بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

# پردازش زبان طبیعی

**نیم‌سال دوم ۰۱-۰۰**

استاد: احسان‌الدین عسگری

تمرین چهارم | طبقه‌بندی متن | مهلت ارسال: ۲۵ خرداد و ۳ تیر

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در تمرین‌هایی که چند ترک دارند، فقط یک نفر از هر گروه در سامانه CW باید ترک مورد نظر گروه را انتخاب کند. امکان تغییر ترک تا قبل از زمان ددلاین انتخاب ترک وجود دارد. البته ذکر این نکته ضروری است که هر ترک محدودیتی برای تعداد افرادی که آن را انتخاب می‌کنند، دارد. بنابرین در اسرع وقت برای انتخاب ترک اقدام کنید.
- در طول ترم امکان ارسال با تاخیر تمرین‌ها بدون کسر نمره تا سقف ۱۲ روز وجود دارد. محل بارگزاری جواب تمرین‌ها بعد از ۳ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۵ درصد از نمره تمرین کسر خواهد شد.
- توجه داشته‌باشید که نوت‌بوکهای شما باید قابلیت بازاجرای ۱۰۰ درصد داشته باشند و در صورت نیاز به نصب یک کتابخانه یا دسترسی به یک فایل، مراحل نصب و دانلود (از یک محل عمومی) در نوت‌بوک وجود داشته باشد.
- تمامی فایل‌های مرتبط به پروژه که حجم کمی دارند باید به شکل فایل زیپ در سامانه CW اپلود شوند. اگر حجم یک فایل زیاد بود (مانند فایل ذخیره شده یک مدل در صورتیکه بیش از ۲۰۰ مگابایت باشد)، تنها همان فایل را در یک محل عمومی، مثل گوگل درایو آپلود بفرمایید و لینک دانلود را در نوت‌بوک و مستندات قرار دهید.
- در پروژه‌های گروهی کافی است که فقط یکی از اعضای گروه پروژه را آپلود کند. اما حتما در گزارش کار نام همه اعضای گروه همراه با شماره دانشجویی آن‌ها آورده شود.
- بخشی از نمره شما به گزارش کار شما اختصاص دارد. در گزارش کار لازم نیست خط به خط کاری را که کرده‌اید توضیح دهید. بلکه باید به شکل کلی ایده‌تان برای حل مساله را شرح دهید. لازم است چند نمونه از خروجی‌های مساله را در گزارش بیاورید و براساس آن رفتار برنامه‌تان را تحلیل کنید. همچنین اگر پارامتری در صورت مساله خواسته شده (مانند دقت، صحت و ...) که در گزارش آورده شود شما باید آن را حساب کنید و در گزارش خود بیاورید.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.

## توضیحات کلی

در این تمرین قرار است که با طبقه‌بندی متن از مرحله آماده‌سازی داده تا انجام آن توسط یادگیری ماشین آشنا شوید. نمره این تمرین دارای دو بخش انفرادی و گروهی است. در بخش انفرادی شما باید به سامانه جمع‌سپاری مراجعه کنید و ۱۰۰ خبر را بر اساس ملاک‌هایی که در صفحات بعد توضیح داده شده است طبقه‌بندی کنید. این کار حداکثر دو ساعت از شما وقت می‌گیرد. دقت شود که مهلت انجام این بخش از تمرین تا ۲۵ خرداد است و این مهلت **نه تمدید می‌شود و نه امکان استفاده از تاخیر مجاز برای آن وجود دارد.**

برای ورود به سامانه جمع‌سپاری، ایمیل و رمز عبور هر نفر در CW به اشتراک گذاشته شده است. برای مشاهده اطلاعات ورود خود به صفحه درس در CW بخش نام کاربری و رمزعبور سامانه جمع‌سپاری مراجعه کنید!

بعد از این که داده‌ها توسط خود دانشجویان طبقه‌بندی شد. داده‌ها در اختیارتان قرار می‌گیرد و هر گروه بر اساس ترکی که انتخاب می‌کند باید براساس یکی از این ملاک‌ها طبقه‌بندی را انجام دهد. برای طبقه‌بندی باید از دو روش استفاده کنید. برای روش اول حتما باید از یک روش مبتنی بر ترنسفورمر از پیش آموزش دیده استفاده کنید (مانند parsbert) و برای روش دوم شما آزادی عمل دارید که از هر روش غیر ترنسفورمر که خواستید استفاده کنید. برای مثال می‌توانید بین روش‌های کلاسیک یا مبتنی بر CNN یا RNN انتخاب کنید.

## فن بیان تیتر

در این تسک هدف پیاده‌سازی یک مدل برای تشخیص کیفیت بیان تیتر یک خبر است. تیتر خبرهایی که می‌توانند با استفاده از کلمات مناسب، باعث شوند خبر به خوبی در ذهن خواننده باقی بماند، دارای فن بیان خوبی هستند.
در گام اول در مجموعه دادگان داده شده براساس کیفیت بیان تیتر با روش گفته شده در بخش توضیحات کلی باید برای هر خبر یکی از دو برچسب بیان متوسط و بیان جذاب انتخاب شود. در گام بعدی براساس داده‌هایی که آماده شده است و براساس توضیحات ابتدای تمرین به پیاده‌سازی مدل دسته‌بندی بپردازید.
برای توضیحات بیشتر به تیتر چند خبر و برچسب مربوط به آن توجه فرمایید:

| خبر | برچسب |
|---|---|
| طلاگیری شناگران شریفی از آب! | بیان جذاب |
| دانشجویان دانشگاه صنعتی شریف در مسابقات شنا مدال طلا کسب کردند. | بیان عادی |
| نجات زمین از گرمایش، حالا یا هیچ‌وقت... | بیان جذاب |
| فرصت محدودی برای رفع مشکل گرمایش زمین باقی‌مانده است. | بیان عادی |

## مثبت منفی بودن خبر

در این تسک هدف پیاده‌سازی یک مدل برای تشخیص تحلیل احساس خبر است. متن خبرها از نظر تحلیل احساسات می‌تواند مثبت، خنثی و یا منفی باشد.

۱. در گام اول در مجموعه دادگان داده شده براساس تحلیل احساسات بیان متن با روش گفته شده در بخش توضیحات کلی باید برای هر خبر یکی از سه برچسب خبر مثبت، خنثی و منفی انتخاب شود.

۲. در گام بعدی براساس داده‌هایی که آماده شده است براساس توضیحات ابتدای تمرین به پیاده‌سازی مدل دسته‌بندی بپردازید.

با توجه به این که مثبت و منفی بودن یک مورد نسبی و قائم به ناظر آن است، با مرجع **منافع ملی ایران و ایرانیان** خبرها را بسنجید. برای توضیحات بیشتر به تیتر چند خبر و برچسب مربوط به آن خبر توجه فرمایید:

| خبر | برچسب |
|---|---|
| ۱. روزهای سبز بورس ادامه دارد.<br>۲. داریا همراه در سال گذشته عملکرد مناسبی داشته است و رضایت مشتریان را کسب کرده است. | مثبت |
| ۱. بازدید حسن اکلیلی، هنرمند مردمی و بازیگر اصفهانی از دفتر خبرگزاری موج<br>۲. یکی از تابلوهای مجموعه «پل واترلو» اثر کلود مونه، نقاش مشهور فرانسوی، ماه آینده در حراجی کریستیز چکش می‌خورد. | خنثی |
| ۱. تازه‌ترین اخبار متروپل: ۳۷ جان‌باخته و ۳۳ مجوز دفن آخرین آمار<br>۲. خبر یک سقوط: هواپیمای تهران_کی‌یف با ۱۷۶ سرنشین سقوط کرد. | منفی |

## جریان‌سازی فکری و فرهنگی خبر

در این تسک هدف پیاده‌سازی یک مدل برای تشخیص جریان‌سازی یک خبر است. خبرهایی که اثرات مهمی دارند و احتمالا باعث ایجاد تحولی فرهنگی، اجتماعی و از این دست بشوند و در آینده خبرهای مشابه دیگری در رابطه با آن‌ها منتشر شود را جریان‌ساز می‌گوییم.
در گام اول در مجموعه دادگان داده شده براساس میزان جریان‌سازی با روش گفته شده در بخش توضیحات کلی باید برای هر خبر یکی از دو برچسب جریان‌ساز یا غیرجریان‌ساز انتخاب شود. در گام بعدی براساس داده‌هایی که آماده شده است براساس توضیحات ابتدای تمرین به پیاده‌سازی مدل دسته‌بندی می‌پردازید.
برای توضیحات بیشتر به تیتر چند خبر و برچسب مربوط به آن خبر توجه فرمایید:

| خبر | برچسب |
|---|---|
| عصبانیت شدید مسی از لاپورتا \| رئیس بارسلونا تهدید شد | غیرجریان‌ساز |
| حضور اقشار مختلف در مراسم اربعین امسال حاکی از اهمیت امام حسین برای مردم حتی پس از دو سال وقفه در مراسمات به دلیل بیماری کویید، می‌باشد | جریان‌ساز |
| مکزیک نخستین مورد آبله میمونی را تایید می‌کند، کارشناسان جهانی خواستار اقدامات بیشتر در برابر بیماری هستند | جریان‌ساز |
| تاکید وزیر آموزش و پرورش به لزوم تحول آموزشی در ایران با ورود متاورس و بلاکچین | جریان‌ساز |
| "پل طبیعت" تهران دو شب متوالی خاموش می‌شود | غیرجریان‌ساز |

## اثر اقتصادی

در این تسک شما باید تاثیرگذازی اقتصادی یک خبر را بررسی کنید. مجموعه‌داده شما یک مجموعه‌داده با برچسب‌های پنج کلاسه خواهد بود که توضیحات هر برچسب به شرح زیر است:

- مستقیم مثبت: خبری که به‌صورت مستقیم روی اقتصاد تاثیر مثبت دارد
- غیرمستقیم مثبت: خبری که به‌صورت غیرمستقیم روی اقتصاد تاثیر مثبت دارد
- خنثی: خبری که هیچ تاثیری رو اقتصاد ندارد
- مستقیم منفی: خبری که به‌صورت مستقیم رو اقتصاد تاثیر منفی دارد
- غیرمستقیم منفی: خبری که به‌صورت غیرمستقیم روی اقتصاد تاثیر منفی دارد

توجه کنید که این تاثیرگذاری صرفا به اخبار اقتصادی محدود نمی‌شود و بسیاری از اخبار اجتماعی یا سیاسی یا دیگر موضوعات خبری نیز می‌توانند به‌لحاظ اقتصادی تاثیرگذار باشند.
پس از برچسب‌زنی شما باید طبق متن تمرین دسته‌بندهای ذکرشده را روی داده خود آموزش دهید و نتایج و معیارهای ارزیابی را گزارش کنید.

| تیتر خبر | پیوند خبر | برچسب |
|---|---|---|
| افراد جدیدی که یارانه می گیرند، چه کسانی هستند؟/ یارانه های جدید کی قابل برداشت می شود؟ | پیوند | مستقیم مثبت |
| افشاگری گاردین از پنهان کاری مقامات آمریکایی در مورد آمار کرونا | پیوند | خنثی |
| آغاز تفکیک حساب‌های بانکی تجاری و شخصی برای مؤدیان مالیاتی | پیوند | غیرمستقیم مثبت |
| افزایش اسلام هراسی در اتریش به زبان آمار | پیوند | خنثی |
| آمریکا ارائه قطعنامه ضد ایرانی در شورای حکام آژانس را تایید کرد | پیوند | غیرمستقیم منفی |
| پژو ۲۰۶ در مرز ۳۶۰ میلیون تومانی/ آخرین قیمت پراید، تیبا، ساینا و دنا | پیوند | مستقیم منفی |