BASF CHALLENGE

Explanation of software design decisions w.r.t the given requirements

Requirements

- Read XML files.
- Extract year, title and abstract felds from each XML file.
- Persist these data in a database.
- Run a NER lib over the abstract text.
- Persist the NER output in the database.
- Deploy the code in a Kubernetes-based microservices architecture.
- Be callable via GraphQL o REST endpoint.
- Upload files and run extraction pipeline.
- Delete the whole database.
- Using Java with Spring Boot.

Decisions

1. Instead of implementing the reading and extraction of information from XML files from scratch, I have used the jaxb library that allows me to parse XML into objects.
2. As a library for the NER I have chosen Stanford CoreNLP because the documentation seemed easier to me.
3. In order to deploy the microservice and the database in Kubernetes, I use Docker to create the containers.
4. I develop a REST API for the knowledge I already have about it, for the ease of identifying errors through HTTP status codes and for the simplicity of exposing resources through URLs
5. For file uploads I use the resources provided by Spring, specifically I will use the org.springframework.web.multipart class that allows managing file uploads through the web.