



دانشگاه اصفهان

تمرین اول درس مبانی پردازش زبان و گفتار  
استاد درس: دکتر حمیدرضا برادران کاشانی  
دستیاران آموزشی: آیین کوپایی – هاجر مظاهری

تاریخ بارگذاری تمرین: ۱۴۰۲/۱۲/۲۲

تاریخ تحویل تمرین: ۱۴۰۳/۰۱/۱۵

## اهداف تمرین:

- ۱- آشنایی با مفاهیم موجود در پیش پردازش متن به دو زبان فارسی و انگلیسی
- ۲- بررسی سه وظیفه خلاصه سازی متن، تحلیل احساسات و ترجمه ماشینی در پردازش زبان طبیعی
- ۳- پیاده سازی الگوریتمی برای تصحیح خطاهای املایی

## بخش اول: پرسش ها

به سوالات زیر پاسخ دهید.

- ۱- ابهام در زبان انسان یکی از دلایل سخت بودن NLP است. ابهام در چهار سطح توضیح داده شده است. این چهار سطح را توضیح دهید و برای هر کدام در زبان فارسی مثال بزنید.
- ۲- الگوریتم Maximum Matching و کاربرد آن را توضیح دهید و مثالی برای آن ارائه دهید.
- ۳- لم سازی<sup>۱</sup> و ریشه یابی<sup>۲</sup> را با ذکر مثال هایی در زبان فارسی توضیح دهید.

## بخش دوم: مقاله خوانی

در مورد هریک از موضوعات زیر تحقیق کنید و توضیح مختصری برای هریک ارائه دهید.

- ۱- Text Summarization
- ۲- Sentiment Analysis
- ۳- Machine Translation

<sup>۱</sup> Lemmatization

<sup>۲</sup> Stemming

یک مقاله در مورد یکی از وظایف داده شده بخوانید و توضیحی برای آن بنویسید. مقاله انتخابی شما باید بین سال‌های ۲۰۲۱ تا ۲۰۲۴ باشد و در یکی از دو کنفرانس زیر پذیرفته شده باشد:

۱- ACL

۲- EMNLP

توضیحات شما باید شامل ایده مقاله انتخابی و مسئله مورد بررسی در این مقاله، مدل پیشنهادی و نتایج مقاله باشد. (حتماً نام مقاله انتخاب شده را بنویسید.)

## بخش سوم: پیش پردازش

قدم اول در وظایف NLP پیش پردازش متن است. در این تمرین با دو مجموعه داده فارسی (hp\_fa.txt) و انگلیسی (hp\_en.txt) کار می‌کنیم. برای مجموعه داده فارسی از کتابخانه Hazm و برای مجموعه داده انگلیسی از کتابخانه nltk استفاده کنید.

۱- مراحل زیر را بر روی مجموعه داده فارسی اعمال کنید.

۱-۱- فضاهای خالی اضافه را حذف کنید، متن را به جملات آن تجزیه کنید و سپس متن را normalize کنید.

۲-۱- جملات را به کلمات آن توکن‌بندی کنید.

۳-۱- علائم نگارشی را حذف کنید.

۴-۱- ابتدا توضیح مختصری در ارتباط با مفهوم و دلیل حذف ایست واژه‌ها<sup>۱</sup> در پیش پردازش متون ارائه نمایید و سپس آنها را از درون متن حذف نمایید.

۵-۱- ایموژی‌های موجود در متن را حذف کنید.

۶-۱- فرآیند لم‌سازی را بر روی متن اعمال کنید.

۲- مراحل زیر را بر روی مجموعه داده انگلیسی اعمال کنید.

۱-۲- فضاهای خالی اضافه را حذف کنید و متن را به جملات آن تجزیه کنید.

۲-۲- حروف بزرگ را به حروف کوچک تبدیل کنید.

۳-۲- جملات را به کلمات آن توکن‌بندی کنید.

۴-۲- اعداد و URL ها را حذف کنید.

---

<sup>۱</sup> Stopword

۲-۵- علائم نگارشی و ایست وازه‌ها را حذف کنید.

۲-۶- ابر کلمات<sup>۱</sup> را برای متن پیش پردازش شده رسم کنید.

### بخش چهارم: تصحیح خطاهای املائی

هدف از این بخش پیاده‌سازی الگوریتمی برای تصحیح خطاهای املائی است. این الگوریتم یک جمله را به عنوان ورودی دریافت می‌کند و بهترین پیشنهادها را برای هر کلمه غلط املائی به عنوان خروجی ارائه می‌دهد. برای این بخش از فایل Vocabulary.txt به عنوان دایره لغات استفاده کنید.

### مراحل انجام کار:

**انتخاب جمله:** اولین جمله از چکیده مقاله انتخابی در بخش دوم تمرین را انتخاب کنید.

**ایجاد جمله غلط املائی:** جمله انتخاب شده را در این [وب سایت](#) کپی کنید و غلط‌های املائی را ایجاد کنید. (برای استفاده از این وب سایت باید از فیلترشکن استفاده کنید)

**استفاده از جمله غلط املائی به عنوان ورودی الگوریتم:** الگوریتم تصحیح خطا را بر روی جمله غلط املائی اجرا کنید و خروجی آن را نمایش دهید.

---

### نکات تحویل

۱- پاسخ خود را در پوشه ای به اسم NLP\_NAME\_FAMILY\_HW1 و در قالب zip بارگذاری نمایید.

۲- این پوشه باید حاوی موارد زیر باشد:

- کد نوشته شده در قالب یک فایل jupyter notebook
- فایل گزارش فنی در قالب یک فایل PDF

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

---

<sup>۱</sup> Wordcloud