باسمه تعالی

گزارش پروژه 2 NLP

مهدی فیض الهی 810199356

1. Spam detection

در این قسمت از روش multinomial naïve bayes استفاده کردم. اسم فایل spam.pyمی­باشد. مراحل کار در کل به این شرح است:

1. اول از همه با استفاده از chardet، encoding فایل csv را به دست آوردم. Encoding، Windows-1252 بود.
2. دو ستون v1و v2 را از dataframe استخراج کردم. و کل داده را به lowercase تغییر دادم.
3. وضعیت ham را کلاس 1 و وضعیت spam را کلاس 2 در نظر گرفتم.
4. در مرحله پیش پردازش، کلمه های stopword و punctuation ها رو پاک کردم.
5. با استفاده از کتابخانه CountVectorizer و TfIdfTransformer و MultinomialNB یک طبقه بند درست کردم و predict کردم.

نتایج:

|  |  |  |
| --- | --- | --- |
| ham | spam |  |
| 33 | 112 | Spam detect |
| 970 | 0 | Ham detect |