Multinominal Naive Bayes ile Türkçe Cümlelerde Pozitiflik Sınıflaması

Kullanılan Yöntem

Projemiz bir text classification projesi olduğu için hem basit hem de efektif bir çözüm olarak Multinominal Naive Bayes kullanmayı ve baştan implente etmeyi seçtik.

Multinominal Naive Bayes ile ilgili Formüller

Bayes Teoremi

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A) \cdot \mathrm{P}(A)}{\mathrm{P}(B)}$$

- A ve B verilen durumlar
- P(A|B), B doğru iken A'nın gerçekleşme ihtimali (Posterior Probability)
- P(B|A), A doğru iken B'nın gerçekleşme ihtimali (Likelihood)
- P(A) (Class Prior Probability) ile P(B) (Predictor Prior Probability) ise A ile B olaylarının gerçekleşme ihtimalleridir.

Bayes Teoreminin Text Classification İçin Kullanılan Hali

$$\mathrm{P}(durum|c\ddot{u}mle) = \prod_{i=1}^{n} \mathrm{P}(kelime_i|durum) \cdot P(durum)$$

$$\mathrm{P}(kelime|durum) = \frac{nc + 1}{n + vocabulary}$$

- nc, durum gerçekleşirkenki girdi olarak verilen kelime sayısı.
- n, durum gerçekleşirkenki bütün kelimelerin sayısı.
- vocabulary, eşsiz kelime sayısı.

Test ve Training İçin Kullanılan Dataset

Geliştirme yaparken Kaggle'da bulunan Mustafa Keskin'e ait olan Turkish Movie Sentiment Analysis Dataset'i kullandık.

Kod Açıklamaları

```
import pandas as pd

stopwords=open("stopwords.txt","r")
stopwordsarray=stopwords.read().splitlines()

dataset = pd.read_csv('./dataset.csv', header=None, names=['Yorum', "Film" ,'Puan

dataset = dataset.tail(-1)
dataset = dataset.drop('Film', 1)

dataset = dataset.sample(frac=.005)
```

Pandas kütüphanesini data frame'leri oluşturmak için kullanıyoruz. Data frame'ler elimizdeki csv dosyasını bir çeşit iki boyutlu array'e dönüştürmemizi sağlıyor ve her feature'a bir isim atamamızı sağlıyor. Elimizdeki datayı tek seferde hafızaya yüklediğimiz için sadece binde beşlik bir kısmını rastgele yükleyerek çalıştırdık projemizi.

```
def map_points(x):
    x = float(x.replace(',', '.'))
    if x < 2.5:
        return False
    else:
        return True

def remove_newline(x):
    x = x.replace('\n', ' ')
    return x

dataset['Puan'] = dataset['Puan'].apply(map_points)
dataset['Yorum'] = dataset['Yorum'].apply(clean)
```

map_points ve remove_newline fonksiyonları adlarından da anlaşılacağı üzere sıra ile kullandığımız dataset'teki continuous değerleri discrete değerlere map etmeye ve "Yorum" column'u üzerindeki boşlukları silmeye yarıyor.