**تمرین سوم درس پردازش زبان‌های طبیعی**

**استاد درس:** دکتر حمیدرضا برادران کاشانی

**دستیاران آموزشی:** آیین کوپایی – یاسین فخار

تاریخ بارگذاری تمرین: ۱۴۰۲/۰۹/۲۴

تاریخ تحویل تمرین: ۱۴۰۲/۱۰/۰۵

دانشگاه اصفهان

## ۱- آشنایی با مفهوم chunking

۱-۱- بررسی کنید مفهوم chunking در پردازش زبان چیست؟

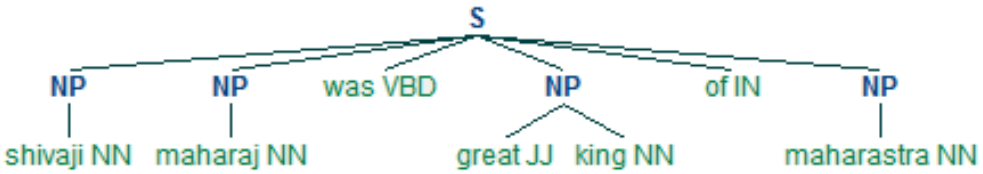
۱-۲- برای جمله زیر، word tokenization و POS tagging را انجام دهید و خروجی را نمایش دهید.

Natural language processing is fun! This text is a sample text.

۱-۳- با استفاده از یک عبارت منظم متشکل از قوانینی که نشان می‌دهد جملات چگونه باید تکه تکه شوند، noun phrase chunking را بر روی نتیجه قسمت ۱-۲ اعمال کنید و نتیجه را در خروجی چاپ کنید (یا در فایل گزارش نمایش دهید). از قاعده زیر برای این کار استفاده کنید.

```
chunk_grammar = r"""
    NP: {<DT>?<JJ>*<NN>}  # NP: Noun Phrase
"""
```

۱-۴- درخت حاصل از قسمت ۱-۳ را به صورت بصری نمایش دهید. خروجی باید از لحاظ بصری مانند تصویر زیر باشد.




۱-۵- verb phrase chunking را برای جمله زیر اعمال کنید و خروجی را نشان دهید. سپس درخت حاصل را به صورت بصری نمایش دهید.

She decided to take a stroll in the park.

## ۲- آشنایی با مفهوم IOB encoding و تشخیص موجودیت‌های نامدار

۲-۱- با استفاده از عبارت منظم داده شده در قسمت ۱-۳ و اعمال آن بر روی مجموعه داده Rock.txt، عمل IOB encoding را بر روی این مجموعه داده اعمال کنید و نتیجه را نمایش دهید.

۲-۲- با استفاده از مدل از پیش آموزش داده شده Stanford برای NER، موجودیت‌های نامدار را از درون مجموعه داده Rock.txt استخراج کنید. سپس نام افراد و مکان‌ها را در خروجی چاپ کنید (یا در فایل گزارش نمایش دهید).

۲-۳- با استفاده از کتابخانه Spacy موجودیت‌های نامدار را از درون مجموعه داده Rock.txt استخراج کنید. سپس نام افراد و مکان‌ها را در خروجی چاپ کنید (یا در فایل گزارش نمایش دهید).

۲-۴- نتیجه حاصل از قسمت ۲-۲ و ۲-۳ را با یکدیگر مقایسه کنید.

## ۳- استخراج کلمات کلیدی و خلاصه سازی متن

۳-۱- stopword ها و علایم نگارشی را از درون مجموعه داده Rock.txt حذف کنید.

۳-۲- با استفاده از روش tf-idf ، ده کلمه کلیدی با بیشترین اهمیت را از درون متن پیش پردازش شده استخراج کنید و نتیجه را در خروجی چاپ کنید (یا در فایل گزارش نمایش دهید).

۳-۳- با استفاده از کلمات کلیدی بدست آمده از مرحله قبل متن را در ۵ جمله خلاصه سازی کنید. نتیجه را در خروجی چاپ کنید (یا در فایل گزارش نمایش دهید).

۳-۴- در مورد الگوریتم TextRank مطالعه کنید و توضیح مختصری از آن ارائه دهید.

۳-۵- متن را یک بار دیگر با استفاده از الگوریتم TextRank در ۵ جمله خلاصه سازی کنید. نتیجه را در خروجی چاپ کنید (یا در فایل گزارش نمایش دهید).

۳-۶- نتیجه قسمت‌های ۳-۳ و ۳-۵ را با یکدیگر مقایسه کنید.

## ۴- ساخت سیستم توصیه گر

مجموعه داده مورد استفاده در این بخش، مجموعه داده TMDB است که حاوی اطلاعات فیلم‌ها است. این مجموعه داده در فایل پیوست با عنوان tmdb_5000_movies.csv قرار داده شده است. در این سیستم توصیه گر قصد داریم با دریافت اسم یک فیلم که در مجموعه داده موجود است، فیلم‌های شبیه به آن را بیابیم. برای این کار از ستون مربوط به ژانر (genres) و ستون کلمات کلیدی (keywords) استفاده می‌کنیم. به این منظور با استفاده از روش tf-idf ، این ویژگی‌ها را به بردار تبدیل کنید تا آماده استفاده در سیستم توصیه گر شوند. سپس از طریق شباهت کسینوسی، میزان شباهت فیلم مورد جستجو را با

تمامی فیلم‌های موجود در مجموعه داده محاسبه کنید و ۵ فیلم با بیشترین شباهت را برگردانید. این سیستم را بر روی فیلم‌های Flywheel ، Mortal Kombat و Frozen اعمال کنید و نتیجه را گزارش دهید.

---

## نکات تحویل

۱- پاسخ خود را در پوشه ای به اسم NLP_NAME_FAMILY_HW3 و در قالب zip بارگذاری نمایید.

۲- این پوشه باید حاوی موارد زیر باشد:

- کد نوشته شده در قالب یک فایل jupyter notebook یا py.
- فایل گزارش در قالب یک فایل PDF

۳- لازم به ذکر است که رعایت قوانین نگارشی حائز اهمیت است.

۴- پاسخ خود را به ایمیل aein.koopaei@gmail.com ارسال کنید.