تکلیف شماره 2

شماره گروه: 8

اعضای گروه: آرمین افتخاری(9622762033)، محمد رضا پوررضا (9612762592)، رضا برزگر طرقبه (9622762384)، سبحان مرادیان(9622762066) و محمد سلیمان بهزاد(9622762453)

لینک "github": https://github.com/rezaBarzgar/IR_HW2.git

تاریخ: 08-10-2020

قسمت اول – تبدیل به txt

در فایل "converter.py" :

ابتدا کتابخانه "docx2txt" را وارد میکنیم سپس با استفاده از آن فایل مورد نظر (paragraph.docx) را تبدیل به متن و با تابع (write.) متن را در فایل (allGroupstext.txt) مینویسیم.

خروجی این قسمت در فایل "allGroupstext " قابل مشاهده است.

در فایل "HW2_gorup8.py" :

برای مراحل بعدی نیاز به کتابخانه های زیر داریم:

1. "hazm" برای ابزارهای پردازش متن فارسی
2. "xlwt " برای نوشتن فایل خروجی با فرمت ".xls" فایل اکسل

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، یک شی از "Normalizer " ایجاد و سپس از:

- ✓ "normalizer.affix_spacing " برای نرمال سازی فاصله‌های بین پسوندها و پیشوندها
- ✓ "normalizer.character_refinement " برای اصلاح کاف و یای عربی، نبدیل اعداد انگلیسی به فارسی
- ✓ "normalizer.punctuation_spacing " برای اصلاح علامه گذاری‌ها در متن و فاصله‌های بین علامه‌ها و قسمت قبل و بعد

خروجی این قسمت در فایل "group8NormalizedText.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 یک شی "Workbook" برای فایل خروجی جدول‌ها ایجاد میکنیم بعد از توکن سازی با استفاده از "POSTagger " هر کدام از توکن ها را تگ میکنیم. اگر آرگومان اندیس "1" مساوی با "V" باشد در لیست فعل‌ها و اگر مساوی با "N" در لیست نام‌ها قرار میدهیم.
- ✓ برای شماره پاراگراف "i + 1 "
- ✓ برای تعداد جملات "len(sentence_tokenizer.tokenize(paragraphs_list[i])) "
- ✓ برای تعداد کلمات "len(token_list) "
- ✓ برای تعداد فعل‌ها "len(verb_list) "
- ✓ برای تعداد اسم‌ها "len(noun_list) "

خروجی این قسمت در فایل "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 لیست پاراگراف‌های ورودی 5 جمله را بصورت رندوم انتخاب میکنیم و در همینجا جملات را توکنایز میکنیم
- ✓ خروجی تابع بالا را به تابع "find_tokens " میدهیم که کلمات داخل هر جمله را توکن سازی کرده و در یک لیست برمیگرداند
- ✓ لیست توکن‌ها را به تابع "risheYab " میدهیم یک شی "Stemmer " برای ریشه‌یابی و یک شی "Lemmatizer " برای بنیابی ایجاد میکنیم، برای فعل‌ها بنیابی و برای بقیه کلمات ریشه یابی میکنیم.

در خروجی این مرحله مشاهده میکنیم که بعضی از کلمات را که " normalizer " کتابخانه هضم به درستی نتوانسته است اصلاح کند در قسمت ریشه‌یابی و بنیابی نیز شناخته نشده و باعث بروز خطا در عملکرد آن شده است.

به عنوان مثال فعل‌های 'میکنند'و 'میدهند' به دلیل اینکه به شکل درست نرمالسازی نشده اند -> "می‌کنند" و '"می‌دهند'" باعث عدم شناخت توسط lemmatizer شده اند.

#sentence

['هست', 'میکنند', 'میدهند', 'تعداد', 'تقریبا', 'هنگامیکه', 'در', 'اتومبیل', 'تلفن', 'همراه', 'خود', 'را', 'به', 'دستگاه', 'پخ', '#'] 'آن', 'وصل', 'و', 'به', 'پادکست', 'گو', '.']

#sentence

['میکند', 'میشوند', 'داشت#دار', 'داشت#دار', '#است', 'کند', 'به', 'عنو', 'مثال', 'وقت', 'کاربر', 'برا', 'خرید', 'یک', 'لپتاپ', '] 'به', 'فروشگاه', 'اینترنت', 'مراجعه', 'پس', 'از', 'انتخاب', 'لپتاپ', 'موردنظر', 'در', 'سبد', 'خرید', 'محصولات', 'مرتبط', 'با', 'این', 'کالا', 'مانند', 'پوش', 'محافظ', '،', 'موس', '،', 'کیف', 'لپتاپ', 'و', '...', 'ه', 'به', 'کاربر', 'پیشنهاد', 'که', 'در', 'این', 'صور', 'کاربر', 'خریدار', 'به', 'عل', 'هیجان', 'که', 'از', 'خرید', 'لپتاپ', 'و', 'صدالبته', 'نیاز', 'که', 'به', 'این', 'اقا', 'جانب', 'وجود', 'ممکن', 'این', 'لواز', 'جانب', 'را', 'ه', 'خریدار', '.']

#sentence

['بوده', 'رساند#رسان', 'همین', 'مسئله', 'نکته', 'برا', 'خیل', 'از', 'کسبو', 'کار', 'تا', 'تبلیغ', 'هدفمند', 'خود', 'را', 'از', 'طریق', '] 'این', 'پادکست', 'یا', 'برنامه', '', 'رادیو', 'آنلاین', 'به', 'گو', 'مخاطبین', '.']

#sentence

['میرود', 'داشت#دار', 'داشت#دار', 'انتظار', 'که', 'صرف', 'هزینه', 'جه', 'انجا', 'این', 'نوع', 'از', 'تبلیغ', 'رشد', 'فوقالعاده', '] 'بالا', 'در', 'سال', '۲۰۲۰', 'میاد', 'به', 'طوریکه', 'از', '۲۷', 'میلیارد', 'دلار', 'در', 'سال', '۲۰۱۹', 'تا', '۳۱', 'میلیارد', 'دلار', 'افزا', 'در', 'سال', '۲۰۲۰', '', '.']

#sentence

['بود#باش', '#است', 'داشت#دار', '#هست', 'اگرچه', 'تأثیر', 'صدا', 'زمان', 'که', 'همراه', 'با', 'تصویر', 'یا', 'ویدئو', 'چندین', 'برابر', '،', 'اما', 'هنوز', 'ه', 'محتوا', 'که', 'فقط', 'صدا', '(', 'مثل', 'پادکست', ')', 'یا', 'تبلیغ', 'رادیو', 'به', 'نوبه', 'خود', 'برا', 'قشر', 'خاص', 'از', 'مرد', 'خود', 'تأثیرگذار', '.']