

# Processamento e Recuperação de Informação

## Information Extraction : An Introduction

Departamento de Engenharia Informática
Instituto Superior Técnico

1º Semestre
2018/2019

# Bibliography



- Bing Liu, Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data. Chapter 9.

# Bibliography - Articles




- AnHai Doan, Raghu Ramakrishnan, and Shivakumar Vaithyanathan. Managing information extraction: state of the art and research directions. In Proceedings of the 2006 ACM SIGMOD international conference on Management of data (SIGMOD '06). http://doi.acm.org/10.1145/1142473.1142595
- William W. Cohen. Information Extraction and Integration: an Overview. http://www.cs.cmu.edu/~wcohen/ie-survey.ppt

# Outline

# Text-based Applications



- Free-text, semi-structured, streaming ...
  - Web pages, email, news articles, call-center text records, business reports, annotations, spreadsheets, research papers, blogs, tags, instant messages (IM), ...
- High-impact applications
  - Business intelligence, personal information management, Web communities and social media, Web search and advertising, scientific data management, e-government, medical records management, ...
- Growing rapidly

# Exploiting Text



Two main directions:

- Information Retrieval
- Information Extraction

# The Task of Information Extraction



Best known as the drummer in Jimi Hendrix's Band of Gypsys, Buddy Miles also had a lengthy solo career that drew from rock, blues, soul, and funk in varying combinations. Born George Miles in Omaha, NE, on September 5, 1947, he started playing the drums at age nine, and joined his father's jazz band the Bebops as a mere 12 year old. As a teenager, he went on to play with several jazz and R&B outfits, most prominently backing vocal groups like Ruby & the Romantics, the Ink Spots, and the Delfonics.

# The Task of Information Extraction



Best known as the drummer in Jimi Hendrix's Band of Gypsys, Buddy Miles also had a lengthy solo career that drew from rock, blues, soul, and funk in varying combinations. Born George Miles in Omaha, NE, on September 5, 1947, he started playing the drums at age nine, and joined his father's jazz band the Bebops as a mere 12 year old. As a teenager, he went on to play with several jazz and R&B outfits, most prominently backing vocal groups like Ruby & the Romantics, the Ink Spots, and the Delfonics.

# The Task of Information Extraction



Best known as the drummer in Jimi Hendrix's Band of Gypsys, Buddy Miles also had a lengthy solo career that drew from rock, blues, soul, and funk in varying combinations. Born George Miles in Omaha, NE, on September 5, 1947, he started playing the drums at age nine, and joined his father's jazz band the Bebops as a mere 12 year old. As a teenager, he went on to play with several jazz and R&B outfits, most prominently backing vocal groups like Ruby & the Romantics, the Ink Spots, and the Delfonics.

# The Task of Information Extraction



Best known as the drummer in Jimi Hendrix's Band of Gypsys, Buddy Miles also had a lengthy solo career that drew from rock, blues, soul, and funk in varying combinations. Born George Miles in Omaha, NE, on September 5, 1947, he started playing the drums at age nine, and joined his father's jazz band the Bebops as a mere 12 year old. As a teenager, he went on to play with several jazz and R&B outfits, most prominently backing vocal groups like Ruby & the Romantics, the Ink Spots, and the Delfonics.

# The Task of Information Extraction



Best known as the drummer in Jimi Hendrix's Band of Gypsys, Buddy Miles also had a lengthy solo career that drew from rock, blues, soul, and funk in varying combinations. Born George Miles in Omaha, NE, on September 5, 1947, he started playing the drums at age nine, and joined his father's jazz band the Bebops as a mere 12 year old. As a teenager, he went on to play with several jazz and R&B outfits, most prominently backing vocal groups like Ruby & the Romantics, the Ink Spots, and the Delfonics.

| **Artist** | **Band** | **Instrument** |
|---|---|---|
| Jimi Hendrix | Band of Gypsys | |
| Buddy Miles | Band of Gypsys | drums |
| Buddy Miles | Bebops | drums |
| Buddy Miles | Ruby & the Romantics | vocal |
| Buddy Miles | Ink Spots | vocal |
| Buddy Miles | Delfonics | vocal |

# Outline

# Many Tasks for Extracting Information



- Named Entity Recognition and Classification
  - E.g. Buddy Miles (person), Band of Gypsys (band), ...
- Named Entity Resolution (i.e., Entity Linking)
  - Buddy Miles = George Miles
- Relationship extraction
  - Buddy Miles played drums in Band of Gypsys
- Among others...

# Different Facets of IE



- Different domains
  - News, scientific papers, the Web, ...
- Different formatting
  - Raw text, web pages, ...
- Different coverage
  - From very domain specific to general open-domain IE
- Different complexity
  - From restricted vocabulary to ambiguous NL
- Different target data models
  - From single records to full relational

# An Example Open-Domain IE Project



**Open Information Extraction**

Hosted by AI2 Created at

**Example Queries:**
- What kills bacteria?
- Who built the Pyramids?
- What did Thomas Edison invent?
- What contains antioxidants?

**Typed Example Queries:**
- What countries are located in Africa?
- What actors starred in which films?
- What is the symbol of which country?
- What foods are grown in which countries?
- What drug ingredients has the FDA approved?

**Argument 1:** what/who | **Relation:** verb phrase | **Argument 2:** what/who | **Corpus:** All | Search

**AI2 proudly announces the launch of Semantic Scholar, an AI-based academic search engine.**

To learn more about Open IE, watch our YouTube video!

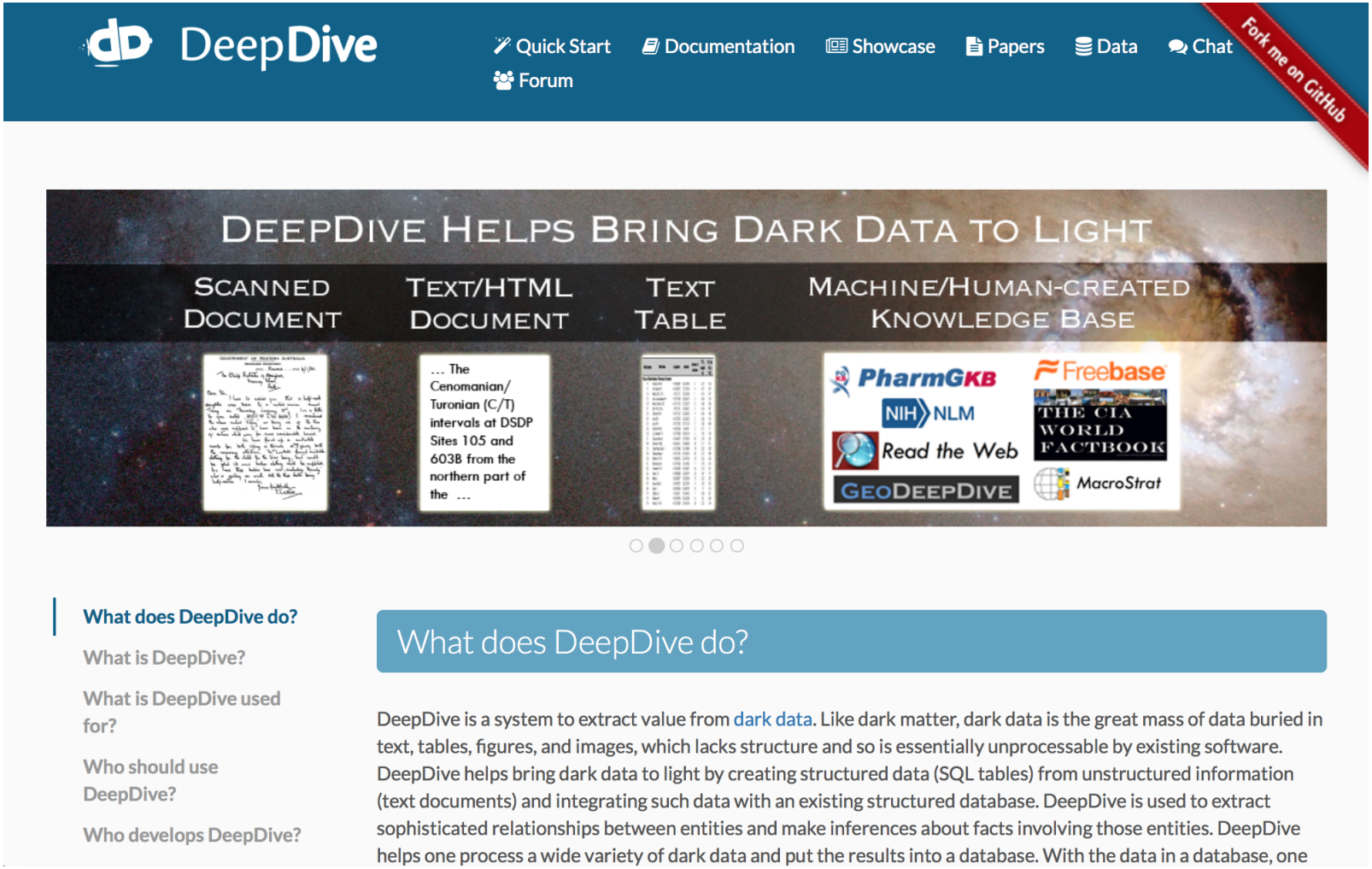
Powered by ReVerb, our Open Information Extractor, yielding over 5 billion extractions from over a billion web pages.

NEW! **Open IE 4.0**, the successor to ReVerb and Ollie, has been released. Download it from GitHub!

http://openie.allenai.org

# An Example IE Toolkit






http://deepdive.stanford.edu/

# Outline

# Different Techniques for IE



- Lexicons
- Rules
- Classifiers
- Sliding Window Classifiers
- Sequential Classification Models
- Finite State Machines

Hand-coded

Machine learning

Multi-step workflows

# Example of Hand-coded Rules



```
##########################################################################
# Regular expressions to construct the pattern to extract conference names
##########################################################################
# These are subordinate patterns
my $wordOrdinals="(?:first|second|third|fourth|fifth|sixth|seventh|eighth|ninth|tenth|eleventh|t
my $numberOrdinals="(?:\\d?(?:1st|2nd|3rd|1th|2th|3th|4th|5th|6th|7th|8th|9th|0th))";
my $ordinals="(?:$wordOrdinals|$numberOrdinals)";
my $confTypes="(?:Conference|Workshop|Symposium)";
my $words="(?:[A-Z]\\w+\\s*)"; # A word starting with a capital letter and ending with 0 or more
my $confDescriptors="(?:international\\s+|[A-Z]+\\s+)"; # .e.g "International Conference ...' or
my $connectors="(?:on|of)";
my $abbreviations="(?:\\([A-Z]\\w\\w+[\\W\\s]*?(?:\\d\\d+)?\\))"; # Conference abbreviations lik
# The actual pattern we search for. A typical conference name this pattern will find is
# "3rd International Conference on Blah Blah Blah (ICBBB-05)"
my $fullNamePattern="((?:$ordinals\\s+$words*|$confDescriptors)?$confTypes(?:\\s+$connectors\\s+
############################## ##################################
# Given a <dbworldMessage>, look for the conference pattern
##########################################################
lookForPattern($dbworldMessage, $fullNamePattern);
##########################################################
# In a given <file>, look for occurrences of <pattern>
# <pattern> is a regular expression
##########################################################
sub lookForPattern {
my ($file,$pattern) = @_;
```

# An Example IE System Based on Rules



The Stanford Natural Language Processing Group

people publications research blog software teaching local

## Software > Stanford TokensRegex

About | Download | Usage | Questions | Mailing lists | Release history

### About

TokensRegex is a generic framework included in Stanford CoreNLP for defining patterns over text (sequences of tokens) and mapping it to semantic objects represented as Java objects. TokensRegex emphasizes describing text as a sequence of tokens (words, punctuation marks, etc.), which may have additional attributes, and writing patterns over those tokens, rather than working at the character level, as with standard regular expression packages. For example, you might match names of people who are painters with a TokensRegex pattern like this:

```
([ner: PERSON]+) /was|is/ /an?/ []{0,3} /painter|artist/
```

http://nlp.stanford.edu/software/tokensregex.html

and

http://nlp.stanford.edu/software/regexner.html

# Most Common Machine Learning Techniques



- Traditional classifiers
  - Nave Bayes, SVM, ...
  - Often adapted to handling sequences of words (i.e., *sliding window approaches*)
- Sequence Classifiers (i.e., *structured predictors*)
  - Hidden Markov Models
  - Structured Perceptrons
  - Conditional Random Fields
  - Recurrent or Convolutional Deep Neural Networks

# IE as Sequence Classification



Information Extraction tasks such as Named Entity (NE) Recognition can be modeled as a sequence classification problem, leveraging a tagging scheme such as B-I-O

- B stands for "beginning" (signifies beginning of a NE)
- I stands for "inside" (signifies that the word is inside a NE)
- O stands for "outside" (signifies that the word is just a regular word, outside of a NE)

Infer the most likely sequence of tags (i.e., classes in a sequential structure), for a given sequence of words.



Questions?