

# Password Strength Classifier(NLP)

October 8, 2022

```
In [4]: import pandas as pd
import numpy as np
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

## Fetching our dataset from root folder

```
In [5]: data=pd.read_csv('./dataset/data.csv',error_bad_lines=False)
data.head()
```

```
b'Skipping line 2810: expected 2 fields, saw 5\nSkipping line 4641: expected 2 fields, saw 5\n
b'Skipping line 264626: expected 2 fields, saw 5\nSkipping line 265028: expected 2 fields, saw
b'Skipping line 525174: expected 2 fields, saw 5\nSkipping line 526251: expected 2 fields, saw
```

```
Out [5]:
```

	password	strength
0	kzde5577	1
1	kino3434	1
2	visi7k1yr	1
3	megzy123	1
4	lamborghini1	1

```
In [6]: data['strength'].unique()
```

```
Out [6]: array([1, 2, 0], dtype=int64)
```

## Checking missing values within the dataset

```
In [7]: data.isna().sum()
```

```
Out [7]: password      1
strength      0
dtype: int64
```

```
In [8]: data[data['password'].isnull()]
```

```
Out [8]:
```

	password	strength
367579	NaN	0

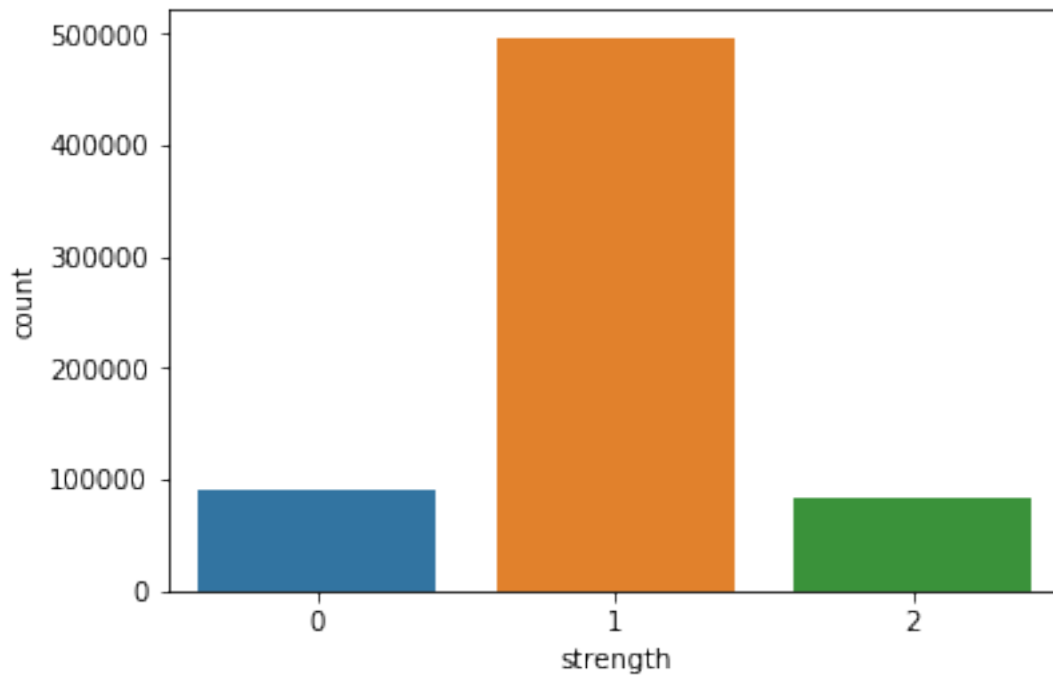
```
In [9]: data.dropna(inplace=True)
```

```
In [10]: data.isnull().sum()
```

```
Out[10]: password    0  
         strength    0  
         dtype: int64
```

```
In [11]: sns.countplot(data['strength'])
```

```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x19bee473f28>
```



```
In [12]: password_tuple=np.array(data)
```

```
In [13]: password_tuple
```

```
Out[13]: array([[ 'kzde5577', 1],  
                [ 'kino3434', 1],  
                [ 'visi7k1yr', 1],  
                ...,  
                [ '184520socram', 1],  
                [ 'marken22a', 1],  
                [ 'fxx4pw4g', 1]], dtype=object)
```

### random shuffling for robustness

```
In [14]: import random
         random.shuffle(password_tuple)

In [15]: x=[labels[0] for labels in password_tuple]
         y=[labels[1] for labels in password_tuple]

In [16]: x
```

```
Out[16]: ['kzde5577',
          'kino3434',
          'kzde5577',
          'visi7k1yr',
          'kino3434',
          'kino3434',
          'AVYq1lDE4MgAZfNt',
          'kino3434',
          'visi7k1yr',
          'universe2908',
          'v1118714',
          'visi7k1yr',
          'AVYq1lDE4MgAZfNt',
          '612035180tok',
          'WUt9IZzE00Q7PkNE',
          'jytifok873',
          'visi7k1yr',
          'lamborghini1',
          'u6c8vhow',
          'jytifok873',
          '612035180tok',
          'jytifok873',
          'kzde5577',
          'WUt9IZzE00Q7PkNE',
          'elyass15@ajilent-ci',
          'jerusalem393',
          '0169395484a',
          'faranumar91',
          'asv5o9yu',
          'intel1',
          'AVYq1lDE4MgAZfNt',
          'visi7k1yr',
          'schalke04',
          '0169395484a',
          'v1118714',
          'AVYq1lDE4MgAZfNt',
          'czuodhj972',
          'kino3434',
          'trabajonet9',
```

'lsdlsd1',  
'jytifok873',  
'WUt9IZzE00Q7PkNE',  
'memjan123',  
'52558000aaa',  
'trabajonet9',  
'olmaz.',  
'go7kew7a2po',  
'g067057895',  
'WUt9IZzE00Q7PkNE',  
'as326159',  
'jerusalem393',  
'lsdlsd1',  
'v1118714',  
'asv5o9yu',  
'ok>bdk',  
'69750381p',  
'TyWM72UNEex8Q8Y',  
'yitbos77',  
'AVYq1lDE4MgAZfNt',  
'g067057895',  
'klara-tershina3H',  
'gaymaids1',  
'yk530mg8',  
'ejeko677',  
'alimagik1',  
'g067057895',  
'fk9qi21m',  
'AVYq1lDE4MgAZfNt',  
'alimagik1',  
'rntprns7',  
'gdfn76',  
'hpqkoxsn5',  
'c3h8bkzr',  
'a2531106',  
'kunyukbabi69',  
'g067057895',  
'calcifer32',  
'trabajonet9',  
'go7kew7a2po',  
'czuodhj972',  
'hodygid757',  
'gill02',  
'k9b8cz6aj2',  
'openup12',  
'juliel009',  
'hodygid757',  
'pHyqueDIyNQ8vmhb',

'ejeko677',  
'olmaz.',  
'bgrvl80',  
'ok>bdk',  
'fk9qi21m',  
'tamanagung6',  
'snolyuj04',  
'gdfn76',  
'lamborghini1',  
'p2share',  
'k9b8cz6aj2',  
'woon12',  
'prisonbreak1',  
'sbl571017',  
'schalke04',  
'cigicigi123',  
'xiau5ff',  
'612035180tok',  
'c3h8bkzr',  
'kzde5577',  
'WUt9IZzE00Q7PkNE',  
'yllime123',  
'universe2908',  
'openup12',  
'ejeko677',  
'ejeko677',  
'yu4cmn',  
'fahad123',  
'intel1',  
'yqugu927',  
'kunyukbabi69',  
'kzde5577',  
'cesarmaio1',  
'memjan123',  
'farrukhcse12',  
'elonex24',  
'icap12',  
'gkrqjs6',  
'olmaz.',  
'hayhayq2',  
'farrukhcse12',  
'fk9qi21m',  
'idofo673',  
'WUt9IZzE00Q7PkNE',  
'asgaliu11',  
'b4NbTxDEyNgG141J',  
'trabajonet9',  
'khmer100.03278&?><Mnb',

'farrukhcse12',  
'yk530mg8',  
'ns2b0727',  
'ejeko677',  
'kunyukbabi69',  
'openup12',  
'b4NbTxDEyNgG141J',  
'woon12',  
'klara-tershina3H',  
'kVczcljg40A25Aeb',  
'oekojWyH120063',  
'megzy123',  
'jUV4dSDQwNwPpA36',  
'farrukhcse12',  
'd04m11',  
'mmm23mm',  
'WUt9IZzE00Q7PkNE',  
'g067057895',  
'intel1',  
'il0vey0u',  
'zoobike04',  
'klara-tershina3H',  
'patri1973',  
'gill02',  
'p@sslng2diword',  
'barra23',  
'929865yt',  
'damyvo114',  
'3vszncp4',  
'mohantra1',  
'ok>bdk',  
'schalke04',  
'aquhih220',  
'jonothe poop1',  
'j09000',  
'fk9qi21m',  
'sarahi1628',  
'damyvo114',  
'kswa2mrv',  
'mazdarx7',  
'ejeko677',  
'pHyqueDIyNQ8vmhb',  
'damyvo114',  
'v1118714',  
'5gzj5uf',  
'rntprns7',  
'p@sslng2diword',  
'4osxw4r',

'universe2908',  
'puegwajy416',  
'faranumar91',  
'demon10',  
'ns2b0727',  
'omakiva153',  
'123net123',  
'mickael12',  
'yqugu927',  
'sasuke4',  
'123net123',  
'puegwajy416',  
'52558000aaa',  
'zoobike04',  
'afs34214',  
'jalingo1',  
'ikanez886',  
'marita1',  
'elyass15@ajilent-ci',  
'kayal123',  
'p2share',  
'jerusalem393',  
'xiau5ff',  
'qn5xpg3k00',  
'147963asd',  
'yitbos77',  
'prisonbreak1',  
'jerusalem393',  
'sarahi1628',  
'Iamthelegend1!',  
'vehat387',  
'potatobus150',  
'4fqa52vecr',  
'ga98SIzk0NwhiZaE',  
'jerusalem393',  
'exitos2009',  
'exitos2009',  
'juliana19',  
'metopelo1623',  
'elyass15@ajilent-ci',  
'kswa2mrv',  
'6tequila6',  
'kzde5577',  
'gill02',  
'bellsuki1',  
'peluchin4',  
'virush1n1',  
'bellsuki1',

'patri1973',  
'kayal123',  
'hodygid757',  
'283671gus',  
'lamborghini1',  
'owote852',  
'snolyuj04',  
's9830950044',  
'3y6iwef2g6',  
'sanki1',  
'olmaz.',  
'xiau5ff',  
'owote852',  
'10Erjrlmebup0n',  
'cockw0mble',  
'2akira2',  
'faranumar91',  
'wibi182d',  
'alchimie79',  
'ok>bdk',  
'ekufite742',  
'holamundo1',  
'planes123',  
'aquhih220',  
'k1k2k3k4k5k6',  
'juliana19',  
'faranumar91',  
'www32223222',  
'26522876p',  
'kikeq102',  
'hayhayq2',  
'pato221182',  
'alchimie79',  
'ajyrew547',  
'tia150979',  
'p2share',  
'lzhzad1989',  
'czuodhj972',  
'exitos2009',  
'gandhi8513',  
'5gzj5uf',  
'gill102',  
'ga98SIzk0NwhiZaE',  
'ycqtdgso3',  
'lrhxmevb620',  
's0xwym7h',  
'il0vey0u',  
'mustang337',



'mustang337',  
'ejeko677',  
'gandhi8513',  
'asgalieu11',  
'sasuke4',  
'123maxbala',  
'djngeyut2707',  
'TyWM72UNEex8Q8Y',  
'a0972986650',  
'mickael12',  
'52558000aaa',  
'witek1709',  
's9830950044',  
'as326159',  
'TyWM72UNEex8Q8Y',  
'intel1',  
'calcifer32',  
'owote852',  
'barra23',  
'pikey231',  
'bgrvl80',  
'123net123',  
'il0vey0u',  
'jytifok873',  
'u6c8vhow',  
'tia150979',  
'idofo673',  
'ebacuro434',  
'memjan123',  
'mathilde54550',  
'oekojWyH120063',  
'929865yt',  
'atigi839',  
'yut0838828185',  
'mickael12',  
'1597535youssi',  
'DTUQG5jU5MwmR1L9',  
'4fqa52vecr',  
'woon12',  
'vgnfs495vp',  
'denise18',  
'lamborghini1',  
'xp;ysmybst',  
'caramelo9',  
'alchimie79',  
'schalke04',  
'v1118714',  
'caramelo9',

'holamundo1',  
'kuntz80',  
'wo9aiwangyan',  
'123net123',  
'memjan123',  
'wibi182d',  
'QWERTY0011',  
'gtlek',  
'gill02',  
'yitbos77',  
'q0pv0fk',  
'vgnfs495vp',  
'RqsuUsDYxNgr8T40',  
'y0unus',  
'ram@!sita15392',  
'hayhayq2',  
'patty94',  
'wo9aiwangyan',  
'juany57',  
'nello11',  
'0169395484a',  
'alimagik1',  
'gaymaids1',  
'6tequila6',  
'tahseen75',  
'pazzini24',  
'mazdarx7',  
'UF1Z2WjE5Mg26R1K',  
'z3ro1sm',  
'taurofive16',  
'OVKWoODkw0Ac0pZK',  
'sebo82',  
'aquhih220',  
'marita1',  
'zidadoh258',  
'koabcswzt3',  
'tahseen75',  
'sasuke4',  
'rntprns7',  
'4fqa52vecr',  
'portales1',  
'prisonbreak1',  
'sebo82',  
'ppnyadam09',  
'vuqADUSatAJ0800',  
'ubojig109',  
'yk530mg8',  
'bugatti01',

'tomas7896',  
'poseidon2011',  
'lsdlsd1',  
'6yy6yy',  
'tahseen75',  
'0169395484a',  
'GGmm26120904..',  
'change201',  
'a2486315',  
'obstacle25',  
'portales1',  
'sasuke4',  
'czuodhj972',  
'examy624',  
'GGmm26120904..',  
'j09000',  
'megzy123',  
'2fakjv',  
'wibi182d',  
'aosmaxd0',  
'visi7k1yr',  
'sknq7m0',  
'x0004534',  
'kjkjkj1',  
'hpqkoxsn5',  
'jonothe poop1',  
'kikeq102',  
'denise18',  
'sasuke4',  
'trabajonet9',  
'sydney213',  
'universe2908',  
'ok>bdk',  
'yut0838828185',  
'mohantra1',  
'deryxi704',  
'weicat12',  
'kayal123',  
'znbl5tj1',  
'lrhxm evb620',  
'idof0673',  
'v10rica',  
'a0972986650',  
'seng987321',  
'3CgRg8DA1NQY1iEj',  
'64whbrb351',  
'2021848709.',  
'hasan18',

'asdasdf1',  
'sanki1',  
'a0972986650',  
'krumbul123',  
'mayur@8netinfotech',  
'alimagik1',  
'1ngaymuadong',  
'witek1709',  
'0870330135a',  
'6975038lp',  
'oscar69',  
'junaid5',  
'killer5',  
'd4xQ3LjUwMQFVCYQ',  
'fbjurcd961',  
'bellsuki1',  
'a2531106',  
'123477889a',  
'finisterra1',  
'hosna1368',  
'52558000aaa',  
'meopvywk628',  
'laedbchsx687',  
'xp;ysmybst',  
'mickael12',  
'4165000yakub',  
'folashade1',  
'afs34214',  
'e667794c1d',  
'taiga0088',  
'sebax2013',  
'123nicole',  
'mohantra1',  
'elabadmin1386',  
'parent777',  
'2akira2',  
'RqsuUsDYxNgr8T40',  
'Iamthelegend1!',  
'kyxvuf137',  
'abizar08',  
'1234159hero',  
'spl51190595',  
'ejeko677',  
'hlQ8gDTExMQWkeda',  
'746xitEGiqObog',  
'kd19cl53',  
'd04m11',  
'cockw0mble',

'w9209640',  
'p2share',  
'kinga22',  
'www32223222',  
'kyodai666',  
'elonex24',  
'wasanun13',  
'10Erjrlmebup0n',  
'sarahi1628',  
'fahad123',  
'g3rappa',  
'oscar69',  
'taulant123',  
'gkrqjs6',  
'just1n0k',  
'weicat12',  
'warriors08',  
'lamborghini1',  
'AVYq1lDE4MgAZfNt',  
'elabadmin1386',  
'edcmki90',  
'portales1',  
'clave08',  
'7942vikas',  
'1597535youssi',  
'ok>bdk',  
'studenko123',  
'ahibyg892',  
'jsm159167',  
'xlxlxl777',  
'ginger972',  
'4fqa52vecr',  
'snolyuj04',  
'peluchin4',  
'6tequila6',  
'holamundo1',  
'ginger972',  
'847XagYxUHUXOW',  
'sw10d014',  
'bgrvl80',  
'zcsntdmhe098',  
'zedika521',  
'sebax2013',  
't8IkFRDIxMAFV2JW',  
'woogee04',  
'asdasdf1',  
'tahseen75',  
'sanki1',

'woaini0',  
'roxana1993',  
'we34dar88',  
'novelia21',  
'pHyqueDIyNQ8vmhb',  
'sandra0547',  
'woogee04',  
'weicat12',  
'barboza221294',  
'elyass15@ajilent-ci',  
'u03kz6ez',  
'bugatti01',  
'mickael12',  
'cigicigi123',  
'zoobike04',  
'warriors08',  
'z3ro1sm',  
'yami12',  
'abizar08',  
'matha81',  
'legna13',  
'damyvo114',  
'v1118714',  
'viri13',  
'mike09',  
'kunyukbabi69',  
'yzxwvgbdu503',  
'mustang337',  
'mohantra1',  
'QWERTY0011',  
'oioo9og',  
'hqh2eYjQxQPYIsA',  
'64959rodre',  
'yitbos77',  
'groster152',  
'pHyqueDIyNQ8vmhb',  
'123456ts',  
'3CgRg8DA1NQY1iEj',  
'polo2014',  
'tin030201',  
'pilatyj280',  
'robot425',  
'den019520',  
'carla99',  
'1234159hero',  
'ok>bdk',  
'jytifok873',  
't8IkFRDIxMAFV2JW',

'zgmfnwuq25',  
'kitty555',  
'sysoja794',  
'clave2013',  
'pedronha96',  
'itengf12',  
'tspirits08',  
'ninja420',  
'147963asd',  
'faranumar91',  
'hlQ8gDTExMQWkeda',  
'damyvo114',  
'wuzyci421',  
'ninja420',  
'alchimie79',  
'khurram\_',  
'lollies1989',  
'3vszncp4',  
'nK0yKXTUONQHZE2e',  
'Staterkom20081993',  
'746xitEGiq0bog',  
'ryjypes139',  
'buqodym199',  
'igejasy712',  
'xiau5ff',  
'pekai2004',  
'zu20081965',  
'clumsy0619',  
'jUV4dSDQwNwPpA36',  
'finisterra1',  
'sofietou74',  
'sbnivetha123',  
'sandra0547',  
'upomel180',  
'sebax2013',  
'limichan99',  
'yllime123',  
'a0972986650',  
'kaya1123',  
'woogee04',  
'jEzZXUTE3MgJ4fVk',  
'taiga0088',  
'jUV4dSDQwNwPpA36',  
'nello11',  
'hola45',  
'viri13',  
'yjuqseb416',  
'p2share',

'djinneyut2707',  
'akucinta12',  
'ingaymuadong',  
'tukaxo486',  
'ginger972',  
'TyWM72UNEex8Q8Y',  
'visi7k1yr',  
'rLLh4WDQ20AWbD05',  
'sanki1',  
'balamuc123',  
'wycinu436',  
'gvczfel801',  
'd4xQ3LjUwMQFVCYQ',  
'sysoja794',  
'roxana1993',  
'ydd45ee',  
'kP82iqDMxNgBMxBP',  
'popo0404',  
'jonothe poop1',  
'acgyj188',  
'hqh2eYjQx0QPYIsA',  
'jcav26',  
'b98nwt priyesh',  
'ydd45ee',  
'teste10',  
'ypodahe201',  
'4lgYVfzk1MwuzHcn',  
'2021848709.',  
'cyborged69',  
'juliana19',  
'artom111478',  
'L2i2ZwTg1MQajeBm',  
'robot425',  
'hisnipes1',  
'xtswdypgh936',  
'caramelo9',  
'ypodahe201',  
'whosdabest1',  
'megzy123',  
'847XagYxUHUXOW',  
'gopal8420',  
'regodib479',  
'wasanun13',  
'yami12',  
'a2531106',  
'djinneyut2707',  
'gopal8420',  
'ykfums1',



'jcav26',  
'llahetihw1',  
'kjkjkj1',  
'uziwocy148',  
'balamuc123',  
'vardhan19',  
'xyws951753',  
'pablo321159',  
'jalingo1',  
'qn5xpg3k00',  
'g067057895',  
'9950twofour0',  
'600eretz',  
'lqksuym982',  
'bc5e4vca',  
'nhfdff2512',  
'qn5xpg3k00',  
'52558000aaa',  
'shotiko18',  
'sanki1',  
'patri1973',  
'0112358mayin0',  
'1ngaymuadong',  
'19821010a',  
'ppnyadam09',  
'sofietou74',  
'k9b8cz6aj2',  
'sknq7m0',  
'uou2dae',  
'clyioqzgw42',  
'faisal213',  
'isqizkg1',  
'zendegi2me',  
'nhiannei040',  
'control9',  
'franczuk33',  
'nLIGyhTU1NQTAp6u',  
'DTUQG5jU5MwmR1L9',  
'ixehawojEPe418',  
'teemteem97',  
'XqMB7vDMz0QocAFV',  
'control9',  
'0870330135a',  
'sbnivetha123',  
'lofebop480',  
'utuham322',  
'klara-tershina3H',  
'go7kew7a2po',

'1972vishara',  
'numero2',  
'kah4544875',  
'folashade1',  
'lqksuym982',  
'raykuaz32',  
'j09000',  
'p@sslng2diword',  
'taulant123',  
'krumbul123',  
'clumsy0619',  
'taccy12',  
'1597535yousi',  
'natalia12',  
'hola45',  
'mzhmir786',  
'abizar08',  
'jesmond26',  
'webstudio8',  
'itengf12',  
'cdann123',  
'fnmsdha476',  
'Iamthelegend1!',  
'ns2b0727',  
'speedracer10',  
'upomel180',  
'hasan18',  
'azerty32',  
'Lyp0JUfuLYr0477',  
'kitty555',  
'jonohepoop1',  
'kswa2mrv',  
'kitty555',  
'hosna1368',  
'taccy12',  
'grazi0201',  
'123456rajput',  
'matiz4533',  
'o7ShLdTMONAQRI7E',  
'eVl19ADIXNamU09N',  
'robot425',  
'2010server',  
'fr80ClDY1MAdEhHa',  
'arigato3',  
'pekai2004',  
'ydkmujrze3',  
'as8594505',  
'jsm159167',

'xiau5ff',  
'alimagik1',  
'wjngzro27',  
'byeypb2',  
'pablo321159',  
'yu86640132',  
'pizxmwaos537',  
'khmer100.03278&?><Mnb',  
'sbnivetha123',  
'n501iomf',  
'uxyloga692',  
'peluchin4',  
'speedracer10',  
'7mVOpKTA3MgHy8Jv',  
'utuham322',  
'bafiqkxwu0',  
'mialr325',  
'ezekiel720',  
'4165000yakub',  
'berserk18',  
'yut0838828185',  
'prisonbreak1',  
'alodise603',  
'wbtdrieus345',  
'1jancok',  
'woogee04',  
'shH3t7Tcy0QwKRLt',  
'matiz4533',  
'krumbul123',  
'znbl5tj1',  
'xve33ea',  
'b9m7cxcgc',  
'mayur@8netinfotech',  
'vehat387',  
'byeypb2',  
'wbtdrieus345',  
'djda1203zj',  
'jorge1489',  
'kenneth610',  
'X34y2CzY5MACs6kp',  
'sqsn7a9',  
'gvczfel801',  
'1234159hero',  
'hola45',  
'1k9izx',  
'hisnipes1',  
'zjl0kx03',  
'0169395484a',

'iL1BEmTUyMg8YYbn',  
'yv3hpf',  
'virush1n1',  
'sanjaime1',  
'pekai2004',  
'nikolas369',  
'4lgYVfzk1MwuzHcn',  
'sbaUsoTA10AzuevI',  
'poseidon2011',  
'bjolgvhs69',  
'QWERTY0011',  
'mdaffandi74',  
'iprwxmfk84',  
'QWERTY0011',  
'OVKWoODkw0Ac0pZK',  
'69556236gu',  
'xp;ysmybst',  
'weicat12',  
'snolyuj04',  
'clyioqzgw42',  
'berserk18',  
'atigi839',  
'bozoxik602',  
'1qa2ws3ed4rf',  
'mel008',  
'uuuu123',  
'1972vishara',  
'taulant123',  
'numero2',  
'yuri110995',  
'jsm159167',  
'seeyouagain1',  
'alimagik1',  
'mark11',  
'16731673ir',  
'mayur@8netinfotech',  
'alhama11408',  
'929865yt',  
'speedracer10',  
'a03242241431a',  
'teemteem97',  
'poluxyj32',  
'vardhan19',  
'pablo321159',  
'terrassa6',  
'030005qw',  
'raykuaz32',  
'4165000yakub',

'virush1n1',  
'yuri110995',  
'ukyzopi369',  
'52756652a',  
'njmania114',  
'tamanagung6',  
'nello11',  
'walterivl13',  
'aslpls2009',  
'147963asd',  
'IjUcOtYqAwel725',  
'147963asd',  
'jekkmoeader>',  
'g3rappa',  
'ydd45ee',  
'10Erjrlmebup0n',  
'cerner09',  
'BZVQZBTM1MApRV7s',  
'7l1hu1xa',  
'090mca090',  
'a110804032',  
'kswa2mrv',  
'1w2e3s4l5e6y',  
'elonex24',  
'gdfn76',  
'1w2e3s4l5e6y',  
'jj46azbo',  
'wearehis7',  
'taiga0088',  
'igejasy712',  
'd6VyrkFV6oblXNs5N8cW',  
'seng987321',  
'go7kew7a2po',  
'fnmsdha476',  
'mega0109',  
'kong0074',  
'whosdabest1',  
'x0004534',  
'komorek11',  
'ym2130104',  
'exusop649',  
'macias2010',  
'kciseba24521',  
'0112358mayin0',  
'matiofox08',  
'cztPECjY2MQGFxnN',  
'2021848709.',  
'hqh2eYjQx0QPYIsA',

'webhost0hm000',  
'sbaUsoTA10AzuevI',  
'123maxbala',  
'tuto0378',  
'cyborged69',  
'wuzsadav933',  
'763EBFC5',  
'kashyn777',  
'847XagYxUHUXOW',  
'afs34214',  
'aliki123',  
'0lELoCDI1NAy8u7D',  
'BsKbJHTY4NgesCOs',  
'mehmet4524572',  
'2010server',  
'33kanun03',  
'yqugu927',  
'umetic21',  
'seller1',  
'bjolgvhs69',  
'l4torr3',  
'mdaffandi74',  
'hpqkoxsn5',  
'hodaq103',  
'sanjaime1',  
'brokensong54',  
'RRaa--72783530081984',  
'q4Ger40X7S2GZ3',  
'kong0074',  
'kXzW0ozU2MQ1Jv1h',  
'purpledog1992',  
'oekojWyH120063',  
'kate13',  
'okn9zp9o',  
'examy624',  
'64959rodro',  
'123linux',  
'iprwxfmk84',  
'stalucia66',  
'mpompo1',  
'AVYq1lDE4MgAZfNt',  
'7942vikas',  
'cztPECjY2MQGFxnN',  
'pilatyj280',  
'sknq7m0',  
'frhnsvelhfr1',  
'asdasdf1',  
'lolmdr1',

'williams',  
'4165000yakub',  
'trust123',  
'wycinu436',  
'xknagqo92',  
'zjl0kx03',  
'jonohepoop1',  
'hodaq103',  
'0123one47',  
'uqilyni846',  
'savas123x',  
'7l1hu1xa',  
'owary200',  
'puegwajy416',  
'ppnyadam09',  
'1972vishara',  
'samael666',  
'portales1',  
'nLIGyhTU1NQTAp6u',  
'tiga33',  
'cifinew817',  
'4osxw4r',  
'lsdlsd1',  
'daylit9',  
'pauramos06',  
'cc71797805',  
'BB11227122',  
'vardhan19',  
'paul2007',  
'kikeq102',  
'clave2013',  
'lizbeth12',  
'DTUQG5jU5MwmR1L9',  
'qwekl12',  
'samael666',  
'gpc151192',  
'2akira2',  
'keithcutlip99',  
'yu4cmn',  
'asv5o9yu',  
'xW8-3w7-MFB-CKH',  
'aziz098765',  
'macias2010',  
'wo9aiwangyan',  
'241189dumai',  
'z3ro1sm',  
'a1233210',  
'mustang337',

```
'exusop649',  
...]
```

**creating custom function to split input into chars of list**

```
In [17]: def word_divide_char(inputs):  
         character=[]  
         for i in inputs:  
             character.append(i)  
         return character
```

```
In [18]: word_divide_char('kzde5577')
```

```
Out[18]: ['k', 'z', 'd', 'e', '5', '5', '7', '7']
```

**import TF-IDF vectorizer to convert String data into numerical data**

```
In [19]: from sklearn.feature_extraction.text import TfidfVectorizer
```

```
In [20]: vectorizer=TfidfVectorizer(tokenizer=word_divide_char)
```

**apply TF-IDF vectorizer on data**

```
In [21]: X=vectorizer.fit_transform(x)
```

```
In [22]: X.shape
```

```
Out[22]: (669639, 128)
```

```
In [23]: vectorizer.get_feature_names()
```

```
Out[23]: ['\x05',  
          '\x06',  
          '\x08',  
          '\x10',  
          '\x11',  
          '\x12',  
          '\x16',  
          '\x17',  
          '\x19',  
          '\x1b',  
          '\x1c',  
          '\x1e',  
          ' ',  
          '!',  
          '"',  
          '#',  
          '$',  
          '%',
```



'&',  
'(',  
)',  
'\*',  
'+',  
'-',  
'.',  
'/',  
'0',  
'1',  
'2',  
'3',  
'4',  
'5',  
'6',  
'7',  
'8',  
'9',  
';',  
'<',  
'=',  
'>',  
'?',  
'@',  
'[',  
'\\',  
']',  
'^',  
'\_',  
'`',  
'a',  
'b',  
'c',  
'd',  
'e',  
'f',  
'g',  
'h',  
'i',  
'j',  
'k',  
'l',  
'm',  
'n',  
'o',  
'p',  
'q',  
'r',

's',  
't',  
'u',  
'v',  
'w',  
'x',  
'y',  
'z',  
'{',  
'|',  
'}',  
'~',  
'\x7f',  
'a',  
'g',  
'ı',  
'n',  
'r',  
's',  
'ş',  
's',  
't',  
't',  
'u',  
'z',  
'ı',  
'e',  
'E',  
'',  
'à',  
'á',  
'â',  
'ä',  
'å',  
'æ',  
'ç',  
'é',  
'ê',  
'í',  
'î',  
'ï',  
'ð',  
'ñ',  
'ò',  
'ó',  
'ô',  
'õ',  
'ö',

'æ',  
'û',  
'ú',  
'û',  
'ü',  
'ý',  
'þ',  
'ß',  
'',  
'',  
'',  
'',  
'',  
'']

```
In [24]: first_document_vector=X[0]
         first_document_vector
```

```
Out[24]: <1x128 sparse matrix of type '<class 'numpy.float64'>'
         with 6 stored elements in Compressed Sparse Row format>
```

```
In [25]: first_document_vector.T.todense()
```

[illegible]

[illegible]

[illegible]