دانشکده مهندسی کامپیوتر

دکتر بهروز مینایی

بهار ۱۴۰۱

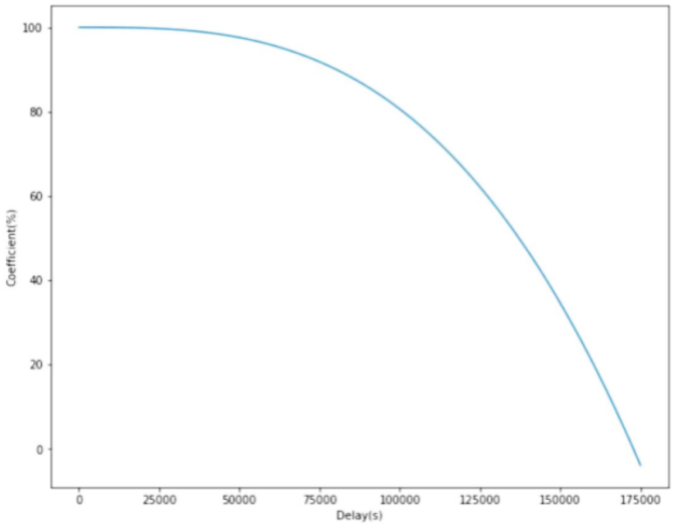
تمرین سری دوم
پردازش زبان و گفتار

مهسا انوریان

تاریخ تحویل: چهارشنبه ۱ اردیبهشت ۱۴۰۱ ساعت ۲۳:۵۹:۵۹

قوانین:

- سوالات این تمرین از مبحث «دسته‌بندی متن» می‌باشد و برای پاسخ به سوالات آن نیاز به دانش نسبی درمورد این مبحث دارید.
- این تمرین شامل ۴ سوال می‌باشد. ۳ سوال تئوری و نوشتاری هستند و ۱ سوال عملی و شامل پیاده‌سازی هستند.
- درصورت وجود هرگونه سوال، در کلاس درس و یا در گروه تلگرامی درس بپرسید. (لطفا پی‌وی پیام ندهید.)
- هرگونه ایده گرفتن از تمرین دیگران و کدهای موجود در اینترنت که موجب تشابه غیرعادی و بالای کد شما با دیگری شود، تقلب محسوب می‌شود. درصورت مشاهده‌ی تقلب، نمره‌ی تمرین برای هر دو دانشجوی متخلف **صفر** منظور خواهد شد.
- لطفا برای انجام تمرین، زمان مناسب اختصاص دهید و انجام آن را به روزهای پایانی موکول نکنید.
- پاسخ ارسالی شما باید علاوه بر کدهای مربوط به هر سوال، شامل یک گزارش در قالب یک فایل PDF باشد که محتوای گزارش مربوطه توضیحات تکمیلی شما درخصوص هر سوال و اسکرین‌شات از نتیجه اجرای کدهای شما باشد.
- تمامی فایل‌های موردنیاز برای تمرین را به صورت یک فایل ZIP با فرمت شماره دانشجویی_نام و نام خانوادگی_HW2 نام‌گذاری کرده و ارسال کنید. (برای مثال HW2_NameFamily_98000000)
- تاخیر در ارسال تمرین‌ها بر اساس نمودار زیر محاسبه خواهد شد. محور افقی نمودار، مقدار تاخیر به ثانیه و محور عمودی، ضریب اعمالی در نمره تمرین است.



سوالات تئوری:

۱. [روش‌های ارزیابی] گونه‌های روش ارزیابی Cross Validation را نام ببرید و کاربرد استفاده از هر کدام را توضیح دهید.

۲. [مدل زبانی] الف) یک مدل زبانی Bigram روی جملات زیر آموزش دهید:

a) <s> B A B A A B B </s>
b) <s> B A B A B A A </s>
c) <s> B A B B A B A </s>

ب) احتمالات زیر را با (و) بدون هموارسازی (smoothing) add-1 محاسبه کنید.

a) P(w1=A|w2=B)=?, Psmooth(w1=A|w2=B)=?
b) P(w1=<s>|w2=B)=?, Psmooth(w1=<s>|w2=B)=?
c) P(w1=A|w2=</s>)=?, Psmooth(w1=A|w2=</s>)=?

۳. [قانون بیز] در یک سیستم بازشناسی ارقام گسسته فارسی۰ تا ۹، پس از آموزش (Training) متوجه شده‌ایم ارقام هفت و هشت دارای سیگنال‌های نسبتا شبیه به هم هستند بطوریکه سیگنال آزمون (Test) رقم هفت به احتمال ۵۰٪ به عنوان عدد هفت و به احتمال ۵۰٪ به عنوان عدد هشت توسط سیستم تشخیص داده می‌شود. همچنین، سیگنال آزمون مربوط به رقم هشت، به احتمال ۳۰٪ به عنوان هفت و به احتمال ۷۰٪ به عنوان عدد هشت قابل تشخیص است. یک سیگنال ناشناس وارد سیستم می‌شود، احتمال تشخیص صحیح اعداد هفت و هشت به ترتیب چقدر است؟ فرض کنید سایر ارقام صحیح تشخیص داده می‌شوند و احتمال رخداد سیگنال ورودی برای همه اعداد یکسان باشد.

سوالات عملی:

۱. [پیاده‌سازی دسته‌بند Naïve Bayes] مجموعه‌داده imdb که شامل ۵۰۰۰ نظر مثبت و منفی است با استفاده از Keras برای شما در فایل نوت‌بوک بارگذاری شده است. ابتدا پیش‌پردازش‌های لازم را بر روی داده را انجام دهید و سپس با استفاده از آن‌ها مدل‌های uni-gram، bi-gram و tri-gram را بسازید و آموزش دهید. در هر کدام از مدل‌های زبانی از laplacian-smoothing و <UNK> استفاده کنید. از داده‌های آزمون (Test) برای ارزیابی هر کدام از دسته‌بندها استفاده کنید و دقت آن‌ها را با معیارهای Accuracy, F1-Score, Precision, Recall بدست آورید. مشخصاً این الگورتیم یکی از ساده‌ترین الگوریتم‌های موجود است. اما این ساده بودن باعث می‌شود که بتوانیم راحت‌تر نتایج و خروجی آن را تحلیل کنیم. فرض‌های اولیه‌ی این الگورتیم باعث می‌شود که توانایی مدل کردن آن کاهش یابد. آیا اثر bag of words را مشاهده می‌کنید؟ در نظر نگرفتن ترتیب آیا باعث شده است که در تشخیص خود دچار اشتباه شود؟ در مواردی که احتمال هر دو کلاس نزدیک به هم است، چه چیزی باعث شده مدل انتخاب درست/اشتباه را انجام دهد؟ فرضیه‌های خود را نمونه ورودی/خروجی مرتبط همراه کنید و آن‌ها را در گزارش خود به طور کامل توضیح دهید و همچنین نتایج بدست آمده را با هم مقایسه کنید.

موفق باشید