CI-Final project

Recommender System

Reyhaneh Kharazmi
97243032

Dr.Malek
January 28, 2022

فاز اول

Preprocessing

در این فاز داده های ستون های *comment , title , advantages* و *disadvantages* را توسط کتابخانه *hazm* نرمال می کنیم و سپس اعداد و *English words* و *stop words* را از آن ها حذف می کنیم.

از کتابخانه *swifter* نیز برای سریع تر شدن این اقدامات استفاده شد.

حال برای انتخاب فیچرهایی که به مدل بدهیم ۴ فیچر *comment , title , advantages* و *disadvantages* را به هم وصل می کنیم و یک ستون جدید به نام *features* را برای دادن مدل در نظر می گیریم.

و با استفاده از روش *TF-IDF* متن ستون *features* را *vectorize* می کنیم.

و پس از آن نیز توسط تابع *Tokenizer* از کتابخانه *keras* متن را *tokenize* می کنیم و *padding* مناسب برای آن قرار می دهیم.

فاز دوم

Model selection

با مطالعه شبکه های RNN و بهره گیری از کتابخانه keras مدل LSTM را پیاده سازی کردیم:

مدل اول bidirectional است و از activation softmax استفاده می کند و optimizer آن نیز Adam است و 20 درصد داده ها را هم به عنوان validation در نظر گرفتیم تا تفاوت accuracy آن ها را مشاهده کنیم.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 300, 100)          4500000

 spatial_dropout1d (SpatialD  (None, 300, 100)         0
 ropout1D)

 bidirectional (Bidirectiona  (None, 1024)             2510848
 l)

 dense (Dense)               (None, 2)                 2050

=================================================================
Total params: 7,012,898
Trainable params: 7,012,898
Non-trainable params: 0
_________________________________________________________________
None
```

Epoch = 6

Batch-size = 64

train-accuracy = 97.89

Val-accuracy = 93.75

برای همین مدل تعداد epoch ها را به ۲۰ رساندیم و نتیجه :

train-accuracy = 99.46

Val-accuracy = 93.71

مدل دوم نیز CNN است.

Activation = relu and softmax

Optimizer = adam

Batch-size = 64

Epoch = 6

Train-acc = 99.77

Val-acc = 92.89

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_2 (Embedding)     (None, 300, 100)          4500000

 conv1d (Conv1D)             (None, 294, 128)          89728

 global_max_pooling1d (Globa  (None, 128)              0
 lMaxPooling1D)

 dense_2 (Dense)             (None, 512)               66048

 dropout (Dropout)           (None, 512)               0

 dense_3 (Dense)             (None, 128)               65664

 dropout_1 (Dropout)         (None, 128)               0

 dense_4 (Dense)             (None, 2)                 258
```

مدل سوم ترکیبی از Lstm و Cnn است

Optimizer = rmsprop

Epoch = 10

Train-acc = 97.17

Val-acc = 94.41