

Information Extraction

Regular Expressions and Beyond

Will Thompson, Ph.D.

February 22, 2019

Table of contents

1. Introduction
2. Regular Expressions
3. Advanced Pattern Matching

Introduction

Natural Language Processing: Use Cases

Some practical use cases for NLP:

- Text exploration (keyphrases, topic models)
- Text classification (including sentiment analysis)
- Information Extraction (unstructured → structured)

Unstructured Data

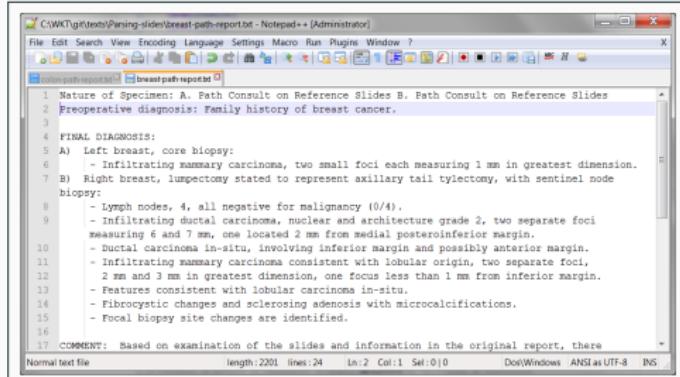
Vast quantities of information are encoded as **unstructured data**, in the form of natural language text.

But it can be hard to make this information available for computational analysis at scale.



Unstructured Data

What a **human** sees:



A screenshot of the Notepad++ application window. The title bar reads "C:\WC\git\tests\Parsing-slides\breast-path-report.txt - Notepad++ (Administrator)". The menu bar includes File, Edit, Search, View, Encoding, Language, Settings, Macro, Run, Plugins, and Window. The toolbar has various icons for file operations. The main text area contains a medical report for a breast biopsy. The report details the nature of the specimen, preoperative diagnosis (family history of breast cancer), and final diagnosis, which includes infiltrating mammary carcinoma, ductal carcinoma in-situ, and lobular carcinoma in-situ. It also lists lymph nodes, features, and other findings. A comment at the end states: "COMMENT: Based on examination of the slides and information in the original report, there". The status bar at the bottom shows "Normal text file", "length:2201", "lines:24", "Col:1", "Sel:0|0", "Dos/Windows", "ANSI as UTF-8", and "INS".

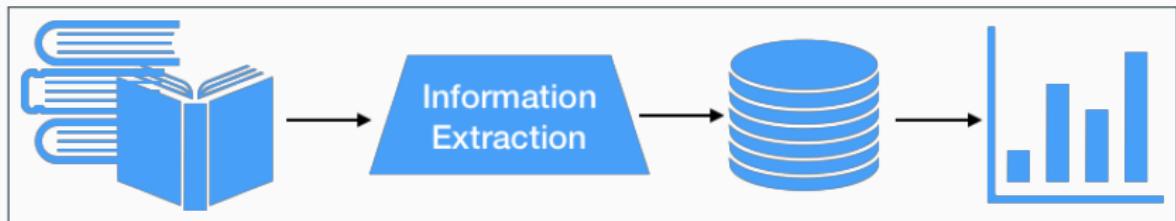
What a **computer** sees:

0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
4e	61	74	75	72	65	20	6f	66	20	53	70	65	63	69	6d
65	6e	3a	20	41	2e	20	50	61	74	68	20	43	6f	6e	73
75	6c	74	20	6f	6e	20	52	65	66	65	72	65	6e	63	65
20	53	6c	69	64	65	73	20	42	2e	20	50	61	74	68	20
43	6f	6e	73	75	6c	74	20	6f	6e	20	52	65	66	65	72
65	6e	63	65	20	53	6c	69	64	65	73	0d	0a	0d	0a	46
49	4e	41	4c	20	44	49	41	47	4e	4f	53	49	53	3a	0d
0a	41	29	20	20	4c	65	66	74	20	62	72	65	61	73	74
2c	20	63	6f	72	65	20	62	69	6f	70	73	79	3a	0d	0a
20	20	20	20	2d	20	49	6e	66	69	6c	74	72	61	74	
69	6e	67	20	6d	61	6d	6d	61	72	79	20	63	61	72	63

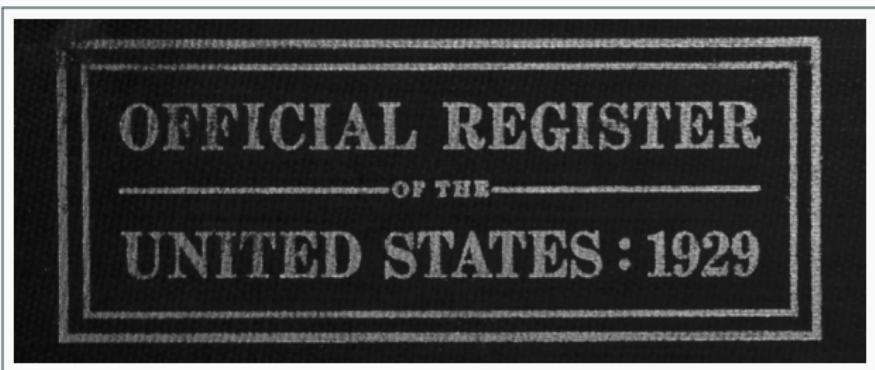
Information Extraction

Information extraction (IE) can be used to convert some of the information stored in text into **structured data**:

1. Search for relevant chunks of information in a corpus
2. Map these chunks to structured representations (potentially including relationships)
3. Store the results in structured format for downstream analysis



Example Use Case



Example Use Case

Official Register of the United States, 1929

7

GOVERNMENT PRINTING OFFICE

NAME	OFFICIAL TITLE	Legal residence		Com-pen-sa-tion
		State	Cong. Dist.	
George H. Carter.....	Public Printer.....	Iowa.....	9th.....	\$10,000
John Greene.....	Deputy Public Printer.....	Mass.....	5th.....	7,500
Mary A. Tate.....	Assistant to the Public Printer.....	Tenn.....	2d.....	4,000
Henry H. Wright.....	Chief clerk.....	N. Y.....	28th.....	4,200
Edward J. Wilver.....	Disbursing clerk.....	Pa.....	16th.....	4,000
William A. Smith.....	Congressional Record clerk.....	D. C.....		4,000
Daniel P. Bush.....	Medical and sanitary officer.....	Nebr.....	1st.....	4,200
James K. Wallace.....	Superintendent of accounts and budget officer.....	Ohio.....	1st.....	5,000
Ernest E. Emerson.....	Purchasing agent.....	Md.....	5th.....	4,600
Edward O. Reed.....	Technical director.....	D. C.....		5,200
Burr G. Williams.....	Chief instructor of apprentices.....	Iowa.....	10th.....	3,600
Ellwood S. Moorhead.....	Production manager.....	Pa.....	6th.....	5,200
Edward A. Huse.....	Night assistant production manager.....	Mass.....	6th.....	4,800
Hermann B. Barnhart.....	Superintendent of printing.....	Ind.....	9th.....	4,400
Bert E. Bair.....	Superintendent of presswork.....	Mich.....	6th.....	4,400
Martin R. Speelman.....	Superintendent of binding.....	Mo.....	4th.....	4,400
Edward G. Whall.....	Superintendent of platemaking.....	Mass.....	12th.....	4,400
William A. Mitchell.....	Superintendent of planning.....	N. C.....	5th.....	4,400
Alfred E. Hanson.....	Superintendent of construction and maintenance.....	Mass.....	14th.....	5,200
Alton P. Tisdel.....	Superintendent of documents.....	Ohio.....	22d.....	5,000
William H. Kervin.....	Storekeeper and traffic manager.....	N. Y.....	39th.....	4,200

LIBRARY OF CONGRESS

Herbert Putnam.....	Librarian.....	Mass.....		\$10,000
Frederick W. Ashley.....	Chief, assistant librarian.....	Ohio.....	22d.....	7,000
Allen R. Boyd.....	Executive assistant.....	Pa.....		4,800

Regular Expressions

Regular Expressions

- Regular expressions (regex) are a compact and expressive mini-language for defining patterns over text
- A regex matches a set of strings
- Regex patterns can contain:
 - Sequences: abc
 - Disjunctions: (a|b|c)
 - Repetitions: [A-Z]*, [0-9]+, a{2,3}b+
 - Ranges: [A-Z], [a-z], [0-9]
 - Classes: \w, \s, .
 - Optionality: colou?rs?
 - Anchors: ^ABC[0-9]+abc\$

<http://web.stanford.edu/~jurafsky/slp3>

Finite State Machines

The top section displays a Finite State Machine (FSM) diagram for a regular expression. The start state is a black circle with an incoming arrow. It transitions to a state with a '^' symbol, which then transitions to a state with '\s'. This is followed by a sequence of states: 'One of [A-Z 0]', a dot, 'One of [a-z]', another dot, 'One of [\s - A-Z 0]', another dot, and 'One of [a-z]'. A bracket labeled 'Group 1' covers the first four states. This is followed by a state with 'r', another state with '\s', and finally a state with 'j r'. The FSM ends at a final state with an outgoing arrow.

The bottom section shows the generated Python code:

```
1 ^\s?([A-Z0]\.?[a-z]*(?:\s[A-Z0]\.?[a-z]*)*(1,3)(?:,\s*jr)?)
```

The result of the regex match is shown in the 'Matches' field:

```
1 | George H. James, jr
```

<https://www.debuggex.com>

<https://jex.im/regulex>

Chomsky Hierarchy

Grammar	Languages	Automaton	Production rules (constraints)*	Examples ^[2]
Type-0	Recursively enumerable	Turing machine	$\alpha \rightarrow \beta$ (where α contains at least one non-terminal)	$L = \{w w \text{ describes a terminating Turing machine}\}$
Type-1	Context-sensitive	Linear-bounded non-deterministic Turing machine	$\alpha A \beta \rightarrow \alpha \gamma \beta$	$L = \{a^n b^n c^n n > 0\}$
Type-2	Context-free	Non-deterministic pushdown automaton	$A \rightarrow \alpha$	$L = \{a^n b^n n > 0\}$
Type-3	Regular	Finite state automaton	$A \rightarrow a$ and $A \rightarrow aB$	$L = \{a^n n \geq 0\}$

Regular Expressions \equiv Regular Languages \equiv Finite State Machines

Demo: Developing Regular Expressions

The screenshot shows the regex101.com interface. On the left, there's a sidebar with 'SAVE & SHARE' (Save Regex, copy, etc.), 'FLAVOR' (set to Python), 'TOOLS' (Code Generator), and a 'SPONSOR' section for Atlassian. The main area has a 'REGULAR EXPRESSION' input field containing `/^\b([0-9]{1,2}(?:d|th|st))`. Below it is a 'TEST STRING' table with several rows of names and titles, each with a green highlighted part corresponding to the regex match. The table includes columns for 'TEST STRING', 'Matched Text', and 'Description'. The rows are:

TEST STRING	Matched Text	Description
4, 200		
James K. Wallace		Superintendent of accounts and budget officer
_ 5, 000		Ohio
Ernest E. Emerson		1st..
5th..... 4, 600		
Edward O. Reed		Purchasing agent
5, 200		Md ..
Burr G. Williams		Technical director
3, 600		D C
Ellwood S. Moorhead		Chief instructor of apprentices
6th..... 5, 200		Iowa.... 10th
Edward A. Huse		
6th..... 4, 800		
Hermann B. Barnhart		Production manager
9th..... 4, 400		Pa
		Night assistant production manager
		Mass..
		Superintendent of printing
		Ind..

Below the table is a 'SUBSTITUTION' section.

<https://regex101.com>
(For fun: <https://regexecrossword.com>)

Demo: Complete Working Example

The screenshot shows a Jupyter Notebook interface with the following details:

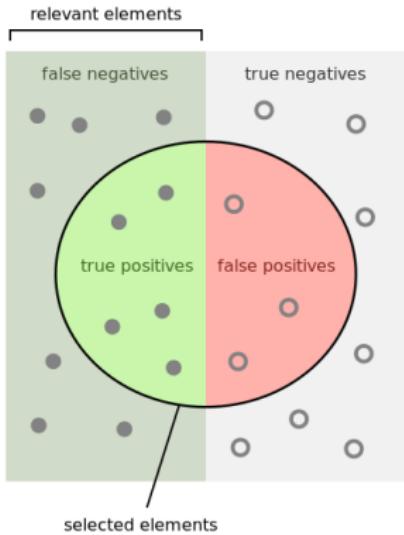
- Tab Bar:** Launcher (selected), extract-records.ipynb, Python 3.
- Cell 4:**

```
[4]: path = pathlib.Path.cwd() / 'excerpt-p17.txt'
```
- Cell 5:**

```
[5]: with open(path, mode='r', encoding='utf-16-le') as fid:  
    lines = [line for line in fid]
```
- Section Header:** Extract Fields
- Cell 6:**

```
[6]: class Record:  
    """  
        A simple data class containing fields for each extracted line  
    """  
    def __init__(self, dept, name, title, state, dist, salary):  
        self.dept, self.name, self.title, self.state, self.dist, self.salary = \  
            dept, name, title, state, dist, salary  
  
    def __str__(self):  
        return f"DEPT: {self.dept}\nNAME: {self.name}\nTITLE: {self.title}\nSTATE: {self.state}\nDIST: {self.dist}"  
  
    def to_dict(self):  
        return {  
            'dept':self.dept,  
            'name':self.name,  
            'title':self.title,  
            'state':self.state,  
            'dist':self.dist,  
            'salary':self.salary}
```

Evaluating Performance



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$


How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$


sensitivity, recall, hit rate, or true positive rate (TPR)

$$\text{TPR} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

specificity or true negative rate (TNR)

$$\text{TNR} = \frac{\text{TN}}{N} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

precision or positive predictive value (PPV)

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

negative predictive value (NPV)

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

false out or false positive rate (FPR)

$$\text{FPR} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR}$$

false discovery rate (FDR)

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}} = 1 - \text{PPV}$$

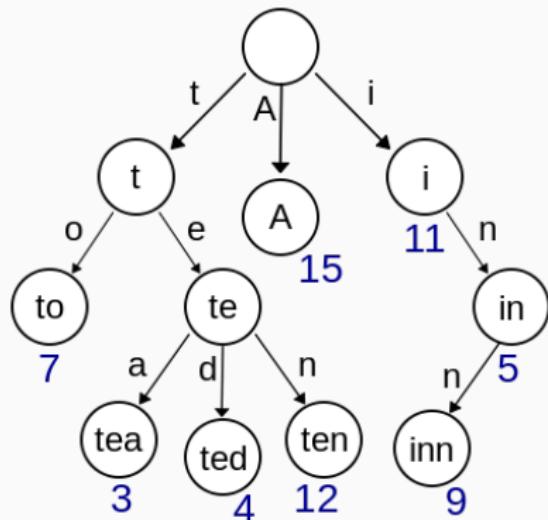
miss rate or false negative rate (FNR)

$$\text{FNR} = \frac{\text{FN}}{P} = \frac{\text{FN}}{\text{FN} + \text{TP}} = 1 - \text{TPR}$$

String Distance Algorithms

I N T E * N T I O N
| | | | | | | |
* E X E C U T I O N

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
#	E	X	E	C	U	T	I	O	N	



<https://phiresky.github.io/levenshtein-demo>

Advanced Pattern Matching

Annotation Patterns

```
def regex = new AnnotationRegex(  
    (DictMatch, [code:INVASIVE])  
    (DictMatch, [code:CANCER])  
    ( (Token, [text:/with|having/])(0,1)  
      (DictMatch, [code:LOBULAR|DUCTAL])  
      (Token, [text:/features/]) )(0,1)  
)
```

- Regular expressions can rapidly become very complex and difficult to modify and maintain
- Solution: **annotation-level regular expressions**
- Create cascades of annotations, each level at a higher level of abstraction

Example: CT Scan



FINDINGS: A 2.1 x 2.6 cm intraparenchymal hematoma is noted in the right high, posterior frontal lobe. There are smaller foci of hemorrhage noted anterior and posterior to this hematoma with another focus of hemorrhage in the right parietal lobe measuring 7 x 9 mm. The intraparenchymal hemorrhages have increased in size when compared to the outside head CT from 11/9/2015. There is associated low-attenuation surrounding the hematoma, compatible with vasogenic edema.

A more ventral region of low-attenuation is noted in the right frontal lobe (series 4, image 26), which is nonspecific and may represent an area of gliosis/encephalomalacia or small vessel ischemic disease. There is no midline shift and the basilar cisterns are preserved.

There is moderate global parenchymal volume loss with corresponding size of the ventricles. Scattered areas of white matter low attenuation are compatible with small vessel ischemic disease. Left-sided lens implant is noted. The visible portions of the paranasal sinuses, as well as the mastoid air cells and middle ear cavities, are clear.

IMPRESSION: Acute intraparenchymal hemorrhages are noted in the right frontal and parietal lobes. Compared to the earliest CT from 11/9/2015 at the outside facility, there has been a mild increase in the size of these hemorrhages. There is no evidence of midline shift or basilar cistern effacement.

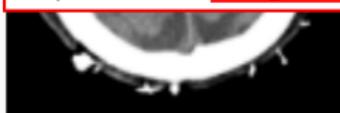
What factors influence use of prophylactic seizure medications in patients with intracerebral hemorrhage?

Example: CT Scan



FINDINGS: A 2.1 x 2.6 cm intraparenchymal hematoma is noted in the right high, posterior frontal lobe. There are smaller foci of hemorrhage noted anterior and posterior to this hematoma with another focus of hemorrhage in the right parietal lobe measuring 7 x 9 mm. The intraparenchymal hemorrhages have increased in size when compared to the outside head

FINDINGS: A 2.1 x 2.6 cm intraparenchymal hematoma is noted in the right high, posterior frontal lobe. There are smaller foci of hemorrhage noted anterior and posterior to this hematoma with another focus of hemorrhage in the right parietal lobe measuring 7 x 9 mm. The intraparenchymal hemorrhages have increased in size when compared to the outside head CT from 11/9/2015. There is associated low-attenuation surrounding the hematoma, compatible with vasogenic edema.



IMPRESSION: Acute intraparenchymal hemorrhages are noted in the right frontal and parietal lobes. Compared to the earliest CT from 11/9/2015 at the outside facility, there has been a mild increase in the size of these hemorrhages. There is no evidence of midline shift or basilar cistern effacement.

What factors influence use of prophylactic seizure medications in patients with intracerebral hemorrhage?

Concepts

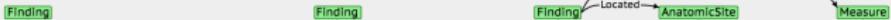
- 1 FINDINGS: A 2.1 x 2.6 cm intraparenchymal Finding AnatomicSite is noted in the right high, posterior frontal lobe.
- 2 There are smaller foci of hemorrhage noted anterior and posterior to this hematoma with another focus of hemorrhage in the right parietal lobe measuring 7 x 9 mm. Finding Finding AnatomicSite
- 3 The intraparenchymal hemorrhages have increased in size when compared to the outside head CT from 11/9/2015. Finding
- 4 There is associated low-attenuation surrounding the hematoma, compatible with vasogenic edema. Finding
- 5 A more ventral region of low-attenuation is noted in the right frontal lobe (series 4, image 26), which is nonspecific and may represent an area of gliosis/encephalomalacia. Finding
- 6 There is no midline shift and the basilar cisterns are preserved. Finding

<u>Concept</u>	<u>Type</u>	<u>Count</u>
right high posterior frontal lobe	Site	1
right parietal lobe	Site	1
hematoma	Finding	3
hemorrhage	Finding	3
vasogenic edema	Finding	1
encephalomalacia	Finding	1
midline shift	Finding	1

Concept Relations



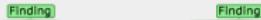
1 FINDINGS: A 2.1 x 2.6 cm intraparenchymal hematoma is noted in the right high, posterior frontal lobe.



2 There are smaller foci of hemorrhage noted anterior and posterior to this hematoma with another focus of hemorrhage in the right parietal lobe measuring 7 x 9 mm.



3 The intraparenchymal hemorrhages have increased in size when compared to the outside head CT from 11/9/2015.



4 There is associated low-attenuation surrounding the hematoma, compatible with vasogenic edema.



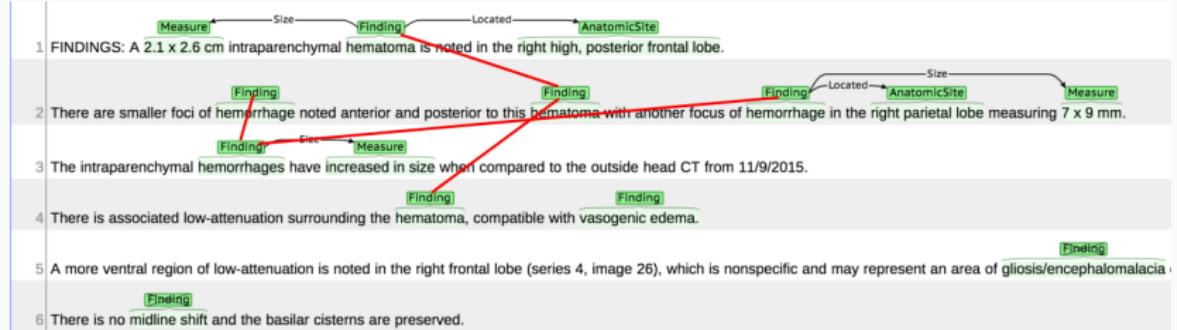
5 A more ventral region of low-attenuation is noted in the right frontal lobe (series 4, image 26), which is nonspecific and may represent an area of gliosis/encephalomalacia.



6 There is no midline shift and the basilar cisterns are preserved.

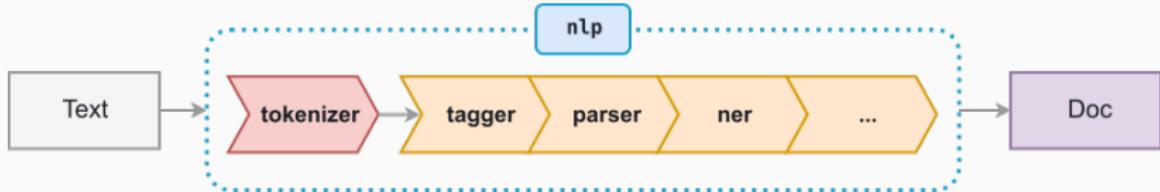
Concept	Type	Status	Site	Size
hematoma	Finding	Asserted	right high posterior frontal lobe	2.1 x 2.6 cm
hemorrhage	Finding	Asserted		
hematoma	Finding	Asserted		
hemorrhage	Finding	Asserted	right parietal lobe	7 x 9 mm
hemorrhage	Finding	Asserted		increased
hematoma	Finding	Asserted		
encephalomalacia	Finding	Possible		
midline shift	Finding	Negated		

Coreference: A bridge too far?



Concept	Type	ID	Status	Site	Size
hematoma	Finding	F1	Asserted	right high post	2.1 x 2.6 cm
hemorrhage	Finding	F2	Asserted		
hematoma	Finding	F1	Asserted		
hemorrhage	Finding	F3	Asserted	right parietal	7 x 9 mm
hemorrhages	Finding	[F2,F3]	Asserted		increased
hematoma	Finding	F1	Asserted		
encephalomalacia	Finding	F4	Possible		
midline shift	Finding	F5	Negated		

Language Processing Pipelines



spaCy

NAME	COMPONENT	CREATES	DESCRIPTION
tokenizer	Tokenizer	Doc	Segment text into tokens.
tagger	Tagger	Doc[i].tag	Assign part-of-speech tags.
parser	DependencyParser	Doc[i].head, Doc[i].dep, Doc.sents, Doc.noun_chunks	Assign dependency labels.
ner	EntityRecognizer	Doc.ents, Doc[i].ent_iob, Doc[i].ent_type	Detect and label named entities.
textcat	TextCategorizer	Doc.cats	Assign document labels.
...	custom components	Doc._.xxx, Token._.xxx, Span._.xxx	Assign custom attributes, methods or properties.

spaCy Architecture

