Loading the packages

In [1]:
```
import os
import pandas as pd

#https://www.sbert.net/docs/pretrained_models.html "to see the new models"

from sentence_transformers import SentenceTransformer
embedder = SentenceTransformer('all-mpnet-base-v2')
```

In [2]:
```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from wordcloud import WordCloud
```

In [3]:
```
# Loading the data
df = pd.read_excel('Topics cluster.xlsx')
df.head()
```

Out[3]:

	#	Nome do artigo	Tópico do artigo	Definição	Description	Description verified	Agrupamento	gru
0	37	A randomized trial of energy cost information ...	Etiquetas de eficiencia energética da União E...	O fornecimento de informações sobre os custos ...	Energy efficiency labels.\n\nProviding informa...	Energy efficiency labels. Providing informatio...	1	
1	29	Analysing of the impacts of the revision of Ec...	Etiquetas de eficiencia energética de maquinas...	O uso de etiquetas escalonadas de maneira efic...	Energy efficiency labels for dishwashers.\n\nT...	Energy efficiency labels for dishwashers: Usin...	1	
2	4	Appliance efficiency standards and labeling pr...	Programa de eficiência e etiquetas energéticas...	A existência do programa de eficiência energét...	Efficiency program and energy labels.\n\nThe e...	Efficiency program and energy labels: The ener...	1	
3	36	Assessing the EU Energy Efficiency Label for A...	Etiqueta de eficiência energética da União Eur...	A apresentação de informações nas etiquetas de...	Energy efficiency label.\n\nThe presentation o...	Energy efficiency label. The presentation of i...	1	
4	26	Consumer Inattention, Heuristic Thinking and t...	Etiqueta de eficiência energética da União Eur...	O uso de informações na etiqueta de eficiência...	Energy efficiency label.\n\nUsing more informa...	Energy efficiency label. Using more informatio...	1	

Clustering

In [4]:
```
corpus = list(df['Description verified'])
```

In [5]:
```
corpus
```

Out[5]:

```
["Energy efficiency labels. Providing information on the energy costs of energy-consuming devi
ces, in addition to the energy class, draws users' attention to the total cost of the products.",
 'Energy efficiency labels for dishwashers: Using scaled labels effectively will enable manufac
turers to develop energy-efficient equipment and consumers to purchase these products.',
 'Efficiency program and energy labels: The energy efficiency program and the development o
f labels provided substantial reductions in residential electricity consumption.',
 'Energy efficiency label. The presentation of information on energy efficiency labels must be
clear and objective to not confuse the consumer, especially regarding technical issues.',
 'Energy efficiency label. Using more information on the energy efficiency label than necessar
y can confuse consumers, causing them to choose less efficient products.',
 'Energy rating label. The label will focus on improving consumer awareness of system energy e
fficiency and facilitating comparison of systems with others.',
 'Label energy rating. The presentation of a well-defined and clear energy classification provid
es the consumer with a better understanding of the energy efficiency of the equipment and, t
hus, the acquisition of the most efficient products.',
 'Energy efficiency label. Adding monetary information increases the likelihood of cost-effect
iveness analysis and promotes the choice of energy-efficient products.',
 'Energy rating label. Using a label that clearly and objectively presents the energy performan
ce of equipment promotes greater consumer acceptance.',
 'Energy rating label and minimum energy performance standard: The improvement in the pres
entation of energy classification labels and using minimum energy performance standards provi
des more efficient equipment with better energy classification.',
 'Energy efficiency label. The presentation of information regarding the cost over the useful l
ife of the certified product leads to the choice of more efficient products when compared to
the standard energy classification label.',
 'Energy Labeling System. Improving the labeling system ranges from improving laws, regulatio
ns, and decrees to increasing social supervision and consumer awareness regarding energy-effi
cient products.',
 'Energy efficiency label. The addition of functional information combined with consumer know
ledge guarantees the acquisition of more efficient products.',
 'Minimum energy performance standard: Using "tighter" energy rating bands can lead to devel
oping more efficient equipment or moving items to lower ratings.',
 'Energy rating label. Improving the presentation of information on energy efficiency labels pr
ovides a better understanding for consumers who purchase more efficient products.',
 'Energy efficiency label. Labels must have clear and objective information not to confuse the
consumer when choosing the product.',
 'Household appliances and choosing energy-efficient appliances:  Households with higher inco
mes consume energy-efficient products. However, politicians could design policies that encour
age reductions in the prices of energy-efficient appliances and provide subsidies to rural hous
eholds for greater use of energy-efficient appliances.',
 'Consumer knowledge and support for the acquisition of efficient products: The dissemination
of information to inform consumers about the efficiency of household appliances and subsidies
is essential to achieve energy efficiency.',
 'Products with an energy efficiency label: Refrigerator consumers are more willing to pay for
better-quality appliances that promote a safe environment.\t',
 'Consumer knowledge of energy-efficient equipment and labels. Consumer knowledge regardin
g the advantages and benefits of choosing products that are energy efficient. Furthermore, u
sing a label with clear and objective information helps when choosing equipment.',
 'Consumer knowledge regarding the information on the energy efficiency label: Using easily u
nderstood information and/or additional information in monetary terms can facilitate consume
r understanding and the propensity to purchase more efficient equipment.',
 "Consumer knowledge regarding the energy efficiency of equipment: The consumer's lack of k
nowledge regarding the energy efficiency of equipment leads to less efficient use.",
 'Consumer knowledge regarding efficient products. Lack of knowledge can create difficulties
in conserving energy for families.',
 'Communication in optimizing energy efficiency for the consumer:  Initiatives that address co
mmunication clearly and effectively for consumers lead them to obtain and consume more effi
cient products.',
 'Consumer knowledge of the efficiency label: The advantages and benefits of choosing energy
-efficient products lead to greater acceptability and willingness to pay more.',
 "Use of efficient products. The energy efficiency labeling program and the three subsidy pro
grams promote households' choice of energy-efficient refrigerators.",
 "Consumer knowledge of energy-efficient equipment and labels: The use of a label that clearly
and objectively presents the equipment's energy performance and provides more accurate inf
```

```
ormation leads the consumer to purchase more efficient equipment.",
 'Consumer knowledge of the Energy Star label: Consumer knowledge of the advantages and be
nefits leads to the choice of products that are energy efficient.',
 'Consumer knowledge regarding the correct use of efficient equipment: By not knowing how t
o use the equipment more efficiently, consumers can considerably increase energy consumptio
n, as in many cases, the label does not show which usage model is the most efficient.',
 'Energy-efficient products. Households demonstrate a favorable preference for labeled appli
ances and an intention to pay more to purchase energy-efficient appliances.',
 'Consumer knowledge about the class A appliance label: Consumer knowledge regarding the ef
ficiency of Class A appliances varies concerning economic class, test level, and other residence
characteristics.',
 'Reduced energy consumption using Energy Star-certified appliances: Certified devices gener
ated energy savings compared to non-certified devices.',
 'Use of efficient products. Applying discounts on Energy Star appliances increases the use of
efficient products.',
 'Minimum energy efficiency standards for appliances: The adoption of minimum standards lea
ds to cuts in electricity-intensive appliances, thus generating the most cost-effective way to a
chieve energy efficiency.',
 'Energy efficiency programs. It provides consumers with reliable products certified for ener
gy efficiency, with Energy Star as the protagonist.',
 'Subsidies (rebates) to promote the purchase of energy-efficient equipment: To avoid the ef
ficiency rebound effect, charging taxes on non-efficient devices can be combined.',
 'Specifications for Energy Star energy-efficient products: The Program aims to develop spec
ifications so that the product is considered energetically sustainable and can earn the certific
ation label.',
 'Energy Star efficiency program. Publicizing the program is essential to achieving energy effi
ciency.',
 'Energy savings from Energy Star-certified appliances: Using certified energy-efficient appli
ances generates Energy and economic savings.',
 'Energy Star-certified products generate energy savings: Using devices certified by the prog
ram generated energy and monetary savings.',
 'Energy Star energy efficiency program. The program provides consumers with the reliability
of certified products for their energy efficiency.']
```

In [6]:
```
corpus_embeddings = embedder.encode(corpus)
```

In [7]:
```
corpus_embeddings
```

Out[7]:
```
array([[ 0.00155502, -0.04275565,  0.00012521, ..., -0.0246107 ,
        -0.02269541, -0.03309743],
       [-0.00982781, -0.02395886, -0.03986001, ..., -0.00622479,
        -0.04117443, -0.03078831],
       [-0.01950175,  0.02895311, -0.0073098 , ...,  0.00183647,
        -0.04639709, -0.03043038],
       ...,
       [-0.01585284,  0.01641727, -0.00586385, ...,  0.01264659,
        -0.03269481, -0.06060567],
       [-0.00913825,  0.02369976, -0.02053741, ..., -0.00137325,
        -0.02708015, -0.05316076],
       [-0.01173451, -0.01453331,  0.01471505, ..., -0.01013431,
         0.0135545 , -0.05425853]], dtype=float32)
```

In [8]:
```
import sklearn.cluster as cluster
```

In [9]:
```
X = corpus_embeddings
X.shape
```

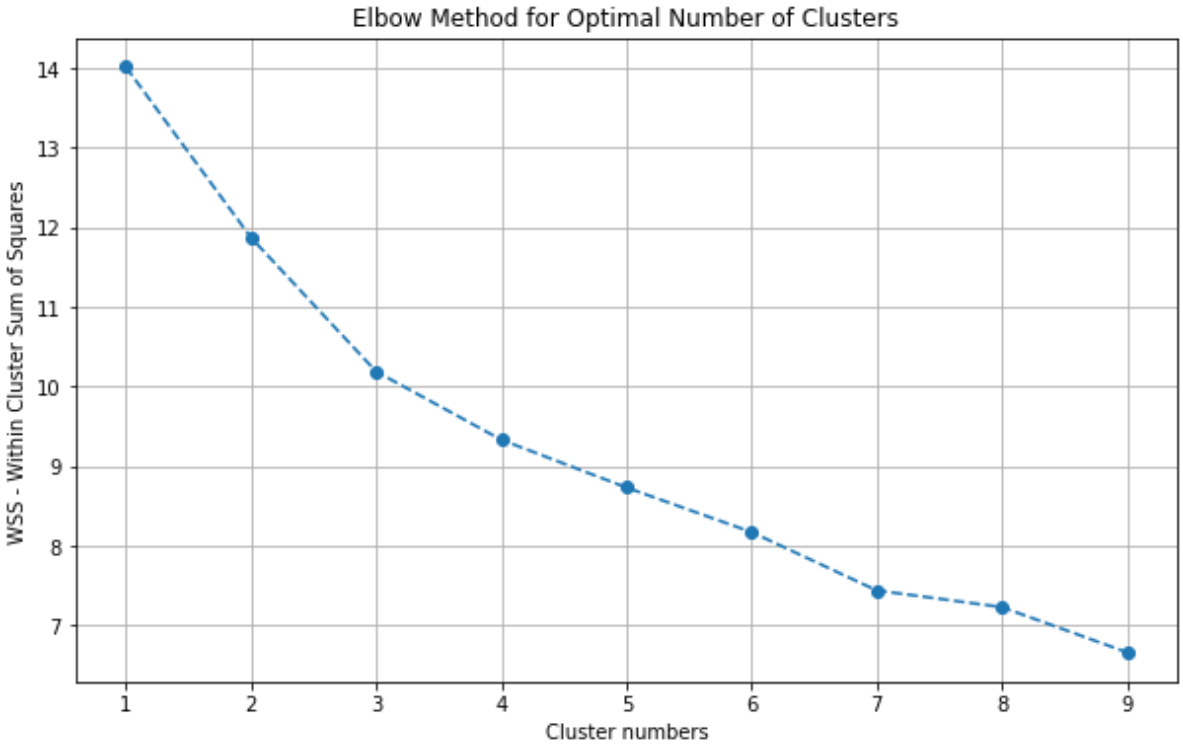
Out[9]:
```
(41, 768)
```

The Clusters - Determinando o número de agrupamentos

In [10]:
```
# Determine the optimal number of clusters using the Elbow method
wcss = []  # within-cluster sum of squares
```

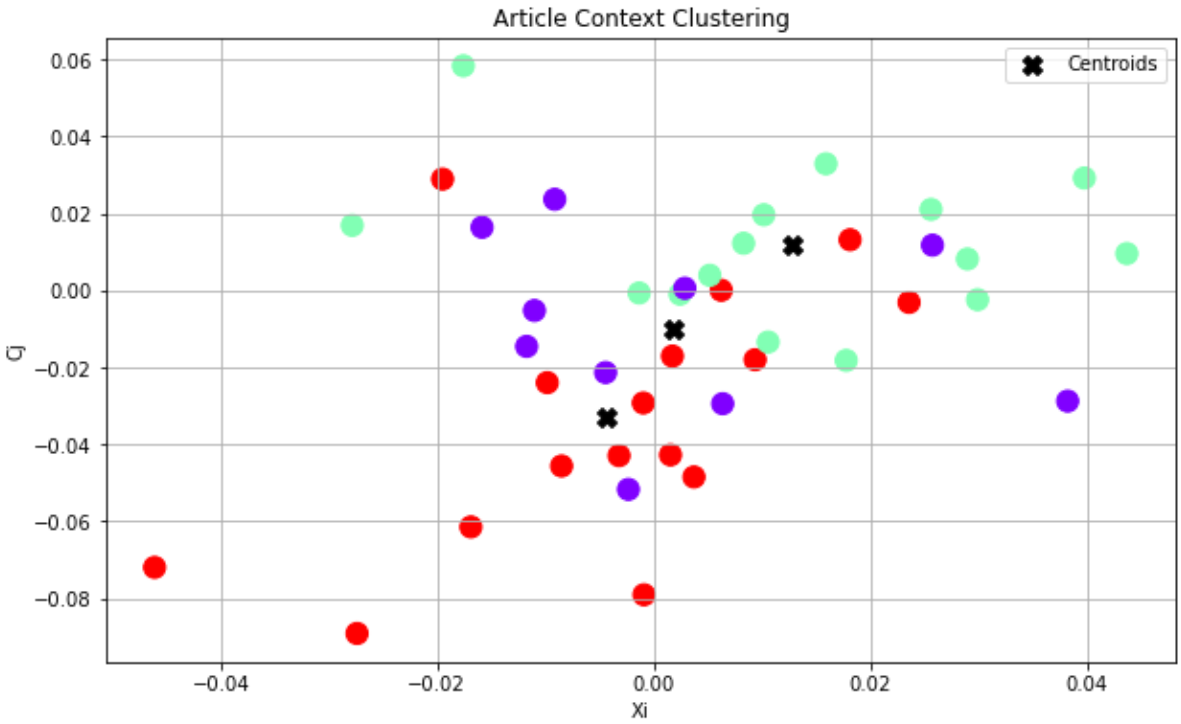
```
cluster_range = range(1, 10)  # test up to 10 clusters

for k in cluster_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

# Plot the Elbow method
plt.figure(figsize=(10, 6))
plt.plot(cluster_range, wcss, marker='o', linestyle='--')
plt.xlabel('Cluster numbers')
plt.ylabel('WSS - Within Cluster Sum of Squares')
plt.title('Elbow Method for Optimal Number of Clusters')
plt.grid(True)
plt.show()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1036: UserWarnin
g: KMeans is known to have a memory leak on Windows with MKL, when there are less chunks t
han available threads. You can avoid it by setting the environment variable OMP_NUM_THREA
DS=1.
  warnings.warn(
```



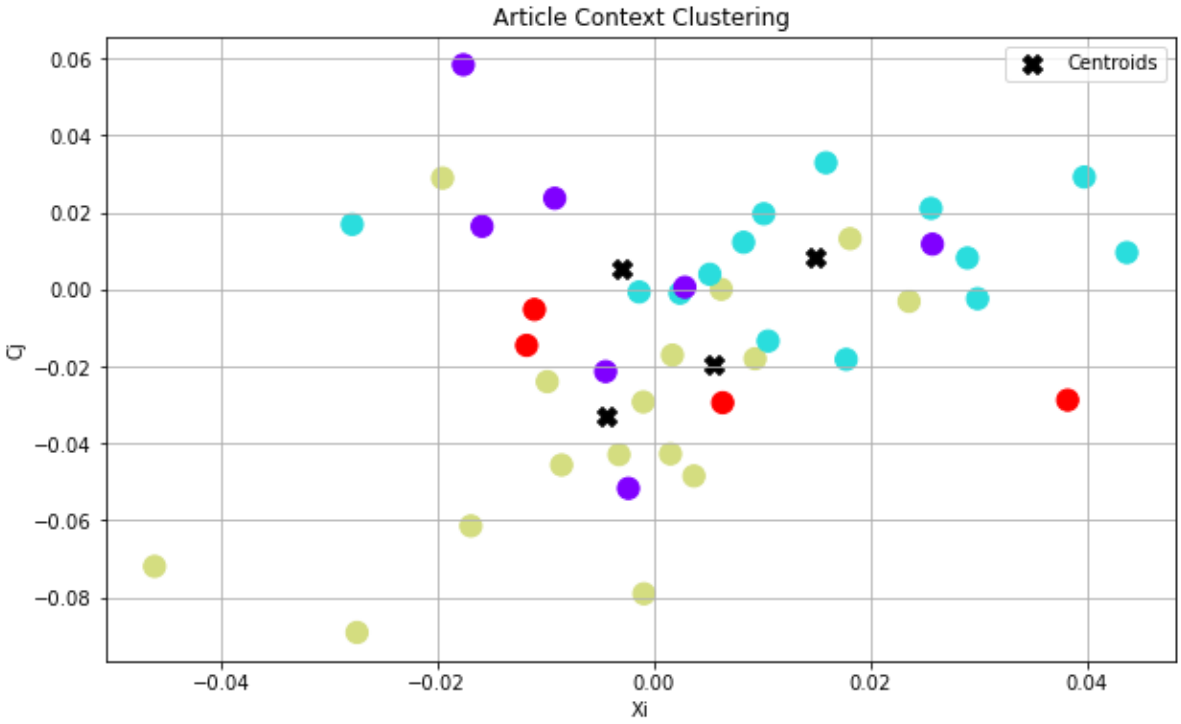
In [11]:

```
# Perform KMeans clustering with 3 clusters
kmeans = KMeans(n_clusters=3, random_state=42, n_init=200)
clusters = kmeans.fit_predict(X)

# Plot the clusters
plt.figure(figsize=(10, 6))
plt.scatter(X[:, 0], X[:, 1], c=clusters, s=125, cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=100, c='black', mark
plt.xlabel('Xi')
plt.ylabel('Cj')
plt.title('Article Context Clustering')
plt.legend()
plt.grid(True)
plt.show()
```



In [12]:

```
# Perform KMeans clustering with 4 clusters
kmeans = KMeans(n_clusters=4, random_state=42, n_init=200)
clusters = kmeans.fit_predict(X)

# Plot the clusters
plt.figure(figsize=(10, 6))
plt.scatter(X[:, 0], X[:, 1], c=clusters, s=125, cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=100, c='black', mark
plt.xlabel('Xi')
plt.ylabel('Cj')
plt.title('Article Context Clustering')
plt.legend()
plt.grid(True)
plt.show()
```



In [13]:
```
# Perform KMeans clustering with 5 clusters
kmeans = KMeans(n_clusters=5, random_state=42, n_init=200)
clusters = kmeans.fit_predict(X)

# Plot the clusters
plt.figure(figsize=(10, 6))
plt.scatter(X[:, 0], X[:, 1], c=clusters, s=125, cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=100, c='black', mark
plt.xlabel('Xi')
plt.ylabel('Cj')
plt.title('Article Context Clustering')
plt.legend()
plt.grid(True)
plt.show()
```

Clustering for 3 clusters

In [14]:
```
clustering_model = KMeans(n_clusters=3, random_state=42, n_init=750)
clustering_model.fit(corpus_embeddings)
cluster_assignment = clustering_model.labels_
```

In [15]:
```
cluster_assignment
```

Out[15]:
```
array([2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

In [16]:
```
cluster_df = pd.DataFrame(corpus, columns = ['corpus'])
cluster_df['cluster'] = cluster_assignment
#cluster_df['cluster_name'] = df['cluster_name']
cluster_df.head()
```

Out[16]:

	corpus	cluster
0	Energy efficiency labels. Providing informatio...	2
1	Energy efficiency labels for dishwashers: Usin...	2
2	Efficiency program and energy labels: The ener...	2
3	Energy efficiency label. The presentation of i...	2
4	Energy efficiency label. Using more informatio...	2

In [17]:
```
# determining the name of the file
file_name = 'Energy_topics_cluster.xlsx'
```

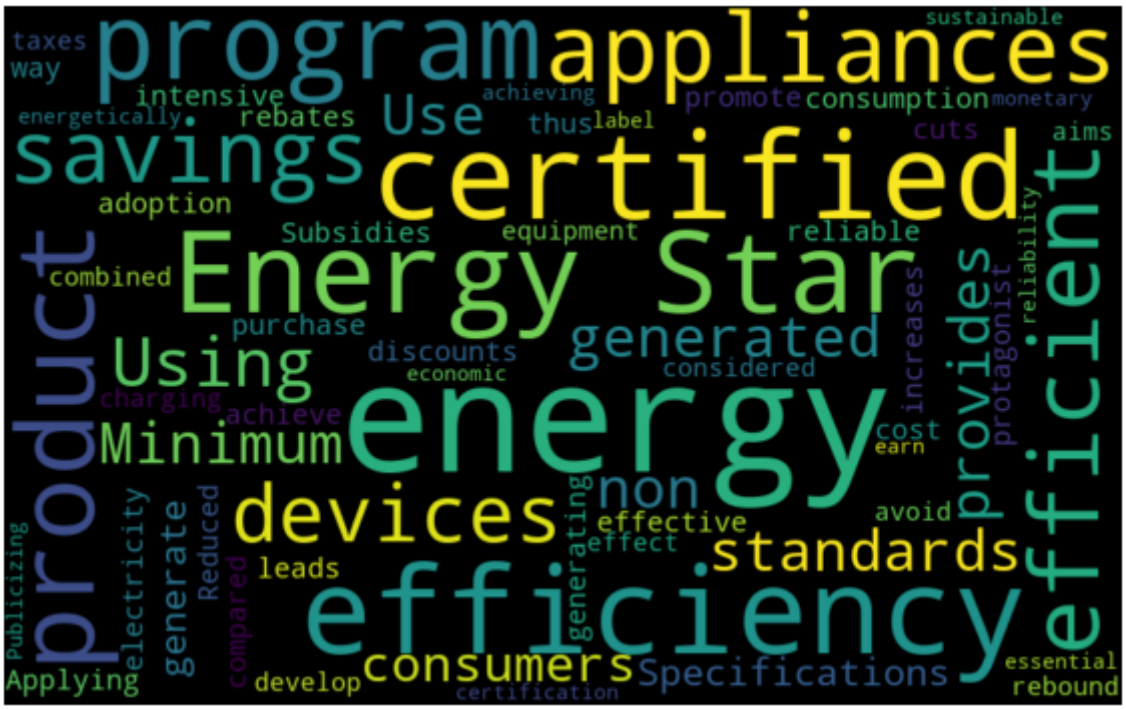
In [18]:
```
# saving the excel
cluster_df.to_excel(file_name)
print
```

Out[18]: <function print>

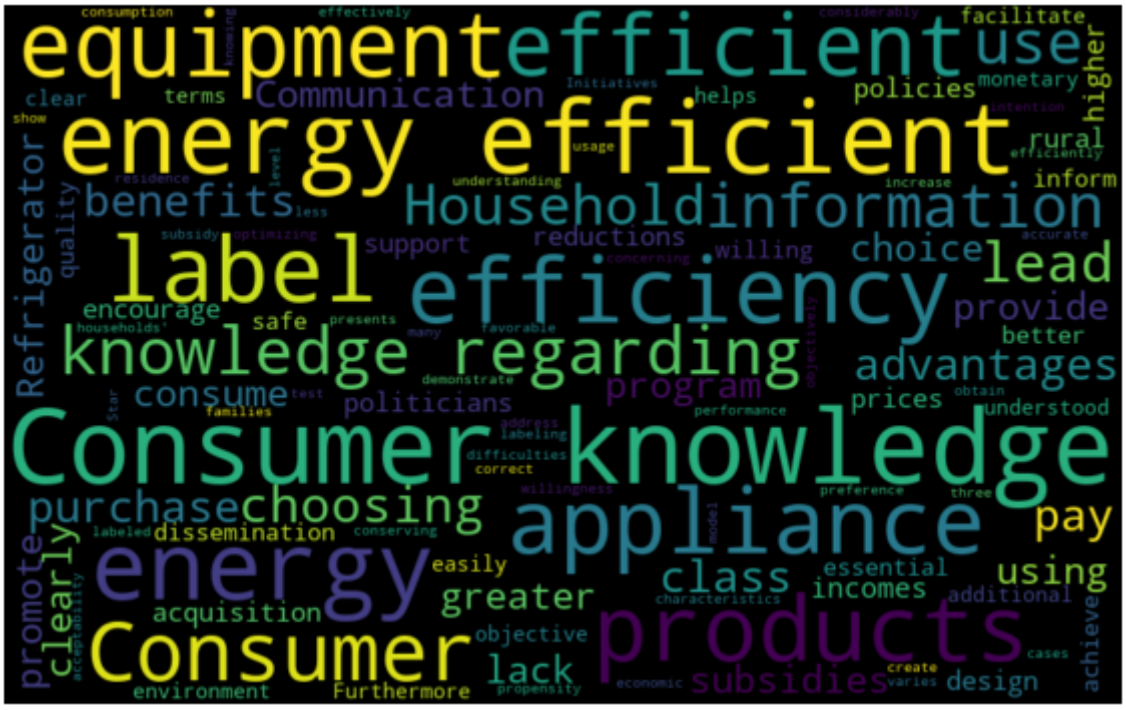
Clusters word clouds

In [19]:
```
#word cloud
def word_cloud(pred_df,label):
    wc = ' '.join([text for text in pred_df['corpus'][pred_df['cluster'] == label]])
    wordcloud = WordCloud(width=800, height=500,
    random_state=21, max_font_size=110).generate(wc)
    fig7 = plt.figure(figsize=(10, 7))
    plt.imshow(wordcloud, interpolation="bilinear")
    plt.axis('off')
```

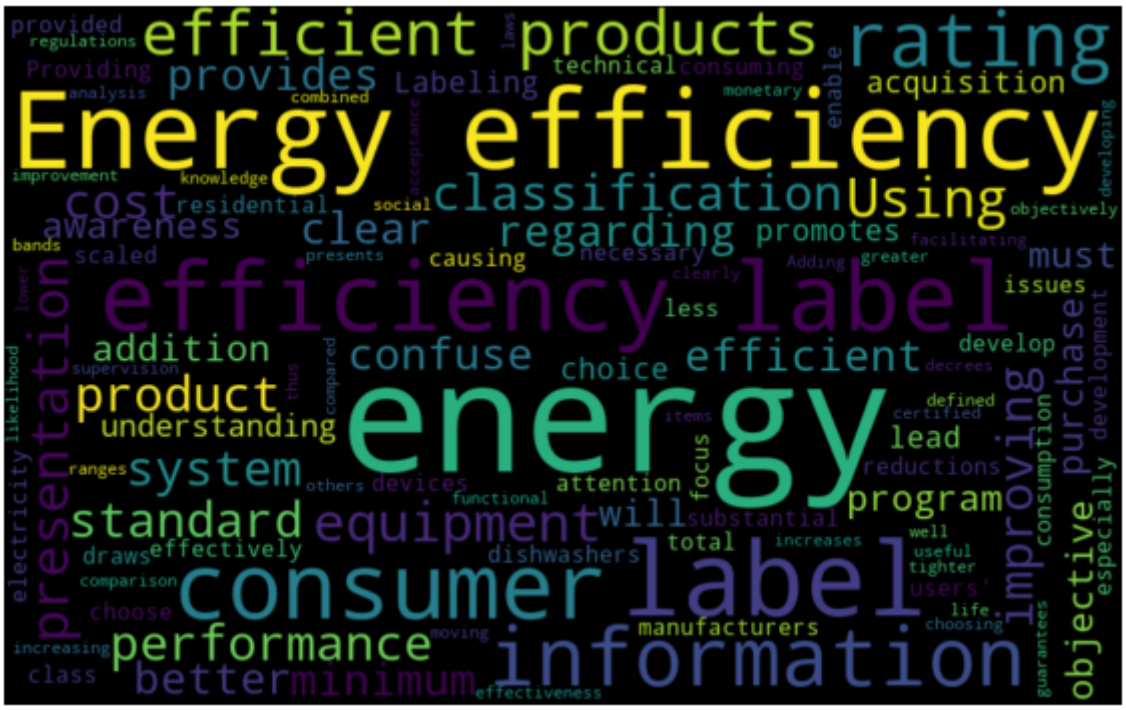
In [20]:
```
word_cloud(cluster_df,0)
```



In [21]:
```
word_cloud(cluster_df,1)
```



In [22]:
```
word_cloud(cluster_df,2)
```



In [ ]: