تمرین اول – درس پردازش زبان طبیعی آماری **دکتر ممتازی** **زمستان ۹۷ – دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر**	

1. درجدول زیر نحوه اعمال نمره منفی برای تاخیر در ارسال تمرین ها آورده شده است:

تاخیر (روز)	میزان نمره منفی
از 1 الی 2	هر روز 5٪
از 3 الی 6	هر روز 10٪

توجه داشته باشید در صورت تاخیر بین 7 تا 14 روز، نمره تمرین از 50٪ محاسبه شده و پس از آن نمره ای تعلق نمی گیرد.

2. هدف از انجام تمرین ها یادگیری عمیق تر مطالب درسی است. در نتیجه هرگونه کپی برداری موجب کسر نمره خواهد شد.

3. تا **ساعت 23:15 روز جمعه ۱۶ فروردین** فرصت دارید تمرین را در مودل بارگذاری کنید. تمام فایل های پیاده سازی را به همراه فایل pdf مربوط به گزارش تمرین را در یک فایل فشرده قرار دهید. نام فایل نهایی را شماره دانشجویی خود (برای مثال HW1_97131024) قرار دهید.

4. زبان برنامه نویسی برای انجام تمرین: پایتون، جاوا و یا متلب در نظر گرفته شده است.

5. برنامه های نوشته شده خوانا باشد و کامنت گذاری مناسب باشد (طوری که روند کار کاملا مشخص باشد).

6. در صورت وجود هر گونه سوال می توانید از طریق ایمیل با تدریس یاران درس در ارتباط باشید:

javadforough@gmail.com و amin.ghsm@aut.ac.ir

مساله: دسته بندی[1] متون

بخش اول: استخراج ویژگی[2]

در این تمرین قصد داریم عمل دسته بندی متون را به روش بیز[3] با استفاده از مدل های Unigram و Bigram انجام دهیم. برای انجام این کار بایستی از کلیه کلمات موجود در پیکره آموزشی به عنوان ویژگی استفاده شود .اما به دلیل حجم محاسبات مورد نیاز ما می‌خواهیم یک مرحله انتخاب ویژگی پیش از انجام دسته بندی صورت گیرد.

بدین منظور، پیکره[4] کوتاه شده همشهری در اختیار شما قرار می گیرد. این پیکره در یک فایل تدوین شده است که هر خط فایل یک سند[5] را شامل می شود. در ابتدای هر خط نیز برچسب[6] آن سند نوشته شده است و با علامت @ از متن جدا شده است.

در این راستا می خواهیم **تنها برای حالت unigram** انتخاب ویژگی با استفاده از الگوریتم information gain صورت گیرد.

خروجی مورد انتظار به منظور گزارش:

- ۲۰۰ عدد از بهترین ویژگی های بدست آمده به همراه امتیاز مربوط به هر یک از ویژگی ها

بخش دوم: دسته بندی

در این بخش برای حالت unigram دسته بندی را به دو صورت زیر انجام دهید:

الف) با استفاده از ویژگی های بدست آمده در مرحله ی قبل

ب) عدم استفاده از انتخاب ویژگی

همچنین برای حالت bigram بدون انجام انتخاب ویژگی عمل دسته بندی صورت گیرد.

لازم به ذکر است که در این مرحله بایستی از هموارسازی[7] به روش الگوریتم Absolute Discounting به منظور حل مشکل الگوهای دیده نشده[8] استفاده شود. این هموارسازی را با استفاده از مقادیر 0.1 ، 0.3 و 0.5 برای δ انجام دهید.

در این بخش شما لازم هست موارد زیر را در گزارش مربوطه ارایه نمایید:

- ماتریس سردرگمی[9] مربوط به تست انجام شده با دسته بند بیز با استفاده از مدل های Unigram و Bigram (همزمان با اعمال Absolute Discounting برای هر سه مقدار δ)
- بررسی تأثیر مقادیر مختلف δ بر روی عملکرد مدل ها

---

[1] Classification
[2] Feature
[3] Bayes
[4] Corpus
[5] Document
[6] Label
[7] Smoothing
[8] Unseen Patterns
[9] Confusion Matrix

- ارزیابی و تحلیل دسته بندی های انجام شده با استفاده از معیارهای[1] زیر برای هر دو دسته بند
  - Precision
  - Recall
  - F1-measure
  - Micro Average و Macro Average برای هر سه معیار بالا

لازم به ذکر است که لینک پیکره همشهری (شامل داده آموزش و داده آزمون) در زیر آورده شده است:
https://www.dropbox.com/s/zkmozli0mmentpy/HAM-Train-Test.zip?dl=0

موفق و پیروز باشید

[1] Metric