پروژه پایانی درس پردازش زبان‌های طبیعی

موضوع پروژه

پروژه طراحی شده برای این درس، بر روی داده‌های خبرهای مربوط به خبرگزاری‌ها خواهد بود.
پروژه در سه بخش متفاوت مورد بررسی قرار خواهد گرفت:
۱ - شناسایی موضوع خبرها (مشخص‌کردن یک تگ برای هر خبر)
۲ - شناسایی موضوع خبرها به صورت مولتی کلاس (مشخص‌کردن یک یا چند تگ برای هر خبر)
۳ - شناسایی منبع خبر (این خبر از چه منبعی است؟)
در بخش اول و دوم باید برای هر خبر یک یا چند تگ مربوط به موضوع آن را مشخص نمایید. مثلاً برای یک خبر خاص، مشخص کنید که آیا این خبر اقتصادی است یا سیاسی و غیره. حالت اول به این صورت خواهد بود که از میان موضوعاتی که برای هر خبر قبلاً مشخص شده است (الگوریتم supervised)، حداقل یک مورد را باید شناسایی کنید که اشتراک موضوعی که مشخص می‌کنید باید با موضوع خود خبر غیر تهی باشد تا امتیاز مربوط به آن خبر را کسب کنید.
در حالت دوم می‌توانید چندین موضوع را برای هر خبر مشخص کنید (حداکثر ۳ تا). که به ازای هر موضوع اشتباه نمره منفی اعمال خواهد شد و به ازای هر موضوع درست نیز امتیاز مثبت لحاظ می‌شود.
در حالت سوم باید مشخص کنید که منبع این خبر از چه خبرگزاری‌ای خواهد بود؟ مثلاً عصر ایران.
هر کدام از این سه بخش را باید در پروژه پایانی ارائه نمایید (الگوریتم‌ها بسیار شبیه به هم خواهند بود اما وظایف کمی متفاوتند).
خروجی هر یک از این سه وظیفه باید با با معیارهای زیر مشخص گردد:

Accuracy
Precision
Recall
F1 measure

پیش‌نیاز و نیازمندی‌ها

پروژه این درس حداکثر به صورت دو نفره خواهد بود.
برای انجام پروژه می‌توانید از هر زبان برنامه‌نویسی دلخواه استفاده کنید (توصیه می‌شود تا از زبان پایتون استفاده کنید).
برخی از ابزارهای معتبر در این زمینه:
- ابزار هضم موجود در پایتون، جاوا و سی‌شارپ
- ابزار NLTK
که می‌توانید از الگوریتم‌های پیش‌پردازش آن‌ها استفاده کنید.
برای پیاده‌سازی پروژه، پیاده‌سازی (از صفر) یکی از روش‌های طبقه‌بندی Naive bayes یا Discriminative الزامی است. در کنار این‌ها می‌توانید از روش‌های دیگری نیز استفاده نمایید. راه‌کاری که با استفاده از این طبقه‌بندها به جواب می‌رسید می‌تواند ابتکاری باشد. دقت کنید که داده‌های شما دارای تگ کلاس هستند و تگ‌های کلاس نیز، تنها از میان تگ‌های موجود در داده‌ها باید انتخاب شوند.
برای ساخت مدل‌های یادگیر می‌توانید داده‌ها را به دو بخش Test و Train تقسیم نمایید یا از روش Cross validation استفاده کنید.

فرمت داده‌ها

داده‌ها به صورت مجموعه از فایل‌های JSONL (فایل‌هایی که هر خط از آن یک داده JSON است) در اختیار شما قرار می‌گیرد. همراه این فایل در مرحله اول یک فایل نمونه JSONL و یک داده JSON از آن به صورت جداگانه قرار داده شده است که می‌توانید به کمک آن پیاده‌سازی خود را آغاز کنید. برای نمونه یک خبر به صورت فرمت JSON به صورت زیر خواهد بود:


```
{
"_id": "5b4f727a020eb20597f401b6",
"NewsAgency": "AsrIran",
"newsCode": "621661",
"newsLink": "http://www.asriran.com/fa/news/621661",
"date": "18 - ۱۳۹۷ تیر ۲۷ - ۲۱:۲۰ :July 2018 تاریخ انتشار",
"newsPath": "صفحه نخست » اقتصادی",
"newsPathLinks": {
"صفحه نخست": "/fa/archive?service_id=1",
"اقتصادی": "/fa/archive?service_id=1&amp;cat_id=4"
},
"title": "قیمت سبد نفتی اوپک یک گام دیگر عقب نشست/ 70 دلار برای هر بشکه",
"rutitr": "",
"subtitle": "",
"body": " قیمت سبد نفتی اوپک دیروز به روند کاهشی خود ادامه داد و در روز سه شنبه به حدود ۷۰ دلار مهر، بر اساس اعلام پایگاه اینترنتی دبیرخانه اوپک، قیمت سبد &amp;nbsp; برای هر بشکه رسید. به گزارش نفتی اوپک دیروز به ۷۰ دلار و ۳۸ سنت رسید. قیمت این شاخص در روز پیش از آن (دوشنبه ۲۵ تیر ماه) ۷۱ دلار و ۹۰ سنت بود. سبد نفتی اوپک شامل انواع نفت خام تولیدی در ۱۵ کشور عضو این سازمان است. همچنین دیروز قیمت شاخص نفت خام دبلیو تی آی به ۶۸ دلار و ۶ سنت و قیمت شاخص نفت خام برنت به ۷۱ دلار و ۸۴ سنت رسید.",
"bodyHtml": " <p> قیمت سبد نفتی اوپک دیروز به روند کاهشی خود ادامه داد و در روز سه شنبه به مهر، بر اساس اعلام پایگاه اینترنتی &amp;nbsp; به گزارش</p> <p>حدود ۷۰ دلار برای هر بشکه رسید دبیرخانه اوپک، قیمت سبد نفتی اوپک دیروز به ۷۰ دلار و ۳۸ سنت رسید. قیمت این شاخص در روز پیش از آن (دوشنبه ۲۵ تیر ماه) ۷۱ دلار و ۹۰ سنت بود سبد نفتی اوپک شامل انواع نفت خام تولیدی</p> <p> همچنین دیروز قیمت شاخص نفت خام دبلیو تی آی به ۶۸</p> <p>در ۱۵ کشور عضو این سازمان است دلار و ۶ سنت و قیمت شاخص نفت خام برنت به ۷۱ دلار و ۸۴ سنت رسید.</p> <div class=\"wrapper\"/> ",
"tags": {
"اوپک": "/fa/tag/1/اوپک",
"نفت": "/fa/tag/1/نفت"
}
}
```


دقت کنید که داده‌های JSON پیش‌پردازش نشده‌اند و شامل فیلدهای بعضا اضافی نیز هستند. همچنین پردازش داده‌ها و تمیزکردن آن‌ها که بخشی از کار طبقه‌بندی است، بر عهده شما است. داده

مربـوط بـه کلاس (طبقـه) هـدف در وظـایف اول و دوم کـه مشـخص‌کردن موضـوع اسـت در فیلـد newsPathLinks قرار دارد امّا باید ابتدا پیش‌پردازش شود (داده نهایی باید تنها شـامل موضـوعات واقعی مثل سیاسی یا اقتصادی باشد و مواردی مانند صفحه نخست نیز حذف شوند) پس از این مرحلـه ممکن است بعضی نمونه‌ها بدون موضوع باشند که باید به شیوه مناسب با آن‌ها برخورد نمایید.
در مورد وظیفه سوم هم فیلد NewsAgency را به عنوان کلاس هدف در نظر بگیرید (در حال حاضـر، داده نمونه تنها شامل یک خبرگزاری است).
* فیلد tags به عنوان داده feature می‌تواند مورد استفاده قرار گیرد و خود به تنهایی کلاس هدف نیست!

در روز تحویل داده تست در اختیار شما قرار می‌گیرد که نتیجه اعمال روش‌های شما بر روی آن به عنوان نتیجه نهایی الگوریتم شما خواهد بود، بنابراین باید داده‌های خود را برای دریافت یک فایل تست آماده کنید تا تنها خروجی مورد نظر را بر روی داده‌های جدید چاپ کند.

با آرزوی موفقیت برای شما
معانی‌جو