INTRO TO TEXT ANALYSIS

From Ngrams to NLP

OUTLINE

modeling language

common NLP research tasks

scattertext

Bulk

LEXICAL FEATURES

- Word Frequency
- Density
- Average Words Per Sentence
- Contexts

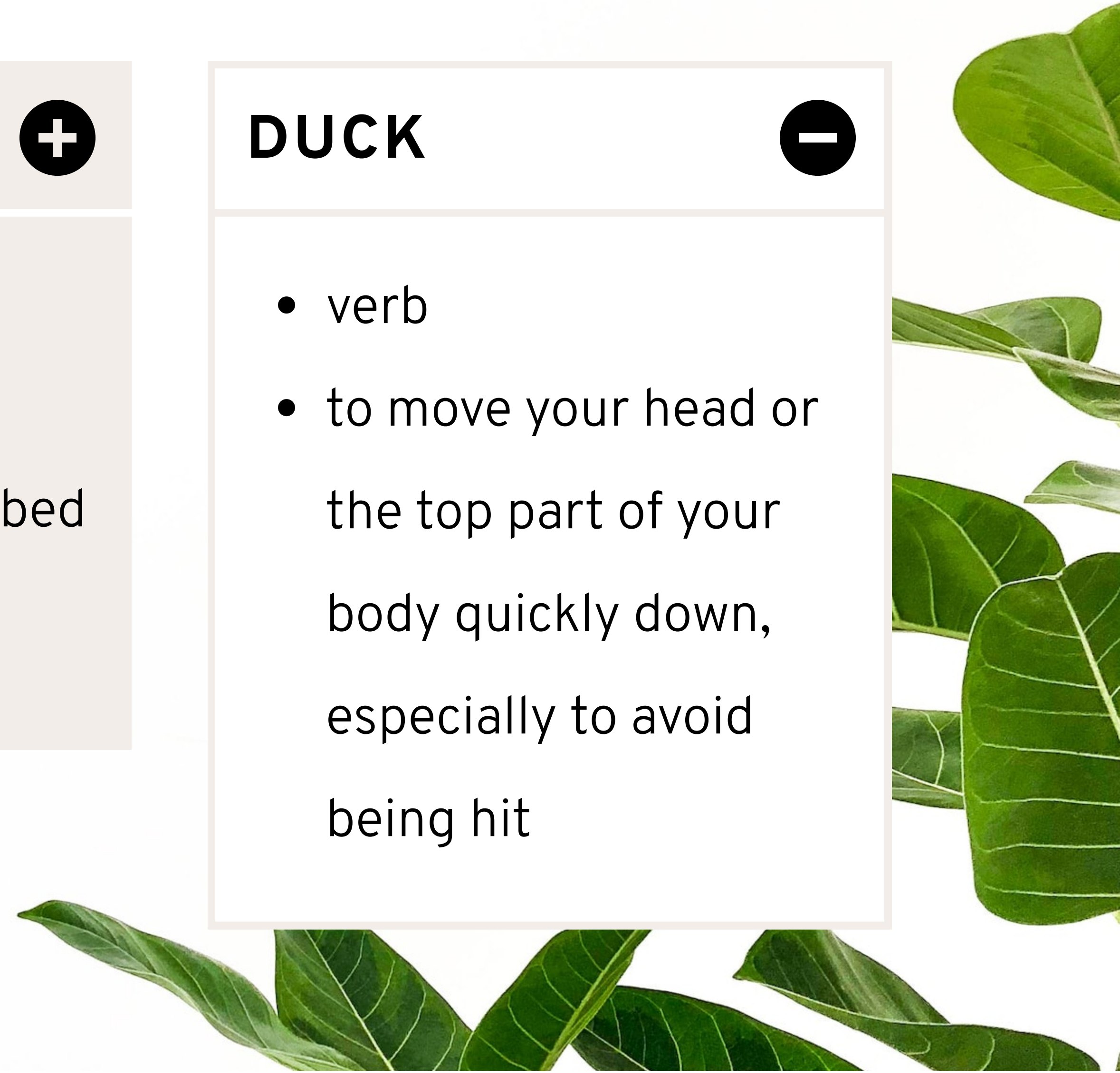
https://books.google.com/ngrams/

DUCK ⊕

- noun
- a bird that lives by water and has webbed feet

DUCK ⊖

- verb
- to move your head or the top part of your body quickly down, especially to avoid being hit



NEGATION

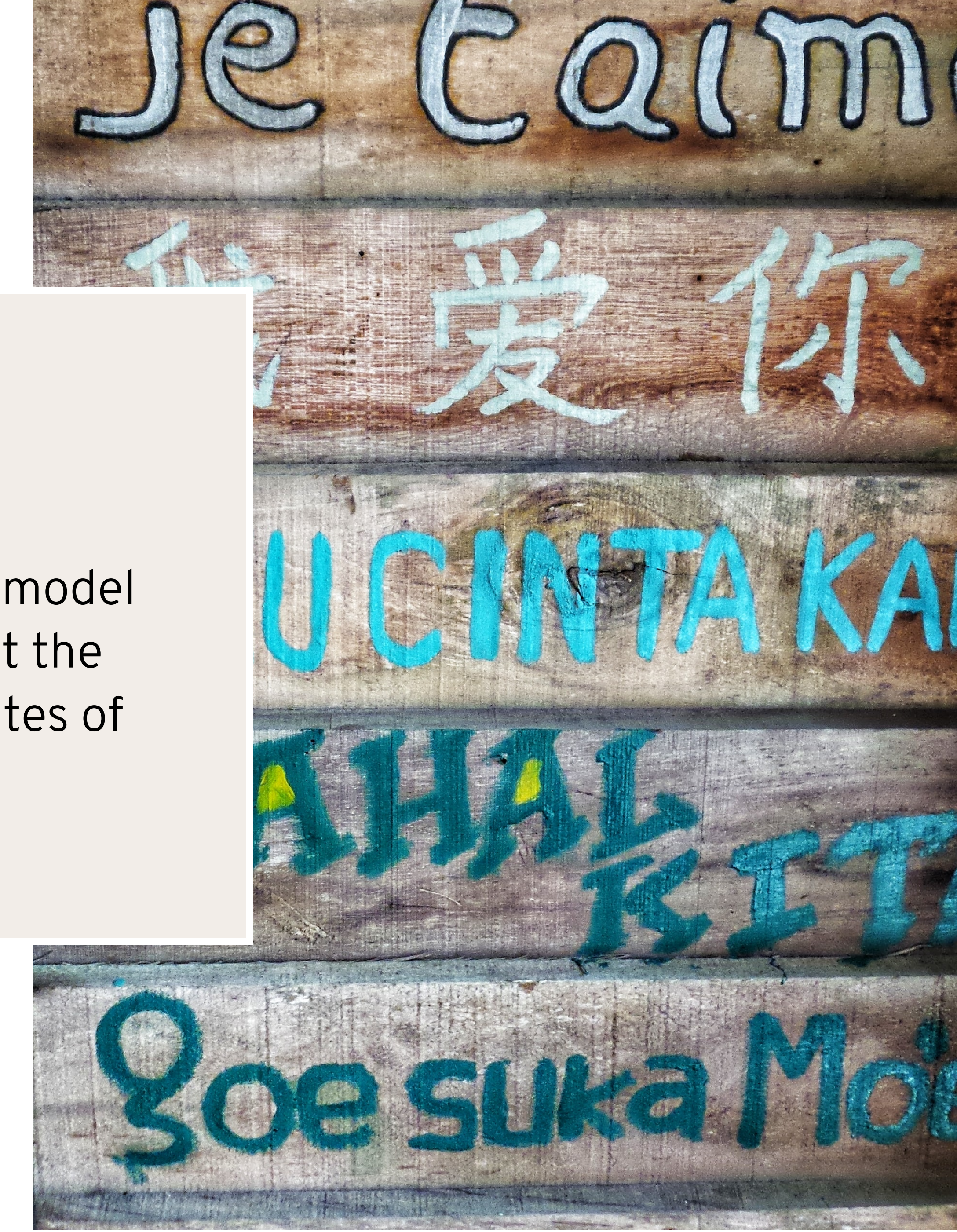
This film isn't about ducks, it's called "Mighty Ducks," and it's about hockey.

CONTEXT

> *“ We shared that blunt with 20 people, and by the time it got to me; All I got was the duck.*
>
> – by theothers August 22, 2012

NATURAL LANGUAGE PROCESSING

NLP uses machine learning to model human language and to predict the linguistic and semantic attributes of text.



https://stanfordnlp.github.io/stanza/

https://course.spacy.io/en/

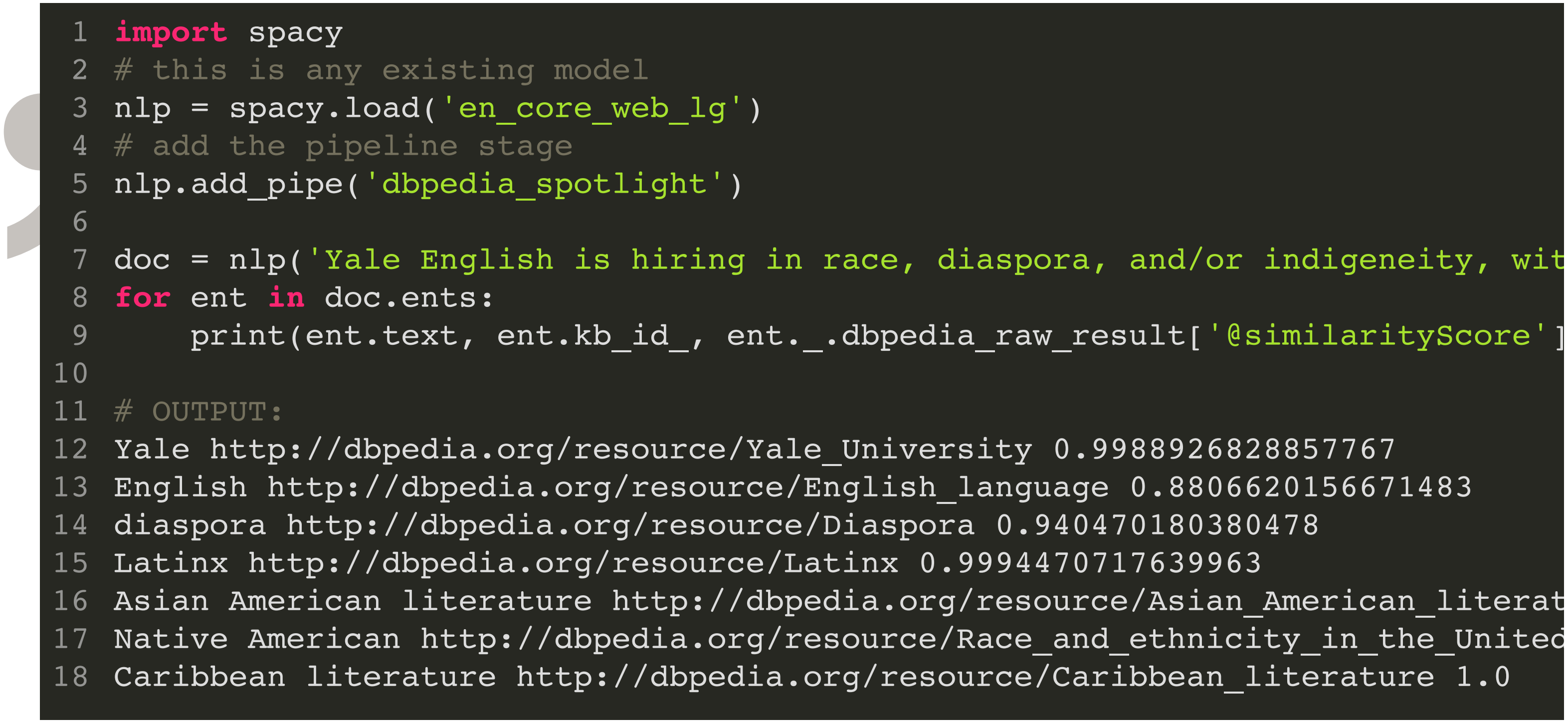
Common NLP Research Tasks

`from raw text to structured data`

Entity Extraction

https://demos.explosion.ai/displacy-ent

Entity Linking



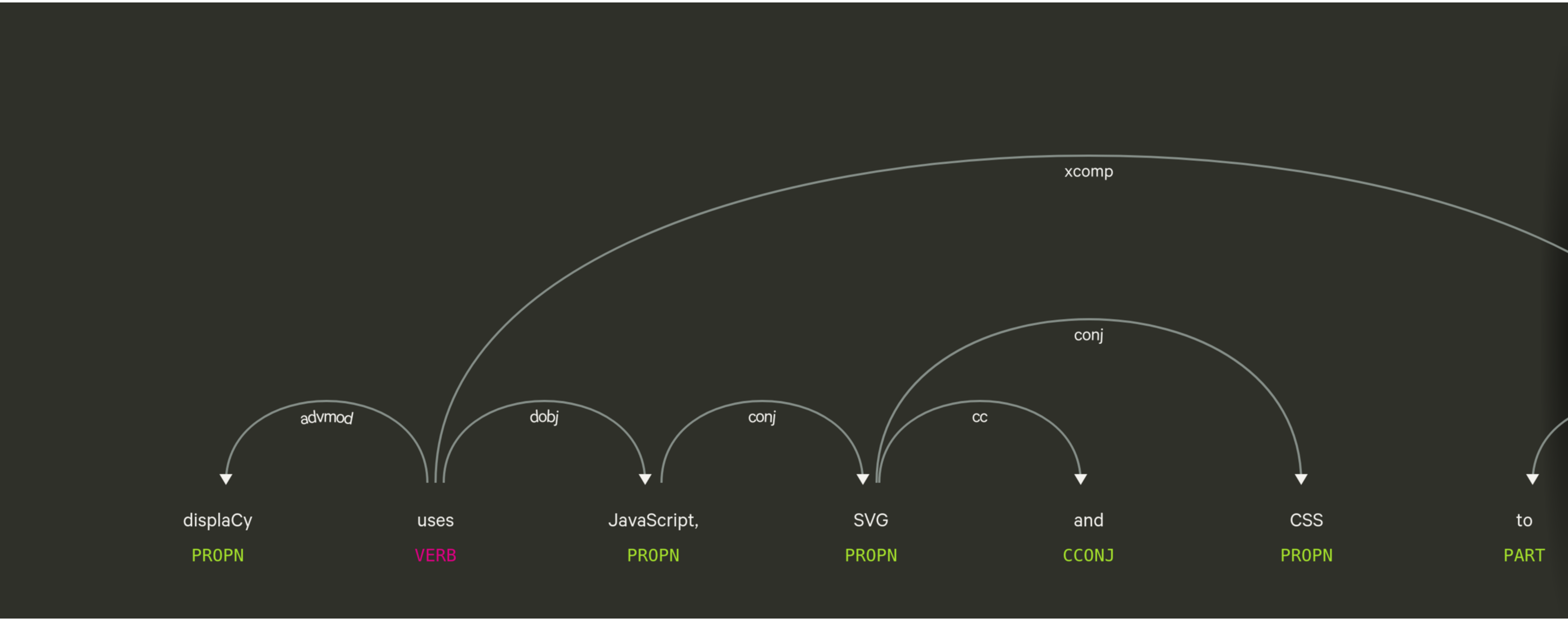
```
import spacy
# this is any existing model
nlp = spacy.load('en_core_web_lg')
# add the pipeline stage
nlp.add_pipe('dbpedia_spotlight')

doc = nlp('Yale English is hiring in race, diaspora, and/or indigeneity, wit
for ent in doc.ents:
    print(ent.text, ent.kb_id_, ent._.dbpedia_raw_result['@similarityScore']

# OUTPUT:
Yale http://dbpedia.org/resource/Yale_University 0.9988926828857767
English http://dbpedia.org/resource/English_language 0.8806620156671483
diaspora http://dbpedia.org/resource/Diaspora 0.940470180380478
Latinx http://dbpedia.org/resource/Latinx 0.9994470717639963
Asian American literature http://dbpedia.org/resource/Asian_American_literat
Native American http://dbpedia.org/resource/Race_and_ethnicity_in_the_United
Caribbean literature http://dbpedia.org/resource/Caribbean_literature 1.0
```

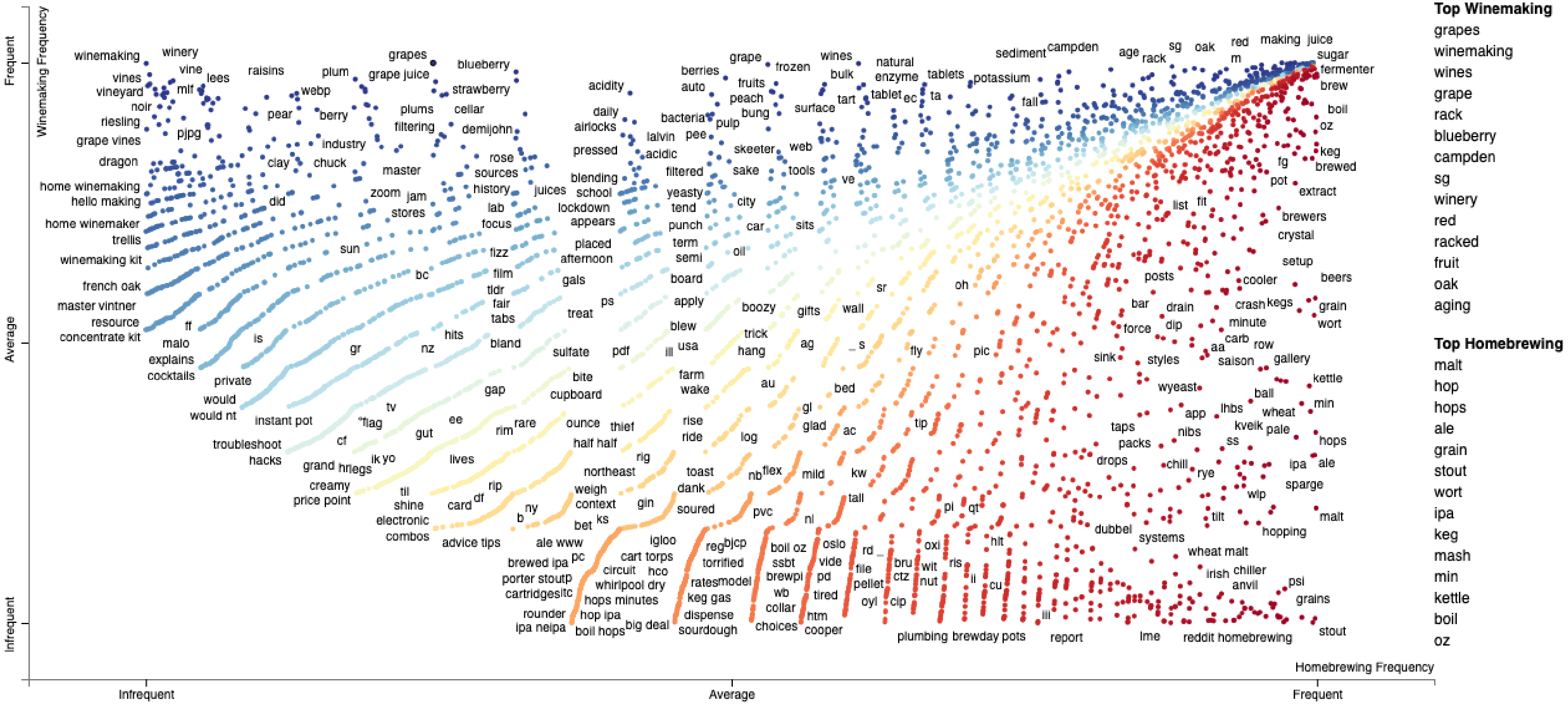
https://github.com/MartinoMensio/spacy-dbpedia-spotlight

Dependency Parsing



Holy NLP! Understanding Part of Speech Tags, Dependency Parsing, and Named Entity Recognition

Categorization



Categorization

https://www.youtube.com/embed/gDk7_f3ovIk?enablejsapi=1

MULTILINGUAL MODELS

Language

Domain

🤗 huggingface.co/models

BOOKNLP

BookNLP is a natural language processing pipeline that scales to books and other long documents (in English), including:

- Part-of-speech tagging
- Dependency parsing
- Entity recognition
- Character name clustering (e.g., "Tom", "Tom Sawyer", "Mr. Sawyer", "Thomas Sawyer" -> TOM_SAWYER) and coreference resolution
- Quotation speaker identification
- Supersense tagging (e.g., "animal", "artifact", "body", "cognition", etc.)
- Event tagging
- Referential gender inference (TOM_SAWYER -> he/him/his)

https://github.com/booknlp/booknlp