

In []:


```
"""Question 2"""
```

In [5]:

```
import numpy as np
import pandas as pd
import random
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from sklearn.metrics.pairwise import euclidean_distances
from sklearn.datasets import make_blobs
from sklearn.cluster import KMeans
import seaborn as sns
import os
```

In [19]:

```
# Download all documents
document = np.load("F:/Annie/CornellMS/Semester 4/Machine Learning/Homework/HW3/Clustering/
vocabulary = pd.read_table('F:/Annie/CornellMS/Semester 4/Machine Learning/Homework/HW3/Clu
titles = pd.read_table('F:/Annie/CornellMS/Semester 4/Machine Learning/Homework/HW3/Cluster
document.shape
```



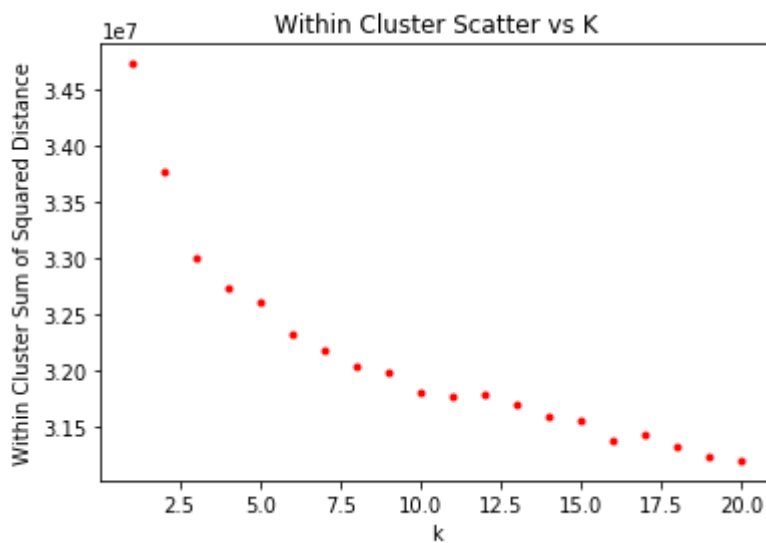
Out[19]:

```
(1373, 5476)
```

In [3]:

```
#a : Cluster the documents using k-means and various values of k
def cluster_k(data,max_k):
    sum_of_sq_dist = []
    for k in range(1,max_k+1):
        kmeans = KMeans(n_clusters=k, random_state=0).fit(data)
        sum_of_sq_dist.append(kmeans.inertia_)
    plt.plot(range(1,max_k+1), sum_of_sq_dist, 'r.')
    plt.xlabel('k')
    plt.ylabel('Within Cluster Sum of Squared Distance')
    plt.title('Within Cluster Scatter vs K')
    plt.show()

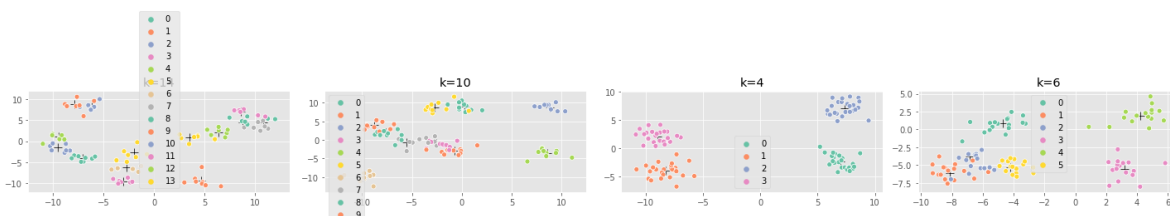
cluster_k(document,20)
```



In [6]:

```
# plot for various values of k
def mean_plots(data):
    plt.style.use('ggplot')
    x=data
    plt.figure(figsize=(20, 3))
    for k in range(4):
        K = random.randint(1,20)
        x, y = make_blobs(centers=K)
        kmeans = KMeans(n_clusters=K, random_state=0).fit(x)
        plt.subplot(1,4,k+1)
        sns.scatterplot(kmeans.cluster_centers[:,0], kmeans.cluster_centers[:,1], marker='x')
        sns.scatterplot(x[:,0], x[:,1], hue=y, palette=sns.color_palette("Set2", n_colors=K))
        plt.title('k={}'.format(K))
    plt.tight_layout()
    plt.show()

mean_plots(document)
```



In [7]:

```
"""We see k=4, gives very reasonable output, with distinct seperation. Hence, the k we sele
```

Out[7]:

```
'We see k=4, gives very reasonable output, with distinct seperation. Hence,
the k we select is 4'
```

In [9]:

```
# calculate how many documents fall in each bins
kmeans_bin = KMeans(n_clusters=4, random_state=0).fit(document)
np.bincount(kmeans_bin.labels_)
```

Out[9]:

```
array([371, 230, 559, 213], dtype=int64)
```

In [16]:

```
# Return the top 10 features
center = np.mean(document, axis=0)
for i in range(4):
    largest = np.argsort(kmeans_bin.cluster_centers_[i] - center)[::-1][:10]
    print( i+1,vocabulary[largest].reshape(1,-1), "\n")
```

```
1 [['protein' 'cell' 'cells' 'expression' 'proteins' 'fig' 'gene'
    'expressed' 'binding' 'specific']]

2 [['fig' 'values' 'period' 'mean' 'lower' 'average' 'estimates' 'range'
    'estimate' 'indicate']]

3 [['says' 'mail' 'researchers' 'scientists' 'years' 'news' 'people'
    'issue' 'year' 'world']]

4 [['energy' 'electron' 'fig' 'density' 'temperature' 'shows' 'structure'
    'measured' 'constant' 'measurements']]
```

In [17]:

```
representing Biology, the forth representing Atomic Chemistry or Physics. The clustering is
```

Out[17]:

```
'We see a distinctive pattern, the first group representing Biology, the forth
representing Atomic Chemistry or Physics. The clustering is bringing out
some unique connection.'
```

In [20]:

```
# Return the top 10 closest documents to each centroid
for i in range(4):
    d = kmeans_bin.transform(document)[: , i]
    ind = np.argsort(d)[:10]
    print(i+1, titles[ind], "\n" )
```

```
1 [['Requirement of NAD and SIR2 for Life-Span Extension by Calorie Restriction in Saccharomyces Cerevisiae']
   ['Suppression of Mutations in Mitochondrial DNA by tRNAs Imported from the Cytoplasm']
   ['Thermal, Catalytic, Regiospecific Functionalization of Alkanes']
   ['Algorithmic Gladiators Vie for Digital Glory']
   ['Reopening the Darkest Chapter in German Science']
   ['Similar Requirements of a Plant Symbiont and a Mammalian Pathogen for Prolonged Intracellular Survival']
   ['Mothers Setting Boundaries']
   ['Turning up the Heat on Histoplasma capsulatum']
   ['Distinct Classes of Yeast Promoters Revealed by Differential TAF Recruitment']
   ['An Arresting Start for MAPK']]
```

```
2 [['Algorithmic Gladiators Vie for Digital Glory']
   ['Reopening the Darkest Chapter in German Science']
   ['Population Dynamical Consequences of Climate Change for a Small Temperate Songbird']
   ['The Formation of Chondrules at High Gas Pressures in the Solar Nebula']
   ['Subducted Seamount Imaged in the Rupture Zone of the 1946 Nankaido Earthquake']
   ['Homogenization of Fish Faunas across the United States']
   ['Tectonic Implications of U-Pb Zircon Ages of the Himalayan Orogenic Belt in Nepal']
   ['Corrections and Clarifications: A Short Fe-Fe Distance in Peroxodiferric Ferritin: Control of Fe Substrate versus Cofactor Decay?']
   ["Corrections and Clarifications: Charon's First Detailed Spectra Hold Many Surprises"]
   ['Corrections and Clarifications: Unearthing Monuments of the Yarmukians']]
```

```
3 [['Algorithmic Gladiators Vie for Digital Glory']
   ['Reopening the Darkest Chapter in German Science']
   ['Information Technology Takes a Different Tack']
   ['National Academy of Sciences Elects New Members']
   ['Archaeology in the Holy Land']
   ['Heretical Idea Faces Its Sternest Test']
   ['Corrections and Clarifications: A Short Fe-Fe Distance in Peroxodiferric Ferritin: Control of Fe Substrate versus Cofactor Decay?']
   ["Corrections and Clarifications: Charon's First Detailed Spectra Hold Many Surprises"]
   ['Corrections and Clarifications: Unearthing Monuments of the Yarmukians']
   ['Divining Diet and Disease from DNA']]
```

```
4 [['The Formation of Chondrules at High Gas Pressures in the Solar Nebula']
   ['Algorithmic Gladiators Vie for Digital Glory']
   ['Thermal, Catalytic, Regiospecific Functionalization of Alkanes']
   ['Reopening the Darkest Chapter in German Science']
   ['Heretical Idea Faces Its Sternest Test']
   ['Information Storage and Retrieval through Quantum Phase']
   ['Synthesis and Characterization of Helical Multi-Shell Gold Nanowires']
   ['A Monoclinic Post-Stishovite Polymorph of Silica in the Shergotty Meteorite']]
```

```
['Quantum Dots as Tunable Kondo Impurities']
['Ambipolar Pentacene Field-Effect Transistors and Inverters']]
```

In [21]:

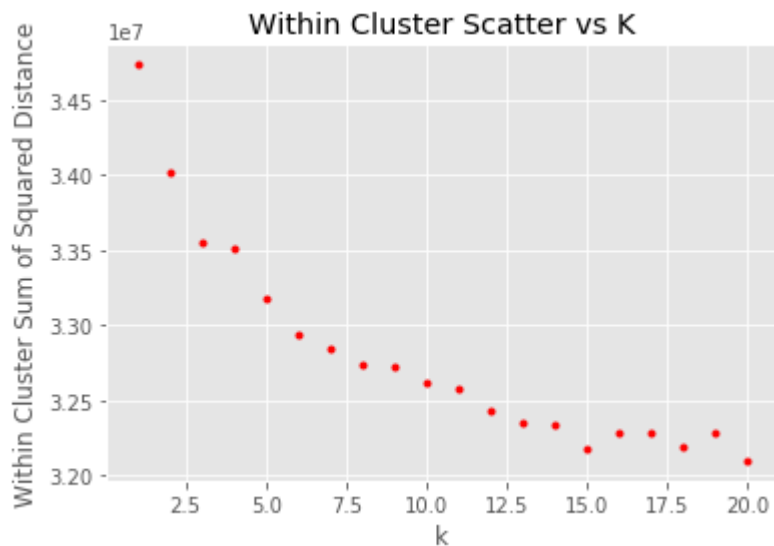
There is a pattern but not very distinct as features. We still see a little Biology clustering in

Out[21]:

'The clustering shown by documents are not very good! It shows multiple repetition for same title. There is a pattern but not very distinct as features. We still see a little Biology clustering in 1 and Physics clustering in 4.'

In [23]:

```
#b : Repeat analysis for term-wise matrix
term = np.load("F:/Annie/CornellMS/Semester 4/Machine Learning/Homework/HW3/Clustering/scie
cluster_k(term, 20)
```



In [26]:

```
# Separation of clusters
mean_plots(term)
```



In [28]:

""Plot with k=3 looks the best for me. k=10 also looks good but intercluster distance is n

Out[28]:

'Plot with k=3 looks the best for me. k=10 also looks good but intercluster distance is not so good. Hence, we go with k=3'

In [34]:

```
# Dividing into bins
kmeans_bin2 = KMeans(n_clusters=3, random_state=0).fit(term_mat)
np.bincount(kmeans_bin2.labels_)
```

Out[34]:

```
array([ 523, 4275,  678], dtype=int64)
```

In [35]:

```
# Top features
center2 = np.mean(term, axis=0)
for i in range(3):
    largest = np.argsort(kmeans_bin2.cluster_centers_[i] - center2)[::-1][:10]
    print( i+1,vocabulary[largest], "\n")
```

```
1 [['adult']
   ['spectrum']
   ['national']
   ['suggested']
   ['mars']
   ['ratios']
   ['provide']
   ['century']
   ['mass']
   ['extended']]
```

```
2 [['mediated']
   ['conserved']
   ['width']
   ['strongly']
   ['past']
   ['environmental']
   ['interactions']
   ['separated']
   ['atom']
   ['contact']]
```

```
3 [['question']
   ['frequency']
   ['people']
   ['high']
   ['coding']
   ['new']
   ['bar']
   ['self']
   ['knowledge']
   ['quantum']]
```

In [36]:

```
, but with only top 10 words it is difficult to predict the groups. Hence, we need more word
```

Out[36]:

```
'The term seems to be clustering, but with only top 10 words it is difficult  
to predict the groups. Hence, we need more words to deduce the exact field o  
f study.'
```

In [32]:

```
# Top titles
center2 = np.mean(term, axis=0)
for i in range(3):
    largest = np.argsort(kmeans_bin2.cluster_centers_[i] - center2)[::-1][:10]
    print( i+1,titles[largest], "\n")
```

```
1 [['Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator of p
53-Induced Apoptosis']
['Positional Syntenic Cloning and Functional Characterization of the Mammal
ian Circadian Mutation tau']
['Central Role for G Protein-Coupled Phosphoinositide 3-Kinase g in Inflamm
ation']
['Kinesin Superfamily Motor Protein KIF17 and mLin-10 in NMDA Receptor-Cont
aining Vesicle Transport']
['Regulated Cleavage of a Contact-Mediated Axon Repellent']
['Role of the Mouse ank Gene in Control of Tissue Calcification and Arthrit
is']
['An Oral Vaccine against NMDAR1 with Efficacy in Experimental Stroke and E
pilepsy']
['Requirement of JNK for Stress-Induced Activation of the Cytochrome c-Medi
ated Death Pathway']
['Function of PI3Kg in Thymocyte Development, T Cell Activation, and Neutro
phil Migration']
['Regulation of STAT3 by Direct Binding to the Rac1 GTPase']]

2 [['National Academy of Sciences Elects New Members']
['NIH, under Pressure, Boosts Minority Health Research']
['Science Survives in Breakthrough States']
['Ground Zero: AIDS Research in Africa']
['Sharp Jump in Teaching Fellows Draws Fire from Educators']
['Africa Boosts AIDS Vaccine R&D']
['A New Breed of Scientist-Advocate Emerges']
['Flushing out Nasty Viruses in the Balkans']
["Stephen Straus's Impossible Job"]
['Bastions of Tradition Adapt to Alternative Medicine']]

3 [['NEAR at Eros: Imaging and Spectral Results']
['The Atom-Cavity Microscope: Single Atoms Bound in Orbit by Single Photon
s']
['Advances in the Physics of High-Temperature Superconductivity']
['The Formation and Early Evolution of the Milky Way Galaxy']
['Subduction and Slab Detachment in the Mediterranean-Carpathian Region']
['The Galactic Center: An Interacting System of Unusual Sources']
["Earth's Core and the Geodynamo"]
['Quantum Criticality: Competing Ground States in Low Dimensions']
["Sediments at the Top of Earth's Core"]
['Internal Structure and Early Thermal Evolution of Mars from Mars Global S
urveyor Topography and Gravity']]
```


In [37]:

s to be Biological Pathway, the second seems to be major break-throughs and third seems a fi

Out[37]:

'The grouping of the title seems better than the previous. We can understand the groups and fields better. For example, the first one seems to be Biological Pathway, the second seems to be major break-throughs and third seems a field with Geography and Physics in it.'

In []: