Blueprint POC AI

Introduction

This project lays out the architecture of the ML2Grow AI project. This project was instantiated in order to leverage the power of AI in current and future LBLOD projects. This blueprint was meant to lay bare this architecture to make potential improvements apparent. [...]

Models

The project uses multiple models throughout the project to achieve various functionalities. These models and their functionalities will be listed out here. [...]

NER - Named Entity Recognition

NER is used to recognize various entities within sentences. For example, using the sentence "I love Berlin.", the model will recognize that Berlin is most likely a location. This can be useful for tagging entities within sentences. So these can be used for linking to related topics and such. [...]

BERTopic

BERTopic is used for identifying topics within a certain article. For example, [...].

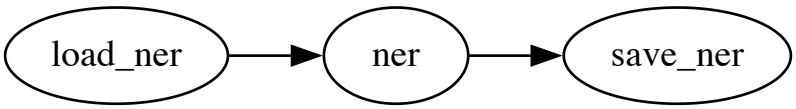
Apache Airflow architecture

Apache Airflow is a framework used to deploy Big Data Networks that can be used amongst multiple projects. There are certain containers that contain various scripts that can be run from CLI. These scripts perform various important tasks such as saving and loading models or data. [...]

DAGs - Directed Acyclic Graphs

DAGs [...] DAGS configure the containers that run the various tasks of the project.

NER



Link to the repository containing the scripts During these tasks. We first load the NER model. Then we perform NER related stuff.

1. Load
   During this script we load the data from the triplestore and export it to a json file. This file can then later be used to perform transformations. This script loads following query

```
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX dct: <http://purl.org/dc/terms/>
```

```
    PREFIX soic: <http://rdfs.org/sioc/ns#>
    PREFIX ext: <http://mu.semte.ch/vocabularies/ext/>


    SELECT DISTINCT ?thing ?text WHERE {
        ?thing a
<http://rdf.myexperiment.org/ontologies/base/Submission>;
prov:generated/dct:hasPart ?part.
        ?part soic:content ?text.
        FILTER NOT EXISTS {  ?thing
ext:ingestedml2GrowSmartRegulationsNer "1" }
    }
```

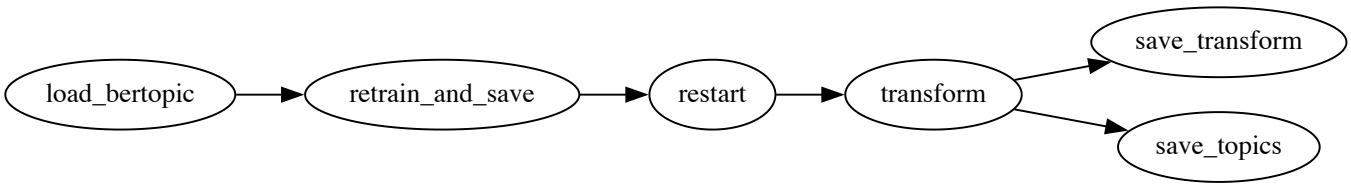
2. NER Afterwards we use the data to run the NER model on. The data gets processed and the results are written to a json file on disk.

3. Save Finally the results are persisted in the triplestore. The results are saved in the triplestore like this.

Predicate	Description
URI	Uri of the subject. Including an UUID
Type	A constant value being ext:Ner. Representing the type of the subject.
start	The start position of the word. Relative to [...].
end	The end position of the word. Relative to [...].
word	The word that was guessed on by the AI model.
entity	Can be either "Location", "Person" or "Organization". This is the value guessed by the AI model.

Additionally we also add a predicate to the file that was used to generate the Ner.

BERTopic Retrain



Link to the repository containing the scripts

1. Load Script that loads the following query. e

2. Retrain & Save

3. Restart

4. Transform

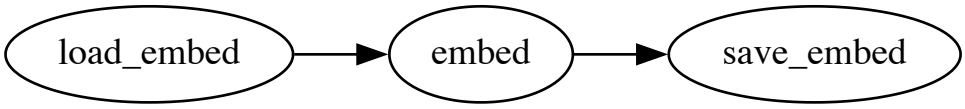
5. Save
    1. Transform
    2. Topics

BERTopic Transform

Link to the repository containing these scripts

1. Load
2. Transform
3. Save

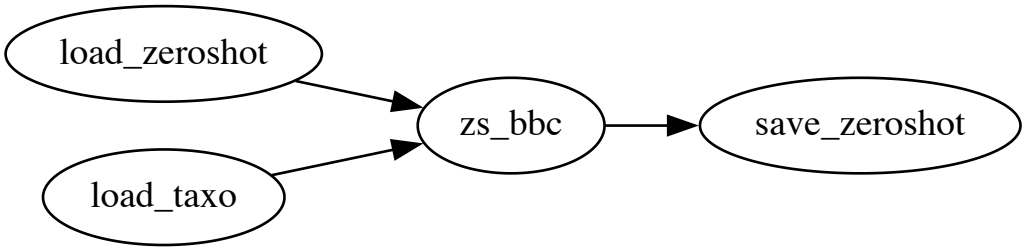
Embed



Link to the repository containing the scripts

1. Load
2. Embed
3. Save

Zeroshot



Link to the repository containing the scripts

1. Load
    1. Zeroshot
    2. Taxo
2. ZS_BBC
3. Save

Data Flow Diagrams

The Data Flow in the project goes a little like this. [...]

Triplestore Mapping

SparQL triplestores are used for persisting most data. Though through Airflow there are also mentions of postgreSQL. The use of PostgreSQL can be an inconvenience for the purpose of linking this project to other LBLOD projects. [...]

Risk Analysis

Conclusion