The objective of this project is to investigate the sentence clustering using Natural Language Processing (NLP).
Explore the NLP approaches for the text file which is uploaded on BB. It consists of around 100 json objects separated by line breaks. Each object has a title and an id. All titles are subcontracts of a single construction project.

So we have to read the file first:

```
data = list()
for line in open('DataP_8.txt','r'):
    data.append(line)
```

Storing it into a list one after the other.

For finding how similar a pair of strings based on which the strings can be added to seperate sets.

```
from difflib import SequenceMatcher

def getRatio(a, b):
    return SequenceMatcher(None, a, b).ratio()
```

As the roots are the same I had set the threshold to 95% anything below it is not considered

```
treshold     = 0.95
minGroupSize = 1

from itertools import combinations
```

As the minimum group size is one and leaving apart the first 18 characters as it is just **{"id": 1, "title":**

```
paired = { c:{c} for c in data }
for a,b in combinations(data,2):
    if getRatio(a[18:38],b[18:38]) < treshold: continue
    paired[a].add(b)
    paired[b].add(a)
```

Finding out groups and storing as sets as it takes only dissimilar items so that a combination is stored only once

```
groups    = list()
ungrouped = set(data)
while ungrouped:
    bestGroup = {}
```

```
    for city in ungrouped:
        g = paired[city] & ungrouped
        for c in g.copy():
            g &= paired[c]
        if len(g) > len(bestGroup):
            bestGroup = g
    if len(bestGroup) < minGroupSize : break  # to terminate grouping early change
minGroupSize to 3
    ungrouped -= bestGroup
    groups.append(bestGroup)
print('Total number of families formed from list :',len(groups))
print('\n')
print('Sample families are \n',groups[0])
print(groups[1])
print(groups[2])
```

```
======== RESTART: C:\Users\atche\Desktop\Study zone\AI\Project 8\p1.py ========
Total number of families formed from list : 32


Sample families are
 {'{"id": 99, "title": "KH Landshut-Achdorf BA 5 \\u2013 120-4080-01 Aufzugsanlage."}\n', '{"id": 75, "title": "KH Landshut-Achdorf BA 5 \\u2013 120-1103-01 Fenster- und Fassadenarbeiten."
}\n', '{"id": 11, "title": "KH Landshut-Achdorf BA 5 \\u2013 120-3061-01 Geb\\u00e4udeautomation 2."}\n', '{"id": 32, "title": "KH Landshut-Achdorf BA 5 \\u2013 120-4013-01 Elektroinstalla
tion mit Beleuchtung."}\n', '{"id": 43, "title": "KH Landshut-Achdorf BA 5 \\u2013 120-3031-01 Sanit\\u00e4r 2."}\n', '{"id": 74, "title": "KH Landshut-Achdorf BA 5 \\u2013 120-1020-01 Roh
bau inkl. Abbruch."}\n', '{"id": 63, "title": "KH Landshut-Achdorf BA 5 \\u2013 120-5031-01 med. Gase 2."}\n'}
{'{"id": 81, "title": "Sanierung und Erweiterung Parkbad Laupheim \\u2013 Dachabdichtungsarbeiten."}\n', '{"id": 98, "title": "Sanierung und Erweiterung Parkbad Laupheim \\u2013 Zimmer- un
d Holzbauarbeiten."}\n', '{"id": 8, "title": "Sanierung und Erweiterung Parkbad Laupheim \\u2013 Rohbau- und Spezialtiefbauarbeiten."}\n', '{"id": 10, "title": "Sanierung und Erweiterung P
arkbad Laupheim \\u2013 Metallbau-/Verglasungsarbeiten."}\n', '{"id": 88, "title": "Sanierung und Erweiterung Parkbad Laupheim \\u2013 Badewassertechnik und MSR-Technik."}\n', '{"id": 1, "
title": "Sanierung und Erweiterung Parkbad Laupheim \\u2013 Edelstahlbecken, Beckenausstattung."}\n', '{"id": 66, "title": "Sanierung und Erweiterung Parkbad Laupheim \\u2013 Heizungsarbei
ten."}\n'}
{'{"id": 51, "title": "Neubau Jugendwohnheim Landshut \\u2013 LV 101 Erd- und Rohbauarbeiten."}\n', '{"id": 49, "title": "Neubau Jugendwohnheim Landshut \\u2013 LV 303 D\\u00e4mmung techni
scher Anlagen."}\n', '{"id": 28, "title": "Neubau Jugendwohnheim Landshut \\u2013 LV 108 Trockenbauarbeiten."}\n', '{"id": 62, "title": "Neubau Jugendwohnheim Landshut \\u2013 LV 107 Metal
lbauarbeiten \\u2013 Sonnenschutz / Schiebel\\u00e4den."}\n', '{"id": 102, "title": "Neubau Jugendwohnheim Landshut \\u2013 LV 401 Elektroinstallation."}\n', '{"id": 91, "title": "Neubau J
ugendwohnheim Landshut \\u2013 LV 113 Estricharbeiten."}\n'}
>>>
```