Table of Contents

Introduction

NLP - To simply put, Natural Language Processing (NLP) is a field which is concerned with making computers understand human language. NLP techniques are applied heavily in information retrieval (search engines), machine translation, document summarization, text classification, natural language generation etc.

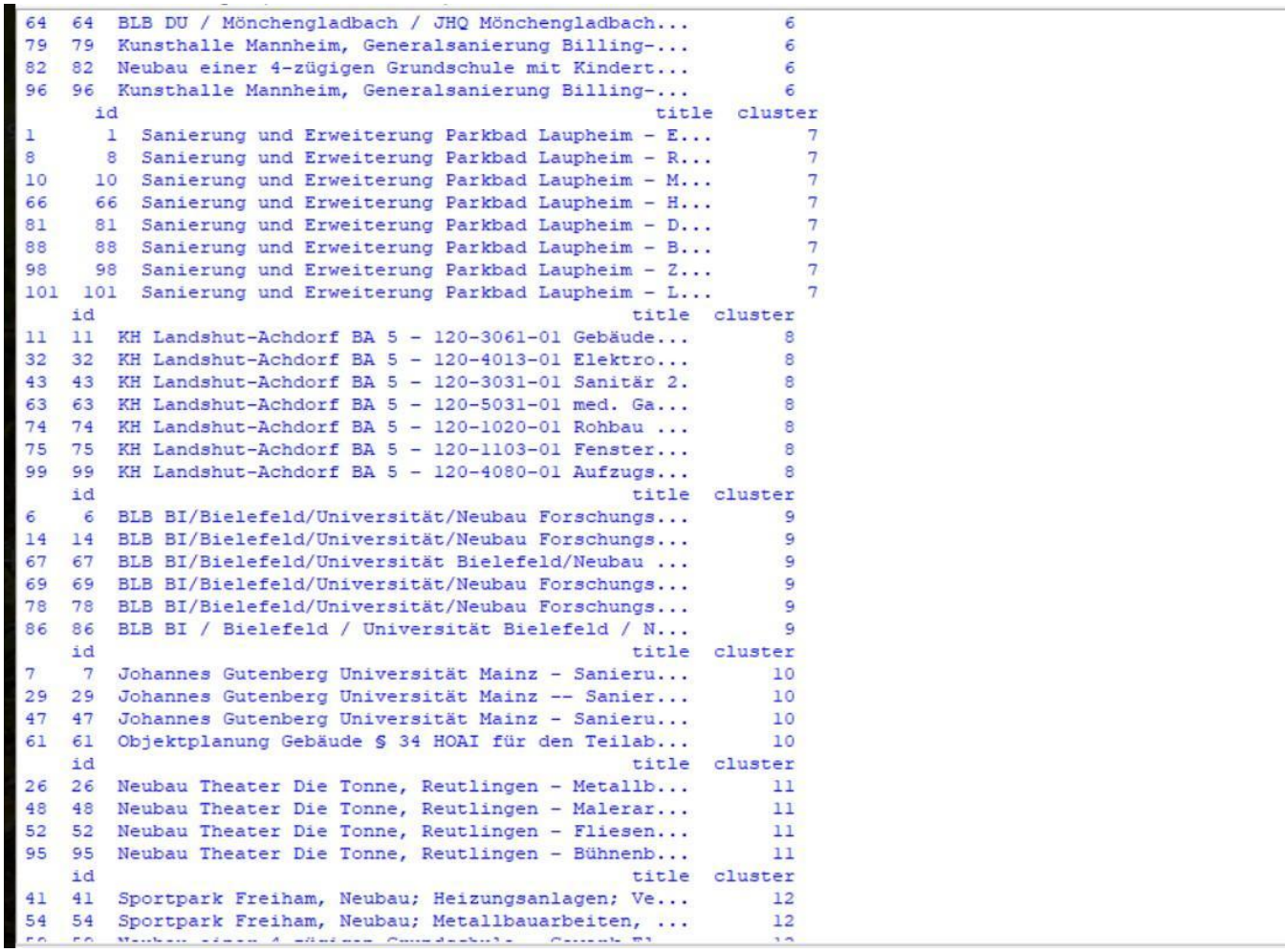
Methodology

➔ **Sentence Clustering using NLP** - Clustering is a process of grouping similar items together. Each group, also called as a cluster, contains items that are similar to each other. Clustering algorithms are unsupervised learning algorithms i.e. we do not need to have labelled datasets. There are many clustering algorithms for clustering including KMeans, DBSCAN, Spectral clustering, hierarchical clustering etc and they have their own advantages and disadvantages. The choice of the algorithm mainly depends on whether or not you already know how many clusters to create. Some algorithms such as KMeans need you to specify the number of clusters to create whereas DBSCAN does not need you to specify. Another consideration is whether you need the trained model to be able to predict clusters for unseen dataset. KMeans can be used to predict the clusters for new dataset whereas DBSCAN cannot be used for new dataset.

➔ **Dataset**

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.

Results



```
64   64  BLB DU / Mönchengladbach / JHQ Mönchengladbach...        6
79   79  Kunsthalle Mannheim, Generalsanierung Billing-...        6
82   82  Neubau einer 4-zügigen Grundschule mit Kindert...        6
96   96  Kunsthalle Mannheim, Generalsanierung Billing-...        6
      id                                              title  cluster
1      1  Sanierung und Erweiterung Parkbad Laupheim - E...        7
8      8  Sanierung und Erweiterung Parkbad Laupheim - R...        7
10    10  Sanierung und Erweiterung Parkbad Laupheim - M...        7
66    66  Sanierung und Erweiterung Parkbad Laupheim - H...        7
81    81  Sanierung und Erweiterung Parkbad Laupheim - D...        7
88    88  Sanierung und Erweiterung Parkbad Laupheim - B...        7
98    98  Sanierung und Erweiterung Parkbad Laupheim - Z...        7
101  101  Sanierung und Erweiterung Parkbad Laupheim - L...        7
     id                                                 title  cluster
11   11  KH Landshut-Achdorf BA 5 - 120-3061-01 Gebäude...        8
32   32  KH Landshut-Achdorf BA 5 - 120-4013-01 Elektro...        8
43   43  KH Landshut-Achdorf BA 5 - 120-3031-01 Sanitär 2.        8
63   63  KH Landshut-Achdorf BA 5 - 120-5031-01 med. Ga...        8
74   74  KH Landshut-Achdorf BA 5 - 120-1020-01 Rohbau ...        8
75   75  KH Landshut-Achdorf BA 5 - 120-1103-01 Fenster...        8
99   99  KH Landshut-Achdorf BA 5 - 120-4080-01 Aufzugs...        8
     id                                                 title  cluster
6     6  BLB BI/Bielefeld/Universität/Neubau Forschungs...        9
14   14  BLB BI/Bielefeld/Universität/Neubau Forschungs...        9
67   67  BLB BI/Bielefeld/Universität Bielefeld/Neubau ...        9
69   69  BLB BI/Bielefeld/Universität/Neubau Forschungs...        9
78   78  BLB BI/Bielefeld/Universität/Neubau Forschungs...        9
86   86  BLB BI / Bielefeld / Universität Bielefeld / N...        9
     id                                                 title  cluster
7     7  Johannes Gutenberg Universität Mainz - Sanieru...       10
29   29  Johannes Gutenberg Universität Mainz -- Sanier...       10
47   47  Johannes Gutenberg Universität Mainz - Sanieru...       10
61   61  Objektplanung Gebäude § 34 HOAI für den Teilab...       10
     id                                                 title  cluster
26   26  Neubau Theater Die Tonne, Reutlingen - Metallb...       11
48   48  Neubau Theater Die Tonne, Reutlingen - Malerar...       11
52   52  Neubau Theater Die Tonne, Reutlingen - Fliesen...       11
95   95  Neubau Theater Die Tonne, Reutlingen - Bühnenb...       11
     id                                                 title  cluster
41   41  Sportpark Freiham, Neubau; Heizungsanlagen; Ve...       12
54   54  Sportpark Freiham, Neubau; Metallbauarbeiten, ...       12
```

Technical References

https://sanjayasubedi.com.np/nlp/nlp-with-python-document-clustering/
https://www.linkedin.com/pulse/nlp-text-analytics-simplified-document-clustering
-parsa-ghaffari/

https://blog.eduonix.com/artificial-intelligence/clustering-similar-sentences-together-using-machine-learning/