پروژه بازیابی اطلاعات

زهراحیدری

داده هایی ک باید در ویکی پدیا سرچ کند را من به این صورت نوشتم.

articles=[ 'European Union','Germany','Iran','Italian food','France', 'Spanish song', 'Pop songs', 'Violin','Piano',

'Guitar',' Setar', 'Iranian poets','Seahorse', 'Giraffe', 'Rhino', 'Psycho', 'AIDS', 'Schizophrenia',

'The hot weather of Paris', 'Persian the literature', 'Learn multiple languages', 'Lifestyle business',

'Concentration',"African Animals",'World famous songs','wake up early'

,'Oil of countries','Rudolf Diesel'

]

خروجی کد اول به صورت زیر است.

loading content: European Union

loading content: Germany

loading content: Iran

loading content: Italian food

loading content: France

loading content: Spanish song

loading content: Pop songs

loading content: Violin

loading content: Piano

loading content: Guitar

loading content: Setar

loading content: Iranian poets

loading content: Seahorse

loading content: Giraffe

loading content: Rhino

loading content: Psycho

loading content: AIDS

loading content: Schizophrenia

loading content: The hot weather of Paris

loading content: Persian the literature

loading content: Learn multiple languages

loading content: Lifestyle business

loading content: Concentration

loading content: African Animals

loading content: World famous songs

loading content: wake up early

loading content: Oil of countries

loading content: Rudolf Diesel

examine content

Out[51]:

['The European Union (EU) is a political and economic union of 27 member states that are located primarily in Europe. The union has a total area of 4,233,255.3 km2 (1,634,469.0 sq mi) and an estimated total population of about 447 million. An internal single market has been established through a standardised system of laws that apply in all member states in those matters, and only those matters, where the states have agreed to act as one. EU policies aim to ensure the free movement of people, goods, services and capital within the internal market; enact legislation in justice and home affairs; and maintain common policies on trade, agriculture, fisheries and regional development. Passport controls have been abolished for travel within the Schengen Area. The eurozone is a monetary union established in 1999, coming into full force in 2002, that is composed of the 19 EU member states that use the euro currency. The EU has often been described as a sui generis political entity (without precedent or comparison) with the characteristics of either a federation or confederation.The union and EU citizenship were established when the Maastricht Treaty came into force in 1993. The EU traces its origins to the European Coal and Steel Community (ECSC) and the European Economic Community (EEC), established, respectively, by the 1951 Treaty of Paris and 1957 Treaty of Rome. The original member states of what came to be known as the European Communities were the Inner Six: Belgium, France, Italy, Luxembourg, the Netherlands, and West Germany. The communities and their successors have grown in size by the accession of 21 new member states and in power by the addition of policy areas to their remit. The latest major amendment to the constitutional basis of the EU, the Treaty of Lisbon, came into force in 2009. In 2020, the United Kingdom became the only member state to leave the EU. Before this, four territories of current member states had left the EU or its forerunners. There are several other countries that are negotiating to join the European Union.\nContaining some 5.8 percent of the world population in 2020, the EU had generated a nominal gross domestic product (GDP) of around US$17.1 trillion in 2021, constituting approximately 18 percent of global nominal GDP. Additionally, all EU countries have a very high Human Development Index according to the United Nations Development Programme. In 2012, the EU was awarded the Nobel Peace Prize

و بقیه متن که به دلیل طولانی شدن گزارش قرار نمیدم...

from sklearn.feature\_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(stop\_words={'english'})

X = vectorizer.fit\_transform(wiki\_lst)

در اینجا ما استپ ورود های کلمات انگلیسی را در متن ویکی پدیا از ان کم میکنیم و سپس بقیه مراحل را انجام میدهیم.

print(X[0,])

خروجی:

(0, 13406) 0.0015850513989957799

(0, 12968) 0.001953610059476498

(0, 2021) 0.0017380170637792576

(0, 10070) 0.0012874937503897148

(0, 8886) 0.0017380170637792576

(0, 4778) 0.001953610059476498

(0, 1833) 0.0015850513989957799

(0, 15210) 0.001466402011572984

(0, 10011) 0.0012874937503897148

(0, 7423) 0.0031701027979915597

(0, 4214) 0.0017380170637792576

(0, 7312) 0.001466402011572984

(0, 2023) 0.0017380170637792576

(0, 9289) 0.0008479340780343432

(0, 14513) 0.0017380170637792576

(0, 20027) 0.0013694584032985394

(0, 11132) 0.0005697137514233399

(0, 15414) 0.0009189350899089962

(0, 2904) 0.0013694584032985394

(0, 4256) 0.001466402011572984

(0, 15541) 0.0005317177517522115

(0, 3303) 0.001953610059476498

(0, 13667) 0.0013694584032985394

(0, 7135) 0.001953610059476498

(0, 12540) 0.0017380170637792576

: :

(0, 861) 0.003907220118952996

(0, 10625) 0.0012874937503897148

(0, 614) 0.001953610059476498

(0, 595) 0.0031701027979915597

(0, 2036) 0.017806615638721206

(0, 18873) 0.005750037685496933

(0, 9029) 0.04246422867135147

(0, 7126) 0.0361241103675467

(0, 9735) 0.18822808412028288

(0, 14755) 0.004239670390171716

(0, 11212) 0.0032935300532767973

(0, 2035) 0.0423789850659774

(0, 18603) 0.036875220771694625

(0, 17697) 0.08095173229880882

(0, 11925) 0.11307786516250336

(0, 624) 0.011009897170996031

(0, 13229) 0.3105231670232915

(0, 6532) 0.01852725746329553

(0, 1758) 0.24671703681302615

(0, 14430) 0.01470296143854394

(0, 10208) 0.06805987222428307

(0, 7093) 0.2986985500904138

(0, 19379) 0.1514176590514192

(0, 7128) 0.23165848487964363

(0, 18606) 0.6907013595261228

کد بعدی به صورت زیر است.

Sum\_of\_squared\_distances = []

K = range(2,10)

for k in K:

km = KMeans(n\_clusters=k, max\_iter=200, n\_init=10)

km = km.fit(X)

Sum\_of\_squared\_distances.append(km.inertia\_)

plt.plot(K, Sum\_of\_squared\_distances, 'bx-')

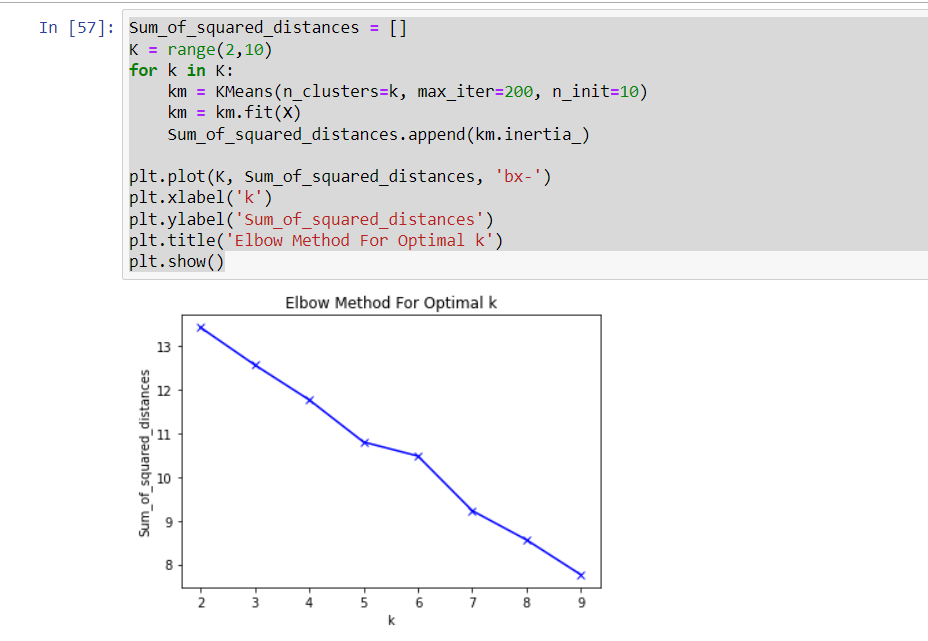
plt.xlabel('k')

plt.ylabel('Sum\_of\_squared\_distances')

plt.title('Elbow Method For Optimal k')

plt.show()

که در اینجا ما با استفاده از روش kmean میخواهیم جدولی بر اساس نزدیک بودن کلمات به همدیگر رو بیان کنیم و ببینیم مجموع فاصله مربعات انها و میزان پراکندگی انها به چه صورت است که خروجی به این صورت نشان داده میشود.



حال با استفاده از قطعه کد بعدی میخواهیم بر اساس موضوع یا تایل انها خوشه بندی کنیم و ببنیم در آن متن چقدر از آن موضوع تکرار شده که به صورت زیر نمایش داده میشود.

(من در اینجا عدد k را 15 قرار دادم که به معنای آن است که 15 خوشه متفاوت برای ما درست کند )

true\_k = 15

model = KMeans(n\_clusters=true\_k, init='k-means++', max\_iter=200, n\_init=10)

model.fit(X)

labels=model.labels\_

wiki\_cl=pd.DataFrame(list(zip(title,labels)),columns=['title','cluster'])

print(wiki\_cl.sort\_values(by=['cluster']))

خروجی:

title cluster

19 Persian the literature 0

11 Iranian poets 0

20 Learn multiple languages 1

15 Psycho 1

14 Rhino 1

18 The hot weather of Paris 1

8 Piano 1

0 European Union 1

5 Spanish song 1

4 France 1

3 Italian food 1

2 Iran 1

1 Germany 1

22 Concentration 2

24 World famous songs 3

21 Lifestyle business 4

12 Seahorse 5

25 wake up early 6

23 African Animals 7

26 Oil of countries 8

27 Rudolf Diesel 9

17 Schizophrenia 10

16 AIDS 11

10 Setar 12

9 Guitar 12

7 Violin 12

6 Pop songs 13

13 Giraffe 14

حال به صورت WordCloud میخواهیم بنویسیم و خروجی بگیریم که خوشه را به ما به صورت ابری با کلمات نزدیک بهم نشان میدهد.

from wordcloud import WordCloud

result={'cluster':labels,'wiki':wiki\_lst}

result=pd.DataFrame(result)

for k in range(0,true\_k):

s=result[result.cluster==k]

text=s['wiki'].str.cat(sep=' ')

text=text.lower()

text=' '.join([word for word in text.split()])

wordcloud = WordCloud(max\_font\_size=50, max\_words=100, background\_color="white").generate(text)

print('Cluster: {}'.format(k))

print('Titles')

titles=wiki\_cl[wiki\_cl.cluster==k]['title']

print(titles.to\_string(index=False))

plt.figure()

plt.imshow(wordcloud, interpolation="bilinear")

plt.axis("off")

plt.show()

و خروجی آن زمانی ک k=15 هست به صورت زیر است.

Cluster: 0

Titles

Iranian poets

Persian the literature



Cluster: 1

Titles

European Union

Germany

Iran

Italian food

France

Spanish song

Piano

Rhino

Psycho

The hot weather of Paris

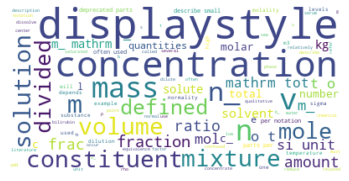
Learn multiple languages



Cluster: 2

Titles

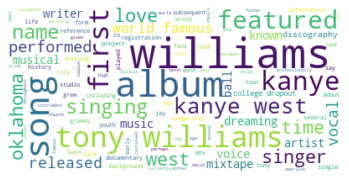
Concentration



Cluster: 3

Titles

World famous songs



Cluster: 4

Titles

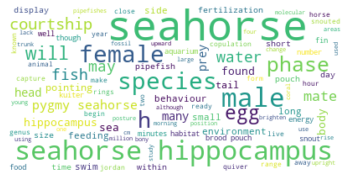
Lifestyle business



Cluster: 5

Titles

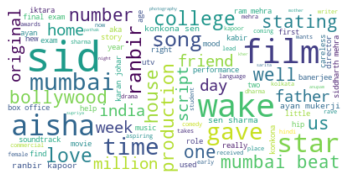
Seahorse



Cluster: 6

Titles

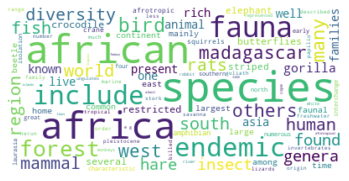
wake up early



Cluster: 7

Titles

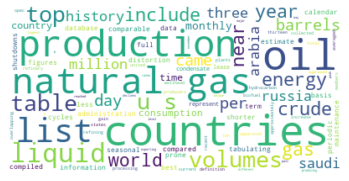
African Animals



Cluster: 8

Titles

Oil of countries



Cluster: 9

Titles

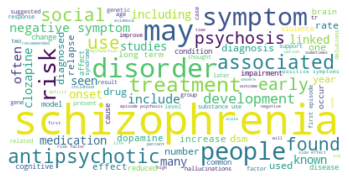
Rudolf Diesel



Cluster: 10

Titles

Schizophrenia



Cluster: 11

Titles

AIDS



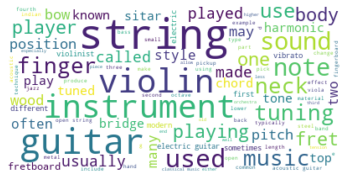
Cluster: 12

Titles

Violin

Guitar

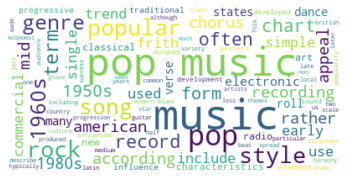
Setar



Cluster: 13

Titles

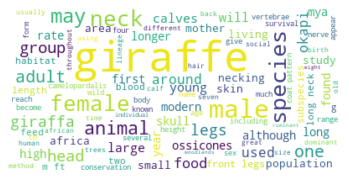
Pop songs



Cluster: 14

Titles

Giraffe



که این بهترین حالتی بود که من برای خوشه بندی

پیدا کردم

اگر k=10 خروجی به صورت زیر است.

Cluster: 0

Titles

Iranian poets

Persian the literature



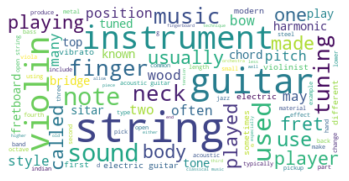
Cluster: 1

Titles

Violin

Guitar

Setar



Cluster: 2

Titles

Lifestyle business



Cluster: 3

Titles

European Union

Germany

Iran

Italian food

France

Spanish song

Piano

Seahorse

Giraffe

Rhino

Psycho

AIDS

Schizophrenia

The hot weather of Paris

Learn multiple languages

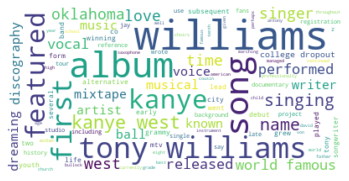
African Animals



Cluster: 4

Titles

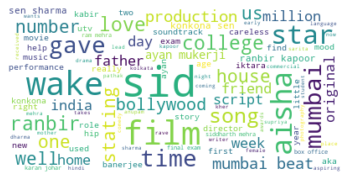
World famous songs



Cluster: 5

Titles

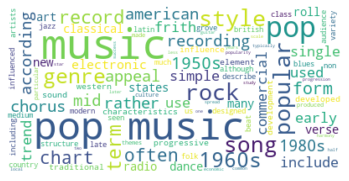
wake up early



Cluster: 6

Titles

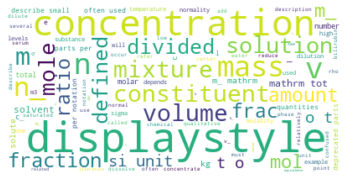
Pop songs



Cluster: 7

Titles

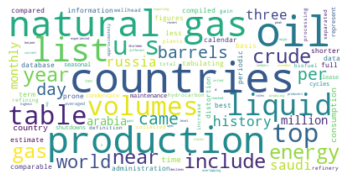
Concentration



Cluster: 8

Titles

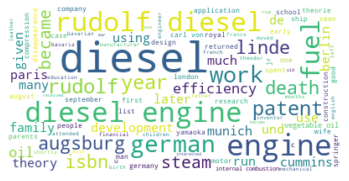
Oil of countries



Cluster: 9

Titles

Rudolf Diesel



و اگر k= 5 باشد خروجی به صورت زیر است.

Cluster: 0

Titles

Oil of countries



Cluster: 1

Titles

Iranian poets

Persian the literature



Cluster: 2

Titles

European Union

Germany

Iran

Italian food

France

Spanish song

Pop songs

Violin

Piano

Guitar

Setar

Seahorse

Giraffe

Rhino

Psycho

AIDS

Schizophrenia

The hot weather of Paris

Learn multiple languages

Lifestyle business

African Animals

wake up early

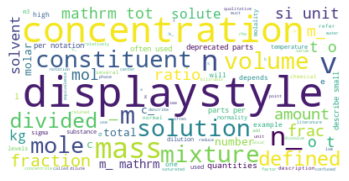
Rudolf Diesel



Cluster: 3

Titles

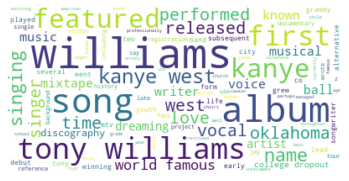
Concentration



Cluster: 4

Titles

World famous songs



و اگر k=3 باشد خروجی به صورت زیر است.

Cluster: 0

Titles

European Union

Germany

Iran

Italian food

France

Spanish song

Pop songs

Violin

Piano

Guitar

Setar

Seahorse

Giraffe

Rhino

Psycho

AIDS

Schizophrenia

The hot weather of Paris

Learn multiple languages

Concentration

African Animals

World famous songs

wake up early

Oil of countries

Rudolf Diesel



Cluster: 1

Titles

Lifestyle business



Cluster: 2

Titles

Iranian poets

Persian the literature

