تکلیف شماره 2

گروه شماره: 8

اعضاي گروه:

## قسمت اول – تبديل به txt

### در فايل "converter.py" :

ابتدا كتابخانه "docx2txt" را وارد ميكنيم سپس با استفاده از آن فايل مورد نظر (paragraph.docx) را تبديل به متن و با تابع (.write) متن را در فايل (allGroupstext.txt) مينويسيم.

خروجي اين قسمت در فايل " allGroupstext" قابل مشاهده است.

### در فايل "HW2\_gorup8.py" :

براي مراحل بعدي نياز به كنابخانه هاي زير داريم:

1. "hazm" براي ابزارهاي پردازش متن فارسي
2. " xlwt" براي نوشتن فايل خروجي با فرمت ".xls" فايل اكسل

## قسمت دوم - پیش پردازشهای لازم ( حذف فضاهای خالی اضافه، تبدیل فاصله به نیم فاصله، ایجاد فاصله مناسب بعد از ویرگول یا نقطه، و...) انجام دهید.

در تابع " normalizing" ابتدا متن را خوانده، يك شي از " Normalizer" ايجاد و سپس از:

* " normalizer.affix\_spacing" براي نرمال سازي فاصله­هاي بين پسوندهها و پيشوندها
* " normalizer.character\_refinement" براي اصلاح كاف و ياي عربي
* " normalizer.punctuation\_spacing" براي اصلاح پرانتز ها در متن

خروجي اين قسمت در فايل " group8text.txt" قابل مشاهده است.

## قسمت سوم - برای هر پاراگراف جدول مجزا ساخته و اطلاعات زیر را در جدول قرار دهید:

## تعداد جملات، تعداد کل کلمات، تعداد فعل ها، تعداد اسم ها

* در تابع " sentences\_count" اول يك شي "Workbook" براي فايل خروجي جدول­ها ايجاد ميكنيم بعد از توكن سازي با استفاده از " POSTagger" هر كدام از توكن ها را تگ ميكنيم. اگر آرگومان انديس "1" مساوي با "V" باشد در ليست فعل­ها و اگر مساوي با "N" در ليست نام­ها قرار ميدهيم.
* براي شماره پاراگراف " i + 1"
* براي تعداد جملات " len(sentence\_tokenizer.tokenize(paragraphs\_list[i]))"
* براي تعداد كلمات " len(token\_list)"
* براي تعداد فعل­ها " len(verb\_list)"
* براي تعداد اسم­ها " len(noun\_list)"

خروجي اين قسمت در فايل "HW2.xls" قابل مشاهده است.

## قسمت چهارم – 5 جمله دلخواه

## همه توکن ها را پیدا کنید. ریشه فعل ها و ریشه کلمات را پیدا کنید. لیست کلماتی که به درستی در خروجی مراحل قبل ظاهر شده اند را پیدا کنید

* درتابع " random\_selection" از ليست پاراگراف­هاي ورودي 5 جمله را بصورت رندوم انتخاب ميكنيم و در همينجا جملات را توكنايز ميكنيم
* خروجي تابع بالا را به تابع " find\_tokens" ميدهيم كه كلمات داخل هر جمله را توكن سازي كرده و در يك ليست برميگرداند
* ليست توكن­ها را به تابع " risheYab" ميدهيم يك شي " Stemmer" براي ريشه­يابي و يك شي " Lemmatizer" براي بنيابي ايجاد ميكنيم، براي فعل­ها بنيابي و براي بقيه كلمات ريشه يابي ميكنيم.