NLP For Economists

Information Extraction

Sowmya Vajjala

Munich Graduate School of Economics - LMU Munich

Guest Course, October 2020

Goals

- Knowing some basic information extraction tasks
- ... without having to learn how to build those models themselves

source: Chapter 5 of our book practicalnlp.ai
another useful book chapter:
`https://web.stanford.edu/~jurafsky/slp3/18.pdf`

Information Extraction

- What is IE? - it is the NLP task of extracting relevant information from text documents.
- What is "relevant" information? : names of people, organizations, relations between them, events, addresses, etc.,
- IE applications: amazon's "reviews that mention" feature, blurb that we see if we search for a popular figure's name on Google, populating a calendar event based on email text in gmail, etc.,

.. ...

Information Extraction Tasks

SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

¿¿

- Who is Luca Mestri?

Information Extraction Tasks

SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

¿¿

- Who is Luca Mestri? needs: Named Entity Recognition and Linking, Relation extraction
- What is the article about?

Information Extraction Tasks

SAN FRANCISCO — Shortly after Apple used a new tax law last year to bring back most of the $252 billion it had held abroad, the company said it would buy back $100 billion of its stock.

On Tuesday, Apple announced its plans for another major chunk of the money: It will buy back a further $75 billion in stock.

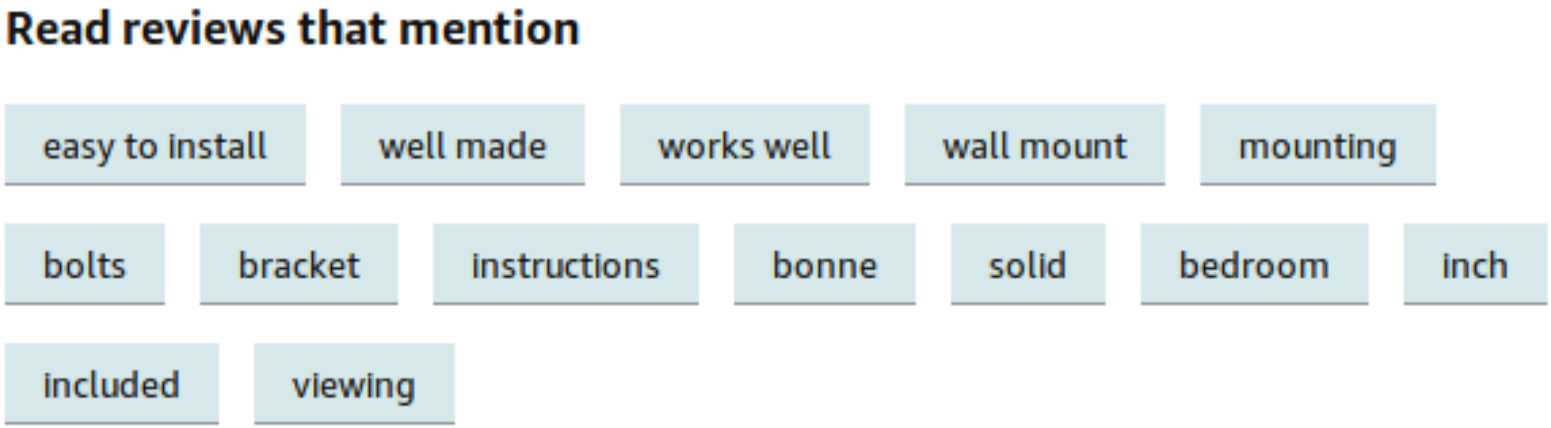
"Our first priority is always looking after the business and making sure we continue to grow and invest," Luca Maestri, Apple's finance chief, said in an interview. "If there is excess cash, then obviously we want to return it to investors."

Apple's record buybacks should be welcome news to shareholders, as the stock price is likely to climb. But the buybacks could also expose the company to more criticism that the tax cuts it received have mostly benefited investors and executives.

¿¿

- Who is Luca Mestri?
  needs: Named Entity Recognition and Linking, Relation extraction
- What is the article about?
  needs: Key phrase extraction, event extraction

Key Phrase Extraction



Named Entity Extraction/Linking

KPE

- Supervised learning approaches require corpora with texts and their respective keyphrases and use engineered features or DL techniques.
- Creating such labeled datasets for KPE is a time- and cost-intensive endeavor.
- Hence, unsupervised approaches that do not require a labeled dataset and are largely domain agnostic are more popular for KPE.
- Recent research has also shown that state-of-the-art DL methods for KPE don't perform any better than unsupervised approaches

KPE Methods

- All the popular unsupervised KPE algorithms are based on the idea of representing the words and phrases in a text as nodes in a weighted graph where the weight indicates the importance of that keyphrase.
- Keyphrases are then identified based on how connected they are with the rest of the graph. The top-N important nodes from the graph are then returned as keyphrases.
- Important nodes are those words and phrases that are frequent enough and also well connected to different parts of the text.
- The different graph-based KPE approaches differ in the way they select potential words/phrases from the text (from a large set of possible words and phrases in the entire text) and the way these words/phrases are scored in the graph.

KPE Implementation: An Example

```
import spacy
import textacy.ke
from textacy import *
#Load a spacy model, which will be used for all further processing.
en = textacy.load_spacy_lang("en_core_web_sm")

#Let us use a sample text file, nlphistory.txt
mytext = open('nlphistory.txt').read()

#convert the text into a spacy document.
doc = textacy.make_spacy_doc(mytext, lang=en)

textacy.ke.textrank(doc, topn=5)

[('successful natural language processing system', 0.02477763226482553),
 ('statistical machine translation system', 0.024691643927525632),
 ('natural language system', 0.020534645561892828),
 ('statistical natural language processing', 0.018614712584248353),
 ('natural language task', 0.015808223689229458)]
```

For more textacy examples:
https://github.com/nishkalavallabhi/practicalnlp/blob/master/Ch5/KPE.ipynb

For textacy documentation: https://chartbeat-labs.github.io/textacy/api_reference.html

NER

- NER refers to the IE task of identifying the entities in a document.
- Entities are typically names of persons, locations, and organizations, and other specialized strings, such as money expressions, dates, products, names/numbers of laws or articles, and so on.
- NER is a prerequisite for being able to do other IE tasks, such as relation extraction or event extraction, question answering etc
- NER is also useful in other applications like machine translation, as names need not necessarily be translated while translating a sentence.

NER Approaches

- Generally dealt as a supervised learning problem, with standard annotated datasets.
- There are several off the shelf tools for general purpose NER.
- However, if we think about what is a named entity, it may mean different things in different disciplines.
- In a new scenario without specialized training data, it is common to:
  - develop a rule based system or
  - "transfer" from a large, existing, general purpose NER or
  - use "active learning" methods, which minimize annotation efforts while maximizing machine learning performance.

NER Example

```
import spacy
nlp = spacy.load("en_core_web_lg")
text_from_fig = "On Tuesday, Apple announced its plans for another major chunk
                 of the money: It will buy back a further $75 billion in stock."
doc = nlp(text_from_fig)
for ent in doc.ents:
    if ent.text:
         print(ent.text, "\t", ent.label_)
```

- Running this code snippet will show Tuesday as DATE, Apple as ORG, and $75 billion as MONEY.

What about the other tasks?

- Relation extraction, named entity linking: best thing to do for you is to look for what is available from providers like IBM.
- There is a free tier with a limit, and there are paid tiers too.
- training own systems - may be not a good idea if you are not into NLP research.
- Event extraction/other forms: first explore regular expressions, and heuristics. (e.g., "PER born in LOC" can be used as a pattern to extract person-birth place relations.)

NEL output from Microsoft Azure service

```
Entities in this document:
San Francisco    Location
https://en.wikipedia.org/wiki/San_Francisco
Facebook         Organization
https://en.wikipedia.org/wiki/Facebook
Alex Jones       Person
https://en.wikipedia.org/wiki/Alex_Jones
InfoWars         Organization
https://en.wikipedia.org/wiki/InfoWars
Louis Farrakhan          Person
https://en.wikipedia.org/wiki/Louis_Farrakhan
Silicon Valley   Location
https://en.wikipedia.org/wiki/Silicon_Valley
Instagram        Organization
https://en.wikipedia.org/wiki/Instagram
us       Location
```

source: https://github.com/practical-nlp/practical-nlp/blob/master/Ch5/06_EntityLinking-AzureTextAnalytics.ipynb

RE output from IBM Watson

```
employedBy
[{'type': 'Person', 'text': 'Nadella'}]
[{'type': 'Organization', 'text': 'Hyderabad Public School', 'disambiguation': {'subtype': ['Commercial']}}]
awardedTo
[{'type': 'Degree', 'text': 'bachelor'}]
[{'type': 'Person', 'text': 'Nadella'}]
educatedAt
[{'type': 'Person', 'text': 'Nadella'}]
[{'type': 'Organization', 'text': 'Manipal Institute of Technology', 'disambiguation': {'subtype': ['Educati
onal']}}]
educatedAt
[{'type': 'Person', 'text': 'Nadella'}]
[{'type': 'Organization', 'text': 'Mangalore University', 'disambiguation': {'subtype': ['Educational']}}]
awardedBy
[{'type': 'Degree', 'text': 'bachelor'}]
[{'type': 'Organization', 'text': 'Manipal Institute of Technology', 'disambiguation': {'subtype': ['Educati
onal']}}]
basedIn
[{'type': 'Organization', 'text': 'Mangalore University', 'disambiguation': {'subtype': ['Educational']}}]
[{'type': 'GeopoliticalEntity', 'text': 'Karnataka'}]
```

source: https://github.com/practical-nlp/practical-nlp/blob/master/Ch5/07_REWatson.ipynb

Concluding notes

- I showed just a few sample IE tasks which are:
  - Easy to use without a lot of background in NLP
  - Potentially useful for any research project you do.
- Being good at regular expressions can be a good starting point for more specific information extraction for a new problem you will encounter in your domain.