک بردار بدست می‌آید که المان اول آن تعداد تکرار کلمه در متن sport و المان دوم آن تعداد تکرار کلمه در متن tech است.)**
f. در نهایت شباهت کسینوسی بین کلمات زیر را محاسبه کرده و جدول زیر را کامل کنید و کلمه‌ای که بیشترین شباهت را به کلمه مرجع دارد مشخص کنید.

کلمه مرجع	کلمات	شباهت کسینوسی
football	sport	
	technology	
	computer	
sport	computer	
	technology	
	basketball	
computer	basketball	
	technology	
	laptop	
website	laptop	
	technology	
	football	

- نتیجه نهایی را در قالب یک فایل Jupyter notebook و یک گزارش PDF در یک فایل فشرده zip که نام آن با فرمت NLP_Name_Family_Ex5 (به جای Name اسم و به جای Family نام‌خانوادگی خودتان را قرار دهید) است قراردهید و به ایمیل abedi.a1997@gmail.com ارسال کنید.

موفق باشید