

به نام خدا



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

پروژه‌ی چهارم درس هوش مصنوعی و کاربردهای آن: پردازش زبان طبیعی

استاد درس: دکتر بهنام روشون فکر

دانیال حمدی – ۹۷۳۱۱۱۱

۱۴۰۰ بهار

۱. تأثیر حذف کلمات پر تکرار و کم تکرار در دقت به دست آمد

برای مقادیر $\lambda_1, \lambda_2, \lambda_3$ ثابت، نتایج مدل را با حذف و بدون حذف این کلمات مقایسه می‌کنیم:

$$\lambda_3: 0.6, \lambda_2: 0.3, \lambda_1: 0.1, \epsilon: 0.5$$

بدون حذف این کلمات:

```
AI_P4 - DanialH@danials-MacBook-Pro-2:~/PycharmProjects/AI_P4 | 80x20 | ttys001 — zsh — 80x20
~/PycharmProjects/AI_P4 └─ DanialH danials-MacBook-Pro-2:s001 └─
└─(14:11:46)─> python3 main.py └─(Tue, Jul 20)─
Using unigram_predict
#True Positives: 144, #False Positives: 37
#True Negatives: 497, #False Negatives: 390
#Accuracy: 0.600187265917603
Precision: 0.2696629213483146, Recall: 0.7955801104972375
F1 Score: 0.40279720279720277

Using bigram_predict
#True Positives: 373, #False Positives: 194
#True Negatives: 340, #False Negatives: 161
#Accuracy: 0.6676029962546817
Precision: 0.6985018726591761, Recall: 0.6578483245149912
F1 Score: 0.6775658492279747

~/PycharmProjects/AI_P4 └─ DanialH danials-MacBook-Pro-2:s001 └─
└─(14:11:48)─>
```

با حذف این کلمات:

```
AI_P4 - DanialH@danials-MacBook-Pro-2:~/PycharmProjects/AI_P4 | 80x20 | ttys001 — zsh — 80x20
~/PycharmProjects/AI_P4 └─ DanialH danials-MacBook-Pro-2:s001 └─
└─(14:12:30)─> python3 main.py └─(Tue, Jul 20)─
Using unigram_predict
#True Positives: 9, #False Positives: 0
#True Negatives: 534, #False Negatives: 525
#Accuracy: 0.5084269662921348
Precision: 0.016853932584269662, Recall: 1.0
F1 Score: 0.03314917127071823

Using bigram_predict
#True Positives: 205, #False Positives: 93
#True Negatives: 441, #False Negatives: 329
#Accuracy: 0.6048689138576779
Precision: 0.3838951310861423, Recall: 0.6879194630872483
F1 Score: 0.49278846153846156

~/PycharmProjects/AI_P4 └─ DanialH danials-MacBook-Pro-2:s001 └─
└─(14:12:31)─>
```

همچنین برای مقادیر زیر خواهیم داشت.

$$\lambda_3: 0.5, \lambda_2: 0.3, \lambda_1: 0.2, \epsilon: 0.3$$

بدون حذف این کلمات:

```
AI_P4 — DanialH@danials-MacBook-Pro-2:~/PycharmProjects/AI_P4 | 80x20 | ttys001 — zsh — 80x20
└─(~/PycharmProjects/AI_P4) ─────────── DanialH@danials-MacBook-Pro-2:s001 ─
  └─(14:16:11)─> python3 main.py
Using unigram_predict
#True Positives: 154, #False Positives: 40
#True Negatives: 494, #False Negatives: 380
#Accuracy: 0.6067415730337079
Precision: 0.2883895131086142, Recall: 0.7938144329896907
F1 Score: 0.4230769230769231

Using bigram_predict
#True Positives: 356, #False Positives: 181
#True Negatives: 353, #False Negatives: 178
#Accuracy: 0.6638576779026217
Precision: 0.6666666666666666, Recall: 0.6629422718808193
F1 Score: 0.6647992530345472

└─(~/PycharmProjects/AI_P4) ─────────── DanialH@danials-MacBook-Pro-2:s001 ─
  └─(14:16:11)─> |
```

با حذف این کلمات:

```
AI_P4 — DanialH@danials-MacBook-Pro-2:~/PycharmProjects/AI_P4 | 80x20 | ttys001 — zsh — 80x20
└─(~/PycharmProjects/AI_P4) ─────────── DanialH@danials-MacBook-Pro-2:s001 ─
  └─(14:16:36)─> python3 main.py
Using unigram_predict
#True Positives: 6, #False Positives: 4
#True Negatives: 530, #False Negatives: 528
#Accuracy: 0.50187265917603
Precision: 0.011235955056179775, Recall: 0.6
F1 Score: 0.022058823529411763

Using bigram_predict
#True Positives: 207, #False Positives: 100
#True Negatives: 434, #False Negatives: 327
#Accuracy: 0.600187265917603
Precision: 0.38764044943820225, Recall: 0.6742671009771987
F1 Score: 0.49227110582639716

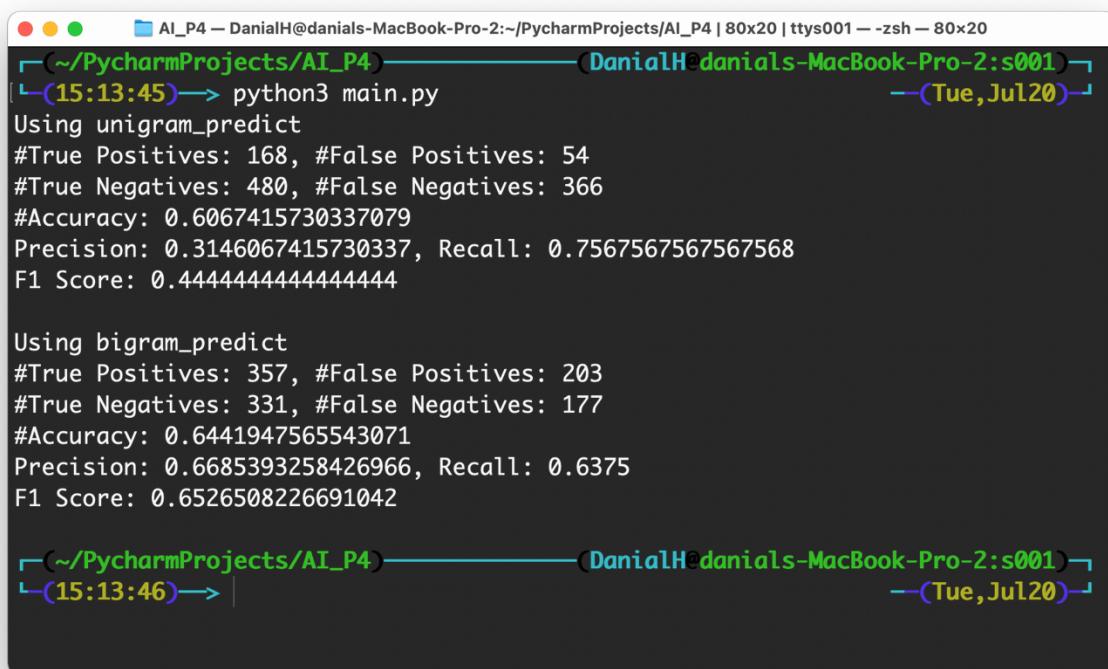
└─(~/PycharmProjects/AI_P4) ─────────── DanialH@danials-MacBook-Pro-2:s001 ─
  └─(14:16:37)─>
```

انتظار داشتیم با حذف کلمات پر تکرار و کم تکرار، به دقت بهتری دست یابیم؛ چرا که کلمات پر تکرار و کم تکرار، عموماً نمی‌توانند نشانه‌ی خوبی برای دسته‌بندی یک جمله باشند. این کلمات عموماً بار خشی دارند، به عنوان مثال کلمه‌هایی مانند *It*, *I*, *a*, *the* از این دسته کلمات هستند.

۲. تأثیر مقدار λ و ϵ در دقت به دست آمده

برای مقادیر مختلف، دقت به دست آمده را بررسی می‌کنیم.

$$\lambda_3: 0.6, \lambda_2: 0.2, \lambda_1: 0.2, \epsilon: 0.3$$



```

AI_P4 — DanialH@danials-MacBook-Pro-2:~/PycharmProjects/AI_P4 | 80x20 | ttys001 — zsh — 80x20
└─(~/PycharmProjects/AI_P4) ─────────── (DanialH@danials-MacBook-Pro-2:s001) ─
  └─(15:13:45) ──> python3 main.py
  Using unigram_predict
  #True Positives: 168, #False Positives: 54
  #True Negatives: 480, #False Negatives: 366
  #Accuracy: 0.6067415730337079
  Precision: 0.3146067415730337, Recall: 0.7567567567567568
  F1 Score: 0.4444444444444444

  Using bigram_predict
  #True Positives: 357, #False Positives: 203
  #True Negatives: 331, #False Negatives: 177
  #Accuracy: 0.6441947565543071
  Precision: 0.6685393258426966, Recall: 0.6375
  F1 Score: 0.6526508226691042

```

$$\lambda_3: 0.7, \lambda_2: 0.2, \lambda_1: 0.1, \epsilon: 0.3$$

```
AI_P4 — DanialH@danials-MacBook-Pro-2:~/PycharmProjects/AI_P4 | 80x20 | ttys001 — zsh — 80x20
└─(~/PycharmProjects/AI_P4) ───────────(DanialH@danials-MacBook-Pro-2:s001)─
  └─(15:14:31)─> python3 main.py
Using unigram_predict
#True Positives: 147, #False Positives: 38
#True Negatives: 496, #False Negatives: 387
#Accuracy: 0.6020599250936329
Precision: 0.2752808988764045, Recall: 0.7945945945945946
F1 Score: 0.4089012517385257

Using bigram_predict
#True Positives: 351, #False Positives: 211
#True Negatives: 323, #False Negatives: 183
#Accuracy: 0.6310861423220974
Precision: 0.6573033707865169, Recall: 0.6245551601423488
F1 Score: 0.6405109489051095

└─(~/PycharmProjects/AI_P4) ───────────(DanialH@danials-MacBook-Pro-2:s001)─
  └─(15:14:49)─> | ───────────(Tue, Jul 20)─
```

$$\lambda_3: 0.8, \lambda_2: 0.1, \lambda_1: 0.1, \epsilon: 0.3$$

```
AI_P4 — DanialH@danials-MacBook-Pro-2:~/PycharmProjects/AI_P4 | 80x20 | ttys001 — zsh — 80x20
└─(~/PycharmProjects/AI_P4) ───────────(DanialH@danials-MacBook-Pro-2:s001)─
  └─(15:15:29)─> python3 main.py
Using unigram_predict
#True Positives: 165, #False Positives: 41
#True Negatives: 493, #False Negatives: 369
#Accuracy: 0.6161048689138576
Precision: 0.3089887640449438, Recall: 0.8009708737864077
F1 Score: 0.44594594594594583

Using bigram_predict
#True Positives: 378, #False Positives: 208
#True Negatives: 326, #False Negatives: 156
#Accuracy: 0.6591760299625468
Precision: 0.7078651685393258, Recall: 0.6450511945392492
F1 Score: 0.675

└─(~/PycharmProjects/AI_P4) ───────────(DanialH@danials-MacBook-Pro-2:s001)─
  └─(15:15:37)─> | ───────────(Tue, Jul 20)─
```

۳. بهترین دقت دست یافته و تحلیل تأثیر پارامترها در آن

بهترین دقت دست یافته برای مدل Bigram برابر ۰.۶۹ بود، که برای دست یافتن به آن، کلمات پر تکرار و کم تکرار حذف شده، و وزن های λ_2 و λ_3 را به نسبت بالاتر گذاشتیم.