Data Anonymization

K.Sai Krishna (IMT2019045)
Divyam Agrawal (IMT2019028)
Prachi Naik (MS2022019)

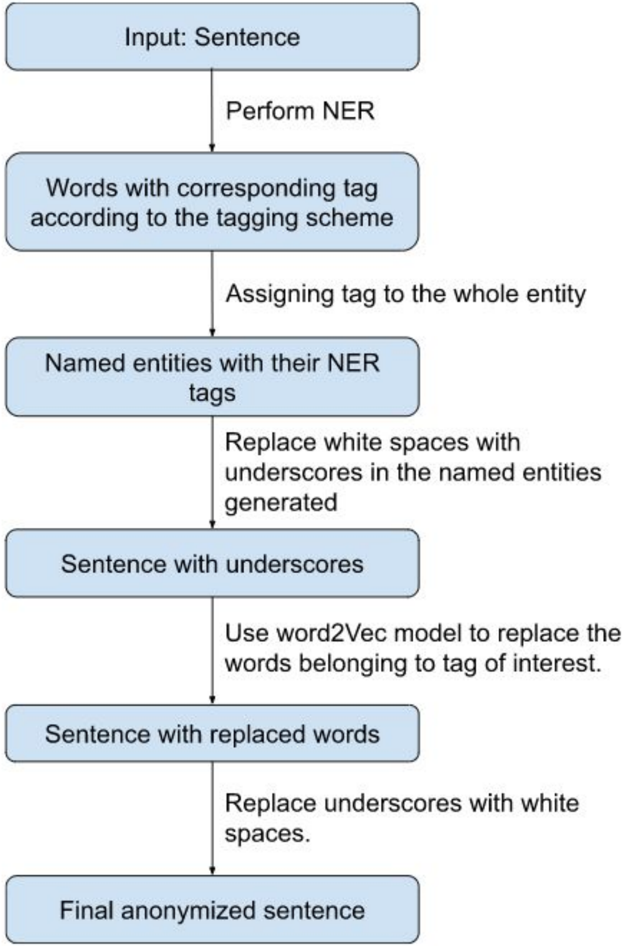
Objective

Create a web application that, after each stage of the data anonymization pipeline, can display the results of the processing that took place. In addition to this, providing the user with the flexibility to choose different models at each stage.

Pre-Mid Term work

- Two pretrained models for NER:
  - Allen NLP
  - SpaCy
- Used Word2Vec model trained on Yago dataset
- Trying to find the replacement model for the data anonymization task.

Flow-Chart



Input: Sentence

Perform NER

Words with corresponding tag according to the tagging scheme

Assigning tag to the whole entity

Named entities with their NER tags

Replace white spaces with underscores in the named entities generated

Sentence with underscores

Use word2Vec model to replace the words belonging to tag of interest.

Sentence with replaced words

Replace underscores with white spaces.

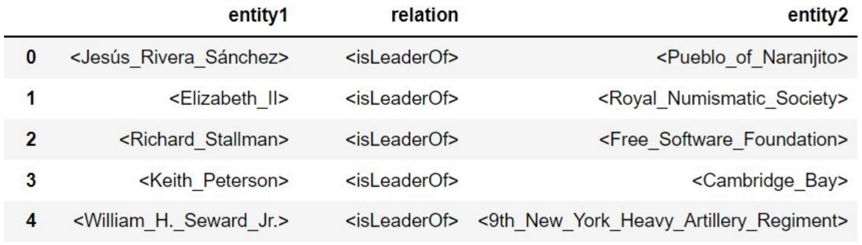
Final anonymized sentence

Code

- Code is divided into 4 parts
    1. Performing NER on the input sentence which returns list of tuples containing (entity, tag)
    2. Replacing white spaces present in entities with underscores and finding the indices of ‘tag of interest’ entities.
    3. Generate most_similar words for the required entity and use look-up table for finding that one replacement word.
    4. Replace with the words found above and then remove underscores from the sentence by replacing them with white space to get final anonymized string.

Replacement Model - 1

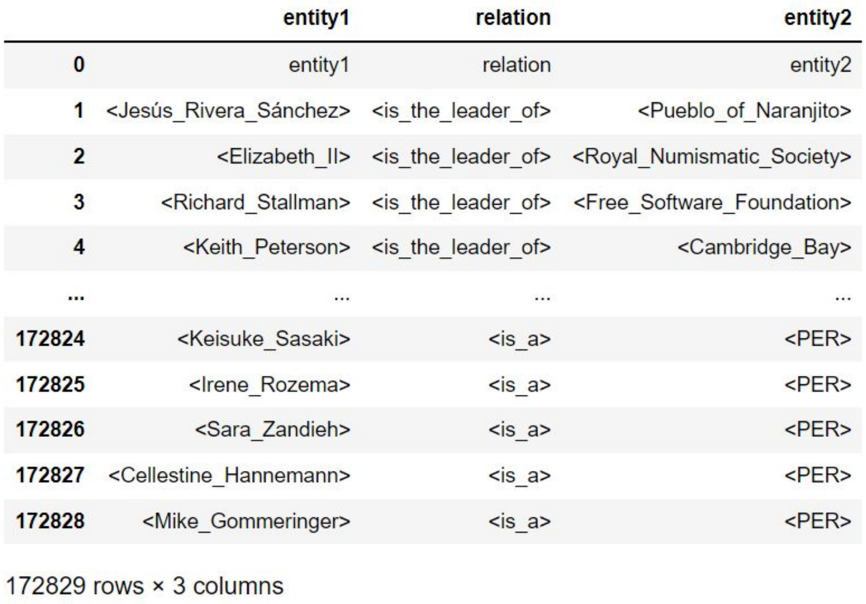
- Trained **Word2Vec** model using the dataset formed by doing ‘join’ on yago_dataset with itself.
- Used **most_similar(word, n)** function to find top-n most similar words of the given word.
- Maintained **look-up** table and chose the replacement word accordingly using the above list.



	entity1	relation	entity2
0	<Jesús_Rivera_Sánchez>	<isLeaderOf>	<Pueblo_of_Naranjito>
1	<Elizabeth_II>	<isLeaderOf>	<Royal_Numismatic_Society>
2	<Richard_Stallman>	<isLeaderOf>	<Free_Software_Foundation>
3	<Keith_Peterson>	<isLeaderOf>	<Cambridge_Bay>
4	<William_H._Seward_Jr.>	<isLeaderOf>	<9th_New_York_Heavy_Artillery_Regiment>

Replacement Model-2

- Created a new dataset by performing NER on yago_dataset.
- Eg: If NER for the word ‘India’ is ‘LOC’, then the list [‘<India>’, ‘<is_a>’, ‘<LOC>’] is added to the original yago_dataset.
- Perform the process done in above model using this newly formed dataframe.

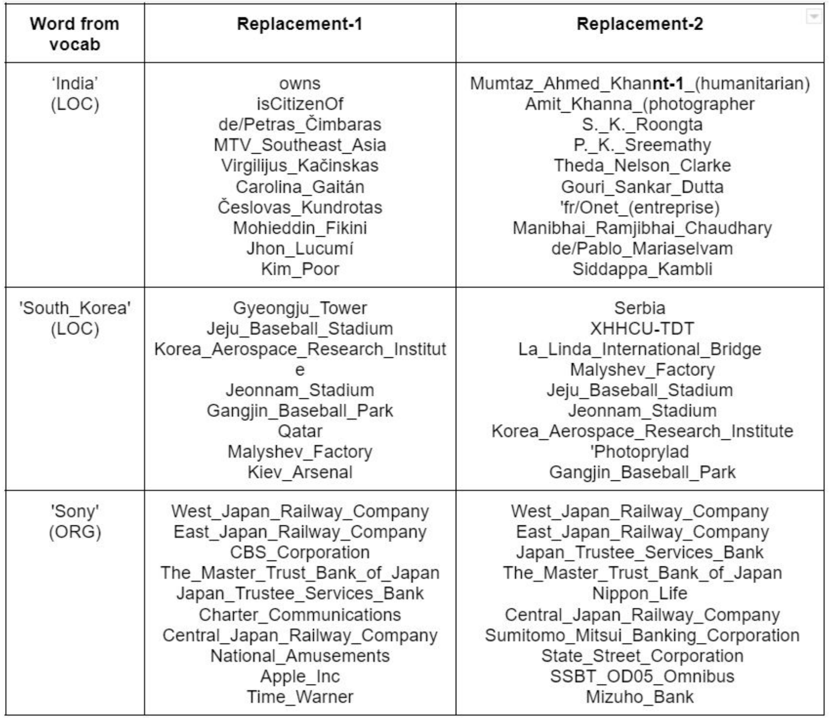


	entity1	relation	entity2
0	entity1	relation	entity2
1	<Jesús_Rivera_Sánchez>	<is_the_leader_of>	<Pueblo_of_Naranjito>
2	<Elizabeth_II>	<is_the_leader_of>	<Royal_Numismatic_Society>
3	<Richard_Stallman>	<is_the_leader_of>	<Free_Software_Foundation>
4	<Keith_Peterson>	<is_the_leader_of>	<Cambridge_Bay>
...	...	...	...
172824	<Keisuke_Sasaki>	<is_a>	<PER>
172825	<Irene_Rozema>	<is_a>	<PER>
172826	<Sara_Zandieh>	<is_a>	<PER>
172827	<Cellestine_Hannemann>	<is_a>	<PER>
172828	<Mike_Gommeringer>	<is_a>	<PER>

172829 rows × 3 columns

Comparison

- In case of ‘South Korea’, we got ‘Serbia’ which also has <’LOC’> tag by using Replacement-2.
- But no improvement observed in ‘India’ and ‘Sony’ entities.
- Might get good results if we train the Word2Vec using whole yaga_dataset which has 132,32,20,606 rows.



Word from vocab	Replacement-1	Replacement-2
‘India’ (LOC)	owns isCitizenOf de/Petras_Čimbaras MTV_Southeast_Asia Virgilijus_Kačinskas Carolina_Gaitán Česlovas_Kundrotas Mohieddin_Fikini Jhon_Lucumí Kim_Poor	Mumtaz_Ahmed_Khan**nt-1**_(humanitarian) Amit_Khanna_(photographer S._K._Roongta P._K._Sreemathy Theda_Nelson_Clarke Gouri_Sankar_Dutta 'fr/Onet_(entreprise) Manibhai_Ramjibhai_Chaudhary de/Pablo_Mariaselvam Siddappa_Kambli
'South_Korea' (LOC)	Gyeongju_Tower Jeju_Baseball_Stadium Korea_Aerospace_Research_Institute Jeonnam_Stadium Gangjin_Baseball_Park Qatar Malyshev_Factory Kiev_Arsenal	Serbia XHHCU-TDT La_Linda_International_Bridge Malyshev_Factory Jeju_Baseball_Stadium Jeonnam_Stadium Korea_Aerospace_Research_Institute 'Photoprylad Gangjin_Baseball_Park
'Sony' (ORG)	West_Japan_Railway_Company East_Japan_Railway_Company CBS_Corporation The_Master_Trust_Bank_of_Japan Japan_Trustee_Services_Bank Charter_Communications Central_Japan_Railway_Company National_Amusements Apple_Inc Time_Warner	West_Japan_Railway_Company East_Japan_Railway_Company Japan_Trustee_Services_Bank The_Master_Trust_Bank_of_Japan Nippon_Life Central_Japan_Railway_Company Sumitomo_Mitsui_Banking_Corporation State_Street_Corporation SSBT_OD05_Omnibus Mizuho_Bank

Future Scope

- Try and enhance the replacement model which presently is using Word2Vec similarity function.
- Some ideas are;
    1. Write a **custom similarity function** which calculates similarity scores among the same tag.
    2. Search for the first occurrence of the word in the most similar words list which has same ‘NER’ tag and replace with that word.

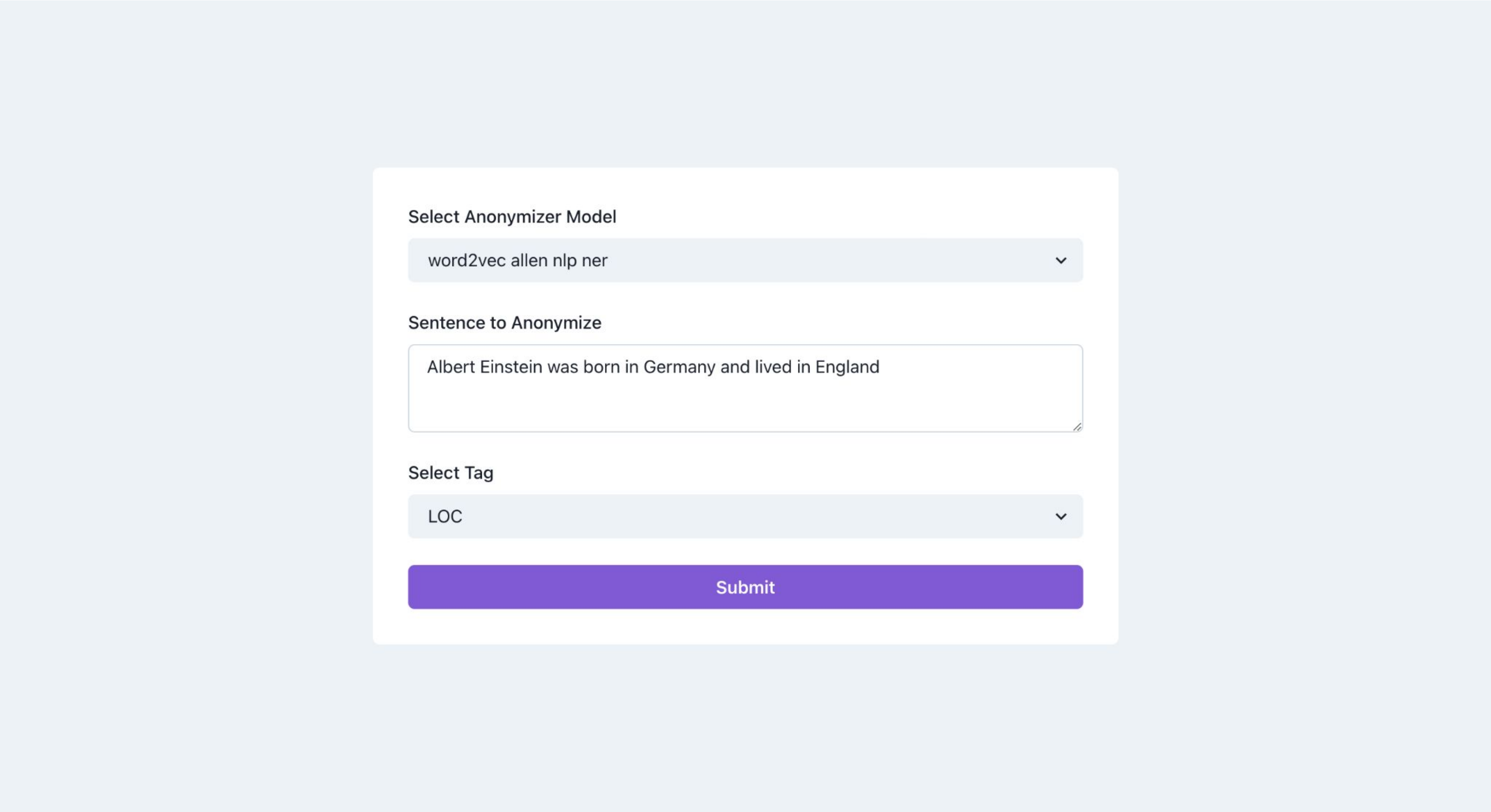
Web Application

Pre-Midterm work

- Created Endpoints for different stages of Pipeline
- Generated template for standardizing ML algorithms to be used with the web application
- Created ML registry to save algorithms.
- Connected NER Models in backend.

Post-Midterm work

- Completed the backend including all the APIs, views and basic testing.
- Created frontend using React for anonymization
- Integrated Allen NLP model and Spacy model and their anonymization part.
- Dynamic fetching of tags of different models.



Select Anonymizer Model

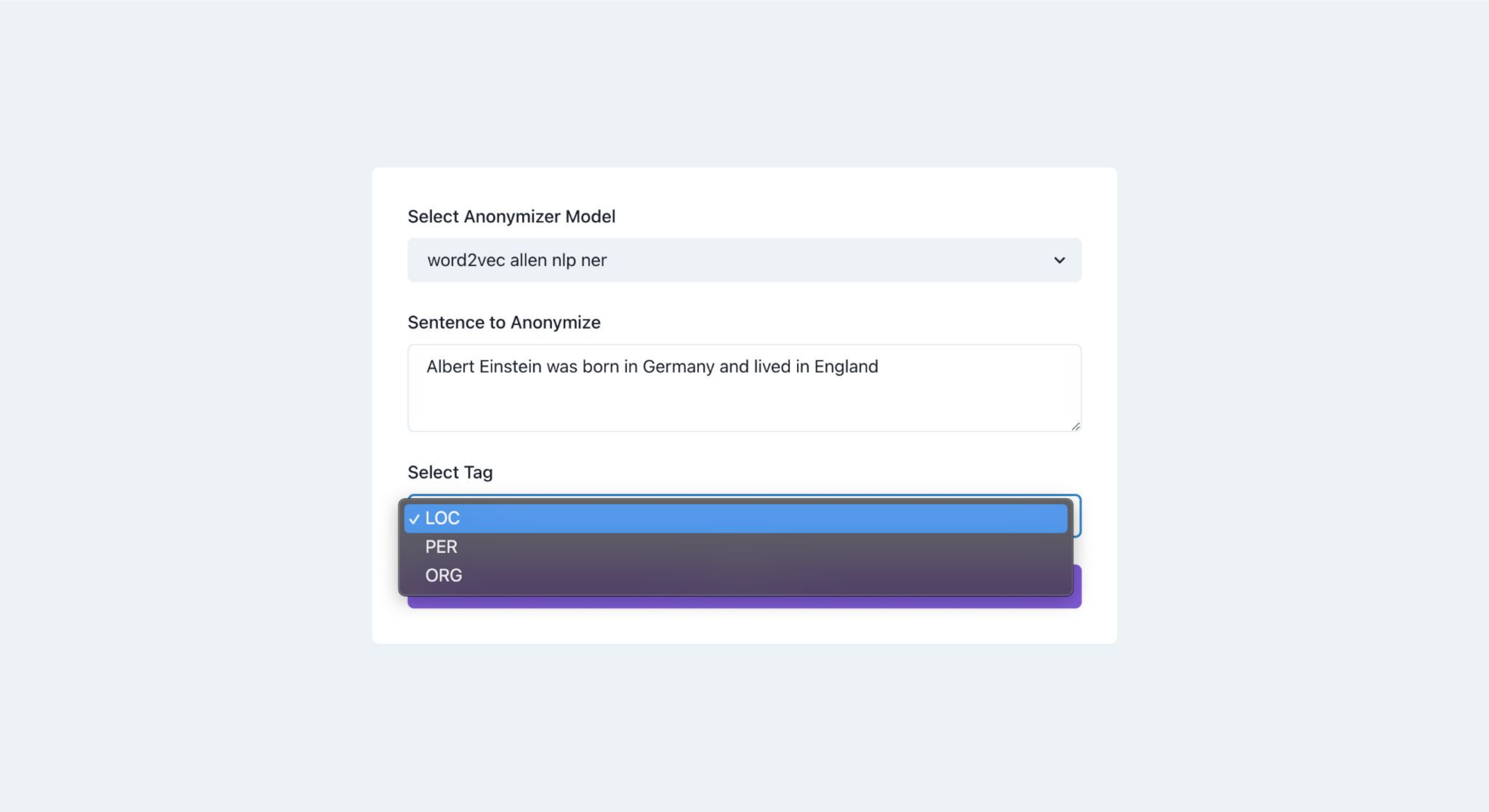
word2vec allen nlp ner

Sentence to Anonymize

Albert Einstein was born in Germany and lived in England

Select Tag

LOC



Select Anonymizer Model

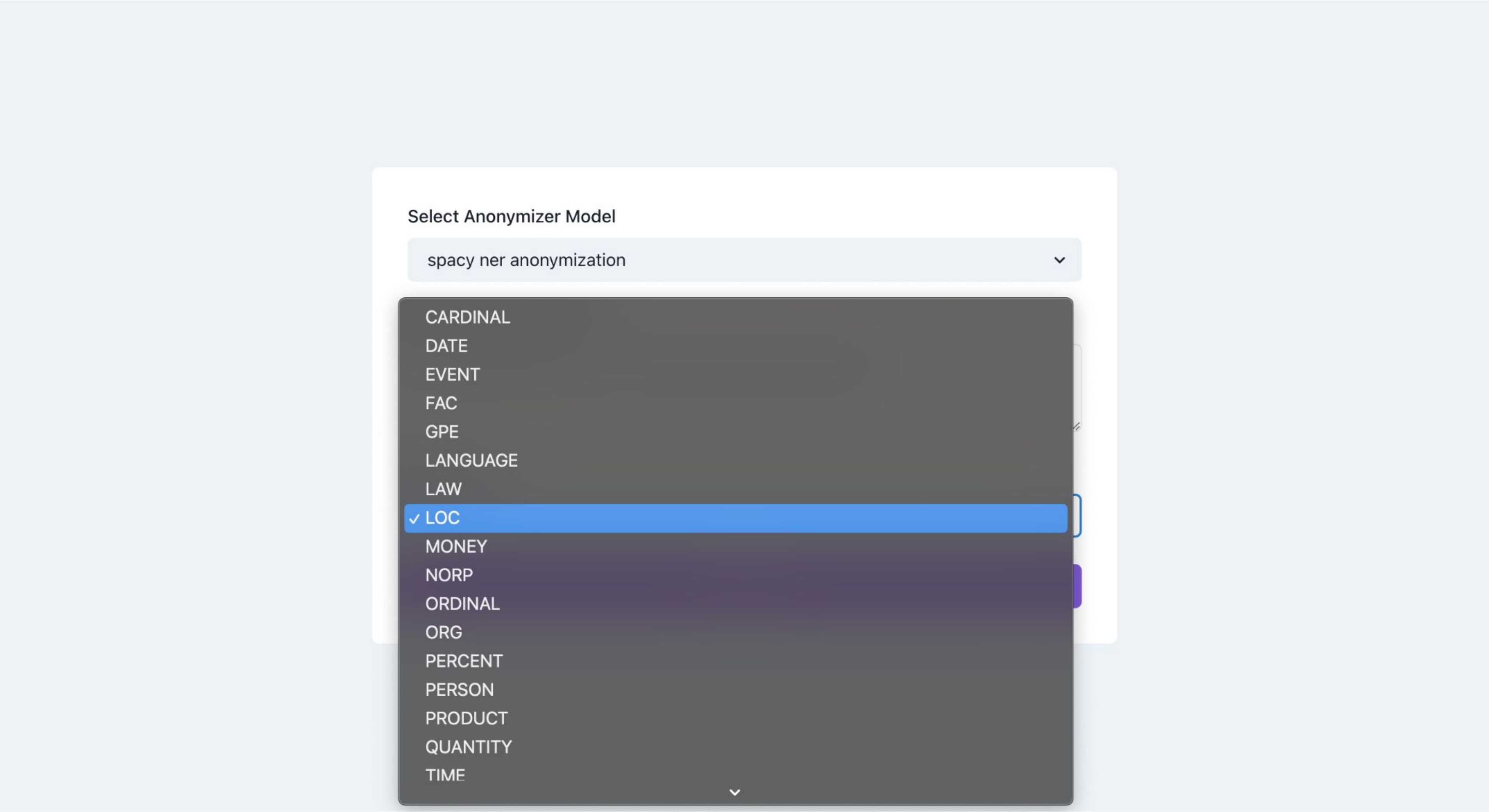
word2vec allen nlp ner

Sentence to Anonymize

Albert Einstein was born in Germany and lived in England

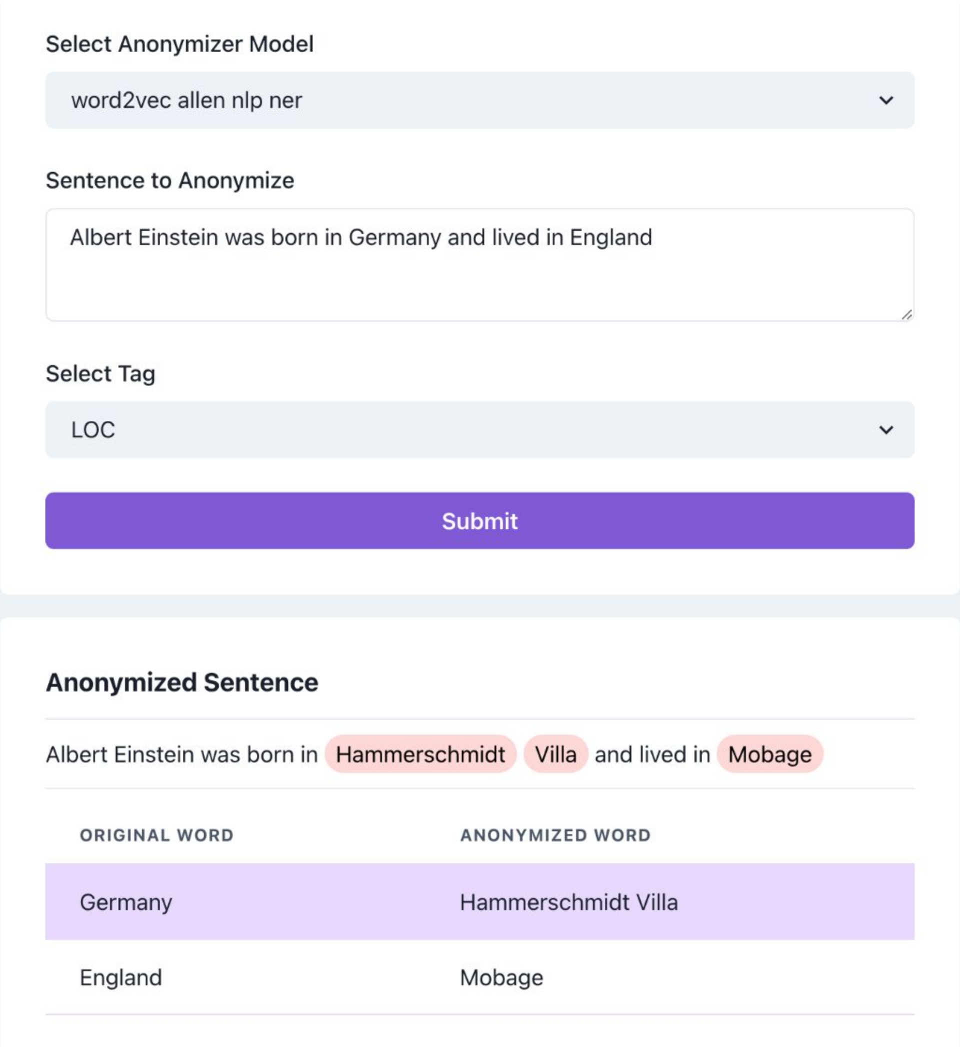
Select Tag

✓ LOC



Select Anonymizer Model

spacy ner anonymization



Select Anonymizer Model

word2vec allen nlp ner

Sentence to Anonymize

Albert Einstein was born in Germany and lived in England

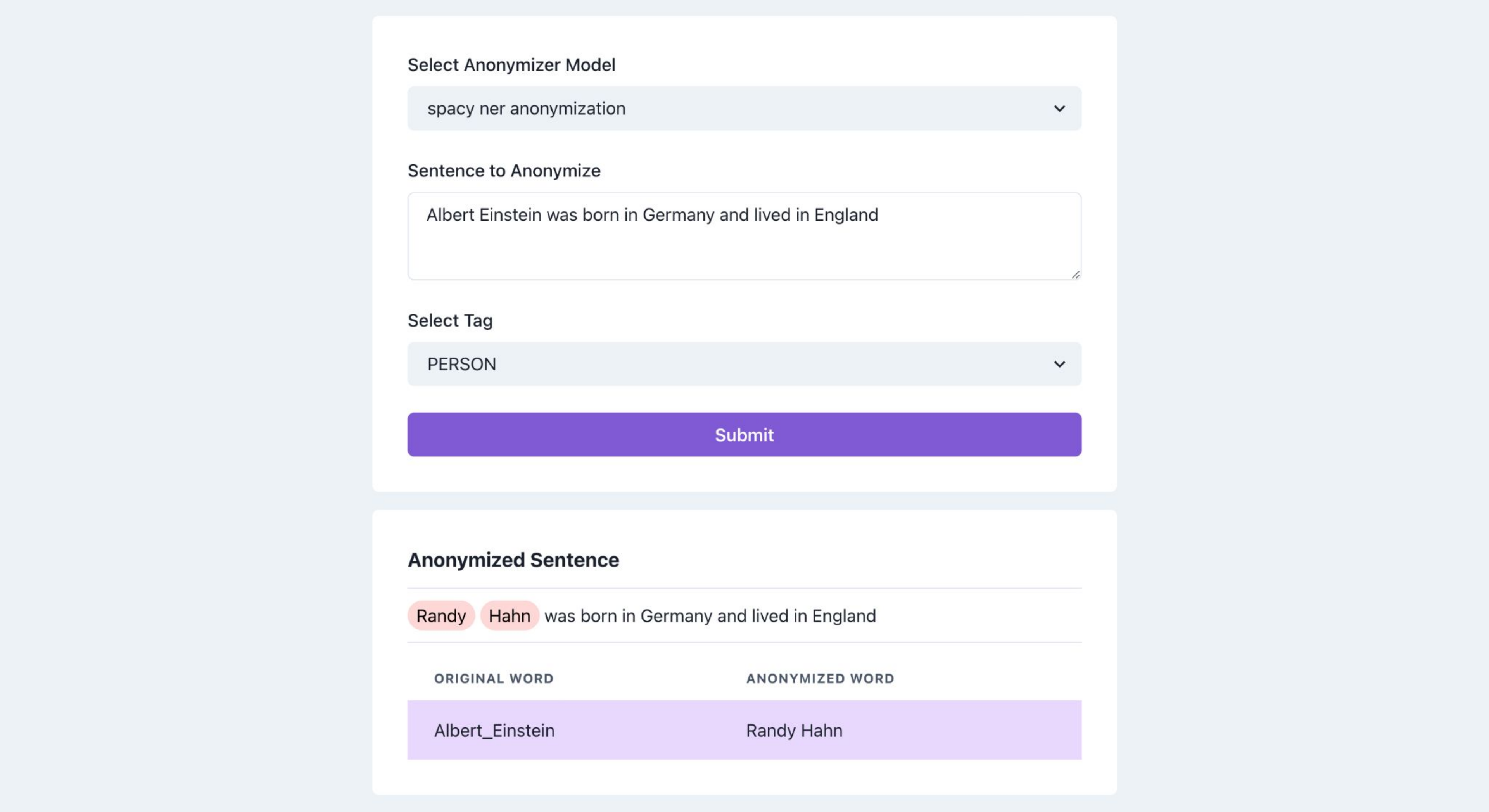
Select Tag

LOC

Anonymized Sentence

Albert Einstein was born in Hammerschmidt Villa and lived in Mobage

ORIGINAL WORD	ANONYMIZED WORD
Germany	Hammerschmidt Villa
England	Mobage



Select Anonymizer Model

spacy ner anonymization

Sentence to Anonymize

Albert Einstein was born in Germany and lived in England

Select Tag

PERSON

Anonymized Sentence

Randy Hahn was born in Germany and lived in England

ORIGINAL WORD	ANONYMIZED WORD
Albert_Einstein	Randy Hahn

Thank You!