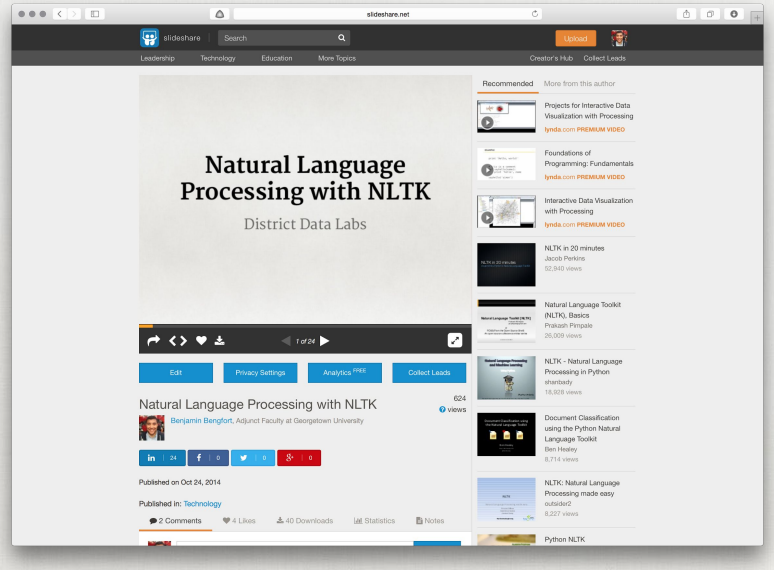
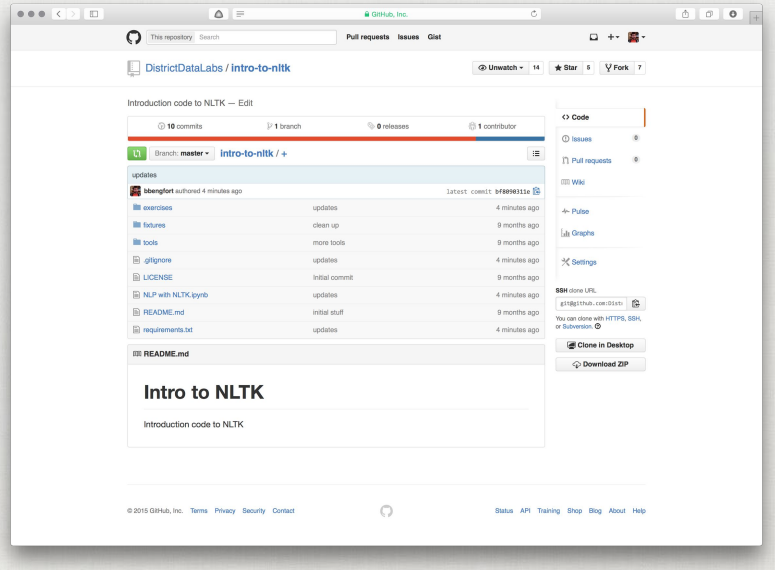
Natural Language Processing with NLTK

District Data Labs

SSID: GuestNet
No Password





Links to these slides

http://bit.ly/intro-to-nltk-slides

http://bit.ly/intro-to-nltk-ipynb

Links to Github Repository

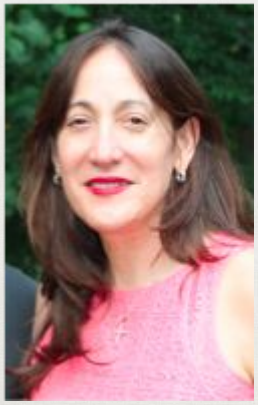
http://bit.ly/intro-to-nltk-code

Links to Various Resources

Benjamin Bengfort

Data scientist and Python programmer; author, father and BBQ specialist.

Twitter: twitter.com/bbengfort
LinkedIn: linkedin.com/in/bbengfort
Github: github.com/bbengfort
Email: benjamin@bengfort.com



Selma Gomez Orr

Summer Intern at District Data Labs and teaching assistant for this course. She will manage chat and Piazza to answer any questions.

Email: selmagomezorr@gmail.com

About the Teaching Assistant

Natural Language Processing

What is NLP?

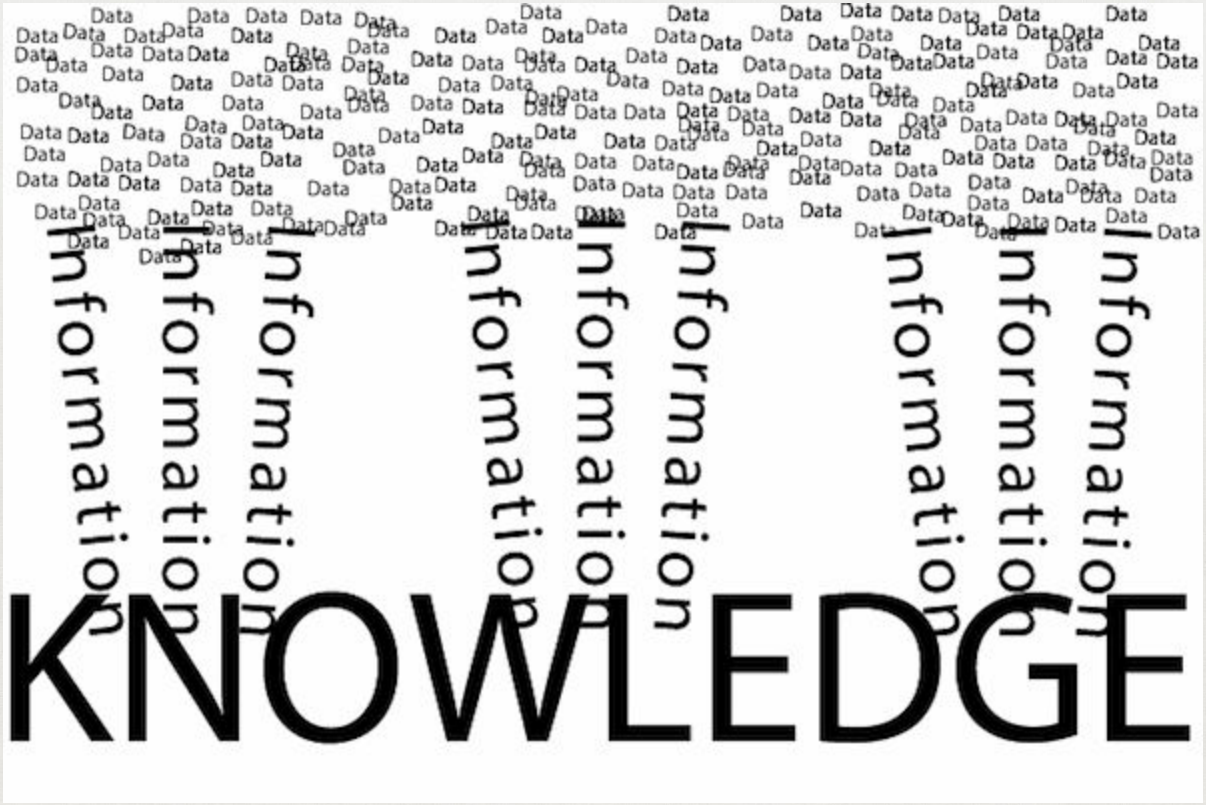
The science that has been developed around the facts of language passed through three stages before finding its true and unique object. First something called "grammar" was studied. This study, initiated by the Greeks and continued mainly by the French, was based on logic. It lacked a scientific approach and was detached from language itself. Its only aim was to give rules for distinguishing between correct and incorrect forms; it was a normative discipline, far removed from actual observation, and its scope was limited.

-- Ferdinand de Saussure

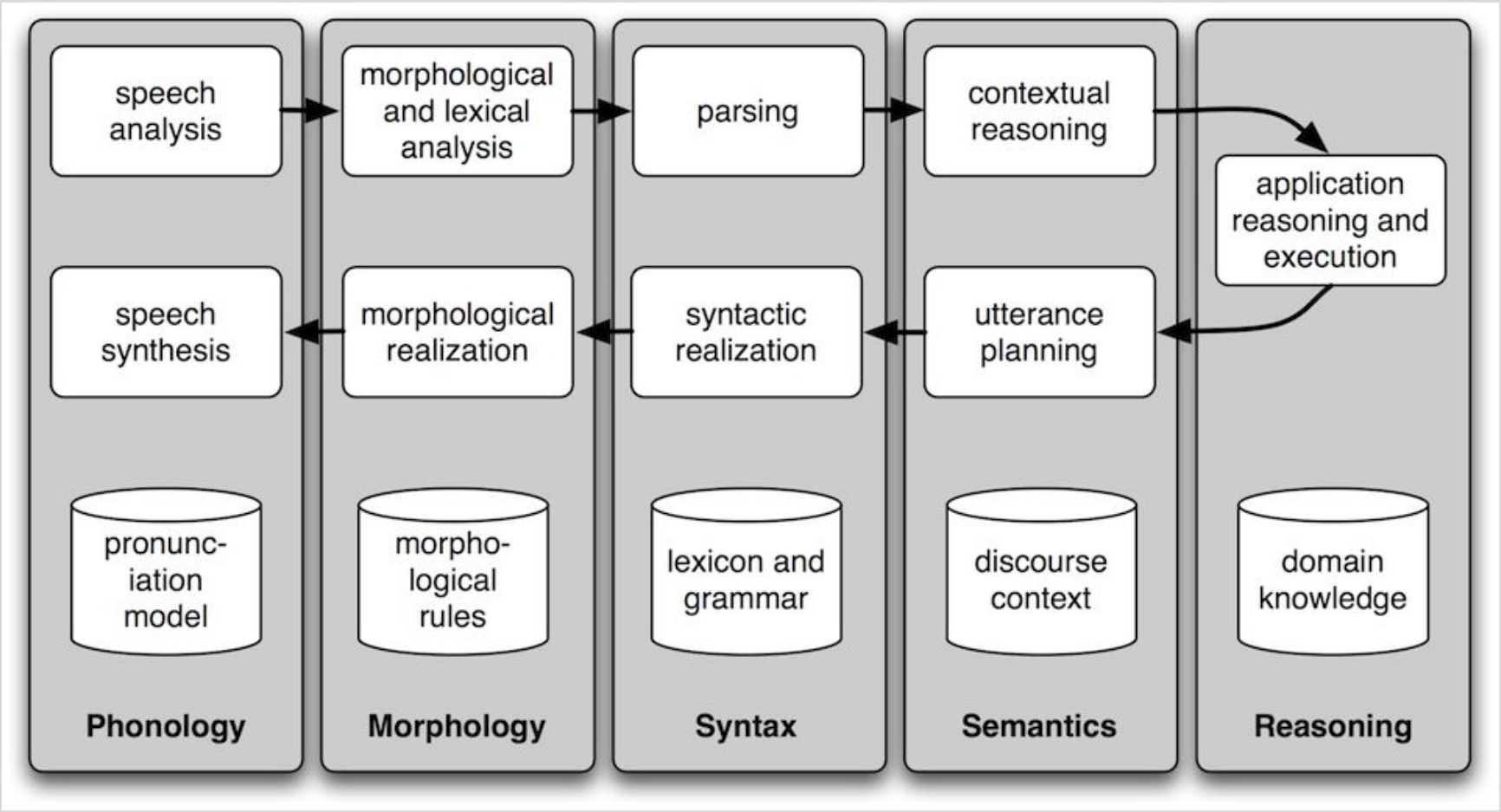
The State of the Art

- Academic design for use alongside intelligent agents (AI discipline)
- Relies on formal models or representations of knowledge & language
- Models are adapted and augment through probabilistic methods and machine learning.
- A small number of algorithms comprise the standard framework.

What is Required?



Domain Knowledge
A Corpus in the Domain



The NLP Pipeline

Morphology

The study of the forms of things, words in particular.

Consider pluralization for English:

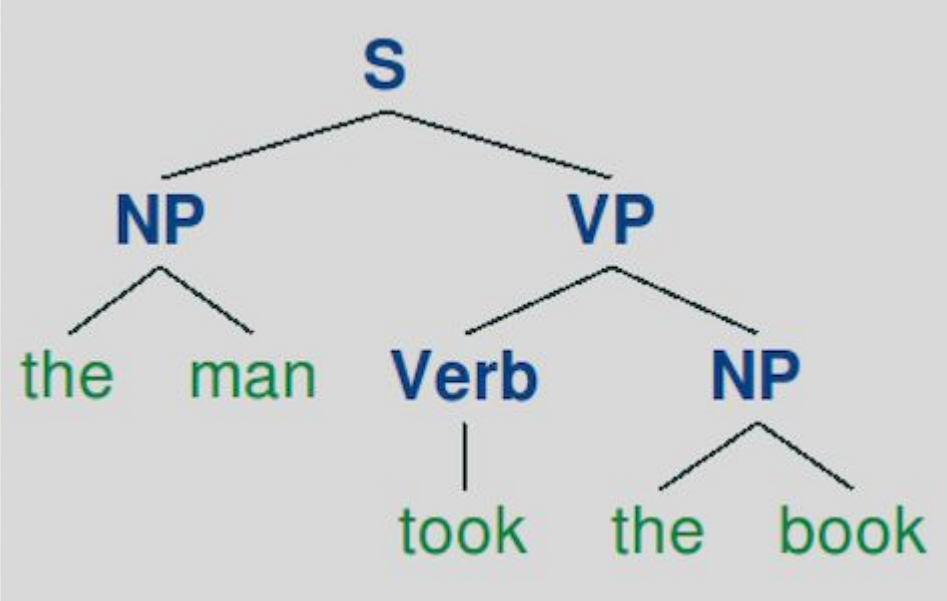
- Orthographic Rules: puppy → puppies
- Morphological Rules: goose → geese or fish

Major parsing tasks:

stemming, lemmatization and tokenization.

Syntax

The study of the rules for the formation of sentences.



Major tasks:

chunking, parsing, feature parsing, grammars

Semantics

The study of meaning.

- I see what I eat.
- I eat what I see.
- He poached salmon.

Major Tasks

Frame extraction, creation of TMRs

NLP Applications

- Yelp Insights
- Winning Jeopardy! IBM Watson
- Computer assisted medical coding (3M Health Information Systems)
- Geoparsing -- CALVIN (built by Charlie Greenbacker)
- Author Identification (classification/clustering)
- Sentiment Analysis (RTNNs, classification)
- Language Detection
- Event Detection
- Google Knowledge Graph
- Named Entity Recognition and Classification
- Machine Translation

Applications are BIG data

- Examples are easier to create than rules.
- Rules and logic miss frequency and language dynamics
- More data is better for machine learning, relevance is in the long tail
- Knowledge engineering is not scalable
- Computational linguistics methodologies are stochastic

The Natural Language Toolkit

What is NLTK?

- Python interface to over 50 corpora and lexical resources
- Focus on Machine Learning with specific domain knowledge
- Free and Open Source
- Numpy and Scipy under the hood
- Fast and Formal

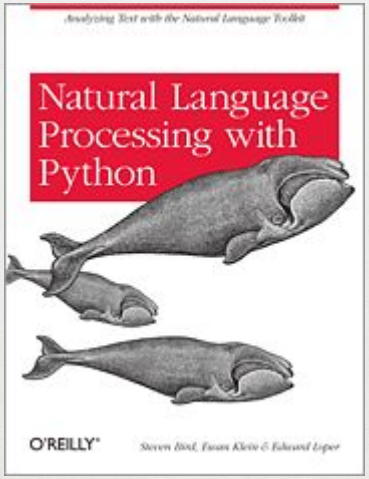
What is NLTK?

Suite of libraries for a variety of academic text processing tasks:

- tokenization, stemming, tagging,
- chunking, parsing, classification,
- language modeling, logical semantics

Pedagogical resources for teaching NLP theory in Python ...

Who Wrote NLTK?



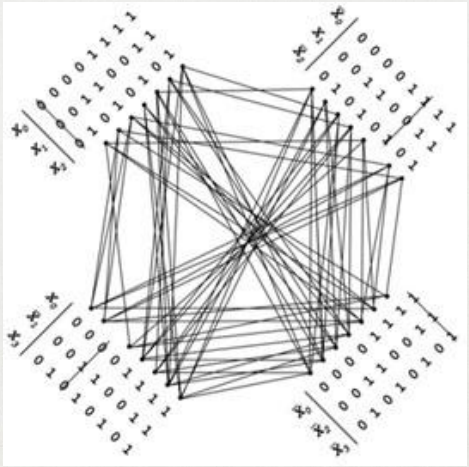
Steven Bird

Associate Professor
University of Melbourne
Senior Research Associate, LDC

Ewan Klein

Professor of Language Technology
University of Edinburgh.

Batteries Included



NLTK = Corpora + Algorithms
Ready for Research!

What is NLTK not?

- Production ready out of the box*
- Lightweight
- Generally applicable
- Magic

*There are actually a few things that are production ready right out of the box.

The Good

- Preprocessing
  - segmentation, tokenization, PoS tagging
- Word level processing
  - WordNet, Lemmatization, Stemming, NGram
- Utilities
  - Tree, FreqDist, ConditionalFreqDist
  - Streaming CorpusReader objects
- Classification
  - Maximum Entropy, Naive Bayes, Decision Tree
  - Chunking, Named Entity Recognition
- Parsers Galore!

The Bad

- Syntactic Parsing
  - No included grammar (not a black box)
- Feature/Dependency Parsing
  - No included feature grammar
- The sem package
  - Toy only (lambda-calculus & first order logic)
- Lots of extra stuff
  - papers, chat programs, alignments, etc.

Other Python NLP Libraries

- TextBlob
- Pattern
- gensim
- MITIE
- guess_language
- Python wrapper for Stanford CoreNLP
- Python wrapper for Berkeley Parser
- readability-lxml
- BeautifulSoup

Our Task

Build a system that ingests raw language data and transforms it into a suitable representation for creating revolutionary applications.

Workshop Tasks

Part One: Demonstrating NLTK

- Working with Included Corpora
- Segmentation, Tokenization, Tagging
- A Parsing Exercise
- Named Entity Recognition Chunker
- Classification with NLTK
- Clustering with NLTK
- Doing LDA with gensim

Workshop Tasks

Part Two: Building an NLP Data Product

- A Deep Look at the Corpus Reader
- A View of a Production System