# پروژه پایانی درس هوش مصنوعی و سیستم های خبره


## محمدحسین ارسلان 98243005


### پرژه شمارهٔ یک

توضیحات: سعی شده است در طول این نوتبوک توضیحات کامل به منظور جایگزین برای گزارش نوشته شود، همچنین در نقاطی سعی شده با چاپ کردن دیتافریم های حاصل و یا لیست ها و متغیرها روند کار را نشان داده و ابهامی به جای نگذارم.

# Import the importants

In [1]:
```
import pandas as pd
import numpy as np
import hazm
from __future__ import unicode_literals
from hazm import *
```

# Reading train data

برای شروع نیاز به یک داده داریم که ما داده آموزشی قرار داده شده در ککل را دریافت میکنیم

In [2]:
```
df = pd.read_csv('train.csv')
```

# Thanks to hazm Library :)

سپس یک ابزار از کتابخانه هضم را به برنامه خود اضافه می کنیم و در طول پیش پردازش با آن سر و کار داریم

این ابزار مسوول نرمالایز کردن جلات و حذف نیم فاصله ها و تشخیص یک سری حروف که برای جمع کردن کلمات

به کار می روند است.

همچنین یک تابع از این کتابخانه برای توکن بندی کردن کلمات داریم که در تابع توکنایز خودمان از آن بهره برده ایم.

In [3]:
```
nrm = Normalizer() #some verbs in persian got two parts which are spaced, this module help
```

# Normalize, Tokenize and Cleaning data

اصل کار اینجا و در سلول زیر است چرا که داده ها که گفته های ما هستند باید توکن بندی شوند و به

کمک این توکن ها باید پیشبینی لیبل صورت گیرد. از این رو ما با پیمایشی روی جملات یا کوئری ها

داریم و هربار یک کوئری را به تابع توکنایز می دهیم و در ستونی جدید از دیتافریم آموزش و تست این

خروجی تابع توکنایز که یک لیست از توکن های جمله نرمال شده پاس داده شده است قرار می دهیم و

مجدد در تابع کا-فولد روی این ستون توکن ها که ستون جدید دیتافریممان است پیمایش کرده و روی هر

لیست موجود در آن تابع تمیزکننده را اجرا می کنیم که هرسری بررسی می کند و فقط کلماتی را در لیست

پاس داده شده نگه می دارد که در لیست اضافه (خودم تعریف کرده ام و یک حالت دیکشنری گونه دارد)

نباشد را در لیست توکن ها نگه می دارد و با این کار داده ها تمیز می شوند و توکن های هر جمله ارزش

معنایی بالایی پیدا می کنند.

In [14]:
```
extra = ['؛','،','؟','.','باشد','است','نیست','هست','.','؟',';','?','.','-','/','_'
         ,'با','از','یا','چه','چیه','کیه','چطوریه','چقدر','کی','این','ان','آن','اون'
         ,'برای','را','رو','تو','من','ما','توی','چرا','هر','و','چقدره','آیا','چند'
        ,'0','1','2','3','4','5','6','7','8','9','به','تا','!']
def tokenize(sentence):
    sentence = nrm.normalize(sentence)
    sentence = word_tokenize(sentence)
    return sentence
def clean(tokens):
    tokens = list(filter(lambda i : extra.count(i) == 0, tokens))
    return tokens
```

# Prediction function

## (Based on naive bayes)

در بخش زیر یک سری تابع که برای محاسبات الگوریتم مورد نظر نیاز است طراحی و پیاده سازی شده است

تابع اصلی این بخش که در تابع کا-فولد هم صدا زده می شود، تابع پیشبینی است، این تابع به ازای هر لیستی

از توکن های سطرهای داده تست صدا زده می شود. یکسری دیتافریم که مخصوص هر لیبل بود را نیز به آن

پاس می دهیم. این تابع در درون خود احتمال اینکه لیست توکن های ما متعلق به هریک از کلاس ها باشند را

به کمک 5 تابع محاسبه احتمال به دست می آورد و پس از مقایسه آن ها بایکدیگر، یک لیبل به ما می دهد که

همان مقدار پیشبینی شده است.

در احتمال هایی که در 5 تابع یادشده باید محاسبه شوند مخرج کسر ها در هر لیبل مقدار ثابتی است و متشکل

از مجموع تعداد توکن های متمایز کل داده آموزشی و مجموع کل توکن های داده های آن لیبل است که برای

جلوگیری از تکرار آن ها را داخل تابع پردیکت محاسبه کرده و به توابع احتمال پاس می. دهیم

In [15]:

```
def probability_label1(p,tokens,l,unique,label_words):
    probe = p
#     prob *= for each token => ((tekrar kalame dar jomlat in label + 1) / (unique words i
    repeats = list(map(lambda i : l.count(i),tokens))

    for i in range(0,len(repeats)):
        probe *= (repeats[i] + 1)/(unique + label_words)
    return probe
def probability_label2(p,tokens,l,unique,label_words):
    probe = p
    repeats = list(map(lambda i : l.count(i),tokens))

    for i in range(0,len(repeats)):
        probe *= (repeats[i] + 1)/(unique + label_words)
    return probe
def probability_label3(p,tokens,l,unique,label_words):
    probe = p
    repeats = list(map(lambda i : l.count(i),tokens))

    for i in range(0,len(repeats)):
        probe *= (repeats[i] + 1)/(unique + label_words)
    return probe
def probability_label4(p,tokens,l,unique,label_words):
    probe = p
    repeats = list(map(lambda i : l.count(i),tokens))

    for i in range(0,len(repeats)):
        probe *= (repeats[i] + 1)/(unique + label_words)
    return probe
def probability_label5(p,tokens,l,unique,label_words):
    probe = p
    repeats = list(map(lambda i : l.count(i),tokens))

    for i in range(0,len(repeats)):
        probe *= (repeats[i] + 1)/(unique + label_words)
    return probe
def predict(sentence,tr,l1,l2,l3,l4,l5):
    ################### now we find probability of class(label) of each sentence ########
    p1 = len(l1['query']) / len(tr['query'])
    p2 = len(l2['query']) / len(tr['query'])
    p3 = len(l3['query']) / len(tr['query'])
    p4 = len(l4['query']) / len(tr['query'])
    p5 = len(l5['query']) / len(tr['query'])
```

```
##########################also we need to find how many unique words we have in all ti
all_sentences = df['query'][0]
all_sentences = " ".join(df['query'][1:len(df['query'])])
all_sentences = tokenize(all_sentences)
all_sentences = clean(all_sentences)
unique_word_tokens = set(all_sentences)
unique = len(unique_word_tokens)


################################### we need all word tokens number in each label's s
l1_sentences = " ".join(l1['query'][0:len(l1['query'])])
l2_sentences = " ".join(l2['query'][0:len(l2['query'])])
l3_sentences = " ".join(l3['query'][0:len(l3['query'])])
l4_sentences = " ".join(l4['query'][0:len(l4['query'])])
l5_sentences = " ".join(l5['query'][0:len(l5['query'])])

l1_sentences = tokenize(l1_sentences)
l1_sentences = clean(l1_sentences)
label1_words = len(l1_sentences)
l2_sentences = tokenize(l2_sentences)
l2_sentences = clean(l2_sentences)
label2_words = len(l2_sentences)
l3_sentences = tokenize(l3_sentences)
l3_sentences = clean(l3_sentences)
label3_words = len(l3_sentences)
l4_sentences = tokenize(l4_sentences)
l4_sentences = clean(l4_sentences)
label4_words = len(l4_sentences)
l5_sentences = tokenize(l5_sentences)
l5_sentences = clean(l5_sentences)
label5_words = len(l5_sentences)

############################## now we merge all tokens into one list for finding each w
label1_tokens,label2_tokens,label3_tokens,label4_tokens,label5_tokens = [],[],[],[],[]

list_1 = list(l1['tokens'])
for i in list_1 :
    label1_tokens.extend(i)

list_2 = list(l2['tokens'])
for i in list_2 :
    label2_tokens.extend(i)

list_3 = list(l3['tokens'])
for i in list_3 :
    label3_tokens.extend(i)

list_4 = list(l4['tokens'])
for i in list_4 :
    label4_tokens.extend(i)

list_5 = list(l5['tokens'])
for i in list_5 :
    label5_tokens.extend(i)

################################### these are constant numbers during naive bayes al

label_1 = probability_label1(p1,sentence,label1_tokens,unique,label1_words)
label_2 = probability_label2(p2,sentence,label2_tokens,unique,label2_words)
label_3 = probability_label3(p3,sentence,label3_tokens,unique,label3_words)
label_4 = probability_label4(p4,sentence,label4_tokens,unique,label4_words)
label_5 = probability_label5(p5,sentence,label5_tokens,unique,label5_words)

maximum = 0
label = 0
```

```
    if label_1> maximum:
        maximum = label_1
        label = 1
    if label_2> maximum:
        maximum = label_2
        label = 2
    if label_3> maximum:
        maximum = label_3
        label = 3
    if label_4> maximum:
        maximum = label_4
        label = 4
    if label_5> maximum:
        maximum = label_5
        label = 5
    return label;
```

# K-fold function

حال تابع کا-فولد را داریم که ورودی اول آن برای تست های کا-فولد مهم نیست چرا که در اصل همان داده تستی است که برای ارسال در سایت ککگل نیاز داریم برای تشخیص اینکه کا-فولد میخواهیم یا تست ککگل، جهت جلوگیری از تکرار کد، یک پارامتر به نام الگوریتم دارم که اگر یک بود یعنی کا-فولد و اگر برابر با صفر بود یعنی تست ککگل دارد انجام می شود. سپس داده دیتافریم که برای کا-فولد معادل کل داده ها است و باید آموزش و تست را انتخاب کنیم ازش و برای تست ککگل برابر است با خود داده آموزش. ابتدا هر کوئری در داده آموزش و تست نرمالایز و توکن بندی می شود و سپس با تمیز کردن توکن ها و تقسیم بندی داده های آموزشی به 5 دیتافریم که هر دیتافریم مرتبط با گفته های یک لیبل است به سراغ انجام الگوریتم بیز ضعیف می رویم.

In [18]:
```
from sklearn.metrics import precision_recall_fscore_support
def k_fold(test_main,df,k,alg):


    if(alg == 1):#we want k-fold
        ################## based on 3-fold way we should #################
        if  (k == 1):
            train_set = df[0:2031]
            test_set = df[2032:3047]
        elif(k == 2):
            train_set = df[0:1015]
            train_set.append(df[2032:3047])
            test_set = df[1016:2031]
        else:
            train_set = df[1016:3047]
            test_set = df[0:1015]
```

```
    else:################## we work on main test set #################
        train_set = df
        test_set = test_main
    ################## we should give indexes to our train set #################
    train_set['index'] = pd.Series(range(0,len(train_set['query']),1)).values
######now we normalize and tokenize all sentences in train set and test set also we add a
    train_set['tokens'] = train_set['query'].apply(tokenize)
    test_set['tokens'] = test_set['query'].apply(tokenize)


################## now we clean the data, which is removing question marks, stop point, c
################## pronouns, some special verbs and others which are included in a list r
    train_set['tokens'] = train_set['tokens'].apply(clean)
    test_set['tokens'] = test_set['tokens'].apply(clean)
    ##### show them after cleaning #######
    print('-------train set after tokenizing and cleaning data is shown below-------')
    display(train_set)
#     train_set.to_csv('train_after_process.csv')
    print('-------test set after tokenizing and cleaning data is shown below-------')
    display(test_set)
################################### at last we divide train samples into 5 label based c
    train_lab1, train_lab2, train_lab3, train_lab4, train_lab5 = pd.DataFrame(), pd.DataFr

    train_lab1 = train_set[train_set['label'] == 1].set_index('label')
    train_lab2 = train_set[train_set['label'] == 2].set_index('label')
    train_lab3 = train_set[train_set['label'] == 3].set_index('label')
    train_lab4 = train_set[train_set['label'] == 4].set_index('label')
    train_lab5 = train_set[train_set['label'] == 5].set_index('label')

    ################## we need a new column named index which starts from 0 to length of
    train_lab1['index'] = pd.Series(range(0,len(train_lab1['tokens']),1)).values
    train_lab2['index'] = pd.Series(range(0,len(train_lab2['tokens']),1)).values
    train_lab3['index'] = pd.Series(range(0,len(train_lab3['tokens']),1)).values
    train_lab4['index'] = pd.Series(range(0,len(train_lab4['tokens']),1)).values
    train_lab5['index'] = pd.Series(range(0,len(train_lab5['tokens']),1)).values

    ################## now we should iterate on test set and predict each sentence's labe
    ################## we pass the current sentence into a function to predict and return
    ################## our procedure is based on modular programming
    test_set['prediction'] = test_set['tokens'].apply(predict,tr = train_set,l1 = train_la


    ##################  at the last part we need to find out percision score by two stand
    rates = precision_recall_fscore_support(test_set['label'],test_set['prediction'])
    print('Fscore details: {}'.format(rates))
    print('Percision for this test is : {}'.format(len(test_set[test_set['prediction'] ==

    return test_set
```

# K-fold tests for K = 1,2,3

سه مرتبه الگوریتم کا-فولد را اجرا می کنیم و هر سری دقت پیشبینی را بررسی می کنیم

سپس جهت بررسی یکبار بدون تمیزکاری و البته نرمالایز کردن داده ها این الگوریتم را

روی مرحله اول که کا معادل یک بود اجرا می کنیم و دقت آن را نیز با حالتی که نرمالایز

و تمیزکاری دیتا کمک کرده اند بررسی کرده ایم.

In [7]:
```
test_1,test_2,test_3 = pd.DataFrame(),pd.DataFrame(),pd.DataFrame()
print('test for 1-fold')
test_1 = k_fold(df,df,1,1)
print('-------dataframe after processing and predicting data is shown below-------')
display(test_1)
print('test for 2-fold')
test_2 = k_fold(df,df,2,1)
print('test for 3-fold')
test_3 = k_fold(df,df,3,1)
```

```
test for 1-fold
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:21: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['index'] = pd.Series(range(0,len(train_set['query']),1)).values
-------train set after tokenizing and cleaning data is shown below-------
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:23: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['tokens'] = train_set['query'].apply(tokenize)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:24: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['tokens'] = test_set['query'].apply(tokenize)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:29: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['tokens'] = train_set['tokens'].apply(clean)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:30: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['tokens'] = test_set['tokens'].apply(clean)
```

| | id | query | label | index | tokens |
|---|---|---|---|---|---|
| **0** | 0 | شرایط حذف ترم چیه؟ | 1 | 0 | [شرایط, حذف, ترم] |
| **1** | 1 | از کجا می تونم با دکتر وحیدی ارتباط برقرار کنم؟ | 2 | 1 | [کجا, می‌تونم, دکتر, وحیدی, ارتباط, برقرار, کنم] |
| **2** | 2 | بوفه برداران تا ساعت چند باز است؟ | 2 | 2 | [بوفه, برداران, ساعت, باز] |
| **3** | 3 | کمترین تعداد واحد چند عدد است؟ | 1 | 3 | [کمترین, تعداد, واحد, عدد] |
| **4** | 4 | سنگ جامد است | 5 | 4 | [سنگ, جامد] |
| **...** | ... | ... | ... | ... | ... |

| | id | query | label | index | tokens |
|---|---|---|---|---|---|
| **2026** | 2026 | انتخاب واحد دورودی ۹۸ چه زمان است؟ | 1 | 2026 | [انتخاب, واحد, دورودی, ۹۸, زمان] |
| **2027** | 2027 | اعضای هیئت علمی دانشکده ریاضی چه کسانی اند؟ | 2 | 2027 | [اعضای, هیئت, علمی, دانشکده, ریاضی, کسانی, اند] |
| **2028** | 2028 | تا حالا مهلت حذف تک درس تمدید شده ؟ | 1 | 2028 | [حالا, مهلت, حذف, تک, درس, تمدید, شده] |
| **2029** | 2029 | شماره صندلی های امتحانات را کجا ببینم؟ | 1 | 2029 | [شماره, صندلی‌های, امتحانات, کجا, ببینم] |
| **2030** | 2030 | بررسی و غربالگری سلامت دانشجویان به صورت سالان... | 4 | 2030 | [بررسی, غربالگری, سلامت, دانشجویان, صورت, سالا...] |

2031 rows × 5 columns

```
-------test set after tokenizing and cleaning data is shown below-------
```

| | id | query | label | tokens |
|---|---|---|---|---|
| **2032** | 2032 | کجا میتونم رمز پرتالم رو عوض کنم؟ | 1 | [کجا, میتونم, رمز, پرتالم, عوض, کنم] |
| **2033** | 2033 | سنگین ترین دانشجوی دانشکده کیست؟ | 5 | [سنگین‌ترین, دانشجوی, دانشکده, کیست] |
| **2034** | 2034 | چطور درست بردارم؟ | 1 | [چطور, درست, بردارم] |
| **2035** | 2035 | چطور عضو کتاب خونه بشم؟ | 3 | [چطور, عضو, کتاب, خونه, بشم] |
| **2036** | 2036 | .کولر اتاق 004 کار نمیکنه | 4 | [کولر, اتاق, ۰۰۴, کار, نمیکنه] |
| **...** | ... | ... | ... | ... |
| **3042** | 3042 | ...برخی استادها هیچ اهمیتی به اینکه دانشجو جز درس | 4 | ...برخی, استادها, هیچ, اهمیتی, اینکه, دانشجو, جز] |
| **3043** | 3043 | چند تا درس میشه حذف کرد | 1 | [درس, میشه, حذف, کرد] |
| **3044** | 3044 | جدید ترین ویرایش کتاب هریس که موجوده چیه؟ | 3 | [جدیدترین, ویرایش, کتاب, هریس, که, موجوده] |
| **3045** | 3045 | شرایط مهمان شدن در دانشکده ما چیست؟ | 1 | [شرایط, مهمان, شدن, در, دانشکده, چیست] |
| **3046** | 3046 | آمفی تئاتر دانشکده کامپیوتر کجاست؟ | 2 | [آمفی, تئاتر, دانشکده, کامپیوتر, کجاست] |

1015 rows × 4 columns

```
Fscore details: (array([0.83387622, 0.65172414, 0.88435374, 0.84090909, 0.71578947]), arra
y([0.88581315, 0.82894737, 0.83333333, 0.78723404, 0.44155844]), array([0.8590604 , 0.7297
2973, 0.85808581, 0.81318681, 0.54618474]), array([289, 228, 156, 188, 154], dtype=int64))
Percision for this test is : 0.7793103448275862
-------dataframe after processing and predicting data is shown below-------
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:55: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['prediction'] = test_set['tokens'].apply(predict,tr = train_set,l1 = train_lab
1,l2 = train_lab2,l3 = train_lab3,l4 = train_lab4,l5 = train_lab5)
```

| | id | query | label | tokens | prediction |
|---|---|---|---|---|---|
| **2032** | 2032 | کجا میتونم رمز پرتالم رو عوض کنم؟ | 1 | [کجا, میتونم, رمز, پرتالم, عوض, کنم] | 2 |
| **2033** | 2033 | سنگین ترین دانشجوی دانشکده کیست؟ | 5 | [سنگین‌ترین, دانشجوی, دانشکده, کیست] | 2 |
| **2034** | 2034 | چطور درست بردارم؟ | 1 | [چطور, درست, بردارم] | 1 |
| **2035** | 2035 | چطور عضو کتاب خونه بشم؟ | 3 | [چطور, عضو, کتاب, خونه, بشم] | 3 |
| **2036** | 2036 | .کولر اتاق 004 کار نمیکنه | 4 | [کولر, اتاق, ۰۰۴, کار, نمیکنه] | 2 |

| | id | query | label | tokens | prediction |
|---|---|---|---|---|---|
| **...** | ... | ... | ... | ... | ... |
| **3042** | 3042 | برخی استادها هیچ اهمیتی به اینکه دانشجو جز ...درس | 4 | [برخی, استادها, هیچ, اهمیتی, اینکه, دانشجو, جز...] | 4 |
| **3043** | 3043 | چند تا درس میشه حذف کرد | 1 | [درس, میشه, حذف, کرد] | 1 |
| **3044** | 3044 | جدید ترین ویرایش کتاب هریس که موجوده چیه؟ | 3 | [جدیدترین, ویرایش, کتاب, هریس, که, موجوده] | 3 |
| **3045** | 3045 | شرایط مهمان شدن در دانشکده ما چیست؟ | 1 | [شرایط, مهمان, شدن, در, دانشکده, چیست] | 1 |
| **3046** | 3046 | آمفی تئاتر دانشکده کامپیوتر کجاست؟ | 2 | [آمفی, تئاتر, دانشکده, کامپیوتر, کجاست] | 2 |

1015 rows × 5 columns

```
test for 2-fold
-------train set after tokenizing and cleaning data is shown below-------
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:21: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['index'] = pd.Series(range(0,len(train_set['query']),1)).values
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:23: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['tokens'] = train_set['query'].apply(tokenize)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:24: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['tokens'] = test_set['query'].apply(tokenize)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:29: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['tokens'] = train_set['tokens'].apply(clean)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:30: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['tokens'] = test_set['tokens'].apply(clean)
```

| | id | query | label | index | tokens |
|---|---|---|---|---|---|
| **0** | 0 | شرایط حذف ترم چیه؟ | 1 | 0 | [شرایط, حذف, ترم] |
| **1** | 1 | از کجا می تونم با دکتر وحیدی ارتباط برقرار کنم؟ | 2 | 1 | [کجا, می‌تونم, دکتر, وحیدی, ارتباط, برقرار, کنم] |
| **2** | 2 | بوفه برداران تا ساعت چند باز است؟ | 2 | 2 | [بوفه, برداران, ساعت, باز] |

| | id | query | label | index | tokens |
|---|---|---|---|---|---|
| **3** | 3 | کمترین تعداد واحد چند عدد است؟ | 1 | 3 | [کمترین, تعداد, واحد, عدد] |
| **4** | 4 | سنگ جامد است | 5 | 4 | [سنگ, جامد] |
| **...** | ... | ... | ... | ... | ... |
| **1010** | 1010 | .ایمیل بهشتی من کار نمی کند | 1 | 1010 | [ایمیل, بهشتی, کار, نمی‌کند] |
| **1011** | 1011 | کی افزایش ظرفیت میزنه به کلاسا ؟ | 1 | 1011 | [افزایش, ظرفیت, میزنه, کلاسا] |
| **1012** | 1012 | لابی دانشکده کجاست؟ | 2 | 1012 | [لابی, دانشکده, کجاست] |
| **1013** | 1013 | تعداد سرویس های بهداشتی طبقات کم است | 4 | 1013 | [تعداد, سرویس‌های, بهداشتی, طبقات, کم] |
| **1014** | 1014 | پوشش افرا د داخل دانشگاه مهمه؟ | 5 | 1014 | [پوشش, افرا, د, داخل, دانشگاه, مهمه] |

1015 rows × 5 columns

```
-------test set after tokenizing and cleaning data is shown below-------
```

| | id | query | label | tokens |
|---|---|---|---|---|
| **1016** | 1016 | برای دانشجویانVPN فراهم کردن | 4 | [برای, دانشجویانVPN ,فراهم, کردن] |
| **1017** | 1017 | فلان استاد ریکورد رو نمیزنه | 4 | [فلان, استاد, ریکورد, نمیزنه] |
| **1018** | 1018 | ترم آینده کی شروع میشه؟ | 1 | [ترم, آینده, شروع, میشه] |
| **1019** | 1019 | چرا نمره درس سیستم عامل برای من اشتباهی رد شده؟ | 1 | [نمره, درس, سیستم, عامل, اشتباهی, رد, شده] |
| **1020** | 1020 | اگه بیش از چند ترم مشروط بشیم چی میشه ؟ | 1 | [اگه, بیش, ترم, مشروط, بشیم, چی, میشه] |
| **...** | ... | ... | ... | ... |
| **2026** | 2026 | انتخاب واحد دورودی ۹۸ چه زمان است؟ | 1 | [انتخاب, واحد, دورودی, ۹۸, زمان] |
| **2027** | 2027 | اعضای هیئت علمی دانشکده ریاضی چه کسانی اند؟ | 2 | [اعضای, هیئت, علمی, دانشکده, ریاضی, کسانی, اند] |
| **2028** | 2028 | تا حالا مهلت حذف تک درس تمدید شده ؟ | 1 | [حالا, مهلت, حذف, تک, درس, تمدید, شده] |
| **2029** | 2029 | شماره صندلی های امتحانات را کجا ببینم؟ | 1 | [شماره, صندلی‌های, امتحانات, کجا, ببینم] |
| **2030** | 2030 | ...بررسی و غربالگری سلامت دانشجویان به صورت سالان | 4 | ...بررسی, غربالگری, سلامت, دانشجویان, صورت, سالا] |

1015 rows × 4 columns

```
Fscore details: (array([0.67987805, 0.63192182, 0.8137931 , 0.80701754, 0.796875  ]), arra
y([0.83208955, 0.82553191, 0.75641026, 0.71134021, 0.31481481]), array([0.74832215, 0.7158
6716, 0.78405316, 0.75616438, 0.45132743]), array([268, 235, 156, 194, 162], dtype=int64))
Percision for this test is : 0.7133004926108374
test for 3-fold
-------train set after tokenizing and cleaning data is shown below-------
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:55: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['prediction'] = test_set['tokens'].apply(predict,tr = train_set,l1 = train_lab
1,l2 = train_lab2,l3 = train_lab3,l4 = train_lab4,l5 = train_lab5)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:21: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
```

```
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['index'] = pd.Series(range(0,len(train_set['query']),1)).values
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:23: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['tokens'] = train_set['query'].apply(tokenize)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:24: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['tokens'] = test_set['query'].apply(tokenize)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:29: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['tokens'] = train_set['tokens'].apply(clean)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:30: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['tokens'] = test_set['tokens'].apply(clean)
```

| | id | query | label | index | tokens |
|---|---|---|---|---|---|
| **1016** | 1016 | برای دانشجویانVPN فراهم کردن | 4 | 0 | [برای, دانشجویانVPN ,فراهم, کردن] |
| **1017** | 1017 | فلان استاد ریکورد رو نمیزنه | 4 | 1 | [فلان, استاد, ریکورد, نمیزنه] |
| **1018** | 1018 | ترم آینده کی شروع میشه؟ | 1 | 2 | [ترم, آینده, شروع, میشه] |
| **1019** | 1019 | چرا نمره درس سیستم عامل برای من اشتباهی رد شده؟ | 1 | 3 | [نمره, درس, سیستم, عامل, اشتباهی, رد, شده] |
| **1020** | 1020 | اگه بیش از چند ترم مشروط بشیم چی میشه ؟ | 1 | 4 | [اگه, بیش, ترم, مشروط, بشیم, چی, میشه] |
| **...** | ... | ... | ... | ... | ... |
| **3042** | 3042 | ...برخی استادها هیچ اهمیتی به اینکه دانشجو جز درس | 4 | 2026 | ...برخی, استادها, هیچ, اهمیتی, اینکه, دانشجو, جز] |
| **3043** | 3043 | چند تا درس میشه حذف کرد | 1 | 2027 | [درس, میشه, حذف, کرد] |
| **3044** | 3044 | جدید ترین ویرایش کتاب هریس که موجوده چیه؟ | 3 | 2028 | [جدیدترین, ویرایش, کتاب, هریس, که, موجوده] |
| **3045** | 3045 | شرایط مهمان شدن در دانشکده ما چیست؟ | 1 | 2029 | [شرایط, مهمان, شدن, در, دانشکده, چیست] |
| **3046** | 3046 | آمفی تئاتر دانشکده کامپیوتر کجاست؟ | 2 | 2030 | [آمفی, تئاتر, دانشکده, کامپیوتر, کجاست] |

2031 rows × 5 columns

```
-------test set after tokenizing and cleaning data is shown below-------
```

| | id | query | label | tokens |
|---|---|---|---|---|
| **0** | 0 | شرایط حذف ترم چیه؟ | 1 | [شرایط, حذف, ترم] |
| **1** | 1 | از کجا می تونم با دکتر وحیدی ارتباط برقرار کنم؟ | 2 | [کجا, می‌تونم, دکتر, وحیدی, ارتباط, برقرار, کنم] |

| | id | query | label | tokens |
|---|---|---|---|---|
| **2** | 2 | بوفه برداران تا ساعت چند باز است؟ | 2 | [بوفه, برداران, ساعت, باز] |
| **3** | 3 | کمترین تعداد واحد چند عدد است؟ | 1 | [کمترین, تعداد, واحد, عدد] |
| **4** | 4 | سنگ جامد است | 5 | [سنگ, جامد] |
| **...** | ... | ... | ... | ... |
| **1010** | 1010 | .ایمیل بهشتی من کار نمی کند | 1 | [ایمیل, بهشتی, کار, نمی‌کند] |
| **1011** | 1011 | کی افزایش ظرفیت میزنه به کلاسا ؟ | 1 | [افزایش, ظرفیت, میزنه, کلاسا] |
| **1012** | 1012 | لابی دانشکده کجاست؟ | 2 | [لابی, دانشکده, کجاست] |
| **1013** | 1013 | تعداد سرویس های بهداشتی طبقات کم است | 4 | [تعداد, سرویس‌های, بهداشتی, طبقات, کم] |
| **1014** | 1014 | پوشش افرا د داخل دانشگاه مهمه؟ | 5 | [پوشش, افرا, د, داخل, دانشگاه, مهمه] |

1015 rows × 4 columns

```
Fscore details: (array([0.75816993, 0.7076412 , 0.8496732 , 0.83425414, 0.77027027]), arra
y([0.88212928, 0.85887097, 0.8496732 , 0.76262626, 0.37254902]), array([0.81546573, 0.7759
5628, 0.8496732 , 0.79683377, 0.50220264]), array([263, 248, 153, 198, 153], dtype=int64))
Percision for this test is : 0.7714285714285715
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:55: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['prediction'] = test_set['tokens'].apply(predict,tr = train_set,l1 = train_lab
1,l2 = train_lab2,l3 = train_lab3,l4 = train_lab4,l5 = train_lab5)
```

## در ادامه تست بدون نورمالایز و البته تمیزشدن دیتا انجام می شود که دقت !!!!!!!! بالاتری پیدا می کند نسبت به حالت قبل

# دلیل خطا در دقت

## بنظرم عدمت توزیع مناسب جمله ها در داده های تست و آموزش می تواند موجب این خطا بشود، نبودن تنوع میان جملات

## می توانست تکرار در لیبل ها را زیاد کند و این امر موجب افزایش دقت در حالتی است که حروف اضافه و غیره حذف نشدن

In [13]:
```
print('test for 1-fold without normalizing and cleaning data')
test_1_withNoNormalizeAndClean = k_fold(df,df,1,1)
print('-------dataframe after processing and predicting data is shown below-------')
display(test_1_withNoNormalizeAndClean)
```

```
test for 1-fold without normalizing and cleaning data
-------train set after tokenizing and cleaning data is shown below-------
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:21: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
```

```
  train_set['index'] = pd.Series(range(0,len(train_set['query']),1)).values
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:23: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['tokens'] = train_set['query'].apply(tokenize)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:24: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['tokens'] = test_set['query'].apply(tokenize)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:29: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  train_set['tokens'] = train_set['tokens'].apply(clean)
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:30: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['tokens'] = test_set['tokens'].apply(clean)
```

| | id | query | label | index | tokens |
|---|---|---|---|---|---|
| **0** | 0 | شرایط حذف ترم چیه؟ | 1 | 0 | [شرایط, حذف, ترم, چیه, ؟] |
| **1** | 1 | از کجا می تونم با دکتر وحیدی ارتباط برقرار کنم؟ | 2 | 1 | ...از, کجا, می, تونم, با, دکتر, وحیدی, ارتباط, ب] |
| **2** | 2 | بوفه برداران تا ساعت چند باز است؟ | 2 | 2 | [بوفه, برداران, تا, ساعت, چند, باز, است, ؟] |
| **3** | 3 | کمترین تعداد واحد چند عدد است؟ | 1 | 3 | [کمترین, تعداد, واحد, چند, عدد, است, ؟] |
| **4** | 4 | سنگ جامد است | 5 | 4 | [سنگ, جامد, است] |
| **...** | ... | ... | ... | ... | ... |
| **2026** | 2026 | انتخاب واحد دورودی ۹۸ چه زمان است؟ | 1 | 2026 | [انتخاب, واحد, دورودی, ۹۸, چه, زمان, است, ؟] |
| **2027** | 2027 | اعضای هیئت علمی دانشکده ریاضی چه کسانی اند؟ | 2 | 2027 | اعضای, هیئت, علمی, دانشکده, ریاضی, چه, ...,کسانی |
| **2028** | 2028 | تا حالا مهلت حذف تک درس تمدید شده ؟ | 1 | 2028 | [تا, حالا, مهلت, حذف, تک, درس, تمدید, شده, ؟] |
| **2029** | 2029 | شماره صندلی های امتحانات را کجا ببینم؟ | 1 | 2029 | [شماره, صندلی, های, امتحانات, را, کجا, ببینم, ؟] |
| **2030** | 2030 | بررسی و غربالگری سلامت دانشجویان به صورت ...سالان | 4 | 2030 | ...بررسی, و, غربالگری, سلامت, دانشجویان, به, صور] |

2031 rows × 5 columns

```
-------test set after tokenizing and cleaning data is shown below-------
```

| | id | query | label | tokens |
|---|---|---|---|---|
| **2032** | 2032 | کجا میتونم رمز پرتالم رو عوض کنم؟ | 1 | [کجا, میتونم, رمز, پرتالم, رو, عوض, کنم, ؟] |
| **2033** | 2033 | سنگین ترین دانشجوی دانشکده کیست؟ | 5 | [سنگین, ترین, دانشجوی, دانشکده, کیست, ؟] |

| | id | query | label | tokens |
|---|---|---|---|---|
| **2034** | 2034 | چطور درست بردارم؟ | 1 | [چطور, درست, بردارم, ؟] |
| **2035** | 2035 | چطور عضو کتاب خونه بشم؟ | 3 | [چطور, عضو, کتاب, خونه, بشم, ؟] |
| **2036** | 2036 | .کولر اتاق 004 کار نمیکنه | 4 | [. ,کولر, اتاق, 004, کار, نمیکنه] |
| **...** | ... | ... | ... | ... |
| **3042** | 3042 | ...برخی استادها هیچ اهمیتی به اینکه دانشجو جز درس | 4 | ...برخی, استادها, هیچ, اهمیتی, به, اینکه, دانشجو] |
| **3043** | 3043 | چند تا درس میشه حذف کرد | 1 | [چند, تا, درس, میشه, حذف, کرد] |
| **3044** | 3044 | جدید ترین ویرایش کتاب هریس که موجوده چیه؟ | 3 | ...جدید, ترین, ویرایش, کتاب, هریس, که, موجوده, چ] |
| **3045** | 3045 | شرایط مهمان شدن در دانشکده ما چیست؟ | 1 | [شرایط, مهمان, شدن, در, دانشکده, ما, چیست, ؟] |
| **3046** | 3046 | آمفی تئاتر دانشکده کامپیوتر کجاست؟ | 2 | [آمفی, تئاتر, دانشکده, کامپیوتر, کجاست, ؟] |

1015 rows × 4 columns

```
Fscore details: (array([0.81875   , 0.68592058, 0.89115646, 0.90163934, 0.76136364]), arra
y([0.90657439, 0.83333333, 0.83974359, 0.87765957, 0.43506494]), array([0.86042693, 0.7524
7525, 0.86468647, 0.88948787, 0.55371901]), array([289, 228, 156, 188, 154], dtype=int64))
Percision for this test is : 0.8029556650246306
-------dataframe after processing and predicting data is shown below-------
C:\Users\mhars\AppData\Local\Temp/ipykernel_2412/2408217947.py:55: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_gu
ide/indexing.html#returning-a-view-versus-a-copy
  test_set['prediction'] = test_set['tokens'].apply(predict,tr = train_set,l1 = train_lab
1,l2 = train_lab2,l3 = train_lab3,l4 = train_lab4,l5 = train_lab5)
```

| | id | query | label | tokens | prediction |
|---|---|---|---|---|---|
| **2032** | 2032 | کجا میتونم رمز پرتالم رو عوض کنم؟ | 1 | [کجا, میتونم, رمز, پرتالم, رو, عوض, کنم, ؟] | 2 |
| **2033** | 2033 | سنگین ترین دانشجوی دانشکده کیست؟ | 5 | [سنگین, ترین, دانشجوی, دانشکده, کیست, ؟] | 2 |
| **2034** | 2034 | چطور درست بردارم؟ | 1 | [چطور, درست, بردارم, ؟] | 1 |
| **2035** | 2035 | چطور عضو کتاب خونه بشم؟ | 3 | [چطور, عضو, کتاب, خونه, بشم, ؟] | 3 |
| **2036** | 2036 | .کولر اتاق 004 کار نمیکنه | 4 | [. ,کولر, اتاق, 004, کار, نمیکنه] | 4 |
| **...** | ... | ... | ... | ... | ... |
| **3042** | 3042 | برخی استادها هیچ اهمیتی به اینکه دانشجو جز ...درس | 4 | برخی, استادها, هیچ, اهمیتی, به, اینکه,] ...دانشجو | 4 |
| **3043** | 3043 | چند تا درس میشه حذف کرد | 1 | [چند, تا, درس, میشه, حذف, کرد] | 1 |
| **3044** | 3044 | جدید ترین ویرایش کتاب هریس که موجوده چیه؟ | 3 | جدید, ترین, ویرایش, کتاب, هریس, که, موجوده,] ...چ | 3 |
| **3045** | 3045 | شرایط مهمان شدن در دانشکده ما چیست؟ | 1 | [شرایط, مهمان, شدن, در, دانشکده, ما, چیست, ؟] | 1 |
| **3046** | 3046 | آمفی تئاتر دانشکده کامپیوتر کجاست؟ | 2 | [آمفی, تئاتر, دانشکده, کامپیوتر, کجاست, ؟] | 2 |

1015 rows × 5 columns

# Final test

حال ما تابع طراحی شده را یکبار با قراردادن متغیر صفر بعنوان آخرین پارامترش صدا میزنیم البته نام کا-فولد دیگر مناسب نیست اما برای جلوگیری از تکرار کد همین تابع را استفاده کرده ام تا روی مجموعه آموزش و تست اصلی جهت ارسال در سایت ککگل کار کنم

:)

In [20]:
```
test_main = pd.read_csv('test.csv')
final = k_fold(test_main,df,0,0)
final.to_csv('result.csv')
print('-------final test after prediction is shown as below-------')
display(final)
```

```
-------train set after tokenizing and cleaning data is shown below-------
```

| | id | query | label | index | tokens |
|---|---|---|---|---|---|
| **0** | 0 | شرایط حذف ترم چیه؟ | 1 | 0 | [شرایط, حذف, ترم] |
| **1** | 1 | از کجا می تونم با دکتر وحیدی ارتباط برقرار کنم؟ | 2 | 1 | [کجا, می‌تونم, دکتر, وحیدی, ارتباط, برقرار, کنم] |
| **2** | 2 | بوفه برداران تا ساعت چند باز است؟ | 2 | 2 | [بوفه, برداران, ساعت, باز] |
| **3** | 3 | کمترین تعداد واحد چند عدد است؟ | 1 | 3 | [کمترین, تعداد, واحد, عدد] |
| **4** | 4 | سنگ جامد است | 5 | 4 | [سنگ, جامد] |
| **...** | ... | ... | ... | ... | ... |
| **3043** | 3043 | چند تا درس میشه حذف کرد | 1 | 3043 | [درس, میشه, حذف, کرد] |
| **3044** | 3044 | جدید ترین ویرایش کتاب هریس که موجوده چیه؟ | 3 | 3044 | [جدیدترین, ویرایش, کتاب, هریس, که, موجوده] |
| **3045** | 3045 | شرایط مهمان شدن در دانشکده ما چیست؟ | 1 | 3045 | [شرایط, مهمان, شدن, در, دانشکده, چیست] |
| **3046** | 3046 | آمفی تئاتر دانشکده کامپیوتر کجاست؟ | 2 | 3046 | [آمفی, تئاتر, دانشکده, کامپیوتر, کجاست] |
| **3047** | 3047 | اسانسور را درست نمیکنید؟ | 4 | 3047 | [اسانسور, درست, نمیکنید] |

3048 rows × 5 columns

```
-------test set after tokenizing and cleaning data is shown below-------
```

| | id | query | tokens |
|---|---|---|---|
| **0** | 0 | چرا آخر ترم درس ها انقدر فشرده میشوند؟ | [آخر, ترم, درس‌ها, انقدر, فشرده, میشوند] |
| **1** | 1 | فرجه این ترم چقدر است؟ | [فرجه, ترم] |
| **2** | 2 | !صندلی های دانشگاه را ابری کنید | [صندلی‌های, دانشگاه, ابری, کنید] |
| **3** | 3 | محل تشکیل امتحان | [محل, تشکیل, امتحان] |
| **4** | 4 | دانشکده زیراکس دارد؟ | [دانشکده, زیراکس, دارد] |
| **...** | ... | ... | ... |
| **757** | 757 | آیا پنج شنبه ها دانشگاه تعطیله؟ | [پنج, شنبه‌ها, دانشگاه, تعطیله] |
| **758** | 758 | آزمایشگاه شبکه کجاست؟ | [آزمایشگاه, شبکه, کجاست] |
| **759** | 759 | ترم تابستان از چه تاریخی آغاز میشود؟ | [ترم, تابستان, تاریخی, آغاز, میشود] |
| **760** | 760 | آلودگی امروز چجوریه؟ | [آلودگی, امروز, چجوریه] |
| **761** | 761 | مزایای ازدواج دانشجویی چیست؟ | [مزایای, ازدواج, دانشجویی, چیست] |

762 rows × 3 columns

-------final test after prediction is shown as below-------

| | id | query | tokens | prediction |
|---|---|---|---|---|
| **0** | 0 | چرا آخر ترم درس ها انقدر فشرده میشوند؟ | [آخر, ترم, درس‌ها, انقدر, فشرده, میشوند] | 4 |
| **1** | 1 | فرجه این ترم چقدر است؟ | [فرجه, ترم] | 1 |
| **2** | 2 | !صندلی های دانشگاه را ابری کنید | [صندلی‌های, دانشگاه, ابری, کنید] | 4 |
| **3** | 3 | محل تشکیل امتحان | [محل, تشکیل, امتحان] | 2 |
| **4** | 4 | دانشکده زیراکس دارد؟ | [دانشکده, زیراکس, دارد] | 2 |
| **...** | ... | ... | ... | ... |
| **757** | 757 | آیا پنج شنبه ها دانشگاه تعطیله؟ | [پنج, شنبه‌ها, دانشگاه, تعطیله] | 2 |
| **758** | 758 | آزمایشگاه شبکه کجاست؟ | [آزمایشگاه, شبکه, کجاست] | 2 |
| **759** | 759 | ترم تابستان از چه تاریخی آغاز میشود؟ | [ترم, تابستان, تاریخی, آغاز, میشود] | 1 |
| **760** | 760 | آلودگی امروز چجوریه؟ | [آلودگی, امروز, چجوریه] | 5 |
| **761** | 761 | مزایای ازدواج دانشجویی چیست؟ | [مزایای, ازدواج, دانشجویی, چیست] | 1 |

762 rows × 4 columns