به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۴-۱۴۰۳

تمرین شماره ۳

تاریخ تحویل:
۱۴۰۳/۰۹/۰۲

---

۱. (۱۰٪) **[پژوهش – مفاهیم بازیابی اطلاعات]** در درس با مفاهیم مختلفی در حوزه بازیابی اطلاعات آشنا شدید. در این بخش می‌خواهیم برخی از اصول و روش‌های رتبه‌بندی در این حوزه را بررسی کنیم.

أ. روش‌های رایج رتبه‌بندی در بازیابی اطلاعات را نام برده و به طور مختصر توضیح دهید.

ب. نقش و اهمیت توابع امتیازدهی (Scoring Functions) در فرآیند رتبه‌بندی را توضیح دهید. برخی از رایج‌ترین توابع امتیازدهی را معرفی کنید.

ت. معیارهای ارزیابی مورد استفاده در بازیابی اطلاعات را توضیح دهید.

ث. درباره‌ی روش‌ها Probabilistic LSA و Latent Dirichlet Allocation تحقیق کنید.

ج. ایمیل‌های اسپم معمولاً از تکنیک‌های مختلف پنهان‌سازی برای عبور از فیلترها استفاده می‌کنند. یکی از روش‌ها، اضافه کردن یا جایگزینی کاراکترها به‌گونه‌ای است که طبقه‌بند متنی مبتنی بر کلمات را شکست دهند. به‌عنوان مثال، عباراتی مانند موارد زیر را در ایمیل‌های اسپم می‌بینید:

رایگآن

دارووووخانه

چگونه می‌توانید ویژگی‌هایی را طراحی کنید که این استراتژی را شکست دهند؟

---

۲. (۱۰٪) **[نظری – مفاهیم بازیابی اطلاعات]**

أ. در درس با مفاهیم مختلفی در حوزه بازیابی اطلاعات آشنا شدید. در این بخش می‌خواهیم برخی از اصول و روش‌های رتبه‌بندی در این حوزه را بررسی کنیم. به جدول فراوانی کلمات برای ۳ سند که به‌عنوان در جدول ۱ نشان داده شده است. وزن‌های tf-idf برای کلمات ماشین، اتومبیل، بیمه و بهترین را برای هر سند محاسبه کنید و از مقادیر idf موجود در جدول ۲ استفاده کنید.

جدول ۱ مقادیر tf

کلمات	Doc1	Doc2	Doc3
ماشین	27	4	24
اتومبیل	3	33	0
بیمه	0	33	29
بهترین	14	0	17

جدول ۲ مقادیر idf

کلمه	df_t	idf_t
ماشین	18165	1.65
اتومبیل	6723	2.08
بیمه	19241	1.62
بهترین	25235	1.5

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۴-۱۴۰۳

تمرین شماره ۳

تاریخ تحویل:
۱۴۰۳/۰۹/۰۲

ب. در یک مجموعه متنی، جمله‌ای به این صورت وجود دارد:
کاربری به دنبال عبارت "سواد رسانه‌ای" می‌گردد. پس از تمیزسازی سنده‌ها (شامل حذف حروف اضافه)، لطفاً میزان تشابه این پرسش را با جملات زیر محاسبه کنید. این محاسبه باید با بهره‌گیری از تحلیل معنایی پنهان (LSA) و استفاده از معیار تشابه کسینوسی انجام شود. لطفاً تمامی مراحل محاسبه را به طور گام به گام و با ارائه جزئیات لازم توضیح دهید.

جملات موجود در سند:

- او با تسلط بر رسانه‌ها به یک تحلیلگر برجسته تبدیل شده است.
- رسانه‌های اجتماعی به محیط‌های جدیدی برای به اشتراک‌گذاری اطلاعات و نظرات بدل شده‌اند.
- در دنیای دیجیتال، آگاهی رسانه‌ای به عنوان یک منبع قدرت شناخته می‌شود.

راهنمایی:
برای تجزیه SVD می‌توانید از تابع SVD در محیط‌های برنامه‌نویسی استفاده کنید. یا از وبسایت‌های آنلاین مانند wolframalpha استفاده کنید.

---

۳. (۲۰٪) **[پیاده‌سازی – محاسبه TF-IDF]** در این سوال قصد داریم TF-IDF را از پایه و بدون استفاده از کتابخانه محاسبه کنیم. این روش به ما کمک می‌کند تا اهمیت هر واژه را در یک سند نسبت به مجموعه‌ای از اسناد اندازه‌گیری کنیم. پیکره در فایل TF-IDF.txt در لینک زیر پیوست شده است:

https://drive.google.com/drive/folders/17dhrvVkmFKRd-SZDDcsFUZfhN3BPucFr?usp=drive_link

برای این کار باید مراحل زیر را طی کنیم.

- مراحل پیش‌پردازش متن: مراحل پیش‌پردازش متن شامل تبدیل حروف بزرگ به کوچک، حذف علائم نگارشی و اعداد، حذف کلمات اضافه، ریشه‌یابی کلمات و نشانه‌گذاری آن‌ها برای آماده‌سازی داده‌ها جهت تحلیل است. (دقت کنید که متن داخل فایل متنی به صورت structred در دسترس شما قرار نگرفته است و باید با استفاده از پایتون، متن مربوطی که در آن است را استخراج کنید.)

پس از اتمام مراحل پیش‌پردازش، می‌توانید مراحل زیر را برای محاسبه‌ی TF-IDF دنبال کنید:

- محاسبه‌ی Term Frequency (TF): برای هر کلمه در یک سند، تعداد دفعات وقوع آن را شمارش کنید و آن را بر تعداد کل کلمات سند تقسیم کنید.
- محاسبه‌ی Document Frequency (DF): برای هر کلمه، تعداد اسنادی را که شامل آن کلمه هستند شمارش کنید.
- محاسبه‌ی Inverse Document Frequency(IDF): از فرمول زیر برای محاسبه‌ی IDF استفاده کنید

$$\text{IDF}(t) = \log\left(\frac{N}{\text{DF}(t)}\right)$$

که در آن N تعداد کل اسناد است و $\text{DF}(t)$ تعداد اسنادی است که شامل کلمه‌ی t هستند.

- محاسبه‌ی TF-IDF: برای هر کلمه در یک سند، TF-IDF را با استفاده از فرمول زیر محاسبه کنید

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t)$$

بررسی و نمایش نتایج: نتایج TF-IDF را برای هر کلمه در هر سند نمایش دهید و بررسی کنید کدام کلمات اهمیت بیشتری دارند.

- محاسبه‌ی نمره انطباق (Matching Score): نمره انطباق ساده‌ترین روش برای محاسبه شباهت است. در این مرحله،

h.veisi@ut.ac.ir

به نام خدا

باید به ازای هر سند، مقادیر TF-IDF توکن‌هایی که در پرسش وجود دارند را جمع‌آوری کنیم. اگر پرسش ورودی "hello world" باشد، باید بررسی کنیم که آیا این کلمات در هر سند وجود دارند یا خیر. اگر کلمه‌ای موجود باشد، مقدار TF-IDF آن به نمره انطباق سند مربوطه اضافه می‌شود. در نهایت، اسناد را مرتب کرده و بهترین k سند را انتخاب می‌کنیم. با استفاده از دیکشنری که حاوی (سند، توکن) است، نمره انطباق را محاسبه کنید. به ازای هر توکن در پرسش، در دیکشنری جستجو کنید و در صورت وجود، شناسه سند و مقدار TF-IDF را در دیکشنری جدید ذخیره کنید. در انتها، k سند برتر را برگردانید.

- شباهت کسینوس (Cosine Similarity) پس از محاسبه نمره انطباق، ممکن است نیاز باشد که شباهت کسینوس را نیز محاسبه کنیم. نمره انطباق ممکن است در پرسش‌های طولانی به درستی عمل نکند، در حالی که شباهت کسینوس با تبدیل اسناد و پرسش به بردارهای TF-IDF و محاسبه زاویه بین آنها، می‌تواند دقت بیشتری داشته باشد. سندها و پرسش را به بردارهای TF-IDF تبدیل کرده و شباهت کسینوس را محاسبه کنید.
- تحلیل: با استفاده از شناسه سند ۲۰۰، محتوای طولانی و کوتاه را برای هر دو نمره انطباق و شباهت کسینوس آزمایش کنید.
- مقایسه با روش‌های کتابخانه‌ای: نتایج خود را با استفاده از کتابخانه‌های موجود مانند scikit-learn مقایسه کنید تا دقت و کارایی روش خود را بسنجید.

---

۴. (۱۰٪) **[پیاده‌سازی – پیدا کردن پارامترهای بهینه]** در این تمرین، شما مدل‌های SVM با استفاده از هسته‌های مختلف و مقادیر مختلف پارامتر C را پیاده‌سازی کرده و تحلیل می‌کنید. هدف این است که تأثیر هسته‌های مختلف و پارامتر C را بر مرز تصمیم و عملکرد مدل مشاهده کنید. برای این تمرین وظیفه تشخیص حس مربوط به کامنت‌های اینستاگرام را در نظر گرفته‌ایم که در فایل CSV در لینک زیر قابل مشاهده است.

https://drive.google.com/drive/folders/17dhrvVkmFKRd-SZDDcsFUZfhN3BPucFr?usp=drive_link

- داده‌ها را به دو بخش‌های آموزش (۸۰٪) و آزمون (۲۰٪) تقسیم کنید.

أ. آموزش و ارزیابی مدل‌های SVM: مدل‌های SVM را با تنظیمات زیر پیاده‌سازی کنید:
- هسته خطی
- هسته چندجمله‌ای (با درجه ۳)
- هسته RBF (گوسی)
- برای هر هسته، مقدار پارامتر تنظیمی C را به صورت زیر تغییر دهید: C = [۰.۱, ۱, ۱۰, ۱۰۰].
- مشاهده و ثبت عملکرد مدل‌ها:
- برای هر جفت (هسته، C)، مدل را روی مجموعه‌ی آموزش داده آموزش دهید و دقت مدل را روی هر دو مجموعه‌ی آموزش و آزمون محاسبه کنید.
- نتایج را به صورت یک جدول با ستون‌های هسته، C، دقت آموزش و دقت آزمون ثبت کنید.

ب. بصری‌سازی مرزهای تصمیم‌گیری:
- برای هر جفت (هسته، C)، مرز تصمیم‌گیری مدل را روی داده‌های آموزش رسم کنید.

ب. تحلیل نتایج: یک تحلیل کوتاه از نتایج ارائه دهید و به موارد زیر بپردازید:
- عملکرد مدل چگونه با هسته‌های مختلف تغییر می‌کند؟

h.veisi@ut.ac.ir

- تغییر مقدار C چگونه بر مرز تصمیم‌گیری و توانایی مدل در تعمیم‌دهی تاثیر می‌گذارد؟
- کدام ترکیب هسته و C بیشترین دقت در تست را دارد؟

مهم‌ترین بخش این سوال تحلیل شما از پارامترهای مختلف آموزش می‌باشد.

---

۵. (۳۰٪) [**پیاده‌سازی – خلاصه سازی‌متن**] خلاصه‌سازی متن یک فرآیند مهم در NLP است که هدف آن استخراج اطلاعات کلیدی و تولید یک نسخه کوتاه‌تر و مفهومی از متن اصلی است. این فرآیند به بهبود دسترسی به اطلاعات، صرفه‌جویی در زمان و کمک به درک بهتر موضوعات پیچیده کمک می‌کند. در این پروژه، شما به بررسی سه رویکرد خلاصه‌سازی خواهید پرداخت:

- تحلیل معنایی نهفته (LSA)
- خلاصه‌ساز Luhn

همچنین مجموعه‌داده‌ی مربوط به این سوال در این لینک قابل پیدا کردن می‌باشد.

۱. **قدم اول: بررسی روش‌ها**
- در مورد دو روش خلاصه‌سازی LSA و Luhn Summarizer مطالعه کنید.
- مفاهیم کلیدی، نقاط قوت و ضعف هر روش را خلاصه کنید.

۲. **قدم دوم: پیاده‌سازی تکنیک‌های خلاصه‌سازی**
- تحلیل معنایی نهفته (LSA): از کتابخانه‌های مناسب مانند Scikit-learn برای پیاده‌سازی این بخش استفاده کنید.
- خلاصه‌ساز Luhn: این روش را با استفاده از کتابخانه‌های NLP مانند NLTK یا SpaCy پیاده‌سازی کنید.

۳. **قدم سوم: ارزیابی**
- از معیار ROUGE برای ارزیابی کیفیت خلاصه‌های تولیدشده توسط هر روش استفاده کنید.
- امتیازهای ROUGE هر تکنیک خلاصه‌سازی را مقایسه کنید.
- در مورد تفاوت‌های عملکرد و کیفیت خلاصه‌های تولیدشده بحث کنید.

---

۶. (۲۰٪) [**پیاده‌سازی – Language Identification**] هدف این بخش، پیاده‌سازی یک سیستم شناسایی زبان است که قادر به شناسایی زبان متون ورودی باشد. این سیستم می‌تواند در کاربردهایی نظیر ترجمه ماشینی، تحلیل متن و سیستم‌های پاسخگو به سوالات مورد استفاده قرار گیرد. مراحل پیاده‌سازی این مدل به ترتیب زیر می‌باشد:

بارگذاری و تحلیل داده:

- از مجموعه داده‌ی در این لینک استفاده کنید.
- پیش‌پردازش: داده‌های متنی ممکن است نیاز به پیش‌پردازش داشته باشند. در این مرحله، می‌توانید از تکنیک‌هایی مانند پاک‌سازی stop wordها، حذف کاراکترهای غیر ضروری و توکن‌سازی استفاده کنید.
- پس از پیش‌پردازش داده‌ها، نیاز داریم که داده‌ها را به دو بخش آموزش و آزمایش تقسیم کنیم.
- Visualization: به بررسی مجموعه داده بپردازید و نمودارهایی مربوط به مجموعه‌داده را رسم کنید (برای مثال

h.veisi@ut.ac.ir

به نام خدا

دانشکده علوم و فنون نوین

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۴-۱۴۰۳

تمرین شماره ۳

تاریخ تحویل:
۱۴۰۳/۰۹/۰۲

توزیع کلاس‌های مختلف مجموعه‌داده، یا توزیع تعداد کلمات در زبان‌های مختلف)

- پس از پیش‌پردازش داده‌ها، نیاز داریم که داده‌ها را به دو بخش آموزش و آزمایش تقسیم کنیم. داده‌ها را به نسبت ۸۰ به ۲۰ به بخش‌های آموزشی و آزمایشی تقسیم کنید.

استخراج ویژگی: برای شناسایی زبان، از روش TF-IDF برای استخراج ویژگی‌ها استفاده خواهیم کرد.

مدل یادگیری ماشین: در این مرحله، مدل‌های مختلف یادگیری ماشین برای شناسایی زبان آموزش خواهیم داد. از الگوریتم‌های Naive Bayes ، Random Forest و SVM استفاده خواهیم کرد. برای بهینه‌سازی هایپرپارامترها، از دو روش گرید سرچ (Grid Search) و رندوم سرچ (Random Search) بهره ببرید تا بهترین هایپرپارامترها را شناسایی کنیم.

نتایج:

- در نهایت، مدل آموزش‌دیده باید ارزیابی شود و پیش‌بینی‌هایی برای متون جدید انجام دهد.
- دقت، ماتریس درهم‌ریختگی و classification report مربوط به هر مدل را نمایش دهید.

h.veisi@ut.ac.ir