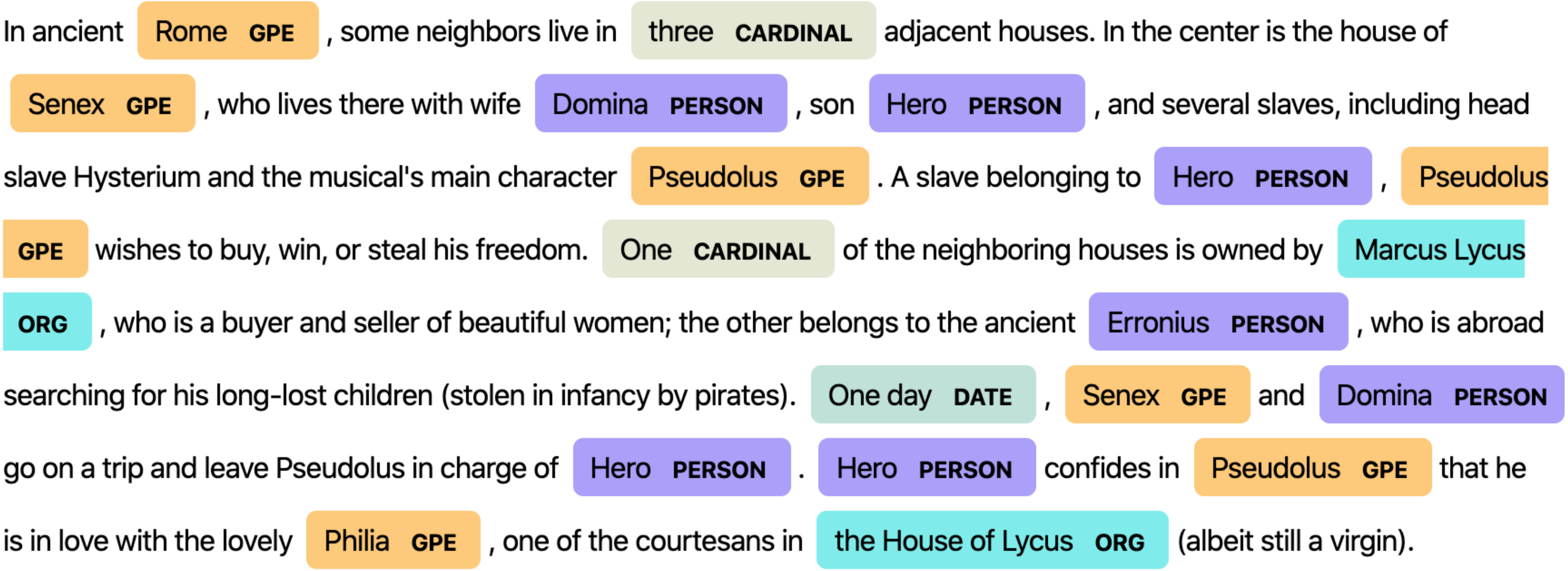
CS 6320: Project 2

Named Entity Recognition

Named Entity Recognition

- Subtask in Information Extraction
- Extract and classify concepts from unstructured text
- Some examples include names of persons, locations, organizations, time mentions, quantities, monetary values, etc.

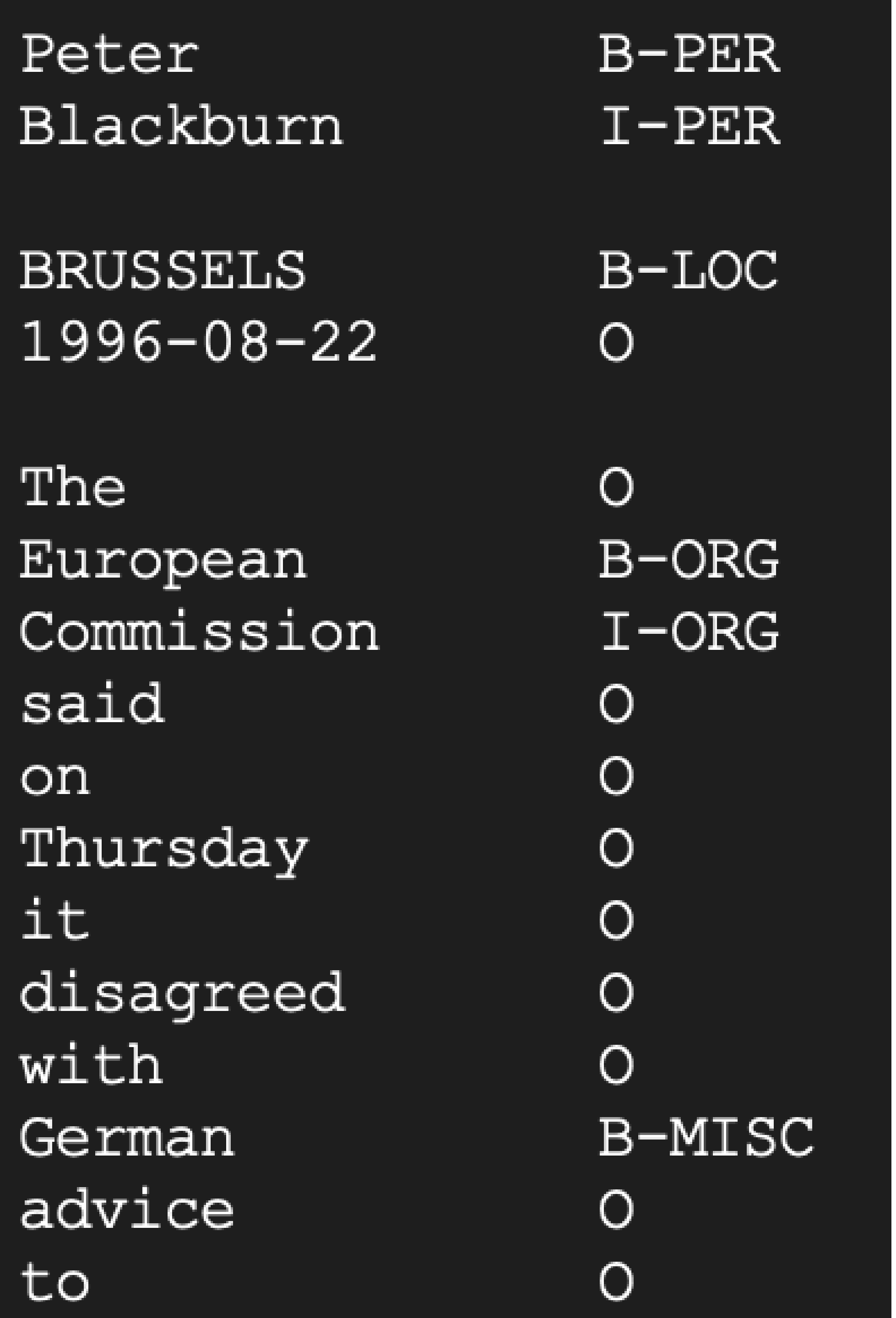
Example



In ancient Rome **GPE** , some neighbors live in three **CARDINAL** adjacent houses. In the center is the house of Senex **GPE** , who lives there with wife Domina **PERSON** , son Hero **PERSON** , and several slaves, including head slave Hysterium and the musical's main character Pseudolus **GPE** . A slave belonging to Hero **PERSON** , Pseudolus **GPE** wishes to buy, win, or steal his freedom. One **CARDINAL** of the neighboring houses is owned by Marcus Lycus **ORG** , who is a buyer and seller of beautiful women; the other belongs to the ancient Erronius **PERSON** , who is abroad searching for his long-lost children (stolen in infancy by pirates). One day **DATE** , Senex **GPE** and Domina **PERSON** go on a trip and leave Pseudolus in charge of Hero **PERSON** . Hero **PERSON** confides in Pseudolus **GPE** that he is in love with the lovely Philia **GPE** , one of the courtesans in the House of Lycus **ORG** (albeit still a virgin).

CoNLL 2003 dataset

- 4 types of concepts: Person, Location, Geo-political entity and Miscellaneous
- CoNLL format
- BIO – Tagging



```
Peter           B-PER
Blackburn       I-PER

BRUSSELS        B-LOC
1996-08-22      O

The             O
European        B-ORG
Commission      I-ORG
said            O
on              O
Thursday        O
it              O
disagreed       O
with            O
German          B-MISC
advice          O
to              O
```

Feature Engineering

- Extract lemmas of all words. The lemma of a word is its root.

Example:

racing -> race   flowers -> flower

unfortunately -> unfortunate

- Get POST for all words. Pass the entire sentence to the method.

Example:

The horse will **race** tomorrow.

Race for outer space

Vocabulary

- Generate a vocabulary of all lemmas from the previous step.
- Add a special UNK token to the vocabulary to handle unseen words seen during testing.
- Represent lemmas and POST as one-hot vectors.

Example:

races/NNS for/IN outer/JJ space/NN

race -> [1 0 0 0]	NNS -> [1 0 0 0]
for -> [0 1 0 0]	IN -> [0 1 0 0]
outer -> [0 0 1 0]	JJ -> [0 0 1 0]
space -> [0 0 0 1]	NN -> [0 0 0 1]

Learning

- Use your favorite machine learning model to predict NER tag for each token.
- Recall Assignment – 2: sklearn provides you with several ML algorithms (Naïve Bayes, Regression, SVMs, Random Forest, etc.)

Example:

For the word ‘races’ from previous example:

Input vector: [1 0 0 0 1 0 0 0]

Output label: O or 0

BIO Tag violations

Token	True label	Predicted label	Post-process# 1
The	O	O	O
University	B-GPE	B-GPE	B-GPE
Of	I-GPE	I-GPE	I-GPE
Texas	I-GPE	I-LOC	I-GPE
At	I-GPE	I-LOC	I-GPE
Dallas	I-GPE	I-LOC	I-GPE

BIO Tag violations

Token	True label	Predicted label	Post-process# 2
The	O	O	O
University	B-GPE	B-GPE	B-LOC
Of	I-GPE	I-GPE	I-LOC
Texas	I-GPE	I-LOC	I-LOC
At	I-GPE	I-LOC	I-LOC
Dallas	I-GPE	I-LOC	I-LOC

Statistics to be reported

- Precision, recall, F-score
- Time taken to make predictions
- Throughput:

size of test CoNLL file / time taken to make predictions