Text Mining
Project/Lab

Behrang QasemiZadeh
behrangatoffice@gmail.com

Information Extraction

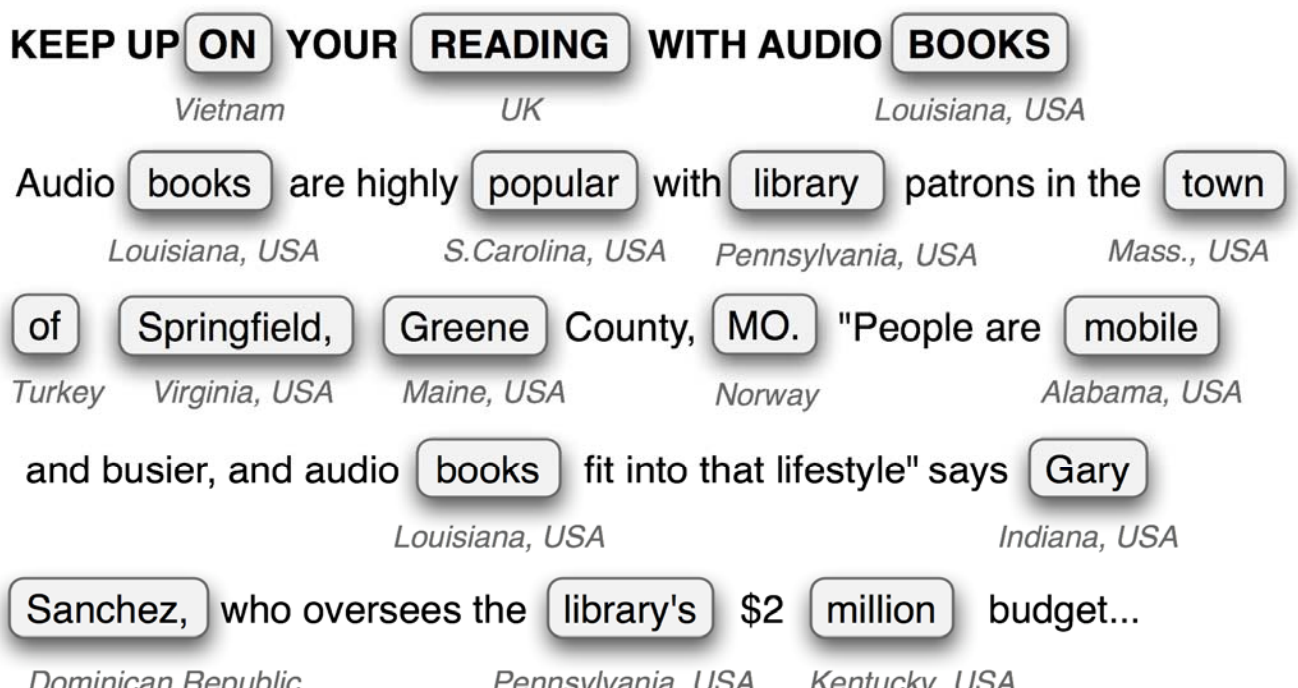
Named Entity Recognition

- Named entities are definite noun phrases that refer to specific types of individuals, such as organizations, persons, dates.
- A **named entity recognition** (NER) system identifies mentions of named entities in text.
- The task is usually done in two steps:
  - First, text boundaries for an NE is identified;
  - Second, the type of NED is recognized.
- Applications:
  - Relation Extraction
  - Enhancing Information Retrieval tasks
  - Question Answering Systems

Named Entity Recognition

- But how to recognize NEs?
  - Using a **gazetteer**?!

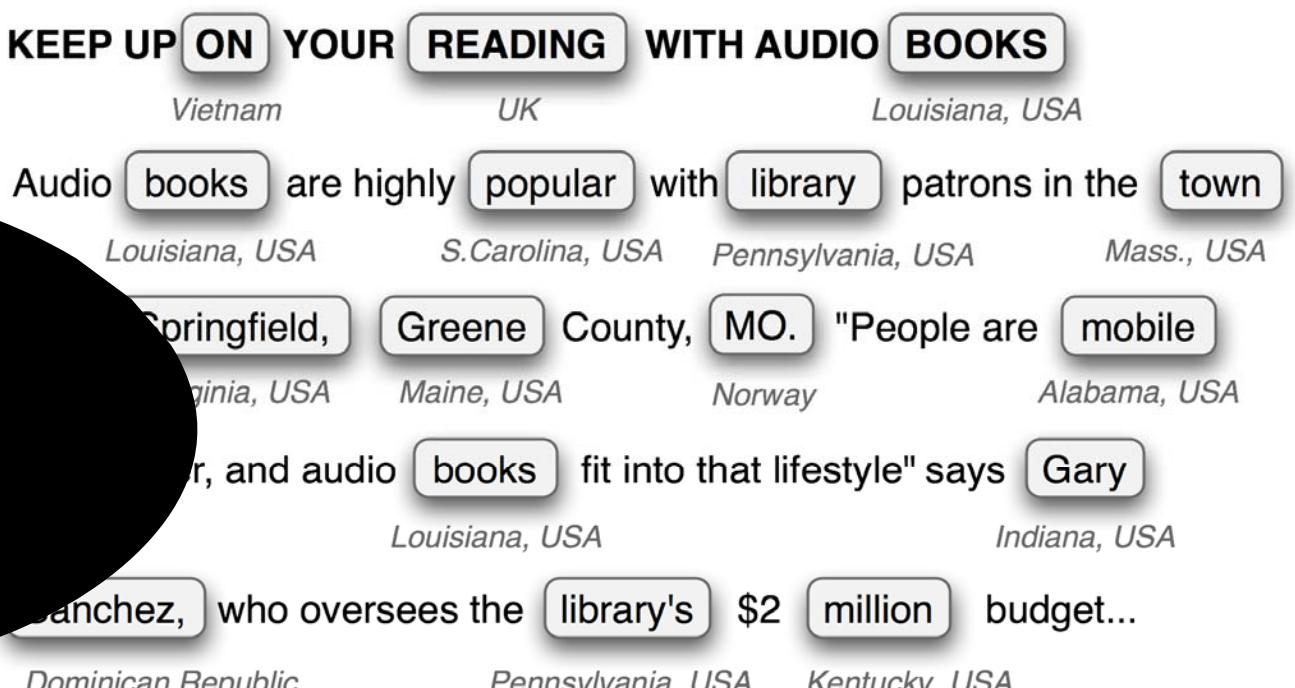
Not a very good idea, ha?!



Named Entity Recognition

- But how to recognize NEs?
  - Using a **gazetteer**?!

What if we want to recognize name of people , organizing, which are dynamically changing?



Named Entity Recognition

- How about a data-driven method:
  - Can we develop a tagger that identify and label chunks with an entity type?
  - Can we use IOB format?
- This was part of the message understanding conference and a number of other evaluation campaigns such as CoNLL:

```
U.N.  NNP   I-NP  I-ORG
official    NN    I-NP  O
Ekeus NNP   I-NP  I-PER
heads VBZ   I-VP  O
for   IN    I-PP  O
Baghdad     NNP   I-NP  I-LOC
.     .     O     O
```

Named Entity Recognition

- We can develop an NER system in a similar way to chunking.
  - Just find the data, perform feature extraction and develop a classifier.
- NLTK comes with a pre-trained NER tagger **`nltk.ne_chunk()`**.

```
>>> print(nltk.ne_chunk(sent))
(S
The/DT
(GPE U.S./NNP)
is/VBZ
...
according/VBG
to/TO
(PERSON Brooke/NNP T./NNP Mossman/NNP)
...
)
```

Named Entity Recognition

- We can develop an NER system in a similar way to chunking.
  - Just find the data, perform feature extraction and develop a classifier.
- NLTK comes with a pre-trained NER tagger **`nltk.ne_chunk()`**.

```
>>> print(nltk.ne_chunk(sent))
(S
The/DT
(GPE U.S./NNP)
is/VBZ
...
according/VBG
to/TO
(PERSON Brooke/NNP T./NNP Mossman/NNP)
...
)
```

You can choose to develop an entity tagger for your project!

Relation Extraction

- Once we have named entities, we may want to identify relationships between them:
  - For example, as used in frame-based knowledge representation systems.
  - In its simple form, given the name of a *Company* and a *Person,* is there a *CEO_OF relationship* between them?

Relation Extraction

- Once we have named entities, we may want to identify relationships between them:
  - For example, as used in frame-based knowledge representation systems.
  - In its simple form, given the name of a *Company* and a *Person,* is there a *CEO_OF relationship* between them?

You can do this for your project too!

Next Session

- If you have not chosen a project title, I will assign you one!
- See the project ideas (http://atmykitchen.info/sites/default/files/documents/proposal.pdf).
- See the instruction there and the break down of the assessment process.
- You perhaps need to read further chapters of the book based on the topic of your project!