تمرین اول – درس پردازش زبان طبیعی آماری دکتر ممتازی ترم زمستان ۱۳۹۷-۱۳۹۶ – دانشکده کامپیوتر، دانشگاه صنعتی امیرکبیر زمان تحویل: ۱۸ فروردین ۹۷	

۱. در جدول زیر نحوه اعمال نمره منفی برای تاخیر در ارسال تمرین‌ها آورده شده است:

تاخیر (روز)	میزان نمره منفی
از ۱ الی ۲	هر روز ۵٪
از ۳ الی ۶	هر روز ۱۰٪

توجه داشته باشید در صورت تاخیر بین ۷ تا ۱۴ روز، نمره تمرین از ۵۰٪ محاسبه شده و پس از نمره‌ای تعلق نمی‌گیرد.

۲. هدف از انجام تمرین‌ها، یادگیری عمیق‌تر مطالب درسی است. در نتیجه هرگونه کپی‌برداری موجب کسر نمره خواهد شد.

۳. تا ساعت **۲۳:۵۵** روز **۱۸ فروردین** فرصت دارید تا تمرین را در مودل بارگذاری کنید. تمام فایل‌های پیاده‌سازی را به همراه فایل، pdf مربوط به گزارش تمرین، در یک فایل فشرده قرار دهید. نام فایل نهایی را شماره دانشجویی خود قرار دهید. (برای مثال HW1_95131105)

۴. زبان برنامه‌نویسی برای انجام تمارین پایتون، جاوا و یا متلب در نظر گرفته شده است.

۵. برنامه‌های نوشته شده خوانا باشد و کامنت‌گذاری مناسب باشد (طوریکه روند کار کاملا مشخص باشد).

۶. در فایل گزارش درباره کد توضیح ندهید! فقط کافی است نتیجه به دست آمده را در گزارش قرار داده و مختصرا آن را تحلیل نمایید (مثلا در این تمرین، بر روی نتایج حاصله تحلیل‌هایی انجام دهید که تفاوت‌های این سه روش را نشان دهد).

۷. در صورت وجود هرگونه سوال می‌توانید از طریق ایمیل با تدریس‌یاران درس در ارتباط باشید:
a.heidarnasab@aut.ac.ir rahbararman@aut.ac.ir

بخش اول

با توجه به مطالب تدریس شده در جلسات آموزشی، یکی از روش‌های نمایش اسناد تبدیل آنها به بردار ویژگی‌ها است. یک ایده در بردار ویژگی استفاده از بردار لغات است. اما با توجه به حجیم بودن این بردار علی رغم بی‌اثر بودن برخی کلمات مانند کلمات تابعی، ایده انتخاب ویژگی در بهبود کیفیت و ارزیابی کارایی هر یک از این ویژگی‌ها مطرح است. بر این اساس در جلسات آموزشی سه الگوریتم Information Gain، Mutual Information و χ-Square برای انتخاب ویژگی‌ها پیشنهاد شد. در این بخش از تمرین شما باید با استفاده از الگوریتم‌های بیان‌شده، انتخاب ویژگی را برای اسناد کلاس‌بندی شده در یک پیکره برچسب‌دار فارسی انجام دهید.

برای این منظور پیکره کوتاه‌شده همشهری در اختیار شما قرار داده می‌شود. این پیکره در یک فایل تدوین شده است که هر خط فایل یک سند را شامل می‌شود. در ابتدای هر خط نیز دامنه (کلاس) آن سند نوشته‌شده است و با علامت @ از متن جدا شده است.

لازم است که شما بر اساس هر یک از سه الگوریتم مذکور یک بردار ویژگی بهینه با حداکثر امتیاز تولید نمایید که در سه فایل مجزا باشند. طول این بردار را ۱۰۰ کلمه در نظر بگیرید. برای هر کلمه از بردار ویژگی ۱۰۰ تایی لازم است که امتیاز اختصاص یافته توسط الگوریتم بیان شود. از طرف دیگر دامنه‌ای که هر یک از کلمات بیشترین امتیاز را در آن کسب کرده به همراه مقدار امتیاز آن نیز برای کلمه مشخص نمایید.

بنابراین خروجی برای هر الگوریتم بصورت زیر خواهد بود:

	Score of algorithm	Main Domain	Score of the main Domain
W_1			
W_2			
.			
.			
.			
W_{100}			

W_i ها کلمات بردار ویژگی هستند.

Score of algorithm امتیاز این کلمه در الگوریتم انتخاب ویژگی است.

Main Domain اسم دامنه‌ای که کلمه W_i در آن بیشترین امتیاز را در آن بدست آورده است.

Score of the main Domain امتیاز دامنه‌ای است که W_i در آن بیشترین امتیاز را بدست آورده است.

لازم به ذکر است که Main Domain و Score of Main Domain برای Information Gain وجود ندارند و طبعا نیاز به محاسبه آن‌ها نیست.

بخش دوم

هدف اصلی از انتخاب ویژگی دستیابی به ویژگی‌هایی است که با کمک آن‌ها بتوان عملیاتی مانند دسته‌بندی و خوشه‌بندی را انجام داد. در این بخش از تمرین لازم است با توجه به بردارهای ویژگی به دست آمده از قسمت قبل عمل دسته‌بندی را با الگوریتم دسته‌بندی SVM انجام دهید. در این بخش نیازی به پیاده‌سازی الگوریتم SVM نیست و می‌توانید از ابزارهای آماده مانند SciKit Learn پایتون یا ابزارهای وابسته به WEKA استفاده کنید. برای این منظور نیاز است ۴ آزمایش مختلف اجرا نمایید. آزمایش اول از ۱۰۰۰ کلمه پرکاربرد در کل اسناد به‌عنوان ویژگی استفاده کنید و هیچ یک از روش‌های مورد استفاده در بخش اول را اجرا نکنید. در سه آزمایش بعد در هر آزمایش یکی از الگوریتم‌های انتخاب ویژگی بخش اول را مورد استفاده قراردهید. در این حالت باید برای هر سند یک بردار ویژگی با توجه به ویژگی‌های انتخابی در قسمت قبل تولید گردد. برای تولید بردار از هر سند به صورت زیر عمل می‌کنیم:

فرض کنید طول بردار برابر ۵ باشد(یعنی ۵ کلمه به عنوان ویژگی انتخاب کرده‌ایم). پس هر سند مانند شکل زیر به صورت یک بردار به طول ۵ بازنمایی می‌شود:

۳	۰	۴	۱	۱۰

فرض کنید در این سند از کلمه (ویژگی) اول ۳ تکرار، کلمه دوم ۰ تکرار، کلمه سوم ۴ تکرار، کلمه چهارم ۱ تکرار و کلمه پنجم ۱۰ تکرار موجود باشد.

مولفه های این بردار شامل تعداد تکرار هر کلمه در سند می‌باشد. سپس این بردار را به تعداد کل کلمات تقسیم کرده تا مقادیر نرمال شوند (بین صفر و یک قرار گیرند).

در این صورت بردار نهایی به صورت زیر خواهد بود:

۳ / ۱۸	۰	۴ / ۱۸	۱ / ۱۸	۱ / ۱۸

به منظور ارزیابی عمل دسته‌بندی، ۵-fold cross Validation را بر روی مجموعه داده‌ای که در اختیار دارید انجام داده و Confusion Matrix حاصل از میانگین ۵ فولد را گزارش کنید.

موفق و پیروز باشید