تمرین اول پردازش زبان طبیعی

توجه:

- **کد باید فقط در زبان پایتون باشد.**
- **استفاده از کتابخانه‌های از پیش آماده مجاز است. در گزارش کتابخانه ها ذکر شوند.**
- **لطفاً علاوه بر ضمیمه کردن کد، نتایج را تحلیل و در فایل گزارش خود ضمیمه کنید.**
- **فایل گزارش به فرمت pdf و به زبان فارسی باشد.**

فایل مجموعه داده IMDB، برای مسئله تشخیص احساسات [1]، را از این لینک دانلود کنید. مجموعه داده IMDB به سه بخش داده آموزشی، اعتبار سنجی و تست تقسیم شده است. برچسب هر نمونه عدد صفر یا یک است، که به ترتیب نشانگر احساسات منفی و مثبت اند. (انتخاب نحوه هموار سازی بر عهده شما است. همچنین اگر پیاده سازی را از صفر انجام می دهید، با لگاریتم احتمالات کار کنید.)

۱. ابتدا برای کاهش هزینه محاسباتی تنها ۵۰۰ کارکتر اول هر نمونه را در نظر بگیرید. سپس پیش پردازش های لازم را بر روی متن خام انجام دهید. (واحد سازی و نرمال سازی، که نرمال سازی حداقل باید شامل کوچک کردن حروف انگلیسی، stemming و حذف stopwords، علائم و اعداد باشد.)

۲. مدل های UniGram، BiGram، TriGram و 4Gram بر روی داده های آموزشی مربوط به هر کلاس (کلاس احساس مثبت و منفی) پیاده سازی کنید (در انتها ۸ مدل خواهید داشت).

۳. برای هر ۸ مدل Ngram، ۵ واحدی که بیشترین تکرار را داشته اند با تعداد تکرار شان را تعیین کنید.

۴. معیار Perplexity را برای هر مدل با استفاده از مجموعه داده اعتبارسنجی مشخص کنید. (هر مدل بر روی کلاس متناظر خود، احساس منفی یا مثبت، ارزیابی شود.) آیا الگویی مشاهده می شود؟ اگر بله در گزارش توضیح دهید.

۵. در این بخش قصد داریم که مجموعه داده دسته بندی کنیم. فرآیند به این صورت است که برای هر نمونه یا استفاده از NGram انتخابی احتمال آن ها را بدست می آوریم، و اگر احتمال کلاس مثبت بیشتر بود، نمونه را به کلاس مثبت انتساب می دهیم و در غیر این صورت به کلاس منفی منتسب می شود.

$$y^* = argmax_y Pr(y|x) \propto argmax_y Pr(x|y)Pr(y) = argmax_y Pr(x|y)$$

هر مدل NGram را بر روی داده اعتبار سنجی بررسی کنید، و گزارش دهید که کدامیک از ۴ نوع Ngram دقت بهتری می دهند. آیا الگویی مشاهده می شود؟ اگر بله در گزارش توضیح دهید.

۶. مدل بهینه را انتخاب کنید و بر روی داده تست پیشبینی را انجام دهید و آن را در فرمت مشابه "y_train" با نام "y_test.txt" ذخیره و ضمیمه کنید.

[1] Sentiment Analysis

```python
import pickle
import numpy as np

with open(data_path/"x_train.pickle","rb") as f:
  x_train = pickle.load(f)

y_train = np.loadtxt(data_path/"y_train.txt",dtype="int32")
```

شکل ۱ : نحوه خواندن داده ها