Text document clustering (Json, Yaml, XML)

Dmitry Parshin

May 2024

Abstract

This project was prepared as a training project after completing a course on NLP.

Project code here: `https://github.com/ink-shtil/nlp_edu_project`.

1 Introduction

Often in large informational-systems there are data channels that require an estimate of the number of types of transmitted documents. This project aims to identify the types of documents passing through the data channel.

1.1 Team

This project was created by one person - Dmitry Parshin. **Dmitry Parshin** prepared this document.

2 Related Work

Skipped.

3 Model Description

Baseline model for this project is Doc2Vec that allows you to get a perfect combination of high performance and minimum system requirements for such a task.

4 Dataset

Data set is a randomly generated text-based documents of differ types: json, xml, yaml. You can choose the number of different documents type and degree of mixin data.

class	parameters
Docs numer	9 * 100 = 900
Doc2Vec	VecSize=3, epochs=100, window=4, $\min_count = 2$
DBSCAN	eps=0.15, $\min_samples = 3$
BIRCH	threshold=0.25

Table 1: Main parameters of the model

5 Experiments

The model should cluster documents with matching number of clusters and types of generated documents

5.1 Metrics

Equality of input parameter = "doctypes" and output clustering groups count.

5.2 Experiment Setup

1. Generating of text documents with some mixins from doc to doc
2. Reading generated docs
3. Extracting docs keys
4. Embedding group of doc keys as tagged data for Doc2Vec model
5. Train Doc2Vec model
6. Clustering with DBSCAN method
7. Clustering with BIRCH method
8. Compare results with generating docs parameters

5.3 Baselines

The key of the experiment is to find best parameters for perfect clustering.

6 Results

Here the results of modeling

```
# "types_count" - count of different txt docs (further clustering should guess this number)
# "docs_per_type" - number of docs for each type (points per cluster)
# "mixin" - degree of mixin "keys" for each document
docs_keys_arr = generate_random_docs(types_count = 9, docs_pre_type = 100, mixin = 1)
```

Figure 1: Input parameters of generated docs

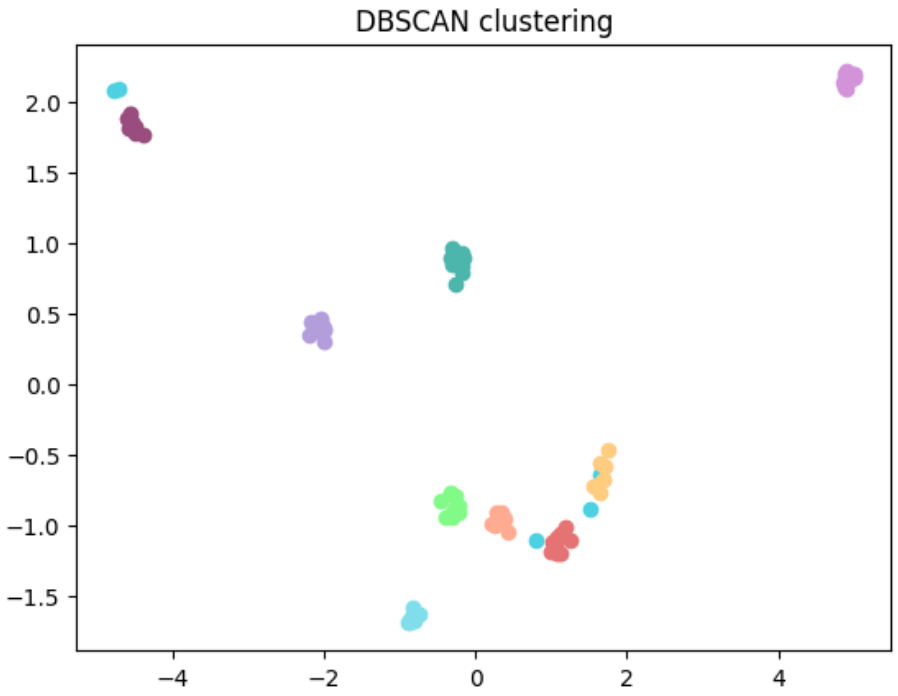

Figure 2: DBSCAN clustering

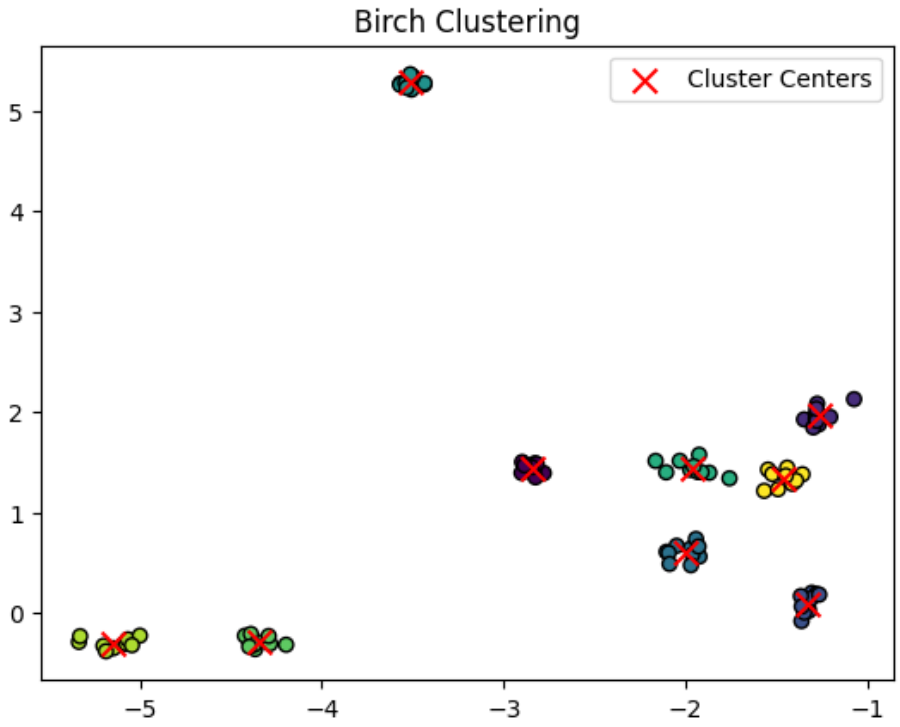

Figure 3: BIRCH clustering

7 Conclusion

As part of the work on the project, the author has generally solved the problem of clustering machine text formats xml, json, yaml. However, the solution has its limitations, in particular, the clustering accuracy decreases when 30 types of documents are reached.