Information Extraction

Pedro Szekely

Information Sciences Institute,

USC Viterbi School of Engineering

Agenda

Information extraction classification

Text extraction techniques

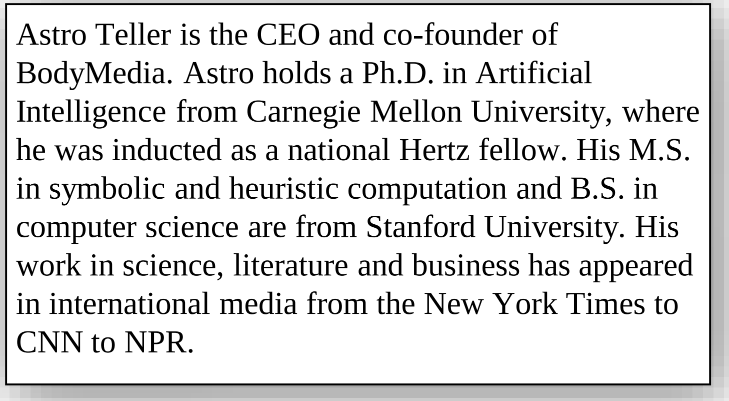
Storing extractions in knowledge graphs

myDIG demo

Summary

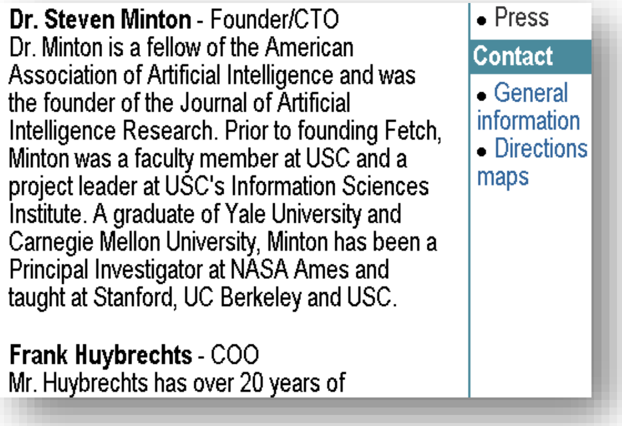
Document Features

Text paragraphs without formatting



Astro Teller is the CEO and co-founder of BodyMedia. Astro holds a Ph.D. in Artificial Intelligence from Carnegie Mellon University, where he was inducted as a national Hertz fellow. His M.S. in symbolic and heuristic computation and B.S. in computer science are from Stanford University. His work in science, literature and business has appeared in international media from the New York Times to CNN to NPR.

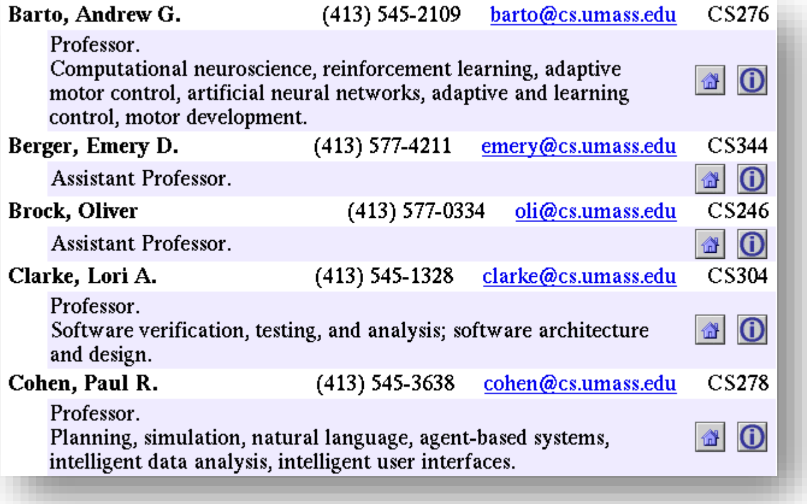
Grammatical sentences plus some formatting & links



Dr. Steven Minton - Founder/CTO
Dr. Minton is a fellow of the American Association of Artificial Intelligence and was the founder of the Journal of Artificial Intelligence Research. Prior to founding Fetch, Minton was a faculty member at USC and a project leader at USC's Information Sciences Institute. A graduate of Yale University and Carnegie Mellon University, Minton has been a Principal Investigator at NASA Ames and taught at Stanford, UC Berkeley and USC.

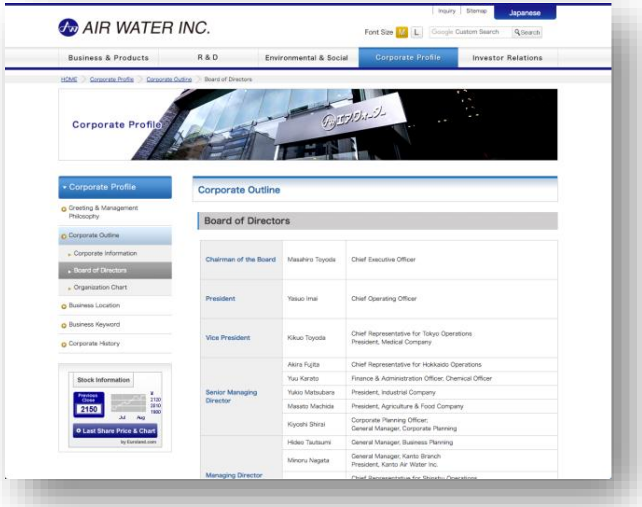
Frank Huybrechts - COO
Mr. Huybrechts has over 20 years of

Non-grammatical snippets, rich formatting & links



Barto, Andrew G.	(413) 545-2109	barto@cs.umass.edu	CS276	Professor. Computational neuroscience, reinforcement learning, adaptive motor control, artificial neural networks, adaptive and learning control, motor development.
Berger, Emery D.	(413) 577-4211	emery@cs.umass.edu	CS344	Assistant Professor.
Brock, Oliver	(413) 577-0334	oli@cs.umass.edu	CS246	Assistant Professor.
Clarke, Lori A.	(413) 545-1328	clarke@cs.umass.edu	CS304	Professor. Software verification, testing, and analysis; software architecture and design.
Cohen, Paul R.	(413) 545-3638	cohen@cs.umass.edu	CS278	Professor. Planning, simulation, natural language, agent-based systems, intelligent data analysis, intelligent user interfaces.

Tables

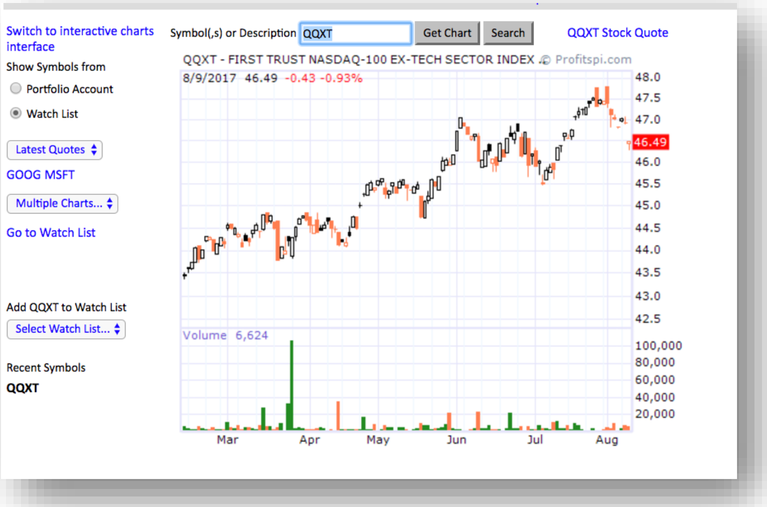


Corporate Outline

Board of Directors

Chairman of the Board	Masahiro Toyoda	Chief Executive Officer
President	Yasuo Imai	Chief Operating Officer
Vice President	Kikuo Toyoda	Chief Representative for Tokyo Operations President, Medical Company
Senior Managing Director	Akira Fujita	Chief Representative for Hokkaido Operations
	Yuu Karato	Finance & Administration Officer, Chemical Officer
	Yukio Matsubara	President, Industrial Company
	Masato Machida	President, Agriculture & Food Company
	Kiyoshi Shirai	Corporate Planning Officer; General Manager, Corporate Planning
Managing Director	Hideo Tsutsumi	General Manager, Business Planning
	Minoru Nagata	General Manager, Kanto Branch President, Kanto Air Water Inc.
		Chief Representative for Shinshu Operations

Charts



Symbol(,s) or Description QQXT

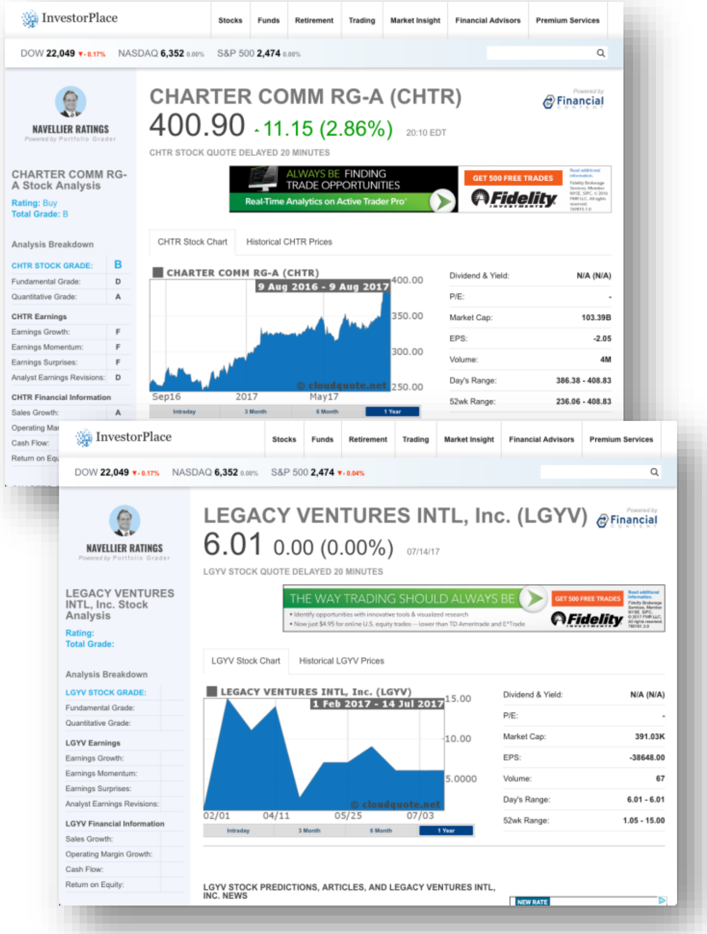
QQXT - FIRST TRUST NASDAQ-100 EX-TECH SECTOR INDEX © Profitspi.com

8/9/2017 46.49 -0.43 -0.93%

Volume 6,624

Scope

Web site specific



CHARTER COMM RG-A (CHTR)

400.90 - 11.15 (2.86%)

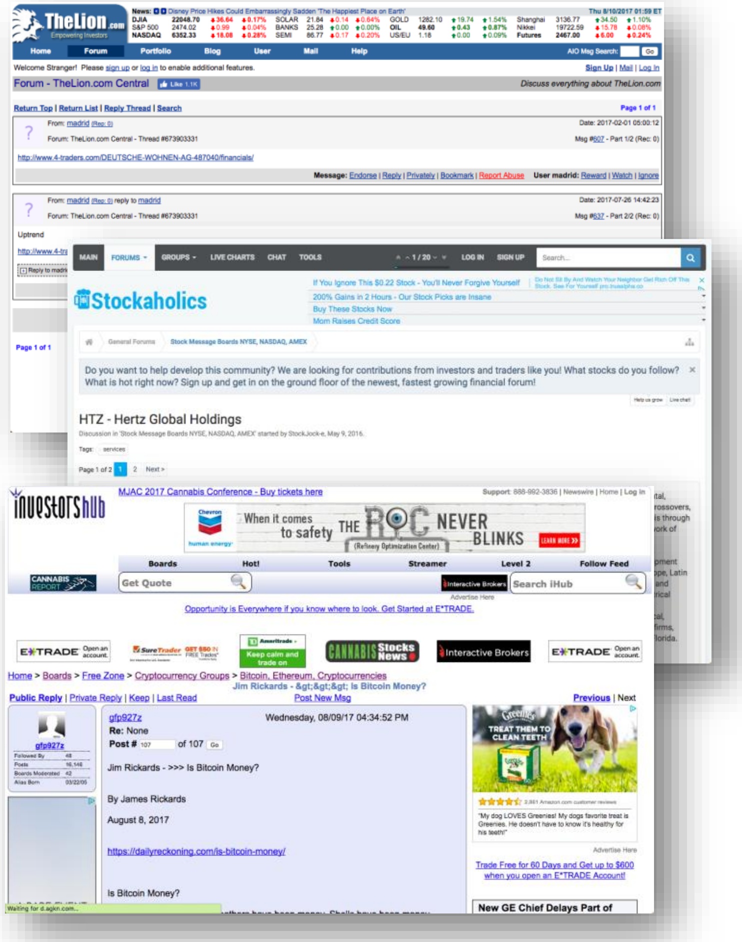
CHTR STOCK QUOTE DELAYED 20 MINUTES

LEGACY VENTURES INTL, Inc. (LGYV)

6.01 0.00 (0.00%)

LGYV STOCK QUOTE DELAYED 20 MINUTES

Genre specific (e.g., forums)



Forum - TheLion.com Central

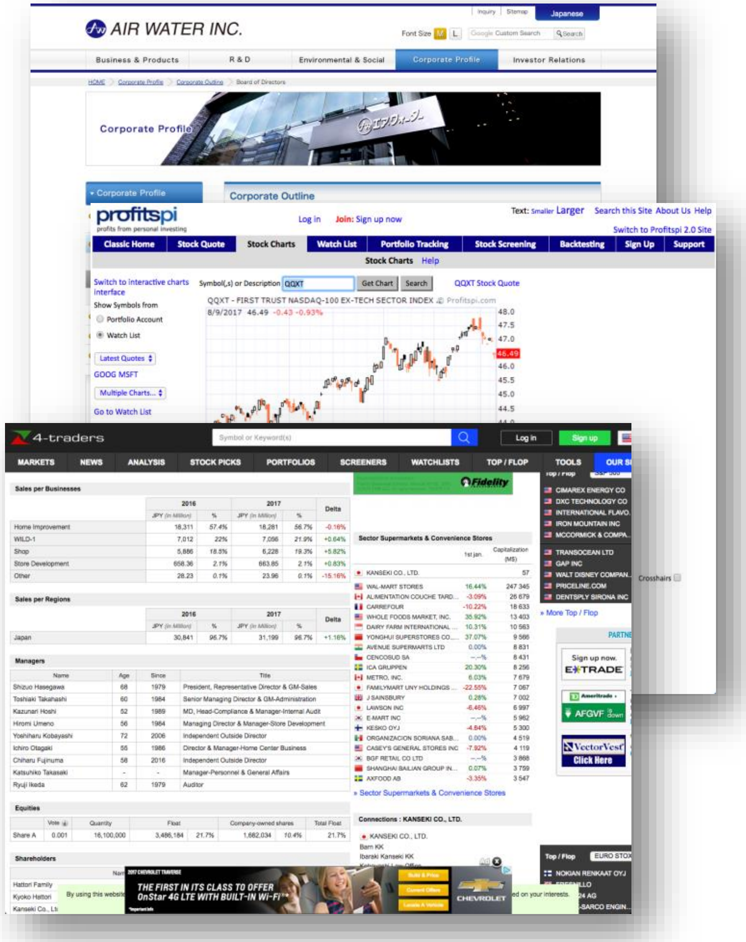
Stockaholics

HTZ - Hertz Global Holdings

Jim Rickards - &gt;&gt;&gt; Is Bitcoin Money?

Wednesday, 08/09/17 04:34:52 PM

Wide, non-specific



AIR WATER INC.

Corporate Profile

profitspi

4-traders

Pattern Complexity

E.g., word patterns

Closed set

U.S. states

He was born in Alabama…

The big Wyoming sky…

Regular set

U.S. phone numbers

Phone: (413) 545-1323

The CALD main office can be reached at 412-268-1299

Complex pattern

U.S. postal addresses

University of Arkansas
P.O. Box 140
Hope, AR 71802

Headquarters:
1128 Main Street, 4th Floor
Cincinnati, Ohio 45210

Ambiguous patterns, needing context and many sources of evidence

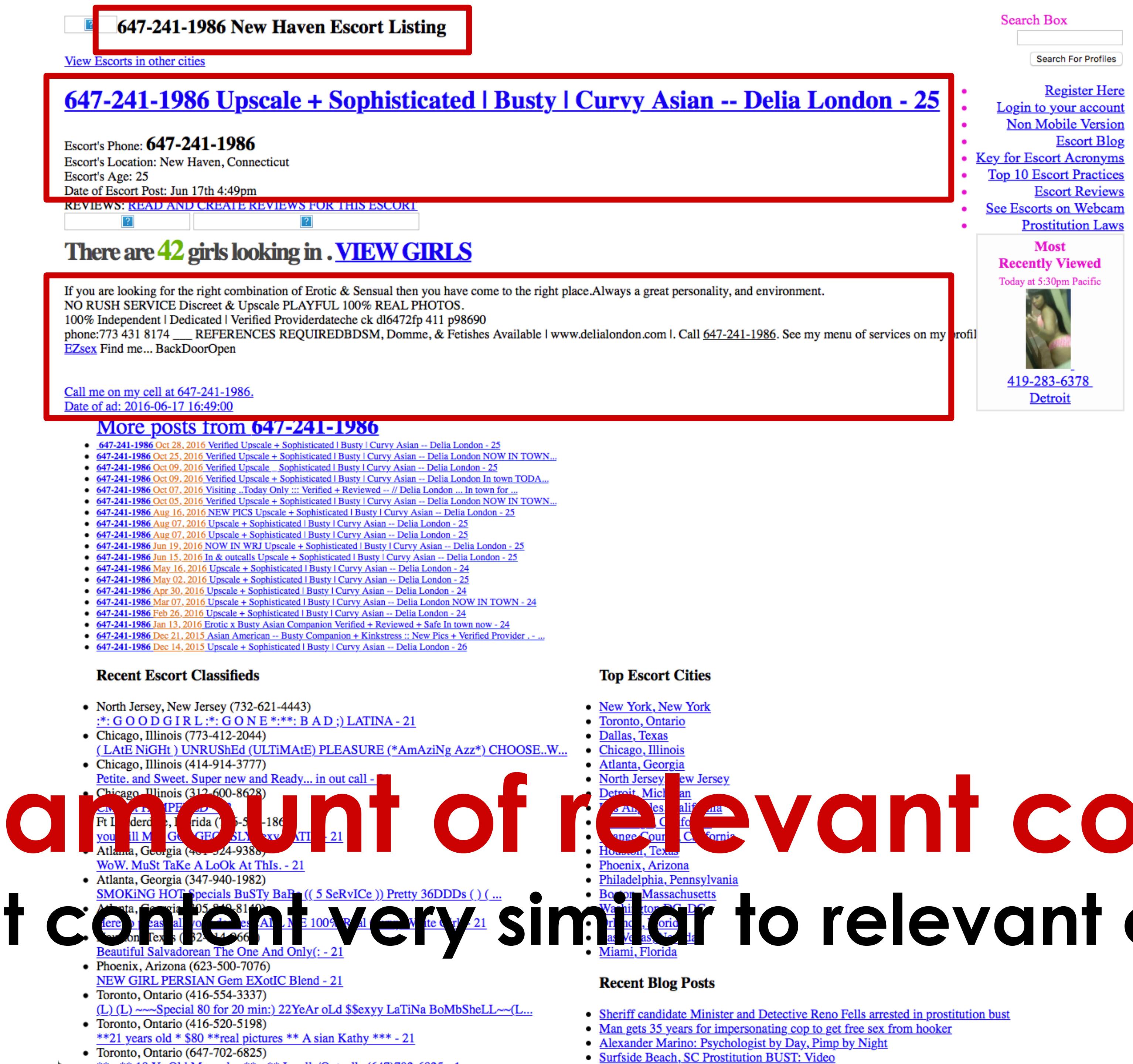
Person names

…was among the six houses sold by Hope Feldman that year

Pawel Opalinski, Software Engineer at WhizBang Labs.

“YOU don't wanna miss out on ME :) Perfect lil booty Green eyes Long curly black hair Im a Irish, Armenian and Filipino mixed princess :) ❤ Kim ❤ 7○7~7two7~7four77 ❤ HH 80 roses ❤ Hour 120 roses ❤ 15 mins 60 roses”

Courtesy of Andrew McCallum



647-241-1986 New Haven Escort Listing

View Escorts in other cities

Search Box

Search For Profiles

- Register Here
- Login to your account
- Non Mobile Version
- Escort Blog
- Key for Escort Acronyms
- Top 10 Escort Practices
- Escort Reviews
- See Escorts on Webcam
- Prostitution Laws

647-241-1986 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25

Escort's Phone: **647-241-1986**
Escort's Location: New Haven, Connecticut
Escort's Age: 25
Date of Escort Post: Jun 17th 4:49pm
REVIEWS: READ AND CREATE REVIEWS FOR THIS ESCORT

There are 42 girls looking in . VIEW GIRLS

If you are looking for the right combination of Erotic & Sensual then you have come to the right place.Always a great personality, and environment.
NO RUSH SERVICE Discreet & Upscale PLAYFUL 100% REAL PHOTOS.
100% Independent | Dedicated | Verified Providerdateche ck dl6472fp 411 p98690
phone:773 431 8174 ___ REFERENCES REQUIREDBDSM, Domme, & Fetishes Available | www.delialondon.com |. Call 647-241-1986. See my menu of services on my profil
EZsex Find me... BackDoorOpen

Call me on my cell at 647-241-1986.
Date of ad: 2016-06-17 16:49:00

Most
Recently Viewed
Today at 5:30pm Pacific

419-283-6378
Detroit

More posts from **647-241-1986**

- **647-241-1986** Oct 28, 2016 Verified Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- **647-241-1986** Oct 25, 2016 Verified Upscale + Sophisticated | Busty | Curvy Asian -- Delia London NOW IN TOWN...
- **647-241-1986** Oct 09, 2016 Verified Upscale _ Sophisticated | Busty | Curvy Asian -- Delia London - 25
- **647-241-1986** Oct 09, 2016 Verified Upscale + Sophisticated | Busty | Curvy Asian -- Delia London In town TODA...
- **647-241-1986** Oct 07, 2016 Visiting ..Today Only ::: Verified + Reviewed -- // Delia London ... In town for ...
- **647-241-1986** Oct 05, 2016 Verified Upscale + Sophisticated | Busty | Curvy Asian -- Delia London NOW IN TOWN...
- **647-241-1986** Aug 16, 2016 NEW PICS Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- **647-241-1986** Aug 07, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- **647-241-1986** Aug 07, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- **647-241-1986** Jun 19, 2016 NOW IN WRJ Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- **647-241-1986** Jun 15, 2016 In & outcalls Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- **647-241-1986** May 16, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 24
- **647-241-1986** May 02, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 25
- **647-241-1986** Apr 30, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 24
- **647-241-1986** Mar 07, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London NOW IN TOWN - 24
- **647-241-1986** Feb 26, 2016 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 24
- **647-241-1986** Jan 13, 2016 Erotic x Busty Asian Companion Verified + Reviewed + Safe In town now - 24
- **647-241-1986** Dec 21, 2015 Asian American -- Busty Companion + Kinkstress :: New Pics + Verified Provider . - ...
- **647-241-1986** Dec 14, 2015 Upscale + Sophisticated | Busty | Curvy Asian -- Delia London - 26

Recent Escort Classifieds

- North Jersey, New Jersey (732-621-4443)
  :*: GOODGIRL :*: GONE *:**: BAD ;) LATINA - 21
- Chicago, Illinois (773-412-2044)
  (LAtE NiGHt) UNRUShEd (ULTiMAtE) PLEASURE (*AmAziNg Azz*) CHOOSE..W...
- Chicago, Illinois (414-914-3777)
  Petite. and Sweet. Super new and Ready... in out call - [illegible]
- Chicago, Illinois (312-600-8628)
  [illegible]
- Ft Lauderdale, Florida ([illegible]-186[illegible]
  you will M[illegible] GORGEOUSLY sexy LATI[illegible] - 21
- Atlanta, Georgia (4[illegible]-324-9388)
  WoW. MuSt TaKe A LoOk At ThIs. - 21
- Atlanta, Georgia (347-940-1982)
  SMOKiNG HOT Specials BuSTy BaBe ((5 SeRvICe)) Pretty 36DDDs ( ) (...
- Atlanta, Georgia ([illegible]05-840-8140)
  Here to please all your desires CALL ME 100% Real [illegible] White Girl - 21
- Houston, Texas ([illegible]32-[illegible]-966[illegible])
  Beautiful Salvadorean The One And Only(: - 21
- Phoenix, Arizona (623-500-7076)
  NEW GIRL PERSIAN Gem EXotIC Blend - 21
- Toronto, Ontario (416-554-3337)
  (L) (L) ~~~Special 80 for 20 min:) 22YeAr oLd $$exyy LaTiNa BoMbSheLL~~(L...
- Toronto, Ontario (416-520-5198)
  21 years old * $80 **real pictures ** A sian Kathy * - 21
- Toronto, Ontario (647-702-6825)

Top Escort Cities

- New York, New York
- Toronto, Ontario
- Dallas, Texas
- Chicago, Illinois
- Atlanta, Georgia
- North Jersey, New Jersey
- Detroit, Michigan
- Los Angeles, California
- [illegible], California
- Orange County, California
- Houston, Texas
- Phoenix, Arizona
- Philadelphia, Pennsylvania
- Boston, Massachusetts
- Washington DC, DC
- Orlando, Florida
- Las Vegas, Nevada
- Miami, Florida

Recent Blog Posts

- Sheriff candidate Minister and Detective Reno Fells arrested in prostitution bust
- Man gets 35 years for impersonating cop to get free sex from hooker
- Alexander Marino: Psychologist by Day, Pimp by Night
- Surfside Beach, SC Prostitution BUST: Video

small amount of relevant content

irrelevant content very similar to relevant content

Practical Considerations

How good (precision/recall) is necessary?

High precision when showing extractions to users

High recall when used for ranking results

How long does it take to construct?

Minutes, hours, days, months

What expertise do I need?

None (domain expertise), patience (annotation), simple scripting, machine learning guru

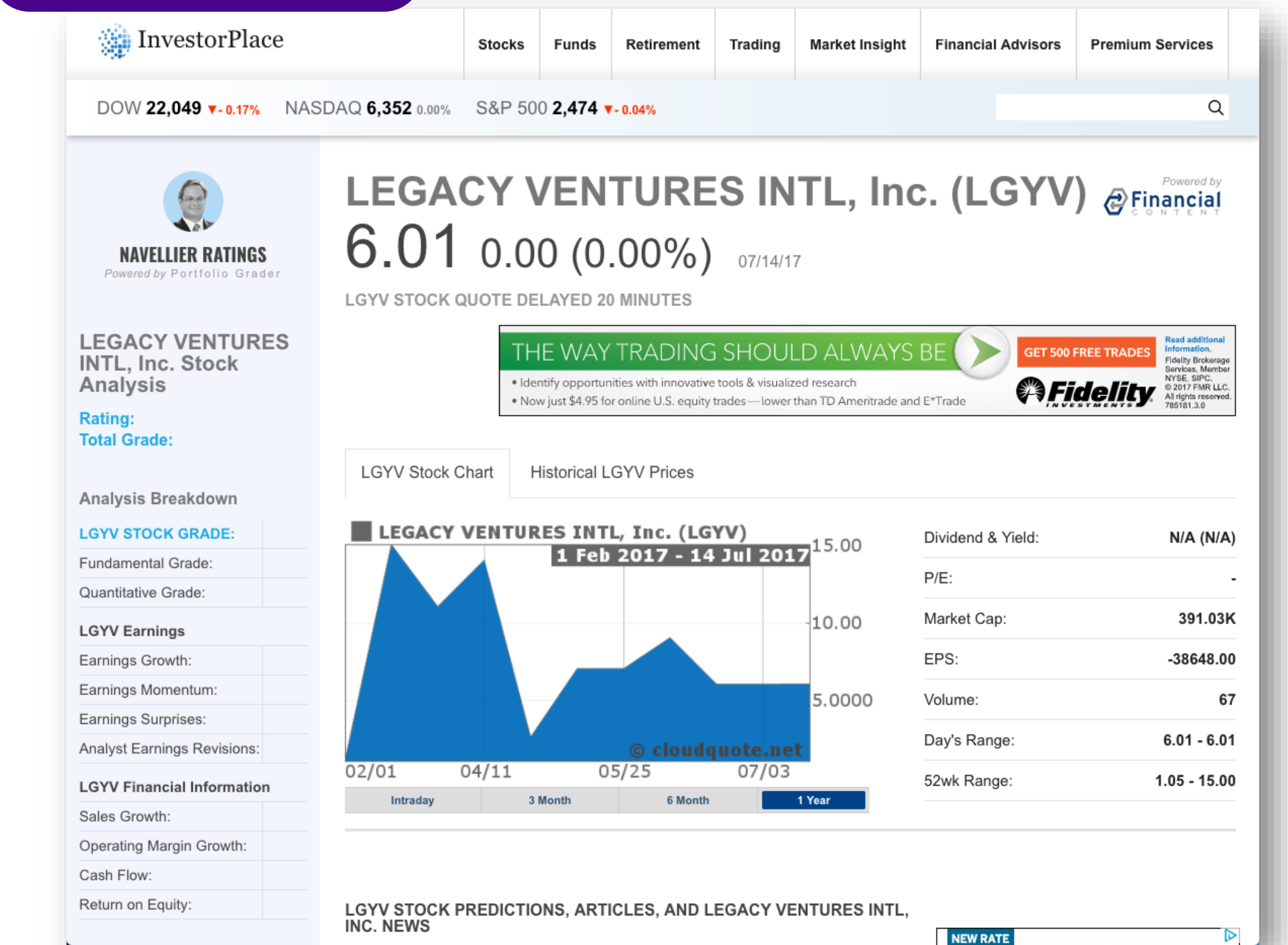
What tools can I use?

Many …

Information Extraction Process

Segmentation

Data Extraction



DOW 22,049 ▼-0.17% NASDAQ 6,352 0.00% S&P 500 2,474 ▼-0.04%

NAVELLIER RATINGS
Powered by Portfolio Grader

LEGACY VENTURES INTL, Inc. Stock Analysis

Rating:
Total Grade:

Analysis Breakdown

LGYV STOCK GRADE:
Fundamental Grade:
Quantitative Grade:

LGYV Earnings
Earnings Growth:
Earnings Momentum:
Earnings Surprises:
Analyst Earnings Revisions:

LGYV Financial Information
Sales Growth:
Operating Margin Growth:
Cash Flow:
Return on Equity:

LEGACY VENTURES INTL, Inc. (LGYV)

6.01 0.00 (0.00%) 07/14/17

LGYV STOCK QUOTE DELAYED 20 MINUTES

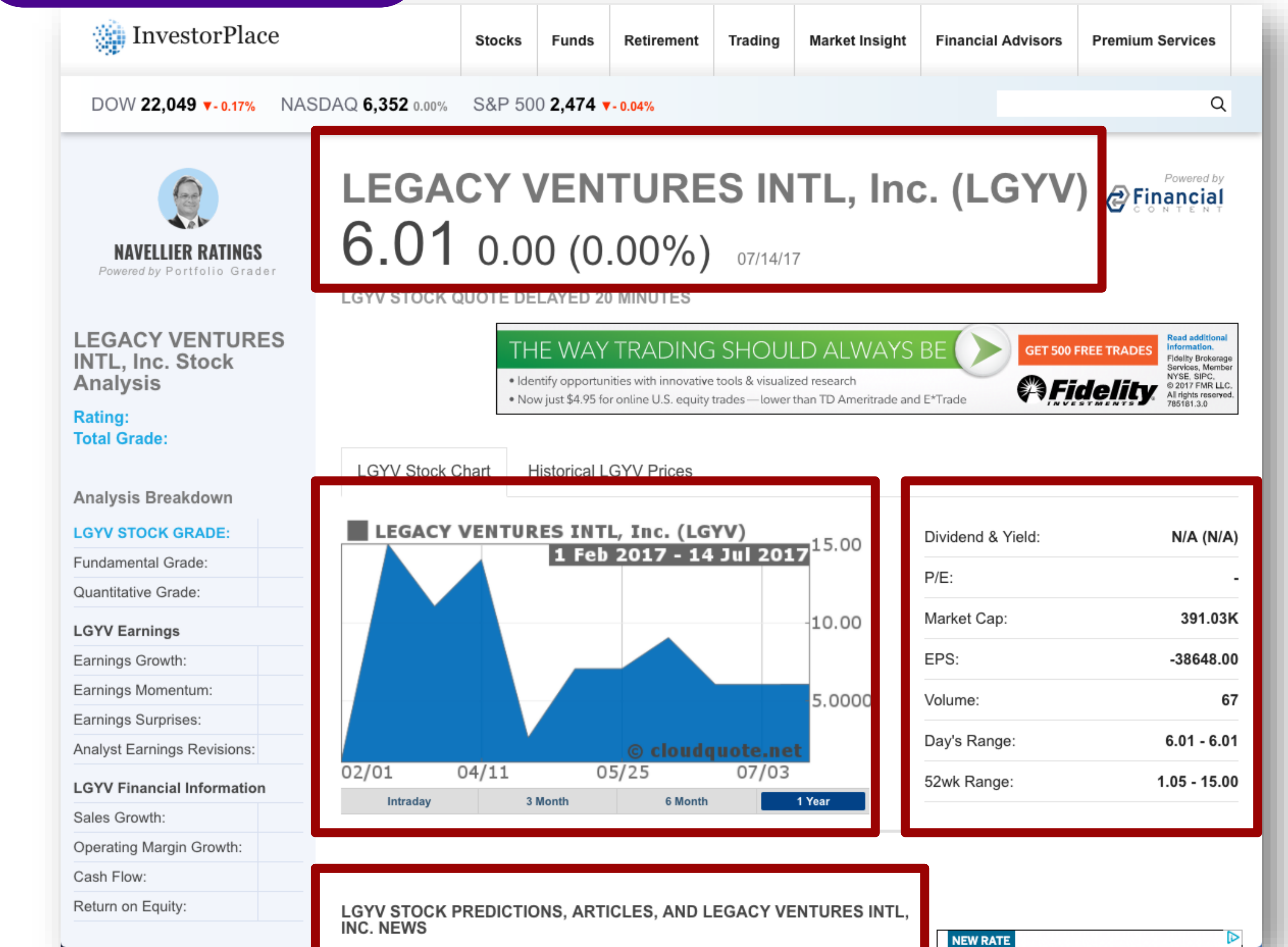
Dividend & Yield:	N/A (N/A)
P/E:	-
Market Cap:	391.03K
EPS:	-38648.00
Volume:	67
Day's Range:	6.01 - 6.01
52wk Range:	1.05 - 15.00

LGYV STOCK PREDICTIONS, ARTICLES, AND LEGACY VENTURES INTL, INC. NEWS

Information Extraction Process

Segmentation

Data Extraction



InvestorPlace

Stocks | Funds | Retirement | Trading | Market Insight | Financial Advisors | Premium Services

DOW 22,049 ▼-0.17% NASDAQ 6,352 0.00% S&P 500 2,474 ▼-0.04%

NAVELLIER RATINGS
Powered by Portfolio Grader

LEGACY VENTURES INTL, Inc. Stock Analysis

Rating:
Total Grade:

Analysis Breakdown

LGYV STOCK GRADE:
Fundamental Grade:
Quantitative Grade:

LGYV Earnings
Earnings Growth:
Earnings Momentum:
Earnings Surprises:
Analyst Earnings Revisions:

LGYV Financial Information
Sales Growth:
Operating Margin Growth:
Cash Flow:
Return on Equity:

LEGACY VENTURES INTL, Inc. (LGYV)

6.01 0.00 (0.00%) 07/14/17

Powered by Financial Content

LGYV STOCK QUOTE DELAYED 20 MINUTES

LGYV Stock Chart | Historical LGYV Prices

Dividend & Yield:	N/A (N/A)
P/E:	-
Market Cap:	391.03K
EPS:	-38648.00
Volume:	67
Day's Range:	6.01 - 6.01
52wk Range:	1.05 - 15.00

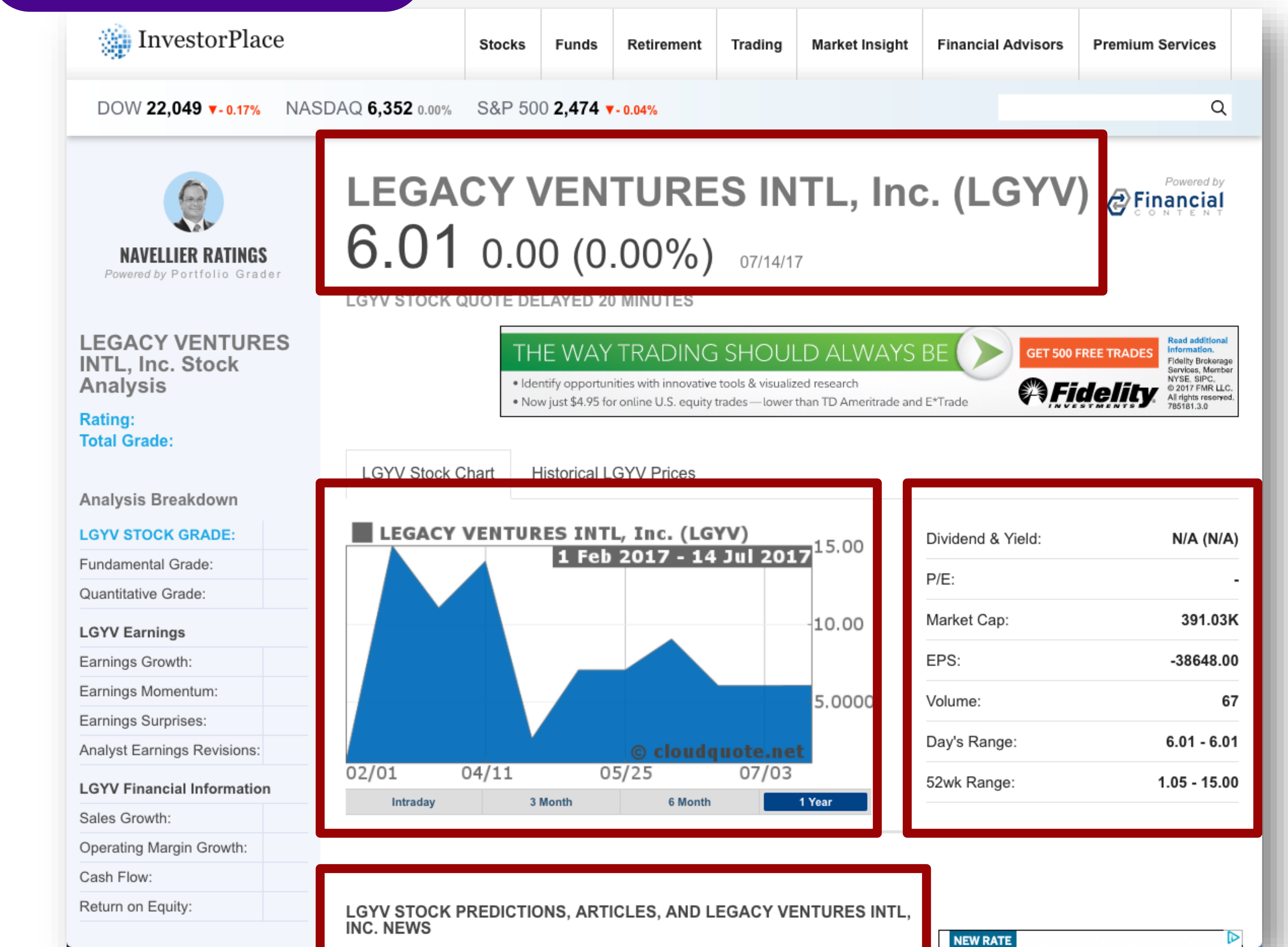
LGYV STOCK PREDICTIONS, ARTICLES, AND LEGACY VENTURES INTL, INC. NEWS

NEW RATE

Information Extraction Process

Segmentation

Data Extraction



InvestorPlace

Stocks | Funds | Retirement | Trading | Market Insight | Financial Advisors | Premium Services

DOW 22,049 ▾-0.17% NASDAQ 6,352 0.00% S&P 500 2,474 ▾-0.04%

NAVELLIER RATINGS
Powered by Portfolio Grader

LEGACY VENTURES INTL, Inc. Stock Analysis

Rating:
Total Grade:

Analysis Breakdown

LGYV STOCK GRADE:
Fundamental Grade:
Quantitative Grade:

LGYV Earnings
Earnings Growth:
Earnings Momentum:
Earnings Surprises:
Analyst Earnings Revisions:

LGYV Financial Information
Sales Growth:
Operating Margin Growth:
Cash Flow:
Return on Equity:

LEGACY VENTURES INTL, Inc. (LGYV)
6.01 0.00 (0.00%) 07/14/17

Powered by Financial Content

LGYV STOCK QUOTE DELAYED 20 MINUTES

LGYV Stock Chart | Historical LGYV Prices

Dividend & Yield:	N/A (N/A)
P/E:	-
Market Cap:	391.03K
EPS:	-38648.00
Volume:	67
Day's Range:	6.01 - 6.01
52wk Range:	1.05 - 15.00

LGYV STOCK PREDICTIONS, ARTICLES, AND LEGACY VENTURES INTL, INC. NEWS

NEW RATE

Name:
Legacy Ventures Intl, Inc.

Stock:
LGYV

Date:
2017-07-14

Market Cap:
391,030

Segmentation

Semi-structured extraction

Table extraction

Main content identification

Custom regular expressions

Segmentation

Semi-structured extraction

Table extraction

Main content identification

Custom regular expressions

→ Text segments

Text Extraction Techniques

Glossary

Regular expressions

Natural language rules

Named entity recognition

Sequence labeling (Conditional Random Fields)

Glossary Extraction

Glossary Extraction

Simple

list of words or phrases to extract

Challenges

Ambiguity: Charlotte is a name of a person and a city

Colloquial expressions: “Asia Broadband, Inc.” vs “Asia Broadband”

Research

Improving precision of glossary extractions using context

Creating/extending glossaries automatically

Regex Extraction

Extraction Using Regular Expressions

Too difficult for non-programmers

regex for North American phone numbers:

```
^(?:(?:\+?1\s*(?:[.-]\s*)?)?(?:\(\s*([2-9]1[02-9]|[2-9][02-8]1|[2-9][02-8][02-9])\s*\)|([2-9]1[02-9]|[2-9][02-8]1|[2-9][02-8][02-9]))\s*(?:[.-]\s*)?)?([2-9]1[02-9]|[2-9][02-9]1|[2-9][02-9]{2})\s*(?:[.-]\s*)?([0-9]{4})(?:\s*(?:#|x\.?|ext\.?|extension)\s*(\d+))?$
```

Brittle and difficult to adapt to unusual domains

unusual nomenclature and short-hands

obfuscation

NLP Rule-Based Extraction

NLP Rule-Based Extraction

Tokenization

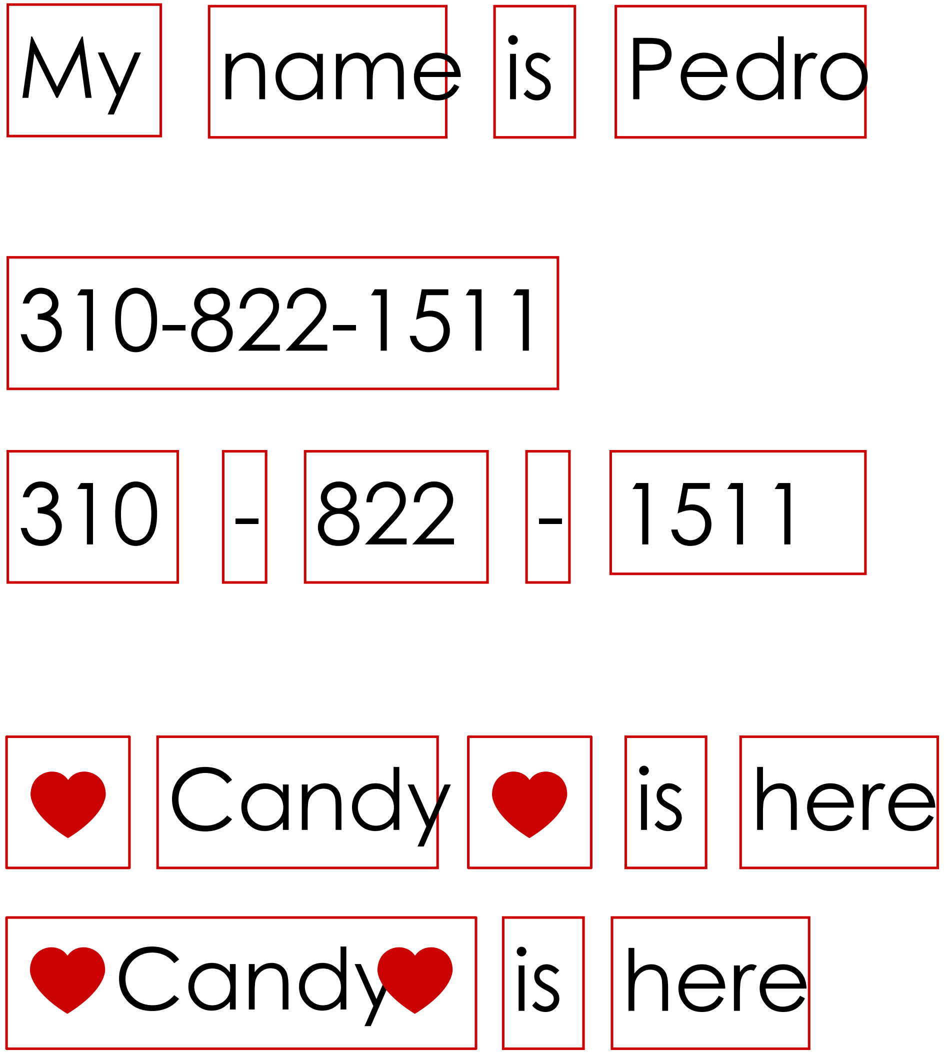
Pattern Matching

Tokenization

My name is Pedro

310-822-1511

❤Candy❤ is here



My | name | is | Pedro

310-822-1511

310 | - | 822 | - | 1511

❤ | Candy | ❤ | is | here

❤Candy❤ | is | here

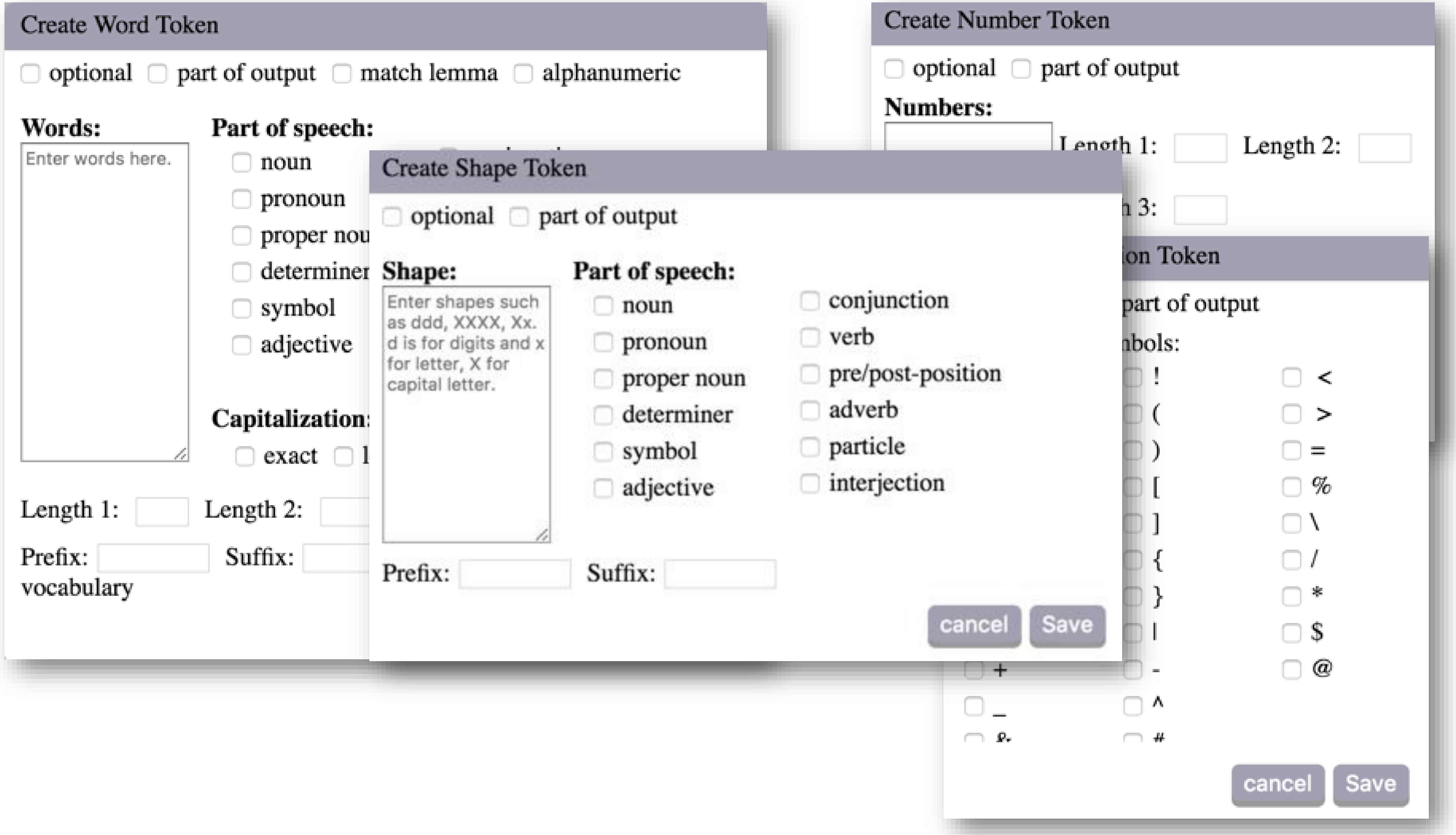
Token Properties

Surface properties

Literal, type, shape, capitalization, length, prefix, suffix, minimum, maximum

Language properties

Part of speech tag, lemma, dependency



Create Word Token

□ optional □ part of output □ match lemma □ alphanumeric

Words:

Enter words here.

Part of speech:

□ noun
□ pronoun
□ proper nou
□ determiner
□ symbol
□ adjective

Capitalization:

□ exact □ l

Length 1: Length 2:

Prefix: Suffix:

vocabulary

Create Number Token

□ optional □ part of output

Numbers:

Length 1: Length 2:

h 3:

Create Shape Token

□ optional □ part of output

Shape:

Enter shapes such as ddd, XXXX, Xx. d is for digits and x for letter, X for capital letter.

Part of speech:

□ noun
□ pronoun
□ proper noun
□ determiner
□ symbol
□ adjective
□ conjunction
□ verb
□ pre/post-position
□ adverb
□ particle
□ interjection

Prefix: Suffix:

cancel Save

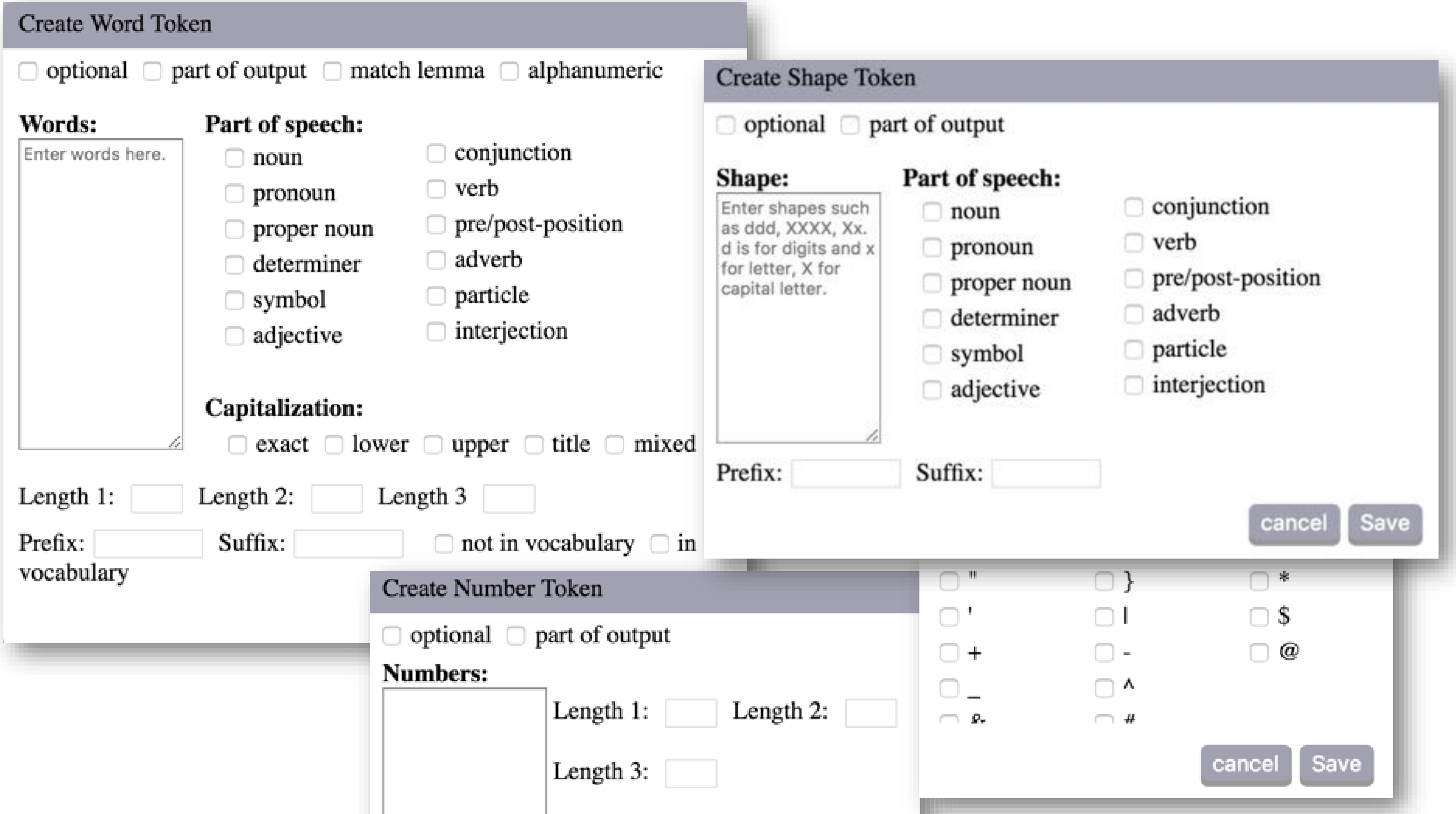
ion Token

part of output

nbols:

□ !
□ (
□)
□ [
□]
□ {
□ }
□ |
□ -
□ ^
□ #
□ <
□ >
□ =
□ %
□ \
□ /
□ *
□ $
□ @
□ +
□ _
□ &

cancel Save



Create Word Token

□ optional □ part of output □ match lemma □ alphanumeric

Words:

Enter words here.

Part of speech:

□ noun □ pronoun □ proper noun □ determiner □ symbol □ adjective

□ conjunction □ verb □ pre/post-position □ adverb □ particle □ interjection

Capitalization:

□ exact □ lower □ upper □ title □ mixed

Length 1: Length 2: Length 3

Prefix: Suffix: □ not in vocabulary □ in vocabulary

Create Shape Token

□ optional □ part of output

Shape:

Enter shapes such as ddd, XXXX, Xx. d is for digits and x for letter, X for capital letter.

Part of speech:

□ noun □ pronoun □ proper noun □ determiner □ symbol □ adjective

□ conjunction □ verb □ pre/post-position □ adverb □ particle □ interjection

Prefix: Suffix:

Create Number Token

□ optional □ part of output

Numbers:

Length 1: Length 2:

Length 3:

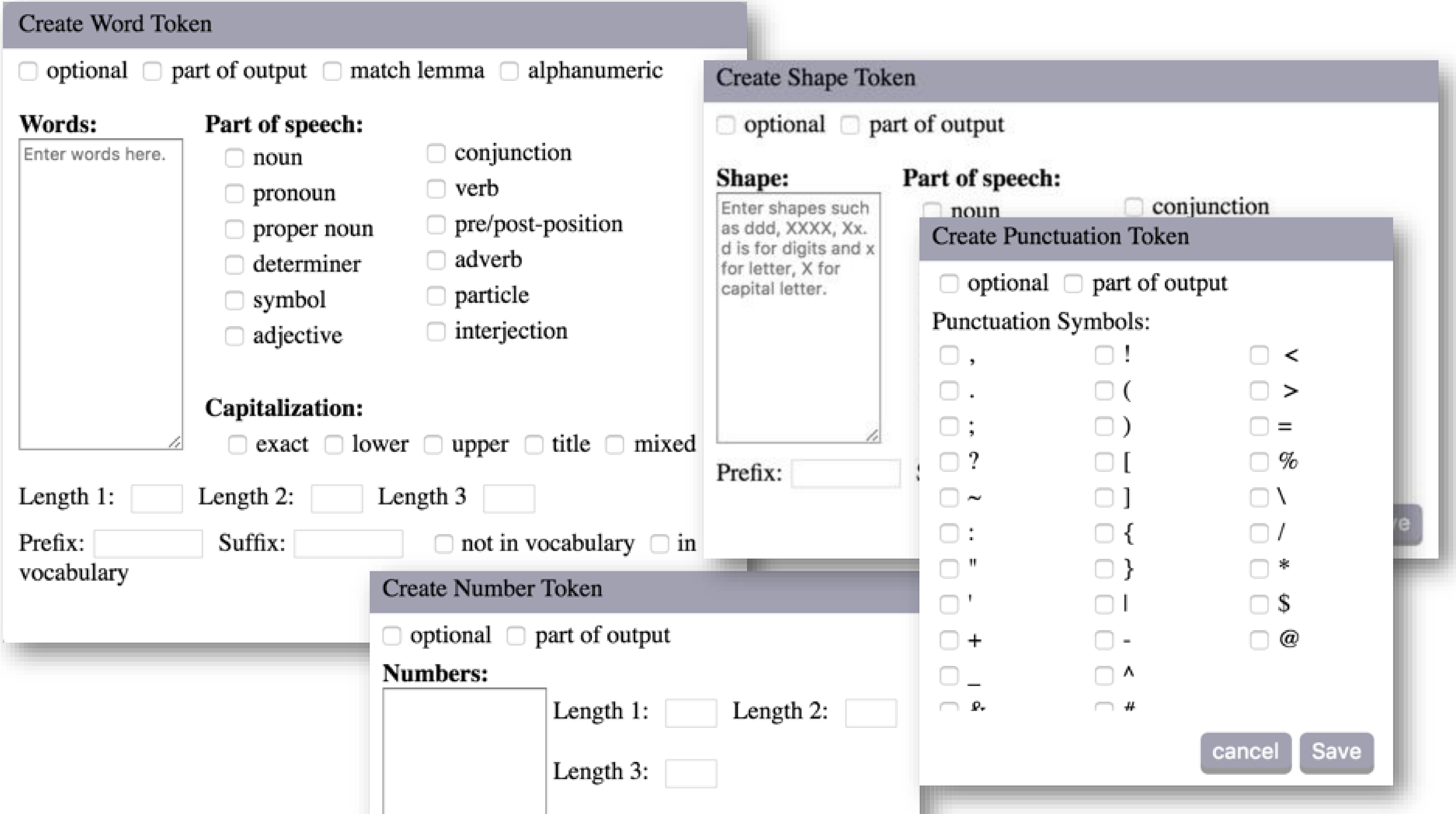
□ " □ } □ *

□ ' □ | □ $

□ + □ - □ @

□ _ □ ^

□ & □ #



Create Word Token

□ optional □ part of output □ match lemma □ alphanumeric

Words:

Enter words here.

Part of speech:

□ noun
□ pronoun
□ proper noun
□ determiner
□ symbol
□ adjective
□ conjunction
□ verb
□ pre/post-position
□ adverb
□ particle
□ interjection

Capitalization:

□ exact □ lower □ upper □ title □ mixed

Length 1: Length 2: Length 3

Prefix: Suffix: □ not in vocabulary □ in vocabulary

Create Shape Token

□ optional □ part of output

Shape:

Enter shapes such as ddd, XXXX, Xx. d is for digits and x for letter, X for capital letter.

Part of speech:

□ noun
□ conjunction

Prefix:

Create Punctuation Token

□ optional □ part of output

Punctuation Symbols:

□ ,
□ .
□ ;
□ ?
□ ~
□ :
□ "
□ '
□ +
□ _
□ &
□ !
□ (
□)
□ [
□]
□ {
□ }
□ |
□ -
□ ^
□ #
□ <
□ >
□ =
□ %
□ \
□ /
□ *
□ $
□ @

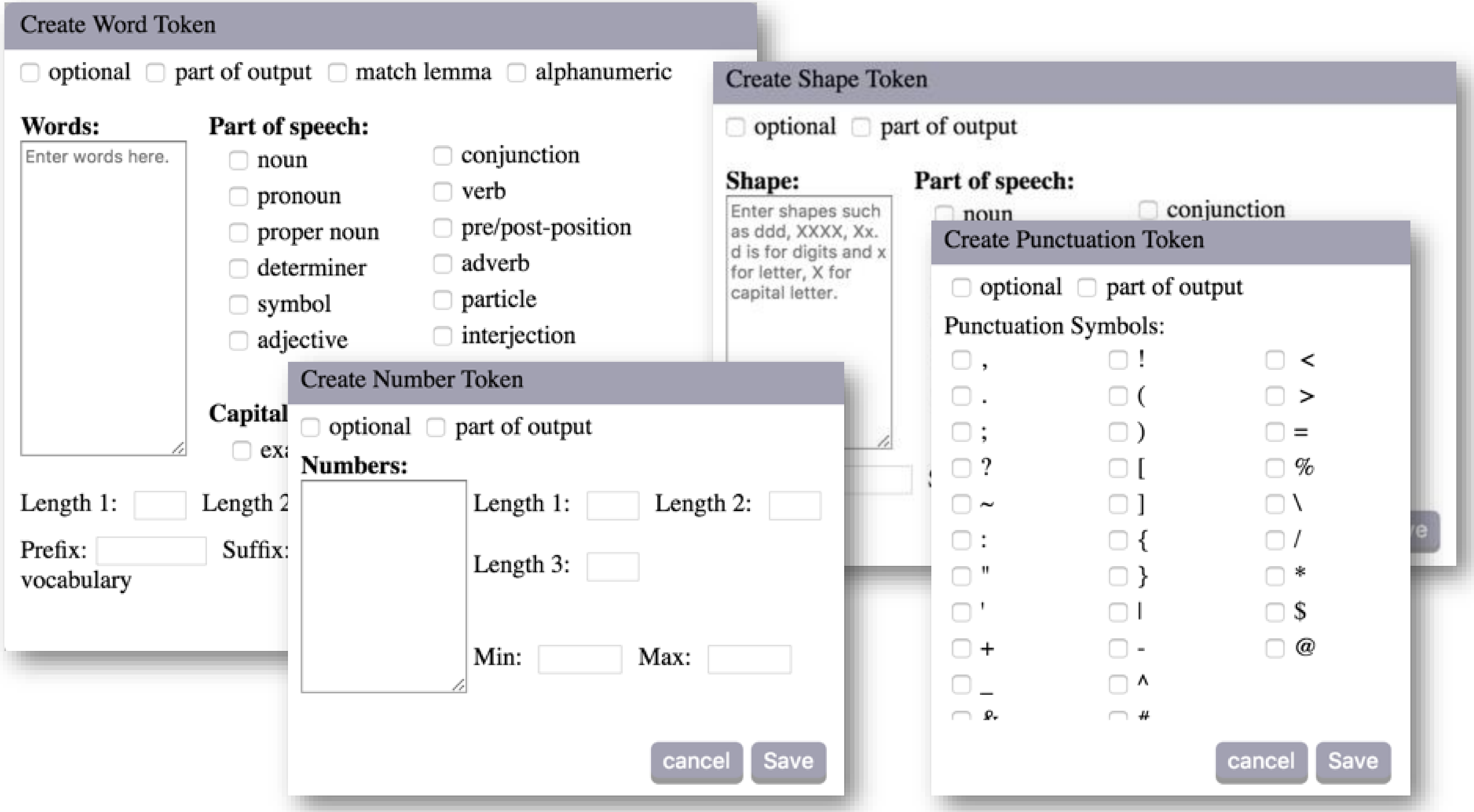
Create Number Token

□ optional □ part of output

Numbers:

Length 1: Length 2:

Length 3:



Create Word Token

□ optional □ part of output □ match lemma □ alphanumeric

Words:

Enter words here.

Part of speech:

□ noun
□ pronoun
□ proper noun
□ determiner
□ symbol
□ adjective
□ conjunction
□ verb
□ pre/post-position
□ adverb
□ particle
□ interjection

Capital

□ exa

Length 1: Length 2

Prefix: Suffix:

vocabulary

Create Shape Token

□ optional □ part of output

Shape:

Enter shapes such as ddd, XXXX, Xx. d is for digits and x for letter, X for capital letter.

Part of speech:

□ noun □ conjunction

Create Number Token

□ optional □ part of output

Numbers:

Length 1: Length 2:

Length 3:

Min: Max:

cancel Save

Create Punctuation Token

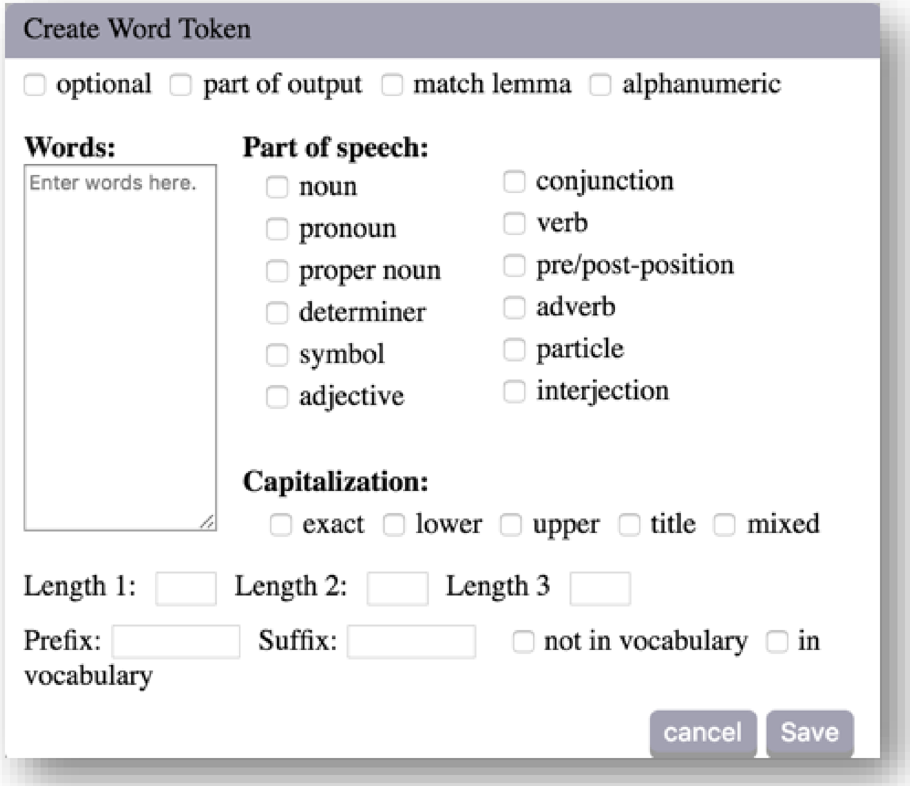
□ optional □ part of output

Punctuation Symbols:

□ ,
□ .
□ ;
□ ?
□ ~
□ :
□ "
□ '
□ +
□ _
□ &
□ !
□ (
□)
□ [
□]
□ {
□ }
□ |
□ -
□ ^
□ #
□ <
□ >
□ =
□ %
□ \
□ /
□ *
□ $
□ @

cancel Save

Token Types



Create Word Token

□ optional □ part of output □ match lemma □ alphanumeric

Words:

Enter words here.

Part of speech:

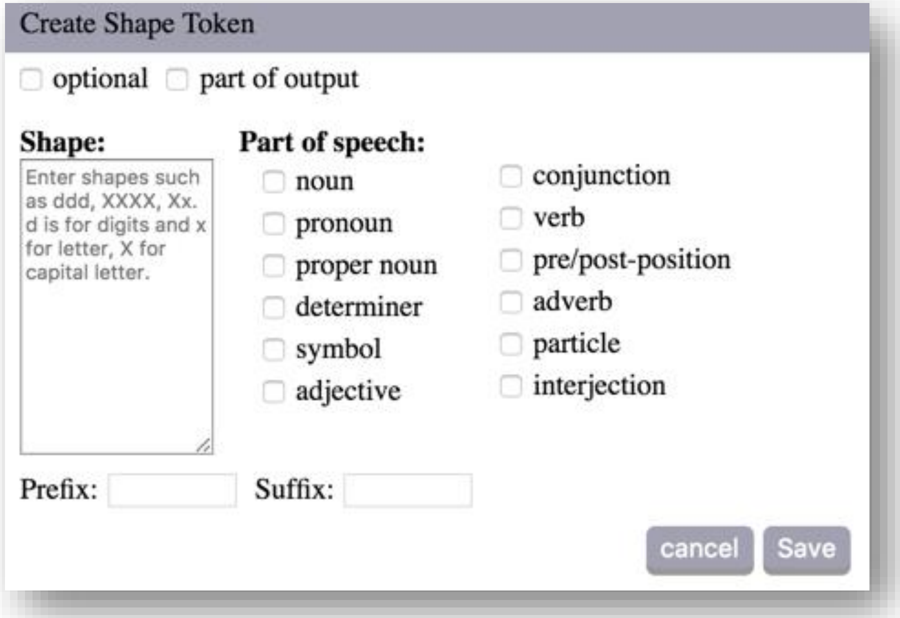
□ noun
□ pronoun
□ proper noun
□ determiner
□ symbol
□ adjective
□ conjunction
□ verb
□ pre/post-position
□ adverb
□ particle
□ interjection

Capitalization:

□ exact □ lower □ upper □ title □ mixed

Length 1: Length 2: Length 3

Prefix: Suffix: □ not in vocabulary □ in vocabulary



Create Shape Token

□ optional □ part of output

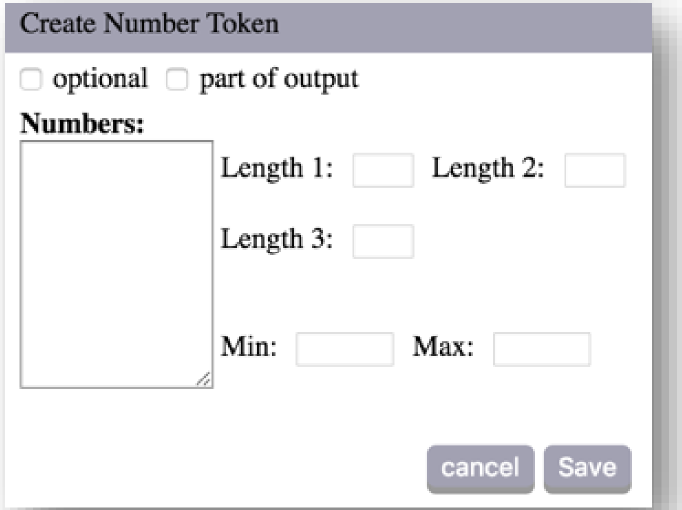
Shape:

Enter shapes such as ddd, XXXX, Xx. d is for digits and x for letter, X for capital letter.

Part of speech:

□ noun
□ pronoun
□ proper noun
□ determiner
□ symbol
□ adjective
□ conjunction
□ verb
□ pre/post-position
□ adverb
□ particle
□ interjection

Prefix: Suffix:



Create Number Token

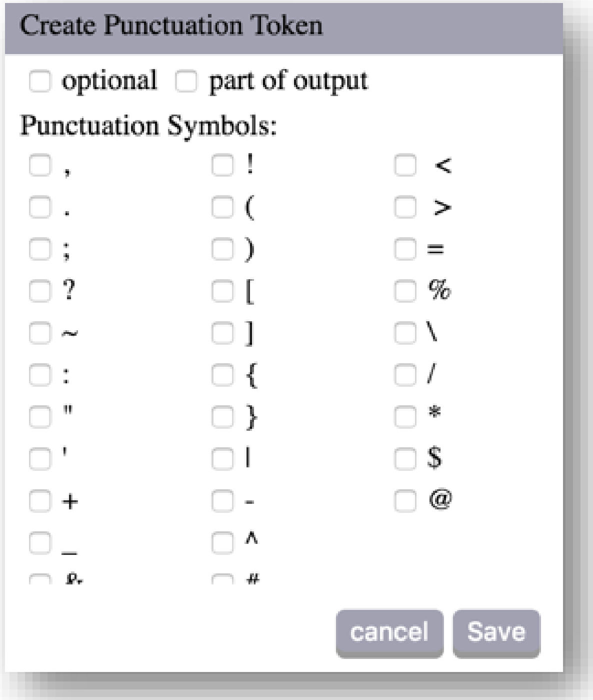
□ optional □ part of output

Numbers:

Length 1: Length 2:

Length 3:

Min: Max:



Create Punctuation Token

□ optional □ part of output

Punctuation Symbols:

□ ,
□ .
□ ;
□ ?
□ ~
□ :
□ "
□ '
□ +
□ _
□ &
□ !
□ (
□)
□ [
□]
□ {
□ }
□ |
□ -
□ ^
□ #
□ <
□ >
□ =
□ %
□ \
□ /
□ *
□ $
□ @

Patterns

Pattern := Token-Spec

[Token-Spec]	Optional
Token-Spec +	One or more
Token-Spec Pattern	

Positive/Negative Patterns

Positive

Generate candidates

Negative

Remove candidates

Output overlaps positive candidates

Positive/Negative Patterns

General **Positive**

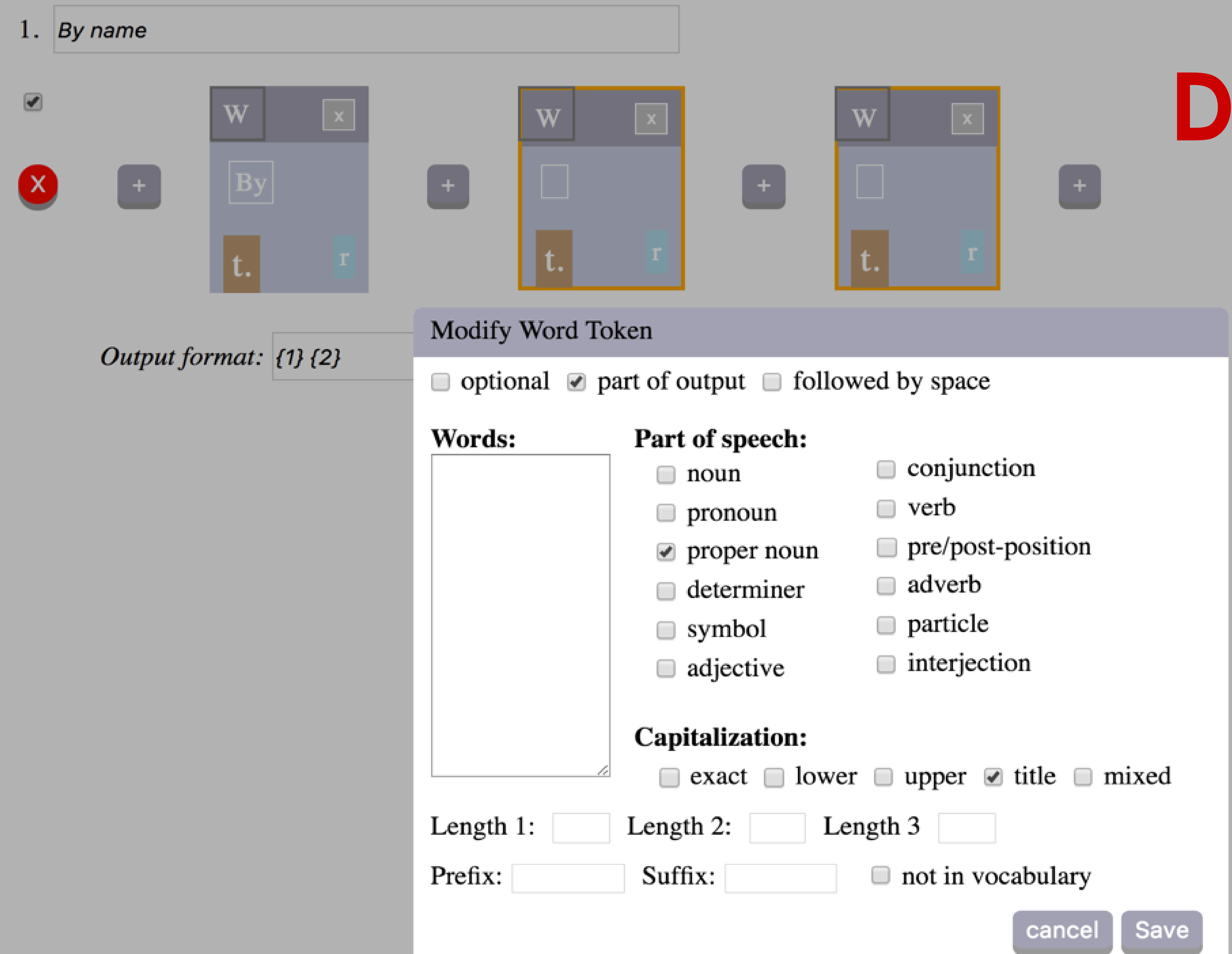
Generate candidates

Specific **Negative**

Remove candidates

Output overlaps positive candidates

DIG Demo



1. *By name*

☑

Output format: {1} {2}

Modify Word Token

□ optional ☑ part of output □ followed by space

Words:

Part of speech:

- □ noun
- □ pronoun
- ☑ proper noun
- □ determiner
- □ symbol
- □ adjective
- □ conjunction
- □ verb
- □ pre/post-position
- □ adverb
- □ particle
- □ interjection

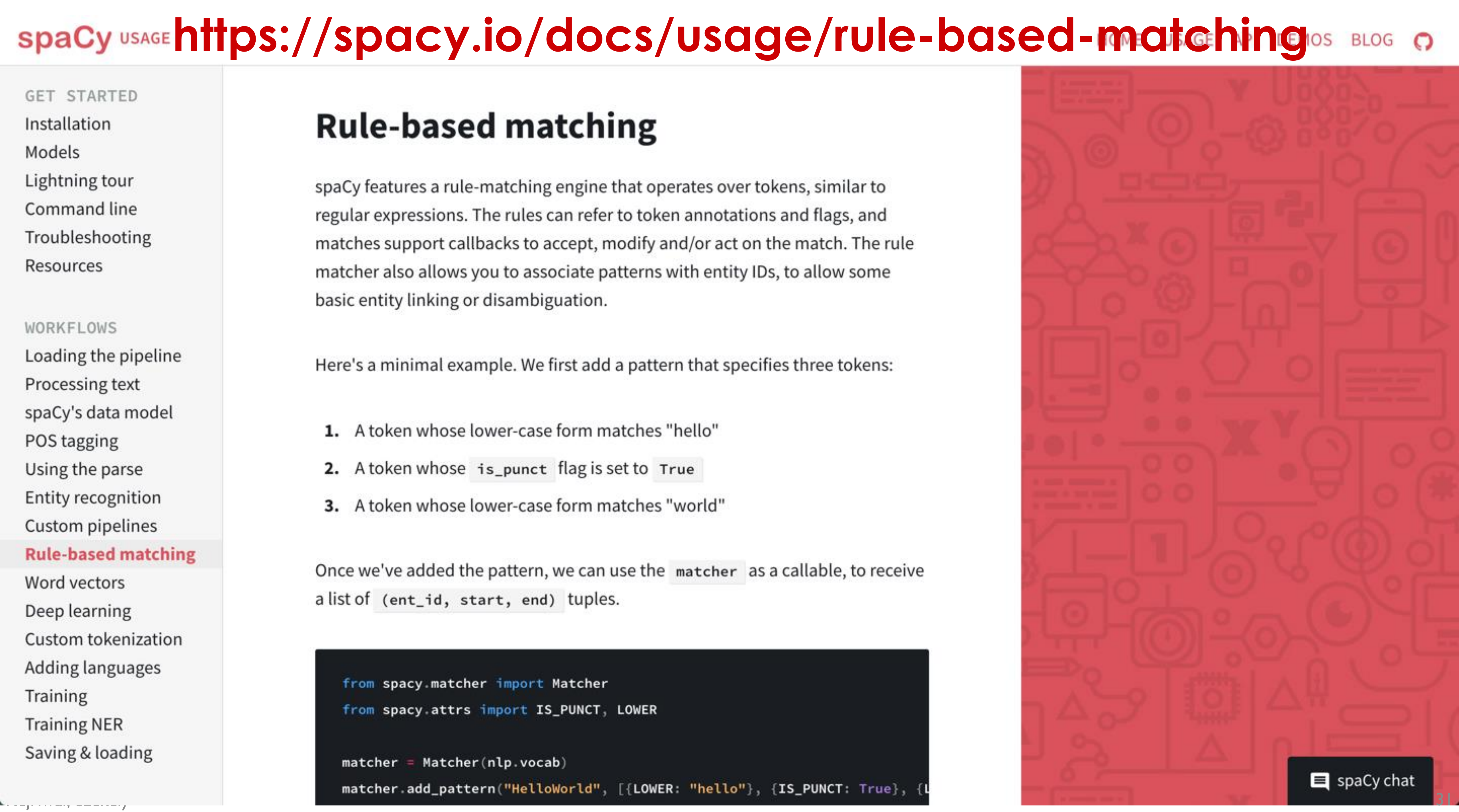
Capitalization:

□ exact □ lower □ upper ☑ title □ mixed

Length 1: Length 2: Length 3

Prefix: Suffix: □ not in vocabulary

https://spacy.io/docs/usage/rule-based-matching



Rule-based matching

spaCy features a rule-matching engine that operates over tokens, similar to regular expressions. The rules can refer to token annotations and flags, and matches support callbacks to accept, modify and/or act on the match. The rule matcher also allows you to associate patterns with entity IDs, to allow some basic entity linking or disambiguation.

Here's a minimal example. We first add a pattern that specifies three tokens:

1. A token whose lower-case form matches "hello"
2. A token whose `is_punct` flag is set to `True`
3. A token whose lower-case form matches "world"

Once we've added the pattern, we can use the `matcher` as a callable, to receive a list of `(ent_id, start, end)` tuples.

```
from spacy.matcher import Matcher
from spacy.attrs import IS_PUNCT, LOWER

matcher = Matcher(nlp.vocab)
matcher.add_pattern("HelloWorld", [{LOWER: "hello"}, {IS_PUNCT: True}, {L
```

Advantages/Disadvantages

Advantages

Easy to define

High precision

Recall increases with number of rules

Disadvantages

Text must follow strict patterns

NLP Rule-Based Extraction

Tokenization for unusual domains

tokenize on white-space, punctuation and emojis

Token properties

literal, part of speech tag, lemma, in/out of dictionary
dependency parsing relationships (advanced)
type (alphanumeric, alphabetic, numeric)
shape (pattern of digits and characters), capitalization, prefix and suffix
number of characters, range (numbers)

Pattern

Sequence of required/optional tokens
positive and negative patterns

Named-Entity Recognizers

Named Entity Recognizers

Machine learning models

people, places, organizations and a few others

SpaCy

complete NLP toolkit, Python (Cython), MIT license

code: https://github.com/explosion/spaCy

demo: http://textanalysisonline.com/spacy-named-entity-recognition-ner

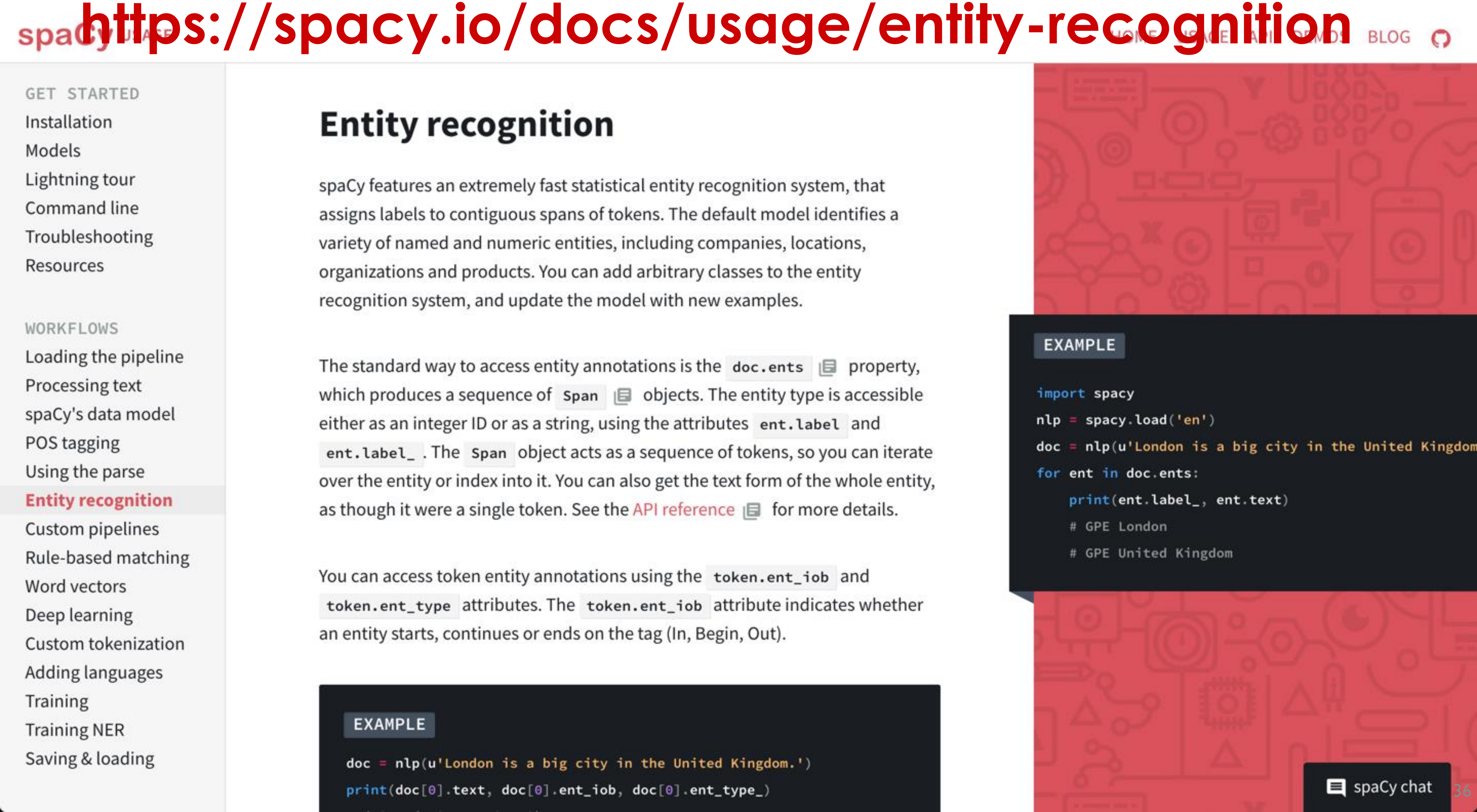
Stanford NER

part of Stanford’s NLP software library, Java, GNU license

code: https://nlp.stanford.edu/software/CRF-NER.shtml

demo: http://nlp.stanford.edu:8080/ner/process

https://spacy.io/docs/usage/entity-recognition



Entity recognition

spaCy features an extremely fast statistical entity recognition system, that assigns labels to contiguous spans of tokens. The default model identifies a variety of named and numeric entities, including companies, locations, organizations and products. You can add arbitrary classes to the entity recognition system, and update the model with new examples.

The standard way to access entity annotations is the `doc.ents` property, which produces a sequence of `Span` objects. The entity type is accessible either as an integer ID or as a string, using the attributes `ent.label` and `ent.label_`. The `Span` object acts as a sequence of tokens, so you can iterate over the entity or index into it. You can also get the text form of the whole entity, as though it were a single token. See the API reference for more details.

EXAMPLE

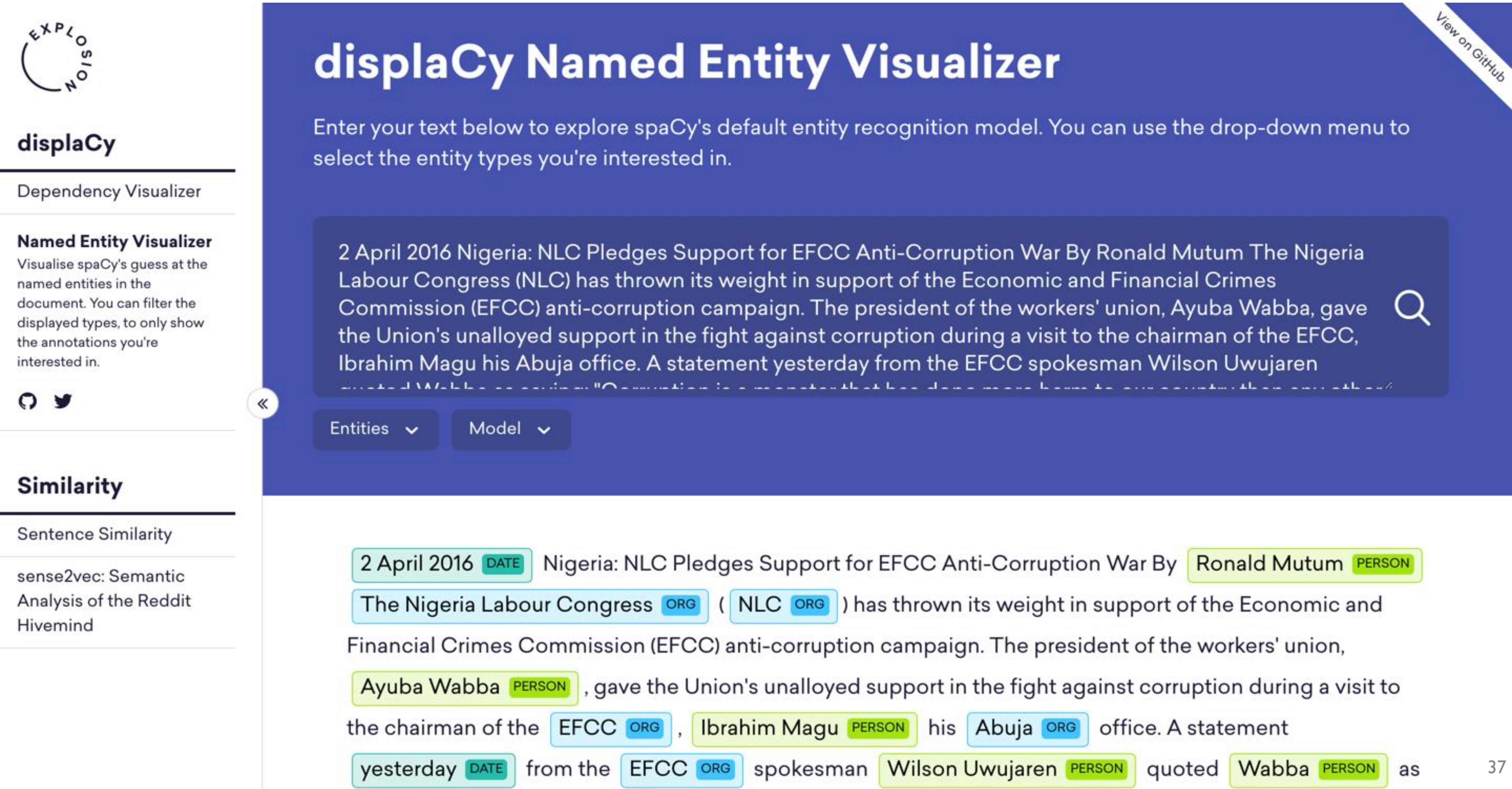
```
import spacy
nlp = spacy.load('en')
doc = nlp(u'London is a big city in the United Kingdom
for ent in doc.ents:
    print(ent.label_, ent.text)
    # GPE London
    # GPE United Kingdom
```

You can access token entity annotations using the `token.ent_iob` and `token.ent_type` attributes. The `token.ent_iob` attribute indicates whether an entity starts, continues or ends on the tag (In, Begin, Out).

EXAMPLE

```
doc = nlp(u'London is a big city in the United Kingdom.')
print(doc[0].text, doc[0].ent_iob, doc[0].ent_type_)
```

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



displaCy

Dependency Visualizer

Named Entity Visualizer
Visualise spaCy's guess at the named entities in the document. You can filter the displayed types, to only show the annotations you're interested in.

Similarity

Sentence Similarity

sense2vec: Semantic Analysis of the Reddit Hivemind

displaCy Named Entity Visualizer

Enter your text below to explore spaCy's default entity recognition model. You can use the drop-down menu to select the entity types you're interested in.

2 April 2016 Nigeria: NLC Pledges Support for EFCC Anti-Corruption War By Ronald Mutum The Nigeria Labour Congress (NLC) has thrown its weight in support of the Economic and Financial Crimes Commission (EFCC) anti-corruption campaign. The president of the workers' union, Ayuba Wabba, gave the Union's unalloyed support in the fight against corruption during a visit to the chairman of the EFCC, Ibrahim Magu his Abuja office. A statement yesterday from the EFCC spokesman Wilson Uwujaren quoted Wabba as saying: "Corruption is a monster that has done more harm to our country than any other

2 April 2016 DATE Nigeria: NLC Pledges Support for EFCC Anti-Corruption War By Ronald Mutum PERSON The Nigeria Labour Congress ORG (NLC ORG) has thrown its weight in support of the Economic and Financial Crimes Commission (EFCC) anti-corruption campaign. The president of the workers' union, Ayuba Wabba PERSON , gave the Union's unalloyed support in the fight against corruption during a visit to the chairman of the EFCC ORG , Ibrahim Magu PERSON his Abuja ORG office. A statement yesterday DATE from the EFCC ORG spokesman Wilson Uwujaren PERSON quoted Wabba PERSON as

Advantages/Disadvantages

Advantages

Easy to use

Tolerant of some noise

Easy to train

Disadvantages

Performance degrades rapidly for new genres, language models

Requires hundreds to thousands of training examples

Conditional Random Fields

Discriminative Vs. Generative

$p(\mathbf{y},\mathbf{x})$

- **Generative Model:** A model that generate observed data randomly
- **Naïve Bayes:** once the class label is known, all the features are independent

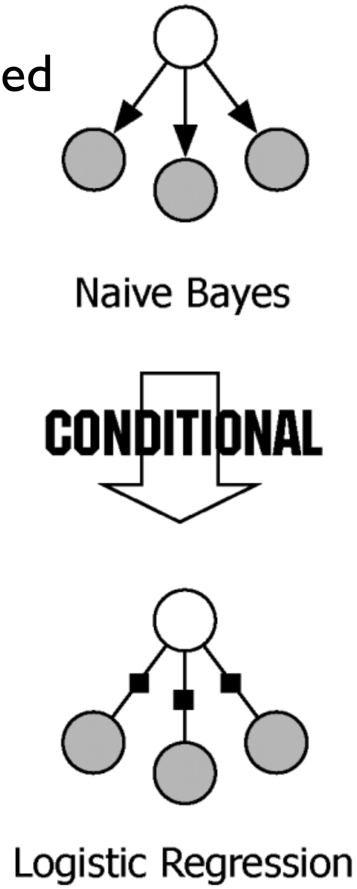
$$p(y,\mathbf{x}) = p(y)\prod_{k=1}^{K} p(x_k|y)$$

- **Discriminative:** Directly estimate the posterior probability; Aim at modeling the “discrimination” between different outputs

$p(\mathbf{y}\mid\mathbf{x})$

- **MaxEnt** classifier: linear combination of feature function in the exponent,

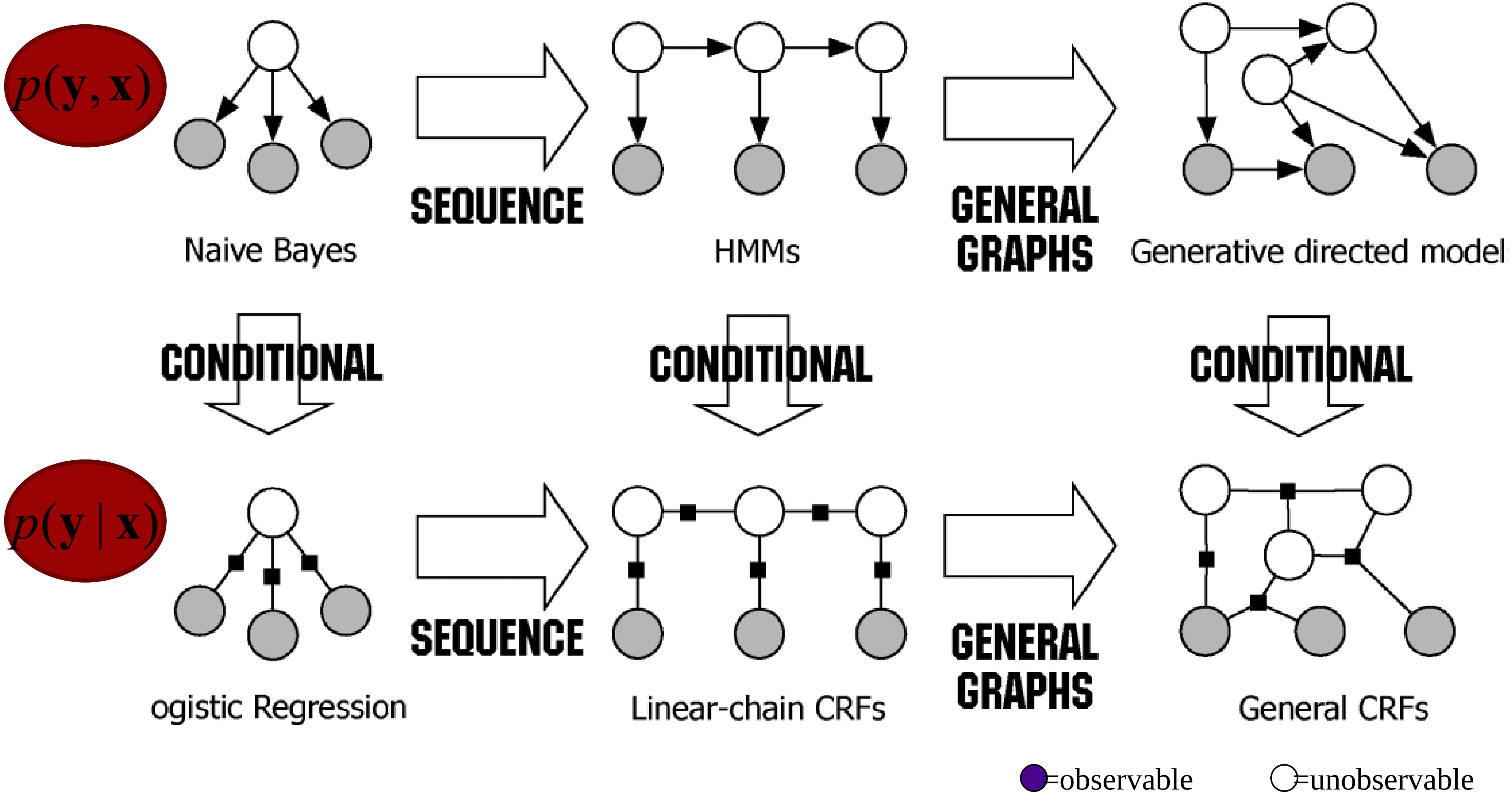
$$p(y|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\left\{\sum_{k=1}^{K}\theta_k f_k(y,\mathbf{x})\right\}$$



Both generative models and discriminative models describe distributions over (y , x), but they work in different directions.

slide by Daniel Khashabi

Discriminative Vs. Generative



slide by Daniel Khashabi

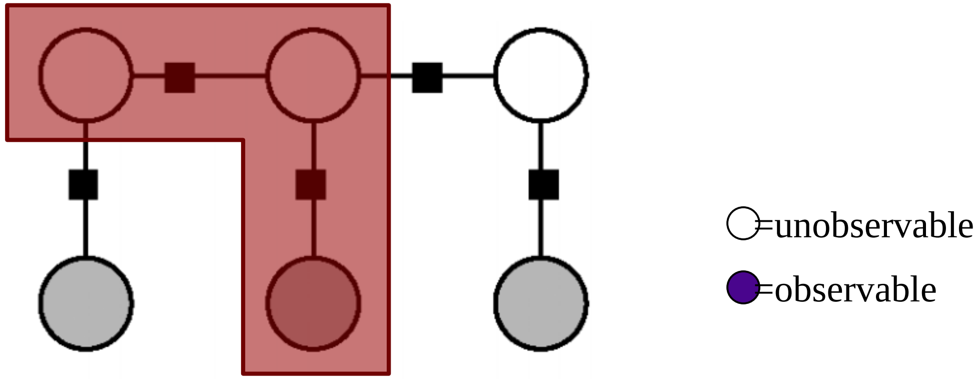
Chain CRFs

- Each potential function will operate on pairs of adjacent label variables

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x})}\exp\left(\sum_j \lambda_j F_j(\boldsymbol{y},\boldsymbol{x})\right)$$

$$F_j(\boldsymbol{y},\boldsymbol{x}) = \sum_{i=1} \boxed{f_j(y_{i-1},y_i,\boldsymbol{x},i),} \Longrightarrow \text{Feature functions}$$

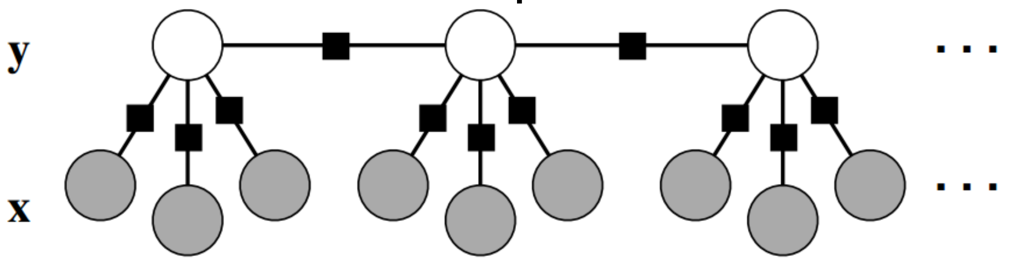
- Parameters to be estimated, λ_j



slide by Daniel Khashabi

Chain CRF

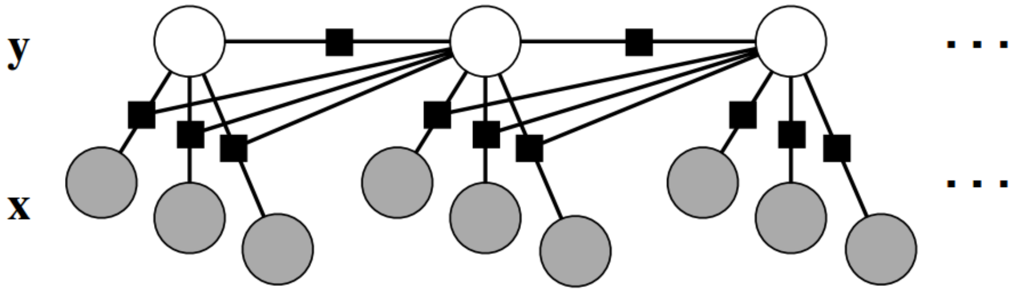
- We can change it so that each state depends on more observations



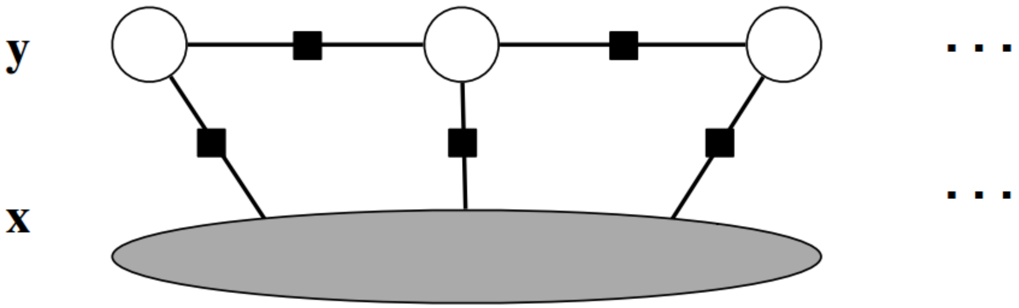
○=unobservable

●=observable

- Or inputs at previous steps



- Or all inputs



slide by Daniel Khashabi

Modeling Problems With CRF

i	X1 (word)	X2 (capitalized)	X3 (POS Tag)	Y (entity)
1	My	1	Possessive Pron	Other
2	name	0	Noun	Other
3	is	0	Verb	Other
4	Pedro	1	Proper Noun	Person-Name
5	Szekely	1	Proper Noun	Person-Name

Modeling Problems With CRF

i	X1 (word)	X2 (capitalized)	X3 (POS Tag)	Y (entity)
1	My	1	Possessive Pron	Other
2	name	0	Noun	Other
3	is	0	Verb	Other
4	Pedro	1	Proper Noun	Person-Name
5	Szekely	1	Proper Noun	Person-Name

Other common features:

lemma, prefix, suffix, length

Modeling Problems With CRF

i	X1 (word)	X2 (capitalized)	X3 (POS Tag)	Y (entity)
1	My	1	Possessive Pron	Other
2	name	0	Noun	Other
3	is	0	Verb	Other
4	Pedro	1	Proper Noun	Person-Name
5	Szekely	1	Proper Noun	Person-Name

feature functions $f_j(x, y_{i-1}, y_i, i)$

Advantages/Disadvantages

Advantages

Expressive

Tolerant of noise

Stood test of time

Software packages available

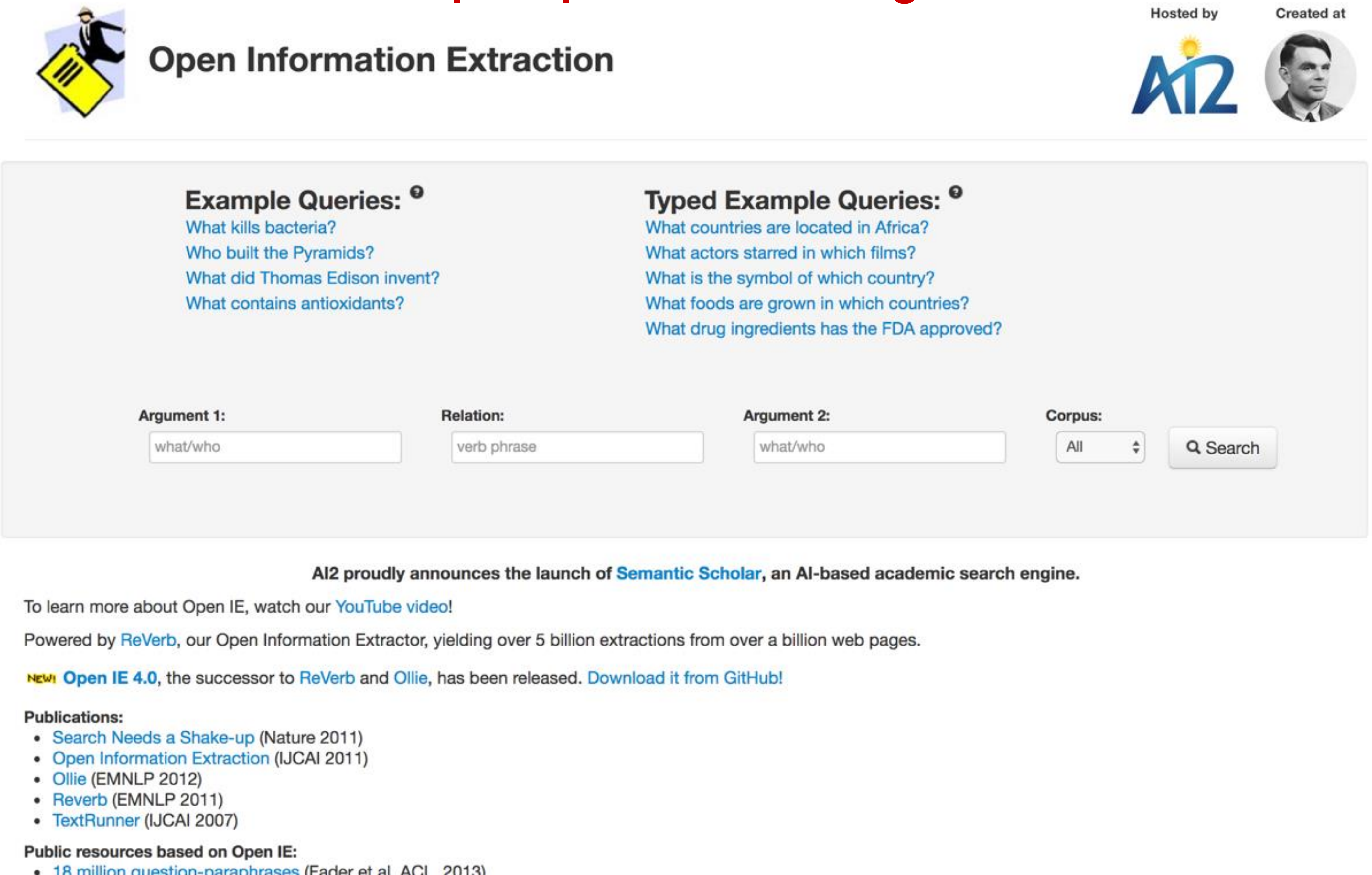
Disadvantages

Requires feature engineering

Requires thousands of training examples

Open Information Extraction

http://openie.allenai.org/



Open Information Extraction

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

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!

Publications:

- Search Needs a Shake-up (Nature 2011)
- Open Information Extraction (IJCAI 2011)
- Ollie (EMNLP 2012)
- Reverb (EMNLP 2011)
- TextRunner (IJCAI 2007)

Public resources based on Open IE:

- 18 million question-paraphrases (Fader et al. ACL, 2013)

Practical IE Technologies

	Glossary	Regex	NLP Rules	Semi-Structured	CRF	NER	Table
Effort	assemble glossary	hours	hours	minutes	O(1000) annotations	zero	O(10) annotations
Expertise	minimal	high, programmer	low	minimal	low-medium	zero	minimal
Precision	medium (ambiguity)	high	high	high	medium-high	medium-high	high
Recall	medium (formatting)	low f(# regex)	medium f(# rules)	high	medium	medium	high

how to **represent** KGs?

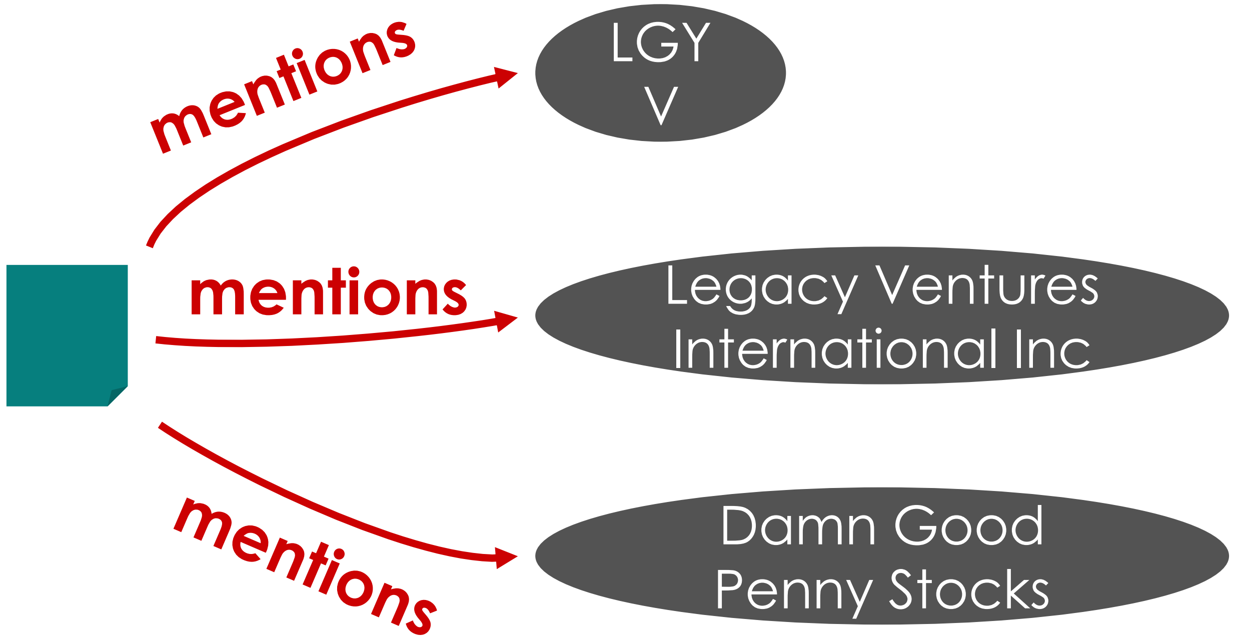
KG Definition

a directed, labeled multi-relational graph
representing facts/assertions as triples

(h, r, t)	head entity, relation, tail entity
(s, p, o)	subject, predicate, object

Simplest Knowledge Graph

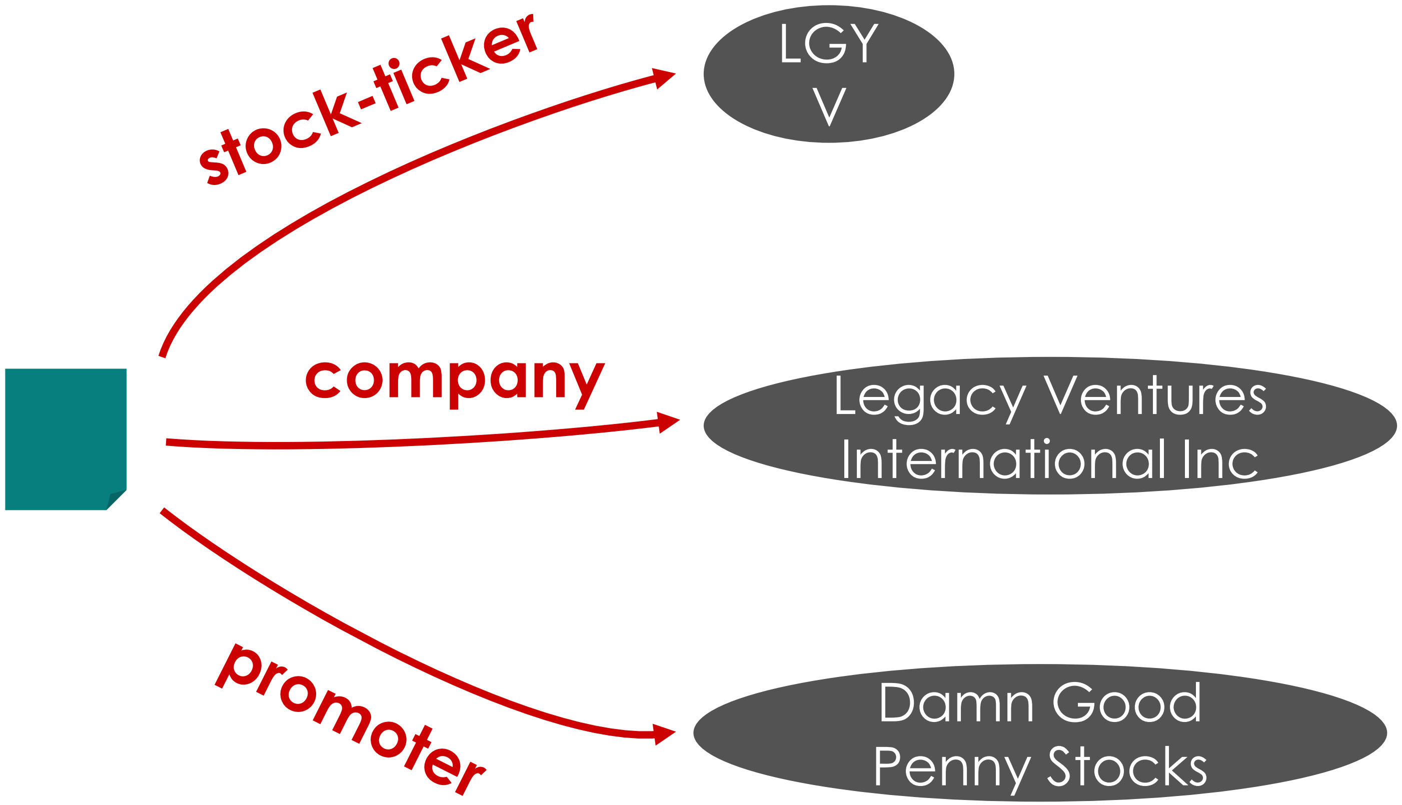
Entities



Easiest to build

Simple, But Useful KG

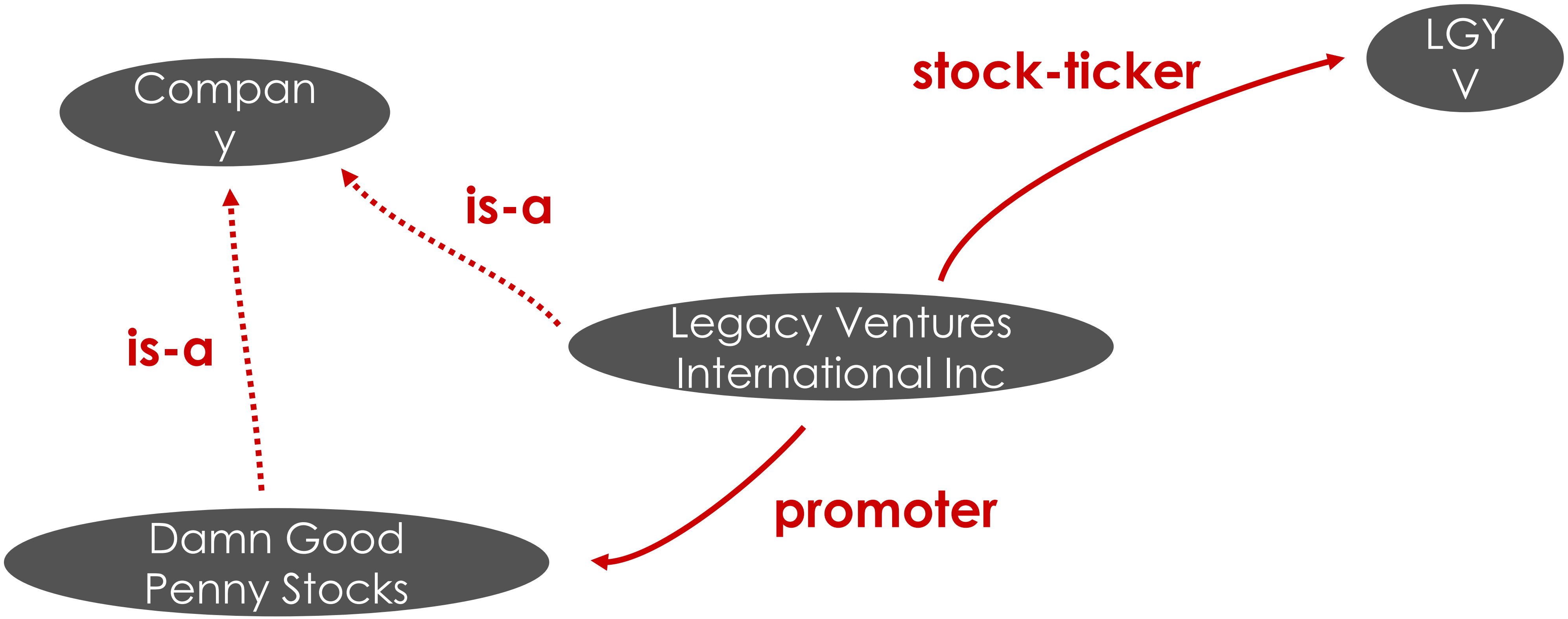
Entities + properties



"Easy" to build

Semantic Web KG (RDF/OWL)

Entities + properties + classes

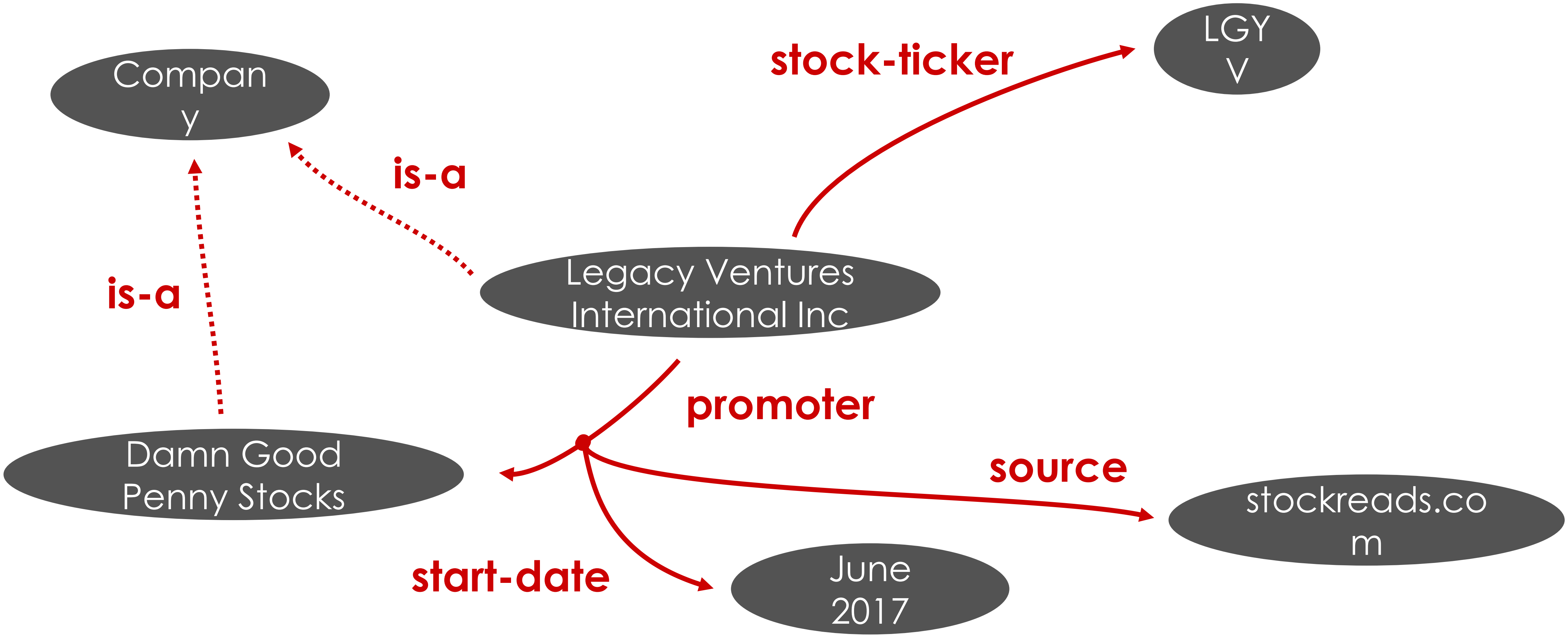


Very hard to build

Kejriwal, Szekely

“Ideal” KG

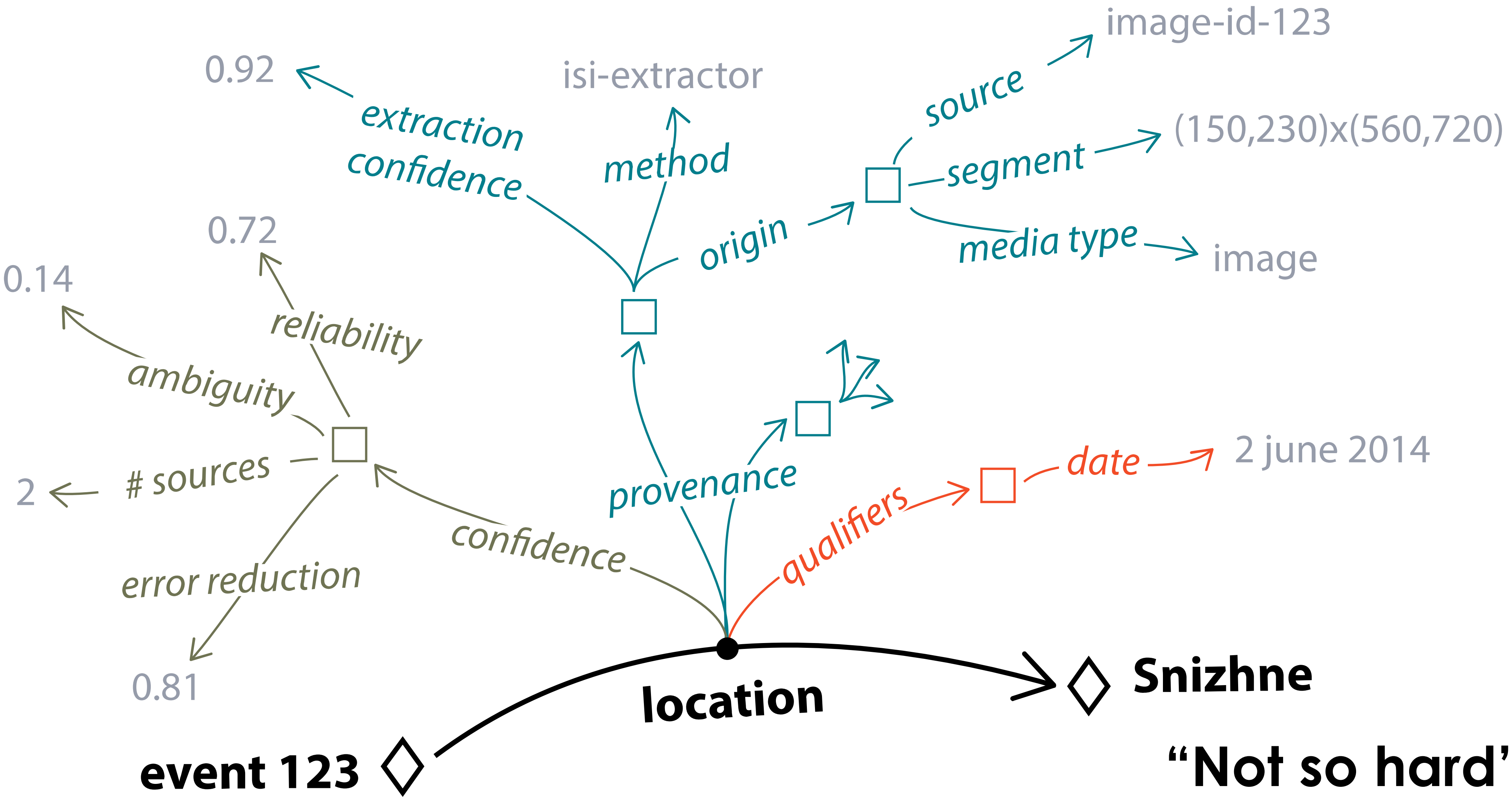
Entities + properties + classes + qualifiers



Very very hard to build

"More Ideal" KG

Entities + properties + provenance + confidence + qualifiers



"Not so hard" to build

Where to Store KGs?

Serializing Knowledge Graphs

Resource Description Framework (RDF)

Database (triple store): AllegroGraph, Virtuoso,
Query: SPARQL (SQL-like)

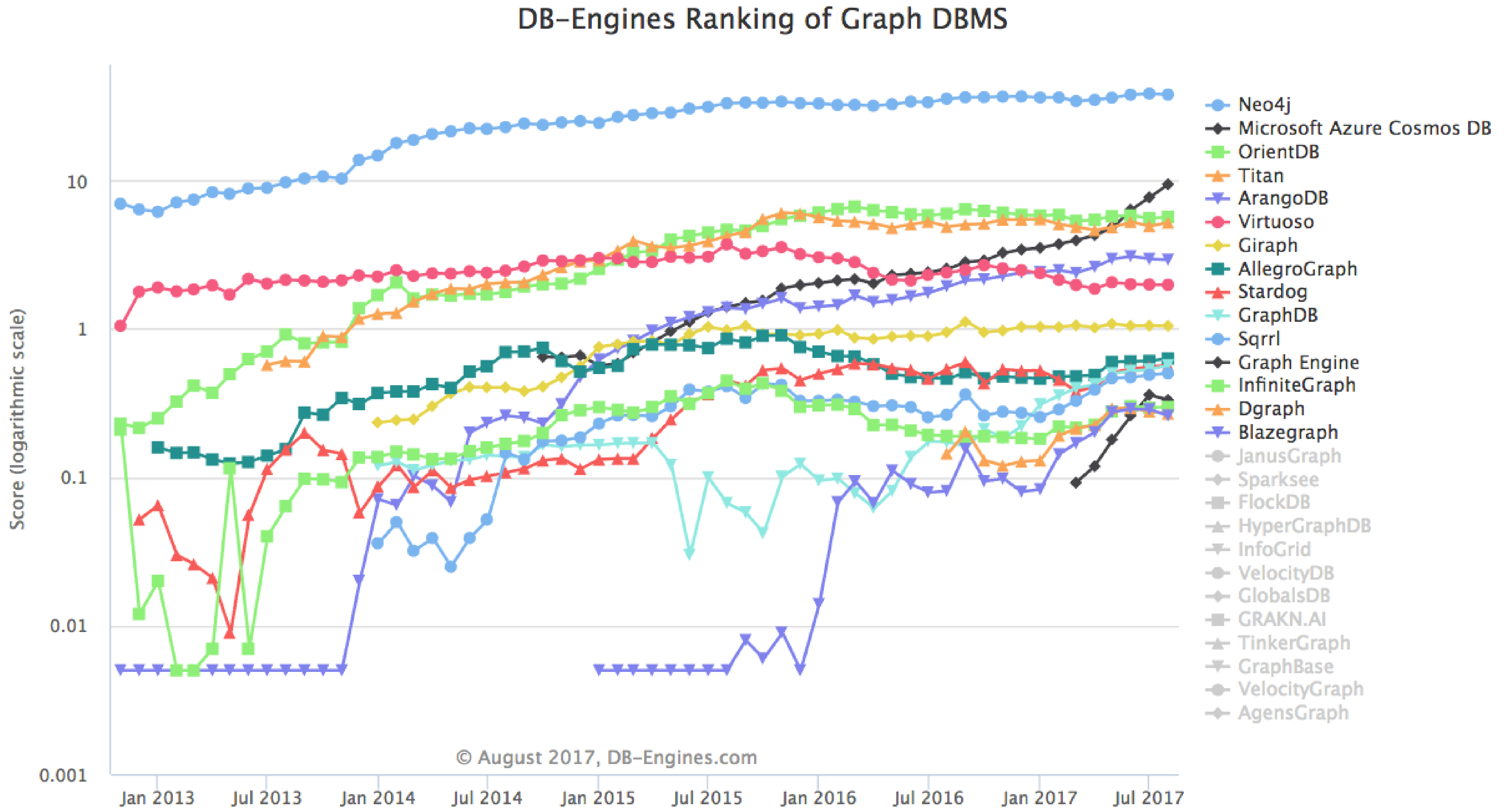
Key-Value, Document Stores

Data model: Node-centric
Databases: Hbase, MongoDB, Elastic Search, …
Query: filters, keywords, aggregation (no joins)

Graph Databases

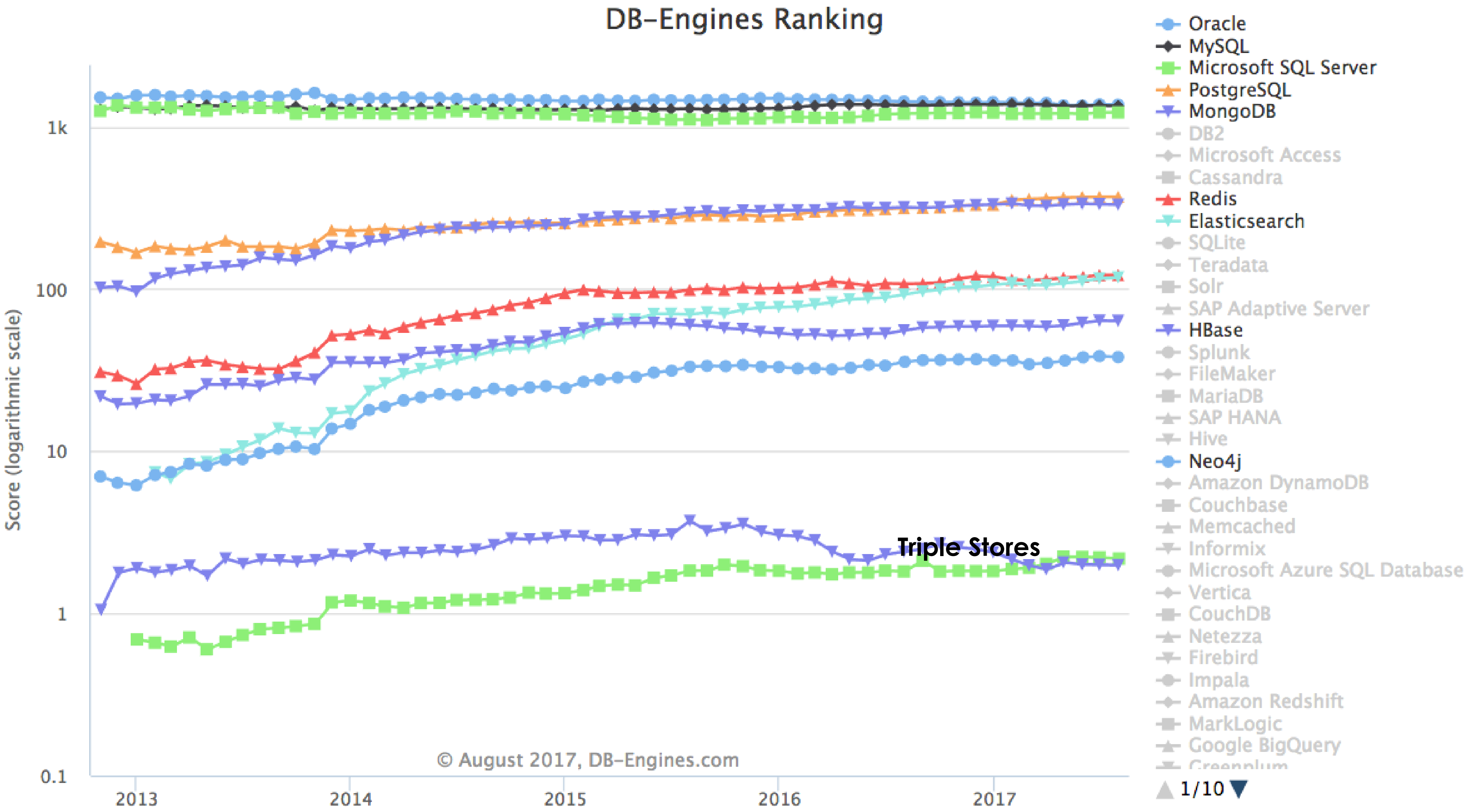
Data model: graph
Databases: Neo4J, Cayley, MarkLogic, GraphDB, Titan, OrientDB, Oracle, …
Query: GraphQL, Gremlin, Cypher

Popularity Ranking Of Graph Databases



https://db-engines.com/en/ranking_trend/graph+dbms

ElasticSearch, MongoDB & Neo4J Have Wide Adoption



https://db-engines.com/en/ranking_trend/graph+dbms

myDIG: A KG Construction Toolkit

Python, MIT license, https://github.com/usc-isi-i2/dig-etl-engine

Enable end-users to construct domain-specific KGs

end users from 5 government orgs constructed KGs in less than one day

Suite of extraction techniques

semi-structured HTML pages, glossaries, NLP rules, NER, tables (coming soon)

KG includes provenance and confidences

enable research to improve extractions and KG quality

Scalable

runs on laptop (~100K docs), cluster (> 100M docs)

Robust

Deployed to many law enforcement agencies

Easy to install

Docker deployment with single "docker compose up" installation

myDIG Demo

Summary

Partition pages into segments

Select technology based on segment features

Do knowledge graph completion (next topic)

Choose representation based on application demands