

# 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

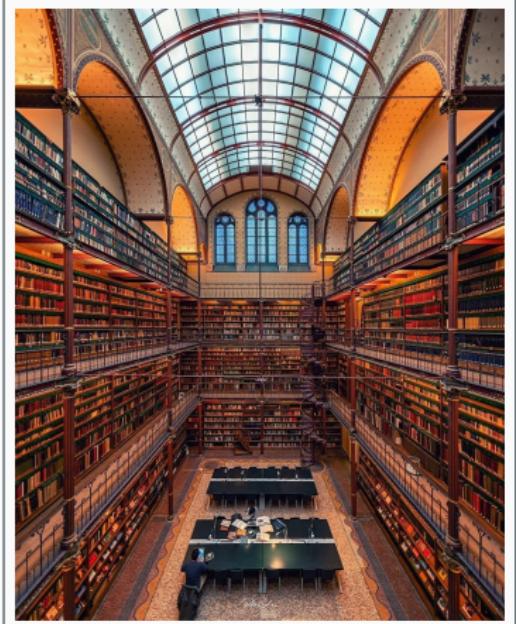
Some practical use cases for NLP:

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

# Unstructured Data

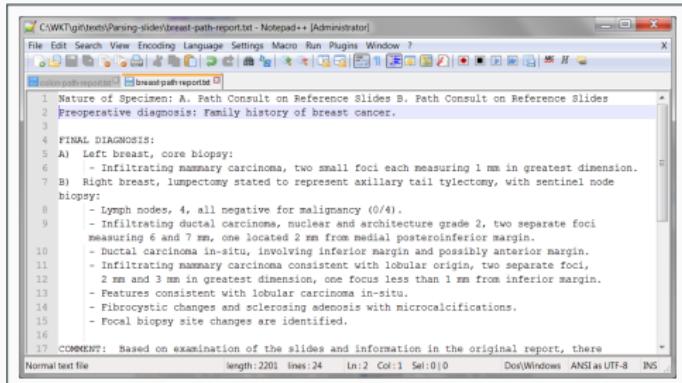
Vast quantities of information are encoded as **unstructured data**, in the form of natural language text.

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:\WKT\git\tests\Parsing-slides\breast-path-report.txt - Notepad++ [Administrator]". The menu bar includes File, Edit, Search, View, Encoding, Language, Settings, Macro, Run, Plugins, Window. The toolbar has various icons for file operations. The status bar at the bottom shows "Normal text file", "length:2201 lines:24 in:2 Col:1 Sel:0:0", "DotWindows", "ANSI as UTF-8", and "INS". The main text area contains a medical report for breast pathology, starting with "Nature of Specimen: A. Path Consult on Reference Slides B. Path Consult on Reference Slides" and "Preoperative diagnosis: Family history of breast cancer". It details findings for both breasts, including infiltrating mammary carcinoma, ductal carcinoma in-situ, lobular carcinoma, and other features. A comment at the end states: "COMMENT: Based on examination of the slides and information in the original report, there".

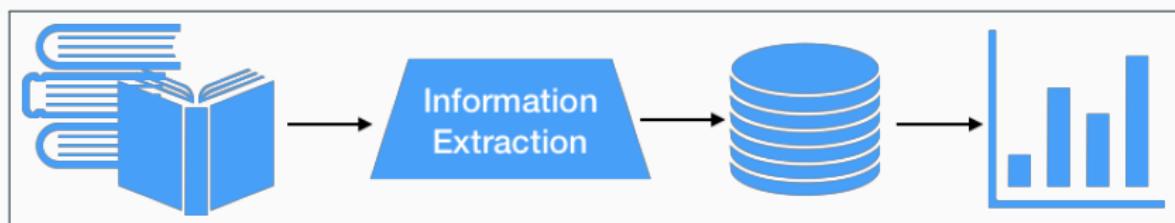
What a **computer** sees:

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

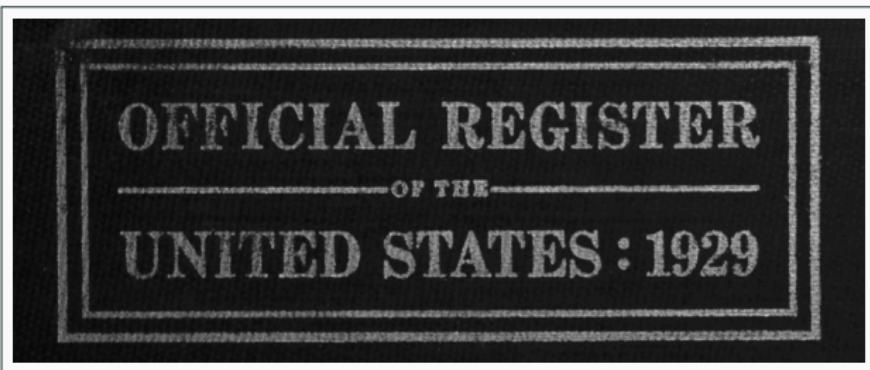
# Information Extraction

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 screenshot shows a finite state machine (FSM) editor interface. At the top, there is a state transition diagram. It starts with a black initial state, followed by a sequence of states connected by arrows. The first arrow has a label '^' above it and a yellow box below it containing '\s'. This is followed by a series of states labeled 'One of [A-Z]' with a yellow box below each. Then there is a state labeled 'One of [a-z]' with a yellow box below it. After this, there is a state labeled 'One of [\s-A-Z]' with a yellow box below it, followed by another 'One of [a-z]' state with a yellow box below it. Finally, there is a state labeled 'One of [r]' with a yellow box below it, followed by a state labeled 'j' with a yellow box below it, and a final state labeled 'x' with a yellow box below it. A pink box labeled 'Group 1' covers the first four states. Below the diagram, there is a text input field containing the regular expression code: `1 ^\s?([A-Z]\.?[a-z]*\?:\s[A-Z]\.?[a-z]*)<(1,3)(?:,\s*x)?>`. Below this, a result section shows the text 'Result: Matches starting at the black triangle slider' and the output '1 |George H. James, jr'. The interface includes dropdown menus for 'Python' and 'Flags', and a 'View Cheatsheet' button.

<https://www.debuggex.com/>

# Chomsky Hierarchy

Grammar	Languages	Automaton	Production rules (constraints)*	Examples <sup>[2]</sup>
Type-0	Recursively enumerable	Turing machine	$\alpha \rightarrow \beta$ (where $\alpha$ 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, 46 matches), 'FLAVOR' (PCRE (PHP), ECMAScript (JavaScript), Python, Golang), 'TOOLS' (Code Generator), and a 'SPONSOR' section for Atlassian. The main area has a 'REGULAR EXPRESSION' input field containing the pattern `\nb([0-9]{1,2}(?:d|th|st))`, which finds 46 matches in the 'TEST STRING'. The test string lists names, titles, and locations, with some parts highlighted in green. A 'SUBSTITUTION' section is at the bottom.

TEST STRING	REGULAR EXPRESSION	RESULTS
James K. Wallace - 5, 000 Ernest E. Emerson 5th..... 4, 600 Edward O. Reed 5, 200 Burr G. Williams 3, 600	<code>\nb([0-9]{1,2}(?:d th st))</code>	46 matches, 11053 steps (~16ms) "gm" [re]
Ellwood S. Moorhead 6th..... 5, 200 Edward A. Huse 6th..... 4, 800 Hermann B. Barnhart 9th..... 4, 400		SWITCH TO UNIT TESTS >

REGULAR EXPRESSION: `\nb([0-9]{1,2}(?:d|th|st))`

TEST STRING:

Match	Value	Description	Location
1	4, 200	Superintendent of accounts and budget officer	Ohio
2	- 5, 000		Md.
3	Ernest E. Emerson	Purchasing agent	
4	5th..... 4, 600	Technical director	D.C.
5	Edward O. Reed		
6	5, 200	Chief instructor of apprentices	Iowa
7	Burr G. Williams		10th
8	3, 600		
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			

SUBSTITUTION:

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

# Demo: Complete Working Example

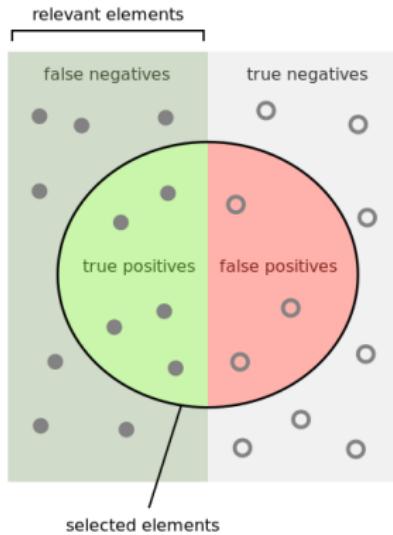
The screenshot shows a Jupyter Notebook interface with the following details:

- Launcher** tab is selected.
- extract-records.ipynb** tab is open.
- Python 3** kernel is selected.
- 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)**

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

**specificity or true negative rate (TNR)**

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

**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)**

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

**false discovery rate (FDR)**

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

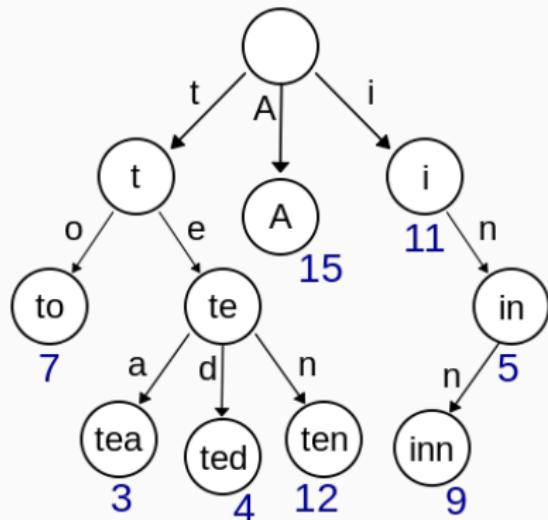
**miss rate or false negative rate (FNR)**

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

# 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

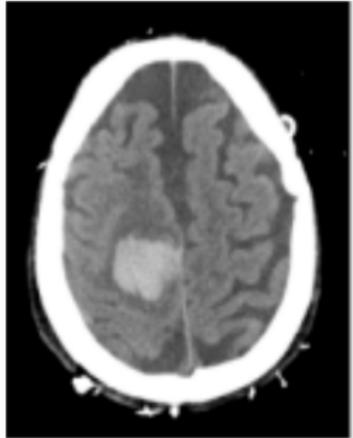
---

## 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 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 **Finding** **AnatomicSite** with another focus of hemorrhage in the right parietal lobe measuring 7 x 9 mm.
- 3 The intraparenchymal **Finding** hemorrhages have increased in size when compared to the outside head CT from 11/9/2015.
- 4 There is associated low-attenuation surrounding the **Finding** hematoma, compatible with **Finding** 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 **Finding** gliosis/encephalomalacia.
- 6 There is no midline shift and the basilar cisterns are preserved.

<b>Concept</b>	<b>Type</b>	<b>Count</b>
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

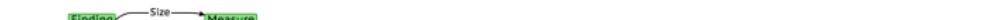
# Concept Relations



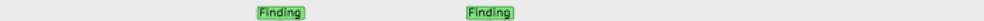
1 FINDINGS: A 2.1 x 2.6 cm intraparenchymal hematoma is noted in the right high, posterior frontal lobe.



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.



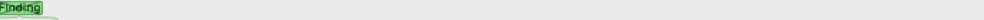
3 The intraparenchymal hemorrhages have increased in size when compared to the outside head CT from 11/9/2015.



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.

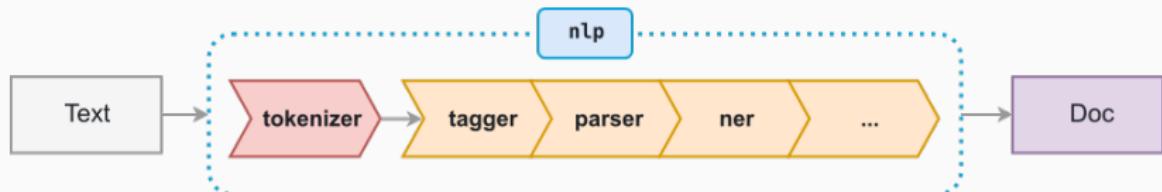
Concept	Type	Status	Site	Size
hematoma	Finding	Asserted	right high posterior	2.1 x 2.6 cm
hemorrhage	Finding	Asserted		
hematoma	Finding	Asserted		
hemorrhage	Finding	Asserted	right parietal lobe	7 x 9 mm
hemorrhage	Finding	Asserted		increased
hematoma	Finding	Asserted		
encephalomalacia	Finding	Possible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

# Coreference: A bridge too far?

- 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.
- 
- ```
graph LR; F1[Findings: F1] -- "Size" --> F2[Findings: F2]; F1 -- "Located" --> AS1[AnatomicSite]; F1 -- "Located" --> F3[Findings: F3]; F1 -- "Etc." --> F5[Findings: F5]; F2 -- "Located" --> AS2[AnatomicSite]; F3 -- "Located" --> AS3[AnatomicSite]; F3 -- "Size" --> M3[Measure]; F5 -- "Negated" --> M5[Measure]
```

| 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 | <a href="#">Tokenizer</a>         | Doc                                                          | Segment text into tokens.                        |
| tagger    | <a href="#">Tagger</a>            | Doc[i].tag                                                   | Assign part-of-speech tags.                      |
| parser    | <a href="#">DependencyParser</a>  | Doc[i].head,<br>Doc[i].dep,<br>Doc.sents,<br>Doc.noun_chunks | Assign dependency labels.                        |
| ner       | <a href="#">EntityRecognizer</a>  | Doc.ents,<br>Doc[i].ent_iob,<br>Doc[i].ent_type              | Detect and label named entities.                 |
| textcat   | <a href="#">TextCategorizer</a>   | Doc.cats                                                     | Assign document labels.                          |
| ...       | <a href="#">custom components</a> | Doc._.xxx,<br>Token._.xxx,<br>Span._.xxx                     | Assign custom attributes, methods or properties. |

# spaCy Architecture

