BITS Pilani
Pilani Campus

ML BASED SOLICITATION IN FEDERAL TRANSCRIPTS

DILIP PRASAD 2020MT12208

PROJECT DETAILS

- For the given German government federal archive written in Germany, implement NLP based search engine.
- This search performed should be more than a simple Regex based or word to word match.
- So, it has to be a semantic based search understanding the meaning and also get results based on synonyms

Problem statement

- There are multiple web pages in the German federal archive website, written in German language.
- Implementing a search based on word/ sentence could be done with a Regex based match, However, this in turn results only with exact match.
- Fetching details based on synonyms or on the basis of context is lacking

NLP Phases

There are 5 phases on any given NLP projects and the same will be performed here

- Lexical & Morphological analysis
- Syntactic analysis (parsing)
- Semantic analysis
- Discourse integration
- Pragmatic analysis

Data cleanup

Below are the common steps performed to clean or normalize the data

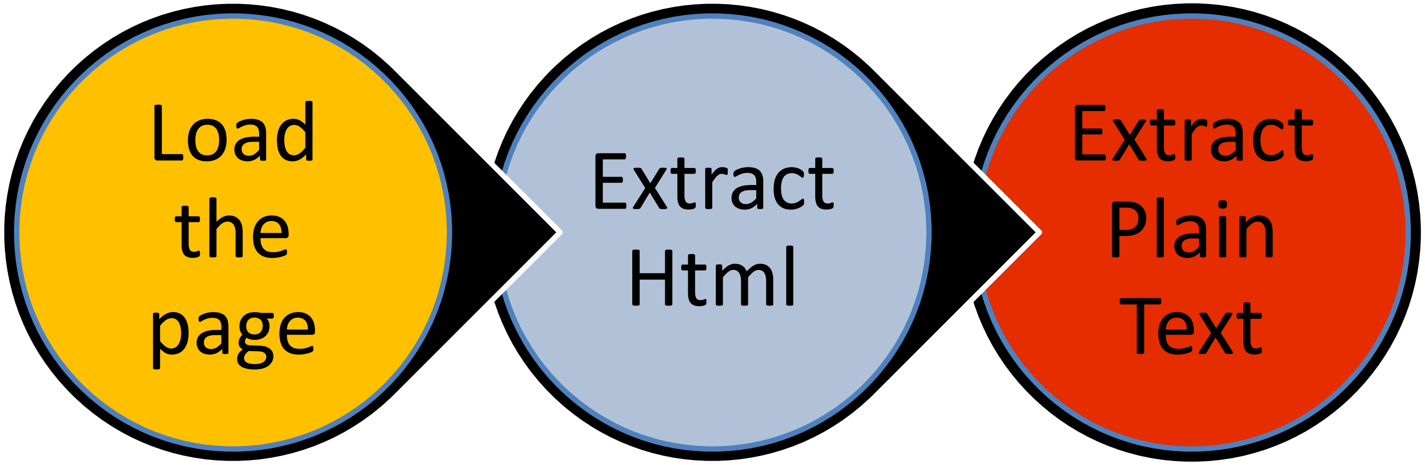
- Tokenizing the string
- Lowercasing
- Removing stop words and punctuation
- Stemming

NLP Tasks

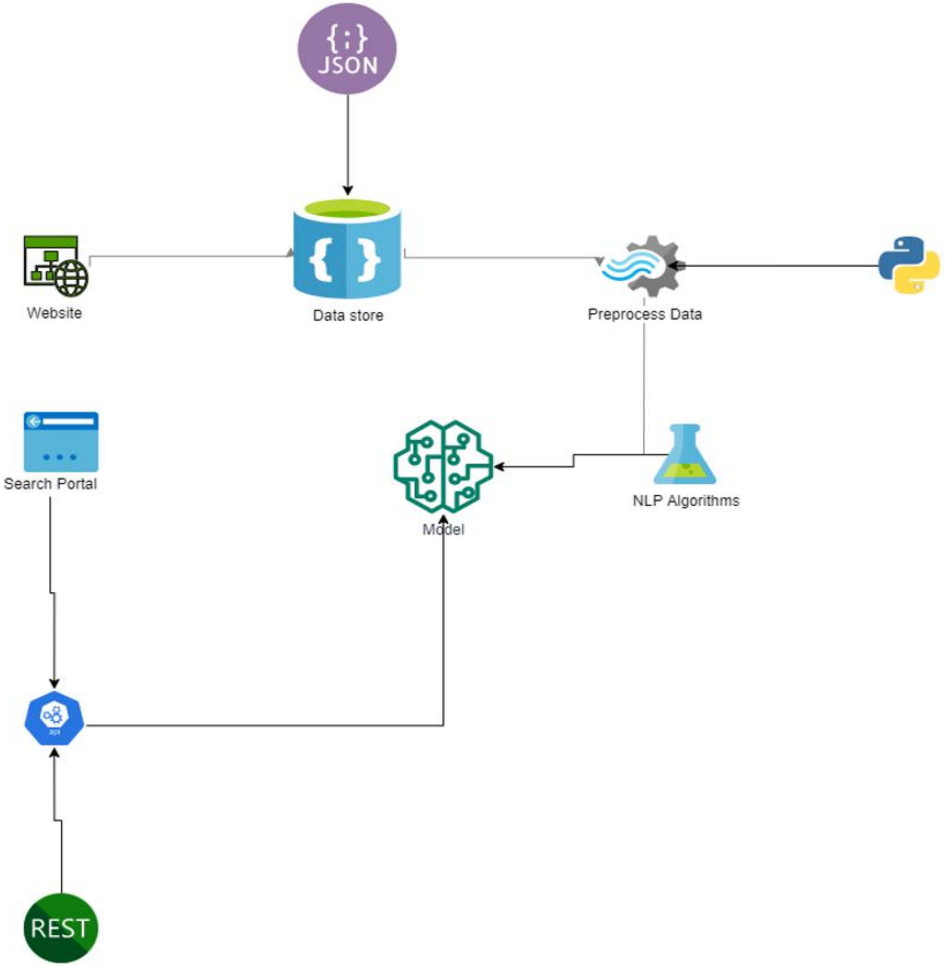
Overview of the tasks performed

- Using a language library
- Building pipeline object
- Using token
- Part of speech tagging
- Understanding token attributes

Web scraping



Architecture Diagram



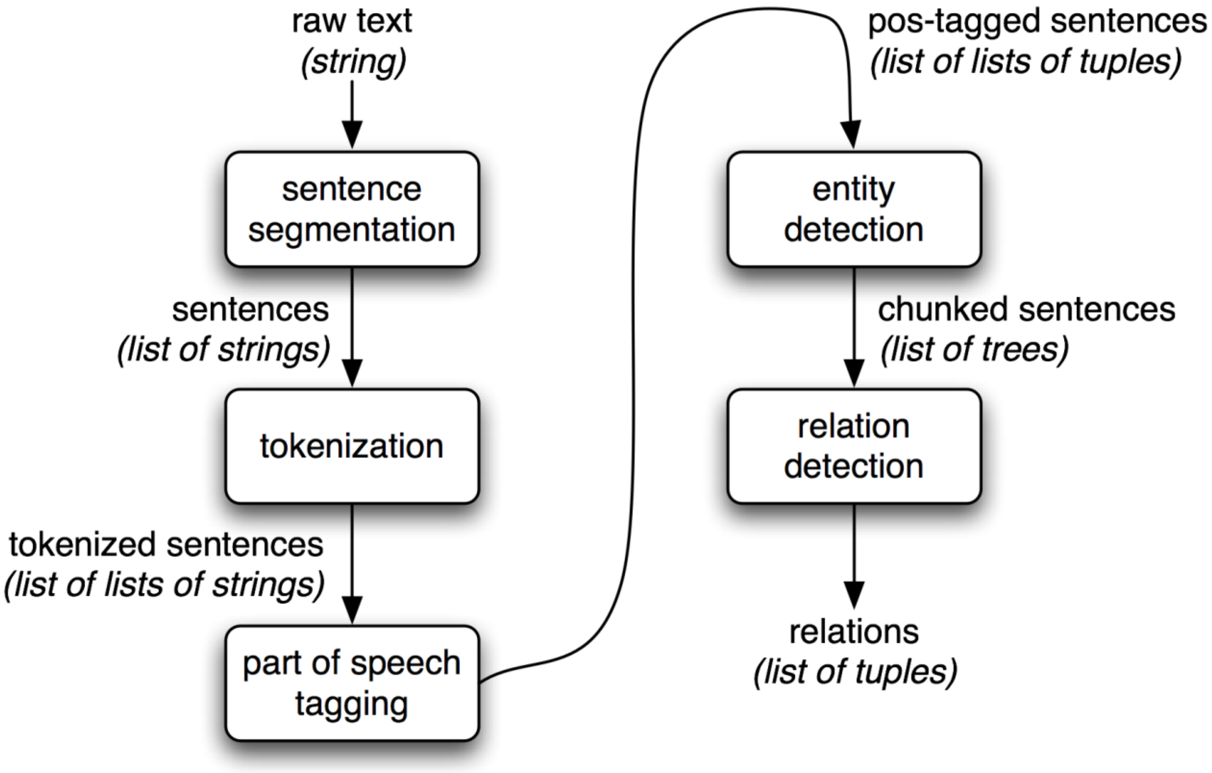
Architecture Explained

- The web crawler module, crawls and finds all the given set of possible links in the entire website.
- Azure Queue functionality is used to maintain the crawled URL temporarily
- Later, the URL Validator module dequeues each of the links validates the same.
- Post that, the same module extracts web text without HTML or Rich media and queues to another queue hosted in azure
- Finally, the NLP modules performs data cleanup and pr processing techniques to build a model

Architecture Explained Cont.

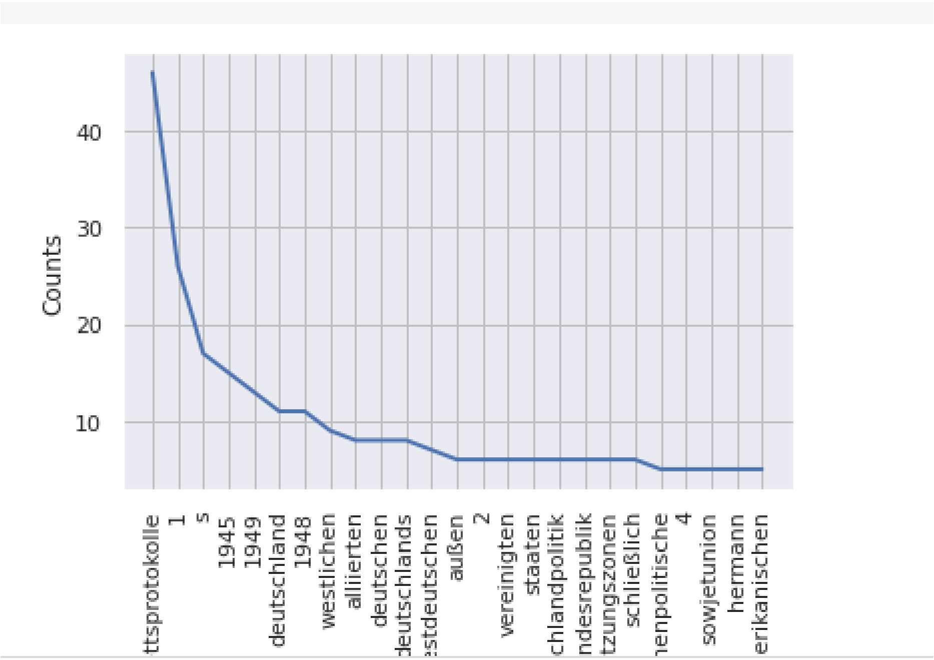
- After creating a module, we make use of it by consuming from an API
- This API is created as a RESTful service to perform REAI operations mostly or CRUD in general.
- To facilitate the search operation, we will be creating a web application in Dot Net core with Docker container.
- Upon performing search, it invokes the API and the API fetches the relevant text from the NLP Model

Info extraction architecture



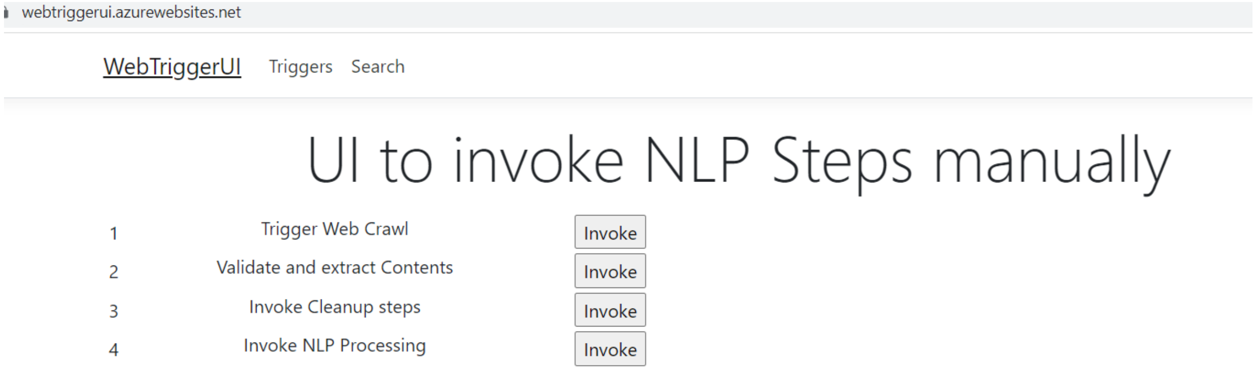
Frequency Distribution

- Sample Frequency distribution



Manual trigger

- Below UI helps us to trigger the various steps in this NL Project



webtriggerui.azurewebsites.net

UI to invoke NLP Steps manually

1 Trigger Web Crawl Invoke

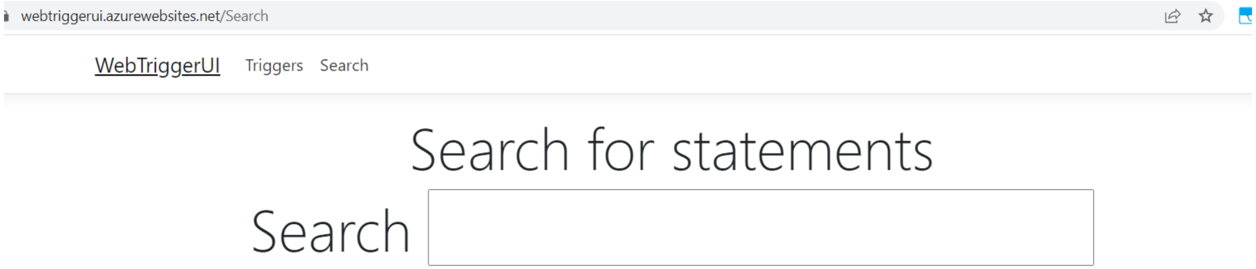
2 Validate and extract Contents Invoke

3 Invoke Cleanup steps Invoke

4 Invoke NLP Processing Invoke

Search Screen

- Below UI helps us to perform search on our model



webtriggerui.azurewebsites.net/Search

WebTriggerUI Triggers Search

Search for statements

Search

Scope for improvement

- Use no server methodologies by making use of azure o demand or AWS functionless offerings
- Communicate between serverless functions to create a cascading experience
- Create a MLOPS pipeline to continuously train and deploy model for enhancements

Conclusion

- Though there are multiple algorithms to handle these scenarios, we have used a simple one available in the market.
- Going forward these algorithms will be replaced with a proprietary algorithm developed in house.

Thank You