NETSL 2016

Metadata Extraction With Python Natural Language Processing

Brendan Short
Team Leader, Content Systems
NEJM Group, Massachusetts Medical Society

Who's this guy?

- Started coding ~4 years ago
- Graduate of DST4L and LJA
- All of which is to say:

Anyone can do this

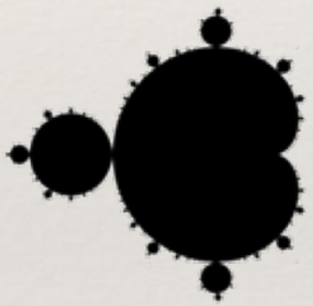
The problem:

- Tons of content across multiple products
- Limited metadata
  - Custom to each product
  - Not interoperable across products
  - Not interoperable with the world

Python + NLP to the rescue!

Python NLTK/TextBlob

- “Out-of-the-box” functionality
- We mostly used it to:
  - identify noun phrases
  - tokenize
  - filter on frequencies
  - create bigrams



Our process

- Retrieve content (Requests!)
- Parse XML data (BeautifulSoup!)
- Process text (NLTK! TextBlob!):
  - extract noun phrases
  - terms with frequency >3
  - filter both against lexicons
  - bigrams with frequency >3
- Write as JSON

Next steps

- Add extracted terms to article metadata
- Integrate MESH RDF into expanded search
- Add other ontologies in future