7. Text Clustering using K-Means

September 22, 2018

Saumay Agrawal
16BCE1151

```
In [1]: from pprint import pprint
```

```
In [2]: # Code for the K-means clustering algorithm

        def vectorize(doc,terms):
            docvector=[]
            count=0
            for i in range(len(terms)):
                docvector.append(0)
                for syn in terms[i]:
                    docvector[i]+=doc.lower().split(" ",500).count(syn.lower())
            return docvector
        def vectorAvg(a):
            center=[0 for i in range(len(a[0]))]
            for vec in a:
                center=list(map(sum,zip(center,vec)))
            n=len(a)
            center[:]=[round(x/n,3) for x in center]
            return center
        def distInitknn(mat,n,k):
            for i in range(n):
                mat.append([0 for j in range(k)])
            return None
        def kMean(docvectors,centroids,k=2,prev=[],n=1):
            print("\niteration: ",n,"\n")
            cluster=[]
            distMat=[]
            print("\ncentroid vectors:\n")
            pprint(centroids)
            print("\ndoc vectors:\n")
            pprint(docvectors)

            veclen=len(docvectors)
            distInitknn(distMat,veclen,k)
```

```
    genDistMat(distMat,docvectors,centroids)
    for i in range(k):
        cluster.append([i])
    if prev==[]:
        for i in range(k,len(docvectors)):
            cluster[i%k].append(i)
    else:
        for i in range(k,veclen):
            cluster[getMinIndex(i,distMat)].append(i)
    newcentroid=[]
    print("cluster formed:\n",cluster)

    for clus in cluster:
        vec=[docvectors[i] for i in clus]
        center=vectorAvg(vec)
        newcentroid.append(list(center))
    print("\nnew centroids:\n")
    pprint(newcentroid)
    print("*"*50)
    if cluster!=prev:
        kMean(docvectors,newcentroid,k,cluster,n+1)
    return None
def calcManDist(a,b):
    dist=0
    for x,y in zip(a,b):
        dist+=abs(x-y)
    return dist
def genDistMat(distMat,docvectors,centroids):
    for i in range(len(docvectors)):
        for j in range(len(centroids)):
            dist=calcManDist(docvectors[i],centroids[j])
            distMat[i][j]=dist
    return None
def getMinIndex(i,mat):
    mindist=float("inf")
    minindex=0
    for j in range(len(mat[i])):
        if mindist>mat[i][j]:
            mindist=mat[i][j]
            minindex=j
    return minindex
```

In [3]:

```
# Code for the clustering of documents in part 1

terms=[['automotive'],['car','cars'],['motorcycles','motorcycle'],['self-drive'],['IoT

doc1='Electric automotive maker Tesla Inc. is likely to introduce its products in Indi
doc2='Automotive major Mahindra likely to introduce driverless cars'
```

```
doc3='BMW plans to introduce its own motorcycles in india'
doc4='Just drive, a self-drive car rental firm uses smart vehicle technology based on I
doc5='Automotive industry going to hire thousands in 2018'
doc6='Famous cricket player  Dhoni brought his priced car Hummer which is an SUV'
doc7='Dhoni led india to its second world cup victory'
doc8='IoT in cars will lead to more safety and make driverless vehicle revolution possi
doc9='Sachin recommended Dhoni for the indian skipper post'

docvectors=[]
doclist=[doc1,doc2,doc3,doc4,doc5,doc6,doc7,doc8,doc9]
for doc in doclist:
    docvectors.append(vectorize(doc,terms))
k=4
kMean(docvectors,docvectors[:k],k)
```

```
iteration:  1


centroid vectors:

[[1, 0, 0, 0, 0, 0, 0],
 [1, 1, 0, 0, 0, 0, 0],
 [0, 0, 1, 0, 0, 0, 0],
 [0, 1, 0, 1, 1, 0, 0]]

doc vectors:

[[1, 0, 0, 0, 0, 0, 0],
 [1, 1, 0, 0, 0, 0, 0],
 [0, 0, 1, 0, 0, 0, 0],
 [0, 1, 0, 1, 1, 0, 0],
 [1, 0, 0, 0, 0, 1, 0],
 [0, 1, 0, 0, 0, 0, 1],
 [0, 0, 0, 0, 0, 0, 1],
 [0, 1, 0, 0, 1, 0, 0],
 [0, 0, 0, 0, 0, 0, 1]]
cluster formed:
 [[0, 4, 8], [1, 5], [2, 6], [3, 7]]

new centroids:

[[0.667, 0.0, 0.0, 0.0, 0.0, 0.333, 0.333],
 [0.5, 1.0, 0.0, 0.0, 0.0, 0.0, 0.5],
 [0.0, 0.0, 0.5, 0.0, 0.0, 0.0, 0.5],
 [0.0, 1.0, 0.0, 0.5, 1.0, 0.0, 0.0]]
**************************************************
```

```
iteration:  2


centroid vectors:

[[0.667, 0.0, 0.0, 0.0, 0.0, 0.333, 0.333],
 [0.5, 1.0, 0.0, 0.0, 0.0, 0.0, 0.5],
 [0.0, 0.0, 0.5, 0.0, 0.0, 0.0, 0.5],
 [0.0, 1.0, 0.0, 0.5, 1.0, 0.0, 0.0]]

doc vectors:

[[1, 0, 0, 0, 0, 0, 0],
 [1, 1, 0, 0, 0, 0, 0],
 [0, 0, 1, 0, 0, 0, 0],
 [0, 1, 0, 1, 1, 0, 0],
 [1, 0, 0, 0, 0, 1, 0],
 [0, 1, 0, 0, 0, 0, 1],
 [0, 0, 0, 0, 0, 0, 1],
 [0, 1, 0, 0, 1, 0, 0],
 [0, 0, 0, 0, 0, 0, 1]]
cluster formed:
 [[0, 4], [1, 5], [2, 6, 8], [3, 7]]

new centroids:

[[1.0, 0.0, 0.0, 0.0, 0.0, 0.5, 0.0],
 [0.5, 1.0, 0.0, 0.0, 0.0, 0.0, 0.5],
 [0.0, 0.0, 0.333, 0.0, 0.0, 0.0, 0.667],
 [0.0, 1.0, 0.0, 0.5, 1.0, 0.0, 0.0]]
**************************************************

iteration:  3


centroid vectors:

[[1.0, 0.0, 0.0, 0.0, 0.0, 0.5, 0.0],
 [0.5, 1.0, 0.0, 0.0, 0.0, 0.0, 0.5],
 [0.0, 0.0, 0.333, 0.0, 0.0, 0.0, 0.667],
 [0.0, 1.0, 0.0, 0.5, 1.0, 0.0, 0.0]]

doc vectors:

[[1, 0, 0, 0, 0, 0, 0],
 [1, 1, 0, 0, 0, 0, 0],
 [0, 0, 1, 0, 0, 0, 0],
 [0, 1, 0, 1, 1, 0, 0],
 [1, 0, 0, 0, 0, 1, 0],
 [0, 1, 0, 0, 0, 0, 1],
 [0, 0, 0, 0, 0, 0, 1],
 [0, 1, 0, 0, 1, 0, 0],
 [0, 0, 0, 0, 0, 0, 1]]
cluster formed:
 [[0, 4], [1, 5], [2, 6, 8], [3, 7]]

new centroids:

[[1.0, 0.0, 0.0, 0.0, 0.0, 0.5, 0.0],
 [0.5, 1.0, 0.0, 0.0, 0.0, 0.0, 0.5],
 [0.0, 0.0, 0.333, 0.0, 0.0, 0.0, 0.667],
 [0.0, 1.0, 0.0, 0.5, 1.0, 0.0, 0.0]]
**************************************************
```

```
In [4]: # Code for text minining from web for part 2

        from bs4 import BeautifulSoup
        import requests
        from string import punctuation

        doclist=['doc1.txt','doc2.txt','doc3.txt','doc4.txt','doc5.txt','doc6.txt','doc7.txt',
        links=['https://www.zigwheels.com/newcars/Tesla',' https://www.financialexpress.com/au
               ' https://en.wikipedia.org/wiki/Toyota_Prius','https://economictimes.indiatimes
               ' https://indianexpress.com/article/india/india-news-india/demonetisation-hits-
               ' https://www.livemint.com/Politics/ySbMKTIC4MINsz1btccBJO/How-demonetisation-a
               ' https://inc42.com/buzz/electric-vehicles-this-week-centre-reduces-gst-on-lith
               'https://www.youthkiawaaz.com/2017/12/impact-of-demonetisation-on-the-indian-ec
               ' https://www.news18.com/news/business/how-gst-will-curb-tax-evasion-1446035.ht

        for i in range(len(doclist)):
            page=requests.get(links[i])
            soup=BeautifulSoup(page.text,'html.parser')
            p_tags=soup.find_all('p')
            text = (''.join(s.findAll(text=True))for s in soup.findAll('p'))
            f=open(doclist[i],'w')
            #f.write("abc")
            gen=[str(y.lower()) for y in text ]
            count=0
            for t in gen:
                f.write(t)
                count=count+1
                if count>=500:
                    break
            f.close()
```

```
In [5]: # Code for text clustering in part 2
```

```
files = ['doc1.txt', 'doc2.txt', 'doc3.txt', 'doc4.txt','doc5.txt', 'doc6.txt', 'doc7.t
terms=[['tesla',"tesla's"], ['electric'], ['car','cars','vehicle','vehicles','automobil
docvectors=[]
for fname in files:
    file=open(fname,'r')
    doclines=file.read().split('.')
    doc=''
    for line in doclines:
        doc+=" "+str(line)
    docvectors.append(vectorize(doc,terms))
kMean(docvectors,docvectors[:k],k)
```

```
iteration:  1


centroid vectors:

[[5, 4, 3, 0, 0, 0, 0],
 [0, 22, 13, 0, 0, 0, 0],
 [0, 1, 7, 0, 0, 0, 0],
 [0, 3, 3, 0, 0, 0, 0]]

doc vectors:

[[5, 4, 3, 0, 0, 0, 0],
 [0, 22, 13, 0, 0, 0, 0],
 [0, 1, 7, 0, 0, 0, 0],
 [0, 3, 3, 0, 0, 0, 0],
 [0, 5, 2, 0, 1, 0, 0],
 [0, 0, 1, 0, 6, 0, 0],
 [0, 0, 4, 0, 0, 14, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 4, 0]]
cluster formed:
 [[0, 4, 8], [1, 5, 9], [2, 6, 10], [3, 7, 11]]

new centroids:

[[1.667, 3.0, 1.667, 0.0, 0.333, 0.0, 0.0],
 [0.0, 7.333, 4.667, 0.0, 2.0, 0.0, 0.0],
 [0.0, 0.333, 3.667, 0.0, 0.0, 4.667, 0.0],
 [0.0, 1.0, 1.0, 0.0, 0.0, 1.333, 0.0]]
**************************************************
```

```
iteration:  2


centroid vectors:

[[1.667, 3.0, 1.667, 0.0, 0.333, 0.0, 0.0],
 [0.0, 7.333, 4.667, 0.0, 2.0, 0.0, 0.0],
 [0.0, 0.333, 3.667, 0.0, 0.0, 4.667, 0.0],
 [0.0, 1.0, 1.0, 0.0, 0.0, 1.333, 0.0]]

doc vectors:

[[5, 4, 3, 0, 0, 0, 0],
 [0, 22, 13, 0, 0, 0, 0],
 [0, 1, 7, 0, 0, 0, 0],
 [0, 3, 3, 0, 0, 0, 0],
 [0, 5, 2, 0, 1, 0, 0],
 [0, 0, 1, 0, 6, 0, 0],
 [0, 0, 4, 0, 0, 14, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 4, 0]]
cluster formed:
 [[0, 4], [1], [2, 6, 11], [3, 5, 7, 8, 9, 10]]

new centroids:

[[2.5, 4.5, 2.5, 0.0, 0.5, 0.0, 0.0],
 [0.0, 22.0, 13.0, 0.0, 0.0, 0.0, 0.0],
 [0.0, 0.333, 3.667, 0.0, 0.0, 6.0, 0.0],
 [0.0, 0.5, 0.667, 0.0, 1.0, 0.0, 0.0]]
**************************************************

iteration:  3


centroid vectors:

[[2.5, 4.5, 2.5, 0.0, 0.5, 0.0, 0.0],
 [0.0, 22.0, 13.0, 0.0, 0.0, 0.0, 0.0],
 [0.0, 0.333, 3.667, 0.0, 0.0, 6.0, 0.0],
 [0.0, 0.5, 0.667, 0.0, 1.0, 0.0, 0.0]]

doc vectors:
```

```
[[5, 4, 3, 0, 0, 0, 0],
 [0, 22, 13, 0, 0, 0, 0],
 [0, 1, 7, 0, 0, 0, 0],
 [0, 3, 3, 0, 0, 0, 0],
 [0, 5, 2, 0, 1, 0, 0],
 [0, 0, 1, 0, 6, 0, 0],
 [0, 0, 4, 0, 0, 14, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 4, 0]]
cluster formed:
 [[0, 4], [1], [2, 6, 11], [3, 5, 7, 8, 9, 10]]

new centroids:

[[2.5, 4.5, 2.5, 0.0, 0.5, 0.0, 0.0],
 [0.0, 22.0, 13.0, 0.0, 0.0, 0.0, 0.0],
 [0.0, 0.333, 3.667, 0.0, 0.0, 6.0, 0.0],
 [0.0, 0.5, 0.667, 0.0, 1.0, 0.0, 0.0]]
**************************************************
```