پروژه پایانی درس بازیابی اطلاعات وب

هدف از این پروژه، دسته بندی متن است. بدین منظور در محیط پایتون، از دسته‌بندهای بانظارت نایوبیز، درخت تصمیم و k نزدیک‌ترین همسایه (و یا یک دسته‌بند دلخواه دیگر) جهت طبقه بندی متن استفاده کنید. در کل برای طبقه‌بندی متن، باید مراحل زیر صورت گیرد:

- (Preprocessing)– پیش پردازش
- (Feature Generation) – تولید ویژگی
- (Feature Weighting) – وزن‌دهی ویژگی
- (Interpretation /Evaluation) – تفسیر و ارزیابی

پیش‌پردازش: در این مرحله، پیش‌پردازش انجام می‌شود تا کلاس‌ها متوازن تر شوند و حجم اطلاعات هم کاهش یابد. در واقع برای کاهش ابعاد و حذف کلمات غیر مفید، از رویکرد حذفکلمات مانع (StopWords) و ریشه‌یابی کلمات (Stemming با استفاده از یکی از الگوریتم‌های ریشه‌یابی) استفاده کنید.

تولید ویژگی: در این پروژه برای نمایش بردار ویژگی از چند گرم‌های (N-gram) آماری استفاده کنید. در این روش n تعداد کلمات انتخاب شده را بیان می‌کند. با توجه به این‌که سه روش تک کلمه‌ای (Unigram)، دو کلمه‌ای (Bigram) و سه کلمه‌ای (Trigram) در بیشتر پژوهش‌ها مورد توجه قرار گرفته است، هر سه رویکرد را در پروژه، اعمال نمایید.

وزن‌دهی ویژگی: برای طبقه‌بندی متن، بهتر است به هر کلمه یک وزن به نسبتِ اهمیتِ آن در جمله اختصاص دهیم. این کار می‌تواند جهت تشخیص کلمات و اهمیتِ آن‌ها به الگوریتم‌های طبقه‌بندی ،کمکِ بیشتری کند. بدین منظور، بعد از طی کردن مراحل قبل (ریشه‌یابی، حذف کلمات مانع و اعمال Ngram)، به محاسبه‌ی وزن ویژگی‌ها (کلمات) می‌پردازیم. به منظور تعیین وزن کلمات، در این پروژه از معیار وزن‌دهی TF-IDF استفاده کنید. این روش، میزان تکرار یک کلمه در یک مستند را در مقابل تعداد تکرار آن در مجموعه کلیه‌ی مستندات در نظر می‌گیرد. با این کار، می‌توان اهمیتِ یک کلمه را بهتر شناسایی کرده و در نهایت ویژگی‌های بهتری را برای اعمال الگوریتم‌های طبقه‌بندی داشته باشیم.

تفسیر و ارزیابی: در سنجش کارآیی مدل دسته‌بندی، از معیارهای صحت (Precision) ، بازخوانی (Recall) و نهایتا معیار ترکیبی F1-Score که میانگین هارمونیک این دو معیار است، استفاده کنید.

انتخاب دیتاست: انتخاب دیتاست، اختیاری است. دقت داشته باشید که فقط این دیتاست باید شرایط ارزیابی را داشته باشد (یعنی دسته‌ها توسط حاشیه نویس، مجزا شده باشند که از آن بتوان برای ارزیابی استفاده کرد). آدرس دیتاست پیشنهادی (smsspamcollection.zip)، در ادامه آورده شده است:

https://archive.ics.uci.edu/ml/machine-learning-databases/00228/

گام‌های ارزیابی روش پیشنهادی

هدف اصلی، یافتن یک مدل طبقه‌بندی است که بیشترین مقدار F1-Score را داشته باشد. بدین منظور برای ارزیابی پروژه، گام‌های زیر را بطور جداگانه ارزیابی کنید:

- **گام اول:** پیش پردازش
- **گام دوم**: گام اول + تولید ویژگی Ngram :
  - **بخش اول:** پیش پردازش + تولید ویژگی Unigram
  - **بخش دوم:** پیش پردازش + تولید ویژگی Bigram
  - **بخش سوم:** پیش پردازش + تولید ویژگی Trigram
- **گام سوم**: گام دوم + گام دوم (بهترین نتیجه از گام دوم یعنی از بین سه بخش مربوط به این گام، آن بخشی که بهترین F1-Score مربوطه را بدست بیآورد) + وزن‌دهی ویژگی (TFIDF)