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('Criteria_description.xlsx')
df.head()
```

Out[3]:

	code	FATORES COLETADOS	Criteria	Criterias definition	Definition	cited	Percentage
0	C001	Communication	Communication	Communication is the biggest challenge for GSD...	Communication. Communication is the biggest ch...	33	0.51
1	C007	Trust	Trust building	Personal or impersonal, including cognitive tr...	Trust building. Personal or impersonal, includ...	28	0.43
2	C003	Culture	Cultural differences among teams	Each culture has its standards, styles and mor...	Cultural differences among teams. Each culture...	27	0.42
3	C017	Coordination	Coordination challenges level	Team coordination is defined as activities req...	Coordination challenges level. Team coordinati...	26	0.40
4	C004	Temporal Issues	Temporal Issues	Temporal issues are related to the time differ...	Temporal issues. Temporal issues are related t...	22	0.34

Clustering

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

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

Out[5]:
```
['Communication. Communication is the biggest challenge for GSD due to the need for adequate and proper ways of communication in general. In addition, the reduced communication frequency with the project team members became a problem due to the need for more i
nformal or face-to-face contact.',
 'Trust building. Personal or impersonal, including cognitive trust, which refers to beliefs about others' competence and reliability. This can lead individuals to engage in less self-protective actions and be more likely to take risks.',
 'Cultural differences among teams. Each culture has its standards, styles and moral principles, which can provoke communication related issues when individual belonging from different cultural background communicate with another one.',
 'Coordination challenges level. Team coordination is defined as activities required to maintain consistency within a work product or to manage dependencies within the workflow. There are many different types of dependencies between task and task holders, thes
e dependencies lead to a need for coordination among stakeholders working on a related set of tasks. When these coordination needs are not satisfied, they will have coordination challenges.',
 'Temporal issues. Temporal issues are related to the time difference between teams that work in several remote locations. Delayed feedback and responses are problematic and restrict the possibility of synchronous interaction, cooperation, and confidential ass
essment. This criterion is related to the geographic dimension.',
 'Knowledge interchange rate. Knowledge interchange rate is a process of exchange of explicit or tacit knowledge between two agents, during which one agent purposefully receives and uses the knowledge provided by another.',
 'English domain. The usage of a different language among distributed team members. In the current years, the English language has been widely used as a professional language at both national and international platforms.',
 'Team issues. Within the global team context there is a clear need for the development of a one team approach. Teamwork is based on team member relationships that facilitate the development of mutual respect and trust. This leads to the development of a cohes
ive motivated team that sees itself as a single unit regardless of its members' location.',
 'Defined of roles and responsibilities. Defined roles and responsibilities are essential to assign the proper responsibility and task to the right person and time and should be clearly defined, articulated, and effectively disseminated for all team members.',
 'Geographical dimension. Geographical distance is the dispersion geographically between team members in remote sites. Communication risk increases whenever geographic distance increases. Therefore, this criterion is related to the geographic dimension.',
 'Degree of cooperation.  Collaboration among distributed teams. Numerous issues directly mitigate against the establishment of cooperation in the global team environment. In these circumstances from the project management perspective cooperation between team
locations must be developed, established, and effectively managed to avoid the reluctance of sharing information.',
 'Effective leadership. The teams may be formed without planning, lacking the required knowledge and skills. Skilled leadership that has the expertise to assess and analyze the impact of demanded changes and will make the right decision at the right time. Lack
of integration planning and lack of management. An effective integration plan is necessary for all Global Software Development projects, especially for large one, to be successful at integration stage.',
 'Availability of human resource. Lack of human resources, knowledge, and skills. Lack of suitable infrastructure for integration and the nonavailability of skilled human resources to solve integration issues in time hinder the integration process. This criter
```

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

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

Out[7]:
```
array([[ 0.06978382,  0.01288685, -0.0278127 , ...,  0.0045306 ,
         0.0103953 ,  0.00848669],
       [-0.01921752,  0.00248105,  0.01927953, ...,  0.0201411 ,
         0.03768687,  0.00922179],
       [ 0.04591352, -0.00671977, -0.03013351, ...,  0.02603393,
         0.05961482,  0.00301218],
       ...,
       [ 0.01086077,  0.03712564, -0.0216443 , ..., -0.03415754,
         0.04323032, -0.00309642],
       [-0.06578007,  0.01272878, -0.03762549, ...,  0.0334544 ,
         0.02336276, -0.01384156],
       [-0.00914138,  0.0748887 , -0.00906282, ..., -0.00305127,
        -0.00660435,  0.00308621]], dtype=float32)
```

In [8]:
```
clustering_model = KMeans(n_clusters=25, random_state=0, n_init=300)
clustering_model.fit(corpus_embeddings)
cluster_assignment = clustering_model.labels_
```

In [9]:
```
cluster_assignment
```

Out[9]:
```
array([ 5, 15,  5, 10,  5, 21,  9, 14, 10,  5, 14, 17, 17, 21, 21,  8, 18,
       23, 20,  6, 11,  9, 23,  1, 14,  5,  8,  8, 23, 10, 16,  5,  9, 23,
        5, 18, 11, 11, 22,  8,  3,  5,  5, 22,  5,  5, 17,  9, 23, 17, 19,
        2,  8, 21,  1, 20, 17, 23,  8,  5,  5,  5,  9,  5, 16,  2,  5, 17,
       11,  5, 18, 18, 17, 10, 21, 12, 14,  8,  8, 14, 22, 17,  5,  5,  2,
       10, 11, 24,  6,  6, 23, 18, 14, 19, 17,  2,  4, 18, 23, 16, 14, 21,
       18, 23,  8,  3,  3,  5, 15, 23,  2,  8, 22, 14,  5,  9, 15, 13,  5,
        9, 14, 10, 10,  7,  0,  9, 23,  0, 17, 14, 14, 14,  6, 11,  9,  9,
       12,  3, 16, 16, 16, 14,  2,  9,  8,  2, 11, 20,  0,  3,  8,  3,  0,
       17,  9,  8, 15,  8, 17, 22, 17, 22, 22, 24, 22,  5,  5,  5,  5, 16,
        8, 14, 21, 21,  9, 19, 23, 17,  2,  1, 17,  8, 15, 10,  5,  9,  4,
       10, 19,  7, 19,  2, 19,  3,  3, 23, 21, 21, 24, 24, 10, 10, 10, 16,
       14, 17,  6,  6,  1,  1, 19, 17, 15, 18, 18,  8, 17, 24, 10, 21,  8,
        2, 24, 24, 11, 20, 23, 20, 20, 20, 20, 20, 20, 14,  0,  3, 20,  0,
        5,  8,  8,  6, 21, 15,  2, 17, 18, 14,  8, 12,  2, 22, 22, 22, 22,
       17,  2,  5,  5, 14,  5,  5, 19,  9, 15,  0, 19,  0, 19,  0, 19,  7,
       19, 19, 19,  2, 19,  4,  5,  2,  1, 20,  6, 19,  6,  6,  4,  7, 13,
       13, 13, 13, 13, 13, 13, 13,  9, 19, 19,  4, 17,  6,  4,  4,  4, 18,
        2, 12, 12,  1,  2,  3, 21, 14, 18,  6,  6,  6,  6])
```

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

Out[16]:

	corpus	cluster	code
0	Communication. Communication is the biggest ch...	5	C001
1	Trust building. Personal or impersonal, includ...	15	C007
2	Cultural differences among teams. Each culture...	5	C003
3	Coordination challenges level. Team coordinati...	10	C017
4	Temporal issues. Temporal issues are related t...	5	C004

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

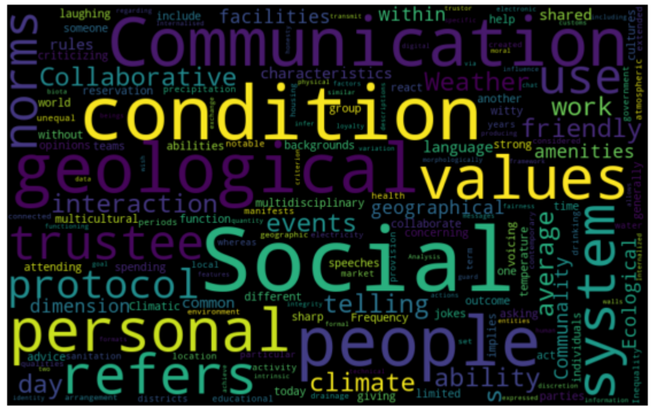
In [18]:
```
# saving the excel
cluster_df.to_excel(file_name)
print
```

Out[18]:
```
<function print>
```

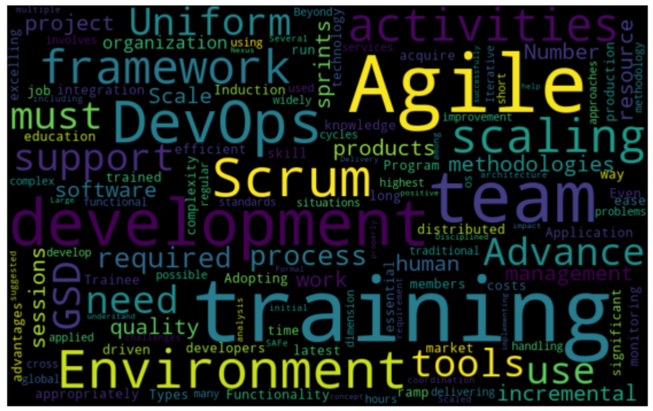
Clusters word clouds

In [11]:
```
#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 [19]:
```
word_cloud(cluster_df,0)
```



In [20]:
```
word_cloud(cluster_df,1)
```

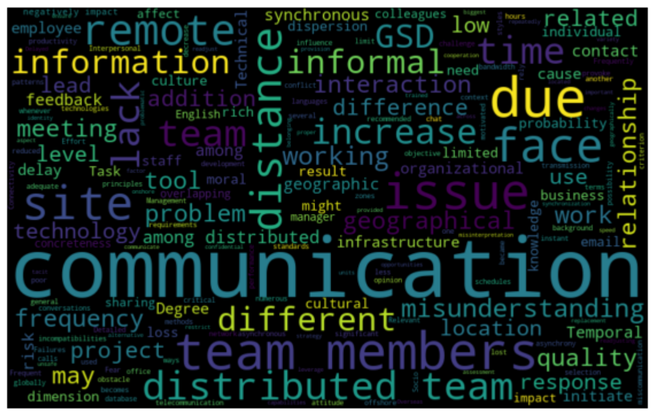


In [21]:
```
word_cloud(cluster_df,2)
```

In [22]:
```
word_cloud(cluster_df,3)
```

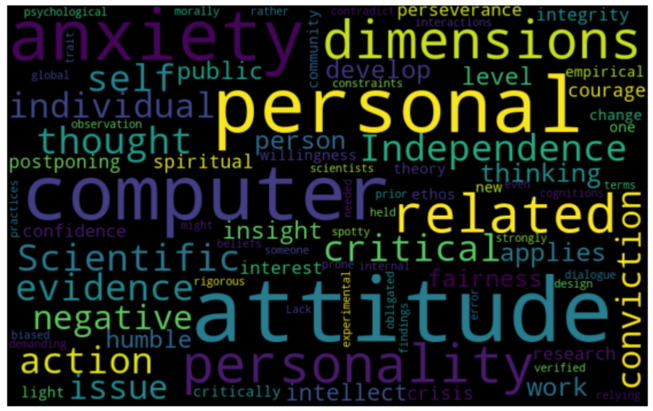
In [23]:
```
word_cloud(cluster_df,4)
```

In [24]:
```
word_cloud(cluster_df,5)
```



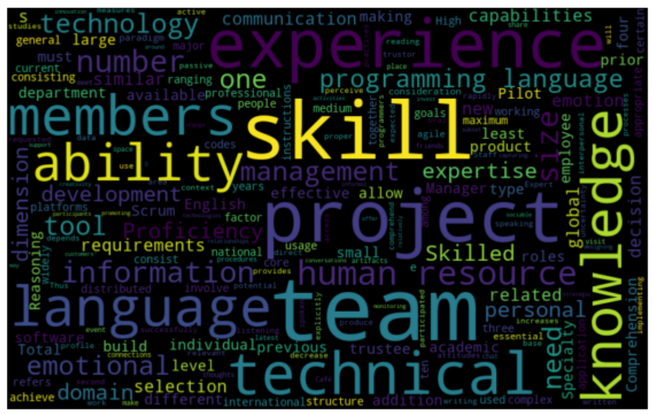
In [25]:
```
word_cloud(cluster_df,6)
```

In [27]:
```
word_cloud(cluster_df,7)
```



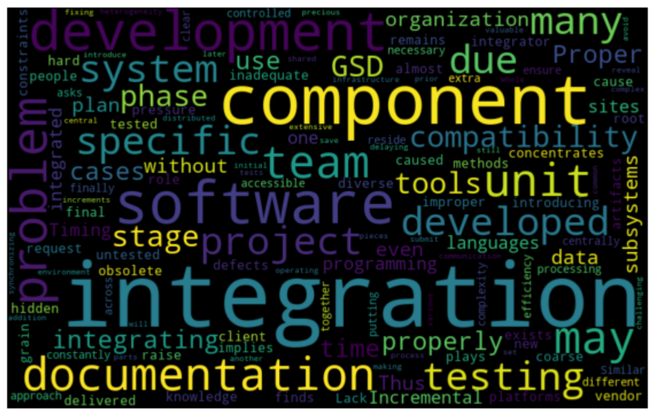
In [28]:
```
word_cloud(cluster_df,8)
```

In [29]:
```
word_cloud(cluster_df,9)
```



In [30]:
```
word_cloud(cluster_df,10)
```

In [31]:
```
word_cloud(cluster_df,11)
```

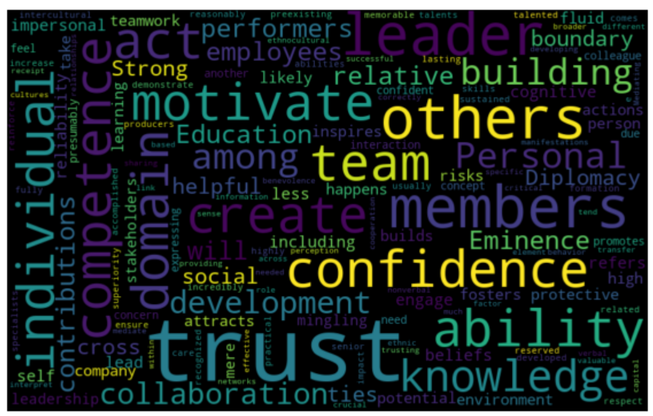


In [32]:
```
word_cloud(cluster_df,12)
```

In [33]:
```
word_cloud(cluster_df,13)
```

In [34]:
```
word_cloud(cluster_df,14)
```

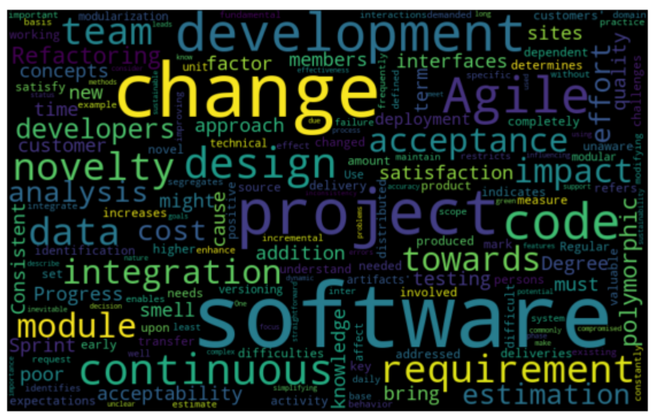
In [35]:
```
word_cloud(cluster_df,15)
```



In [36]:
```
word_cloud(cluster_df,16)
```

In [37]:
```
word_cloud(cluster_df,17)
```

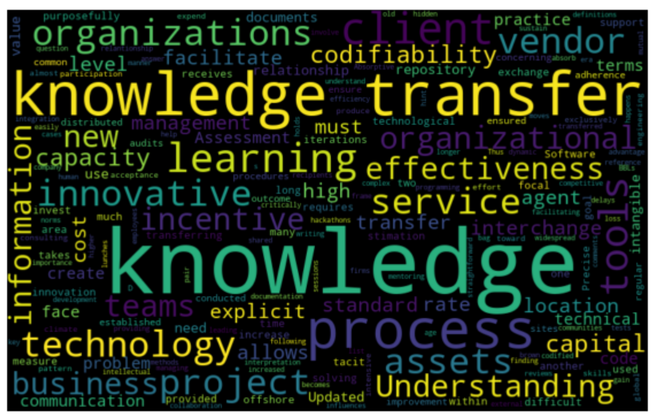
In [38]:
```
word_cloud(cluster_df,18)
```



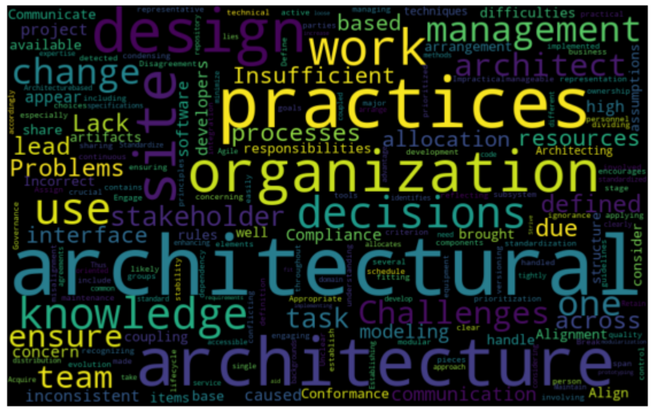
In [39]:
```
word_cloud(cluster_df,19)
```

In [40]:
```
word_cloud(cluster_df,20)
```

In [41]:
```
word_cloud(cluster_df,21)
```

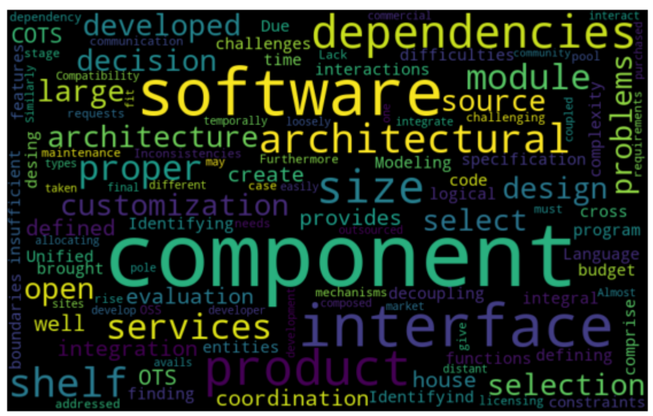


In [42]:
```
word_cloud(cluster_df,22)
```



In [43]:
```
word_cloud(cluster_df,23)
```

In [44]:
```
word_cloud(cluster_df,24)
```



In [ ]:
```

```