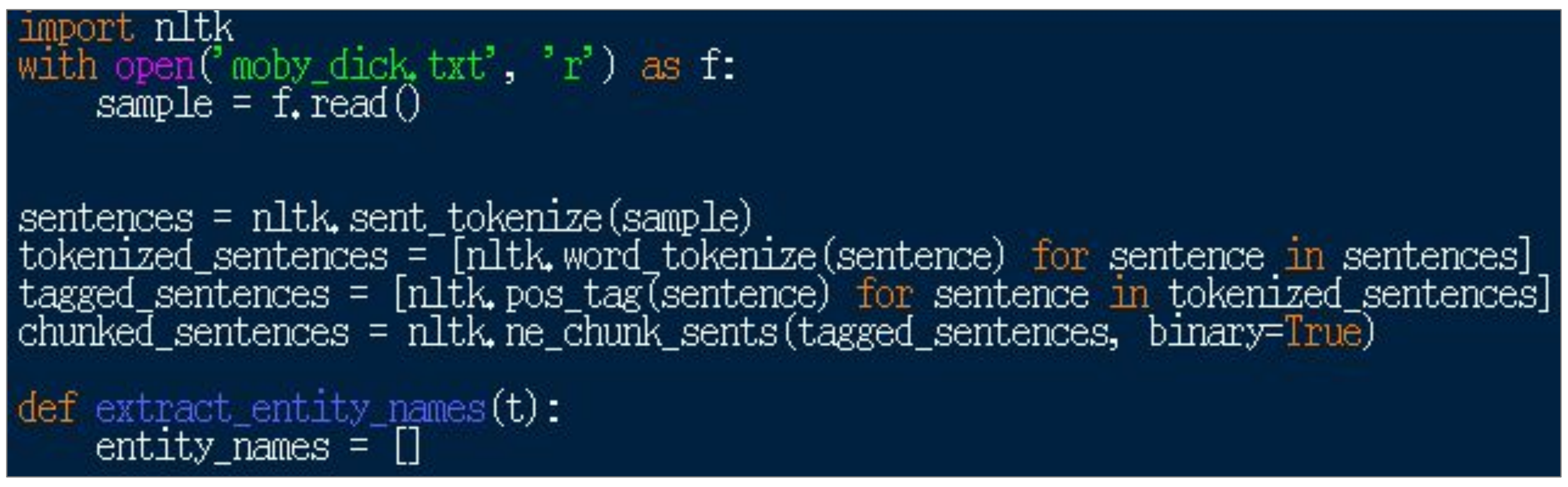
UNIVERSITY *of* WASHINGTON

Enhancing Metadata with Natural Language Processing Tools

Jared Criswell
Sponsor: Charlene Chou, Head of Knowledge Access, NYU

WHAT IS NATURAL LANGUAGE PROCESSING?

Natural language processing (NLP) uses computers to “read” text for analysis and automation purposes. This project shows a few ways info professionals can use NLP tools on items to enhance metadata, allowing for deeper insight with little extra work.



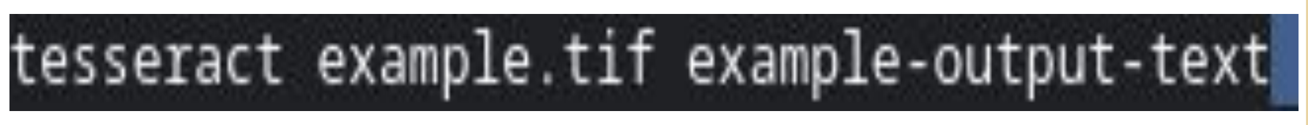
```
import nltk
with open(’moby_dick.txt’, ’r’) as f:
    sample = f.read()


sentences = nltk.sent_tokenize(sample)
tokenized_sentences = [nltk.word_tokenize(sentence) for sentence in sentences]
tagged_sentences = [nltk.pos_tag(sentence) for sentence in tokenized_sentences]
chunked_sentences = nltk.ne_chunk_sents(tagged_sentences, binary=True)

def extract_entity_names(t):
    entity_names = []
```

OCR

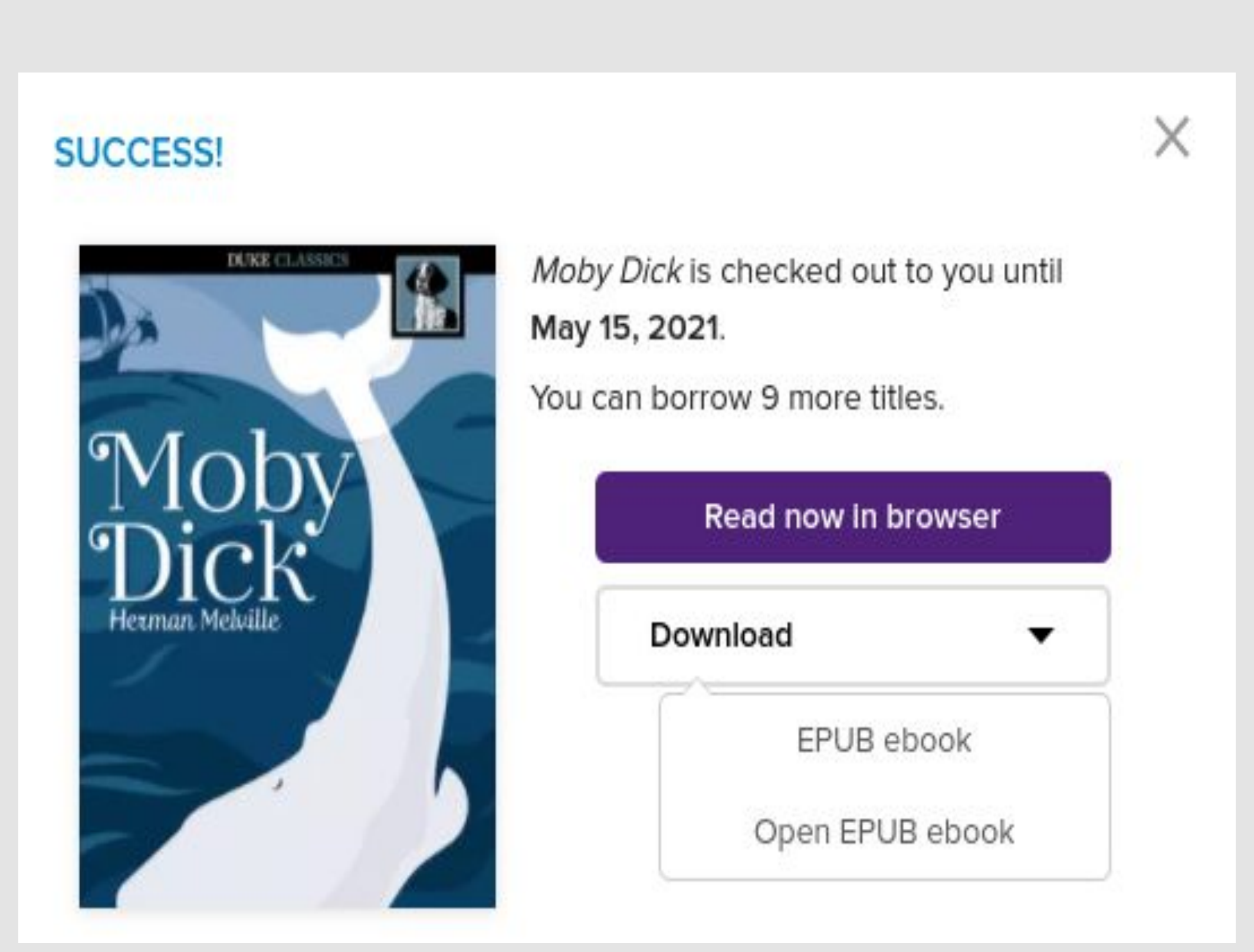
Optical character recognition (OCR) can be used by info professionals to grab text from documents using OCR engines like Tesseract. This text can then be broken down by NLP tools like the Natural Language Toolkit (NLTK).



```
tesseract example.tif example-output-text
```

OPEN EBOOKS

Open eBooks can also easily be downloaded from popular eBook platforms like OverDrive. From there, they can be converted into text files and analyzed.



SUCCESS!

Moby Dick is checked out to you until May 15, 2021.

You can borrow 9 more titles.

Open eBook from OverDrive.

THE PROCESS

> 1: Install Anaconda, a data science platform.

> 2: Download an open eBook and convert it to text with a free online eBook converter, or use optical character recognition (Tesseract) on a high-quality picture for text!

> 3: Use the Natural Language Toolkit (NLTK) to tokenize, tag, and chunk the text.

> 4: Use NLTK to extract named entities from the text for people, places, things, etc. Generally stuff with a capital letter!

> 5: Use NLTK for sentiment analysis on the text to track elements related to emotion!

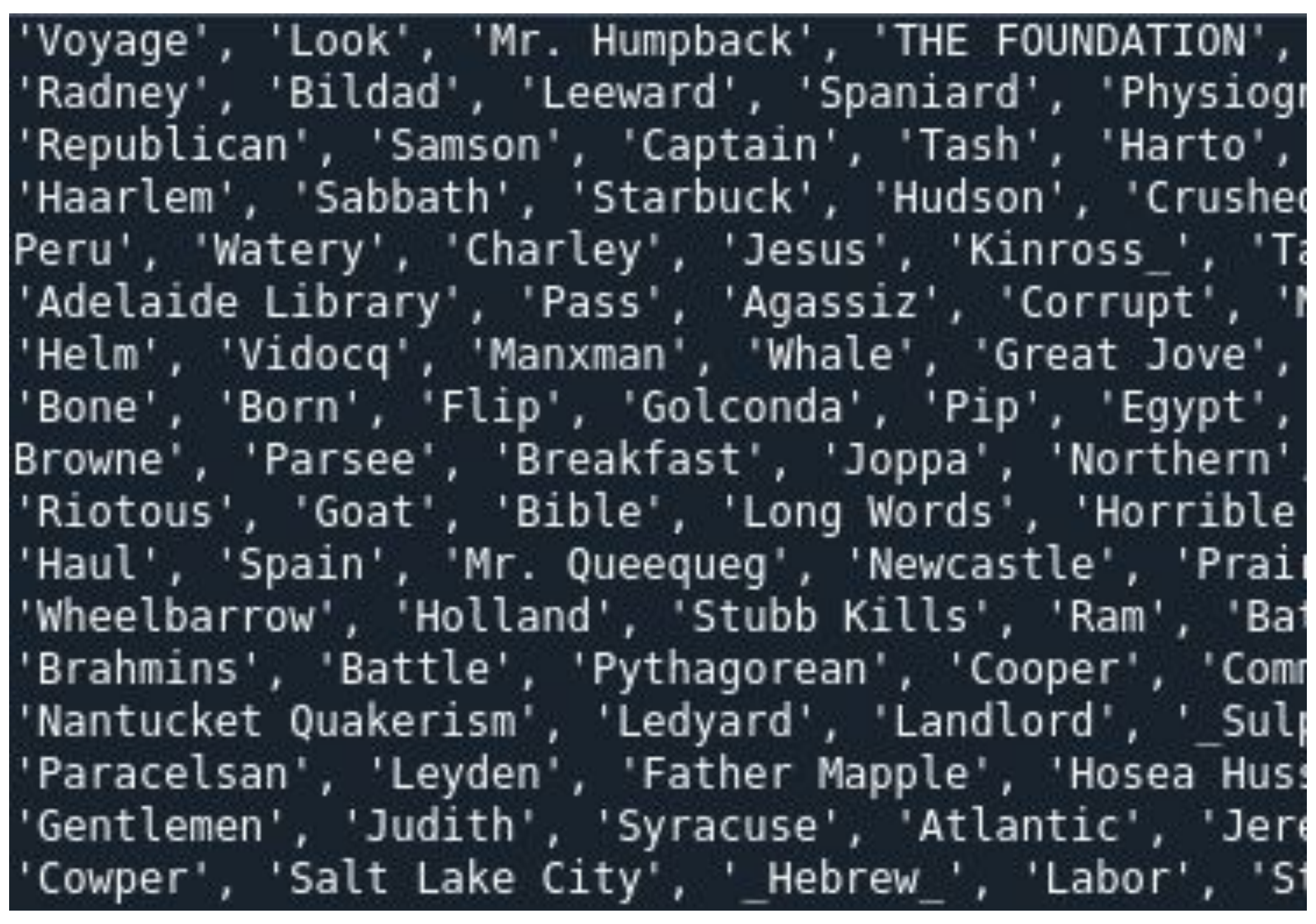
NLTK

Once you have text, process it with NLTK so it can be analyzed! Tokenizing assigns characters a meaning, tagging gives those tokens a part of speech, and chunking makes things like noun phrases out of those tokens and tags.

This phase is very flexible and can use pretrained NLP models like Stanford’s CoreNLP, or the user can even make their own model and test it.

NER

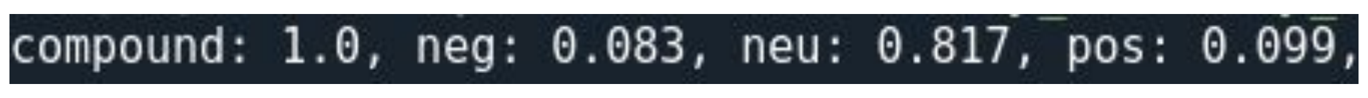
Named entity recognition (NER) can then be used to display unique entities, some of the most important metadata to us human beings!



```
'Voyage', 'Look', 'Mr. Humpback', 'THE FOUNDATION',
'Radney', 'Bildad', 'Leeward', 'Spaniard', 'Physiogr
'Republican', 'Samson', 'Captain', 'Tash', 'Harto',
'Haarlem', 'Sabbath', 'Starbuck', 'Hudson', 'Crushe
Peru', 'Watery', 'Charley', 'Jesus', 'Kinross_', 'T
'Adelaide Library', 'Pass', 'Agassiz', 'Corrupt', '
'Helm', 'Vidocq', 'Manxman', 'Whale', 'Great Jove',
'Bone', 'Born', 'Flip', 'Golconda', 'Pip', 'Egypt',
Browne', 'Parsee', 'Breakfast', 'Joppa', 'Northern'
'Riotous', 'Goat', 'Bible', 'Long Words', 'Horrible
'Haul', 'Spain', 'Mr. Queequeg', 'Newcastle', 'Prai
'Wheelbarrow', 'Holland', 'Stubb Kills', 'Ram', 'Ba
'Brahmins', 'Battle', 'Pythagorean', 'Cooper', 'Com
'Nantucket Quakerism', 'Ledyard', 'Landlord', '_Sul
'Paracelsan', 'Leyden', 'Father Mapple', 'Hosea Hus
'Gentlemen', 'Judith', 'Syracuse', 'Atlantic', 'Jer
'Cowper', 'Salt Lake City', '_Hebrew_', 'Labor', 'S
```

SENTIMENT ANALYSIS

Sentiment analysis, like NLTK’s VADER model, can attempt to analyze how positive, negative, or neutral a text is. What if readers could see that at a glance?



```
compound: 1.0, neg: 0.083, neu: 0.817, pos: 0.099,
```