به نام خدا

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۵-۱۴۰۴

تاریخ تحویل:
۱۴۰۴/۰۹/۱۴

تمرین شماره ۳

۱. (۱۰٪) **[نظری – مفاهیم بازیابی اطلاعات]**

I. 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

II. 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)، 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.

---

۲. (۱۵٪) **[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

h.veisi@ut.ac.ir

به نام خدا

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۵-۱۴۰۴

تمرین شماره ۳

تاریخ تحویل:
۱۴۰۴/۰۹/۱۴

اندازه‌گیری کنیم. پیکره در فایل TF-IDF.txt در لینک زیر پیوست شده است:

https://drive.google.com/file/d/180XLqwGteXI5MhmQHh8lJaMLXl9KWLYT/view?usp=sharing

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

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

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، 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.

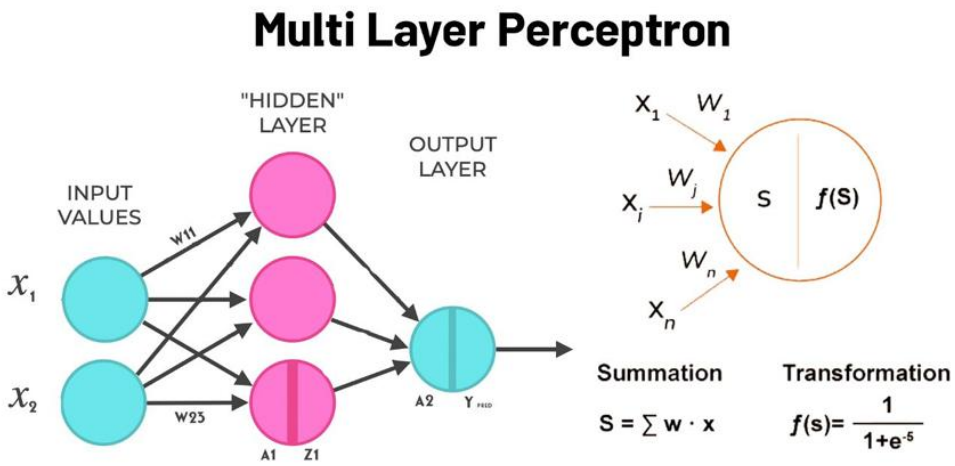
به نام خدا

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)
نیم‌سال اول ۱۴۰۵-۱۴۰۴

تاریخ تحویل:
۱۴۰۴/۰۹/۱۴

تمرین شماره ۳

۳. (۱۵٪) **[نظری - تحلیل شبکه عصبی چندلایه]** در این تمرین هدف آن است که با ساختار شبکه پرسپترون چندلایه (Multi-Layer Perceptron) آشنا شوید و فرآیند انتشار روبه‌جلو (Forward Propagation) ، محاسبه خطا (Loss) و انتشار به‌عقب (Backpropagation) را با استفاده از روش مشتقات زنجیره‌ای و گرادیان نزولی پیاده‌سازی کنید. شبکه زیر را درنظر بگیرید:



$A_1 = W_1 x + b_1$

$Z_1 = \tan h(A_1)$ -> تانژانت هایپر بولیک

$A_2 = W_2 Z_1 + b_2$

$\hat{y} = \sigma(A_2)$

$L(i) = y \log(\hat{y}) + (1-y) \log(1-\hat{y})$

$$J = -\frac{1}{m}\sum_{i=1}^{m} L(i)$$

۱. مشتقات زیر را به روش مشتق‌های زنجیره‌ای حساب کنید:

$\partial J / \partial z_1$
$\partial A_2 / \partial b_1$
$\partial J / \partial W_1$

h.veisi@ut.ac.ir

به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۵-۱۴۰۴

تمرین شماره ۳

تاریخ تحویل:
۱۴۰۴/۰۹/۱۴

II. فرمول‌های به‌روزرسانی پارامترهای W1، W2، b1، b2را با روش gradient descent و با نرخ یادگیری به مقدار α بنویسید. توجه کنید که مشتقات باید محاسبه شوند. می‌توانید از مقادیر مشتق به‌دست‌آمده در قسمت قبل استفاده کنید.

III. با فرض ورودی‌های زیر، شبکه داده‌شده را یک دور آموزش دهید و مقادیر جدید وزن‌ها را به‌دست آورید.

$$x = \begin{bmatrix}1\\2\end{bmatrix}, Y = 1, \alpha = 0.1$$

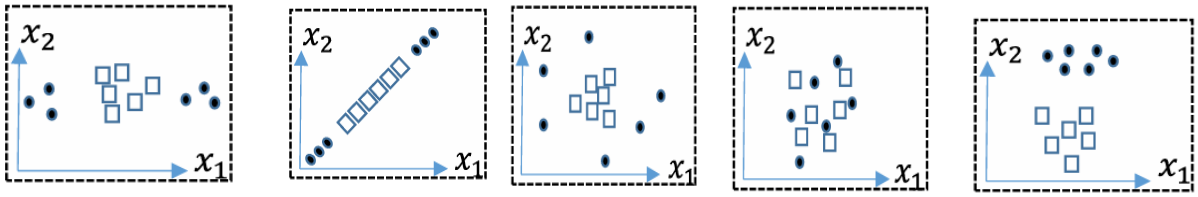
$$W_1 = \begin{bmatrix}0.5 & 1\\1 & 2\end{bmatrix}, b_1 = \begin{bmatrix}0.5\\0.7\end{bmatrix}$$

$$W_2 = [2], b_2 = [0.5]$$

---

۴. (۱۰٪) [**تحلیلی – درک مفهومی از شبکه عصبی مصنوعی**] در یک مسئله دسته‌بندی (Classification) ، توزیع داده‌های ورودی را در یک فضای دوبعدی مطابق حالت‌های زیر نمایش داده‌ایم. در هر مورد، ساده‌ترین شبکه ممکن جهت طبقه‌بندی را معرفی و ترسیم کنید و دلیل انتخاب خود را به‌صورت مفهومی توضیح دهید. در هر شکل، محورهای x_1 و x_2نشان‌دهنده ویژگی‌های ورودی و yخروجی شبکه عصبی است. در هر حالت مشخص کنید که آیا یک شبکه پرسپترون تک‌لایه (Single-Layer Perceptron) کافی است یا نیاز به چندلایه (Multi-Layer Perceptron) وجود دارد.

A) B) C) E) F)



---

۵. (۵۰٪) [**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:

h.veisi@ut.ac.ir

به نام خدا

دانشکده سامانه‌های هوشمند

روش‌های یادگیری ماشین در پردازش زبان طبیعی (۸۳۰۴۳۶۸)

نیم‌سال اول ۱۴۰۵-۱۴۰۴

تمرین شماره ۳

تاریخ تحویل:
۱۴۰۴/۰۹/۱۴

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 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 دو روش TF-IDF و word2vec (می‌توانید از مدل‌های آماده استفاده کنید) 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، MLP (با یک لایه مخفی) و SVM استفاده خواهیم کرد. برای بهینه‌سازی هایپرپارامترها، از دو روش گرید سرچ (Grid Search) و جستجوی تصادفی (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