گزارش پروژه-
پردازش زبان طبیعی

محمدرضا اخگری زیری

۹۶۳۱۰۰۱

زمستان ۱۳۹۸

دسته بندی متون

در این پروژه قصد داریم تا به وسیله‌ی پردازش زبان طبیعی، متون فارسی را دسته‌بندی کنیم.

الگوریتم استفاده شده در این پروژه navie bayse میباشد و مدلهای زبانی استفاده شده، شامل: unigram و bigram میباشد.

در ابتدای کار، اقدام به خواندن محتوای فایل HAM-Train.txt کردیم. در هر خط از این فایل جمله ای مربوط به دسته بندی ای نوشته شده بود. با پردازش این خطوط توانستیم تعداد کلمات و تعداد زوج کلمات را در هر دسته حساب کنیم (فایل های آنها به همراه پروژه قرار داده شده است).

از نکات قابل توجه میتوان به این اشاره کرد که برای بالا بردن دقت علامت <S> را به ابتدای هر جمله اضافه کردم تا در زوج کلمات، کلمات اول جمله نیز حساب شود.

سپس اقدام به حواندن محتوای فایل HAM-Test.txt میکنیم. این فایل، برای آزمایش الگوریتم ماست. برای آزمایش هر خط را از فایل خوانده و در ابتدای آن نماد شروع <S> را میگذاریم. حال باید برای دسته بندی های مختلف ای که از قسمت قبل یافتیم، احتمال وجود تک کلمات و زوج کلمات را حساب کنیم.

برای هموار سازی از ترکیب خطی این احتمالات استفاده شده و با فرمول زیر حساب میشود:

$$p(i|i-1) = \lambda p_i + (1-\lambda)p(i|i-1)$$

مقدار λ سه مقدار متفاوت ۰,۲، ۰,۵ و ۰,۹ داده شد. و نتیجه آنها به همراه پروژه قرار داده شده است ولی بهترین حالت ۰,۵ شد.

برای پیدا کردن مقادیر احتمال از الگوریتم naive bayse استفاده شده است، که در این الگوریتم به دلیل کوچک بودن الگوریتم ها از ضرب آنها خودداری کردیم و مجموع log های آنها را حساب میکنیم. کلماتی که در training set قرار ندارند را با احتمال -۱۰۰۰ در نظر میگیریم.

برای ارزیابی نهایی پروژه از سه متغیر گفته شده در سوال استفاده میکنیم که به شرح زیر است:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, recall = \frac{TP}{TP + FN}, F = 2 * r * \frac{p}{r + p}$$

در صورت نزدیک بودن به یک یعنی پیاده سازی بهتری داشتیم.