Information Extraction

Regular Expressions and Beyond

Will Thompson, Ph.D.

February 22, 2019

Table of contents

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 $\rightarrow$ structured)

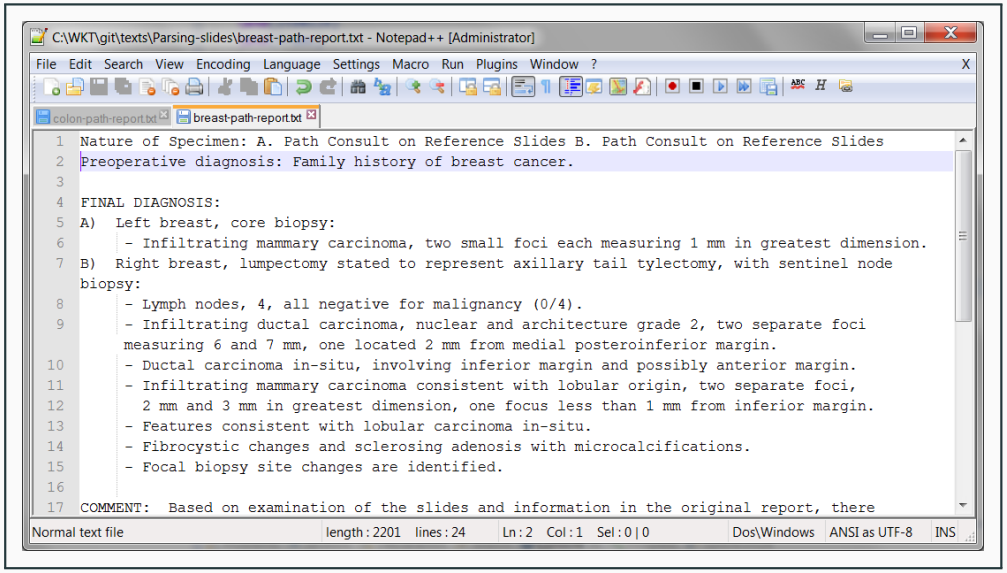
Unstructured Data

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:



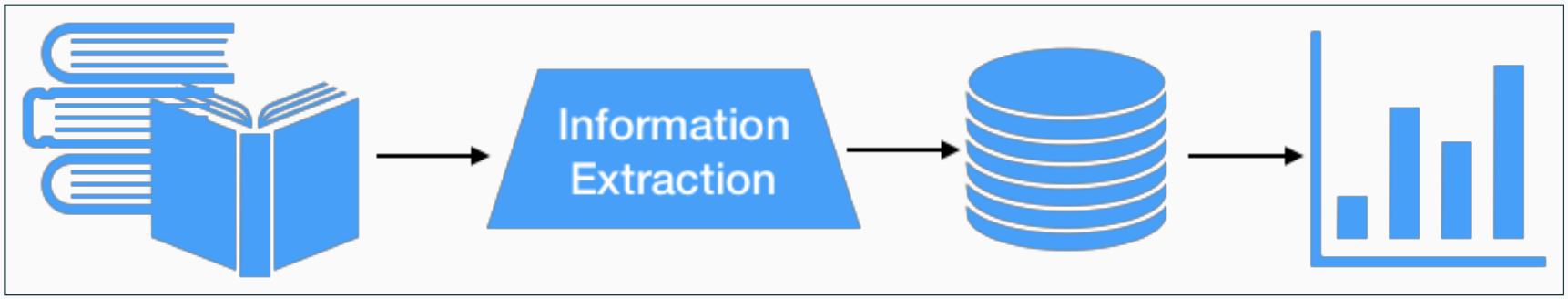
What a computer sees:

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	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

GOVERNMENT PRINTING OFFICE

NAME	OFFICIAL TITLE	Legal residence		Compensation
		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

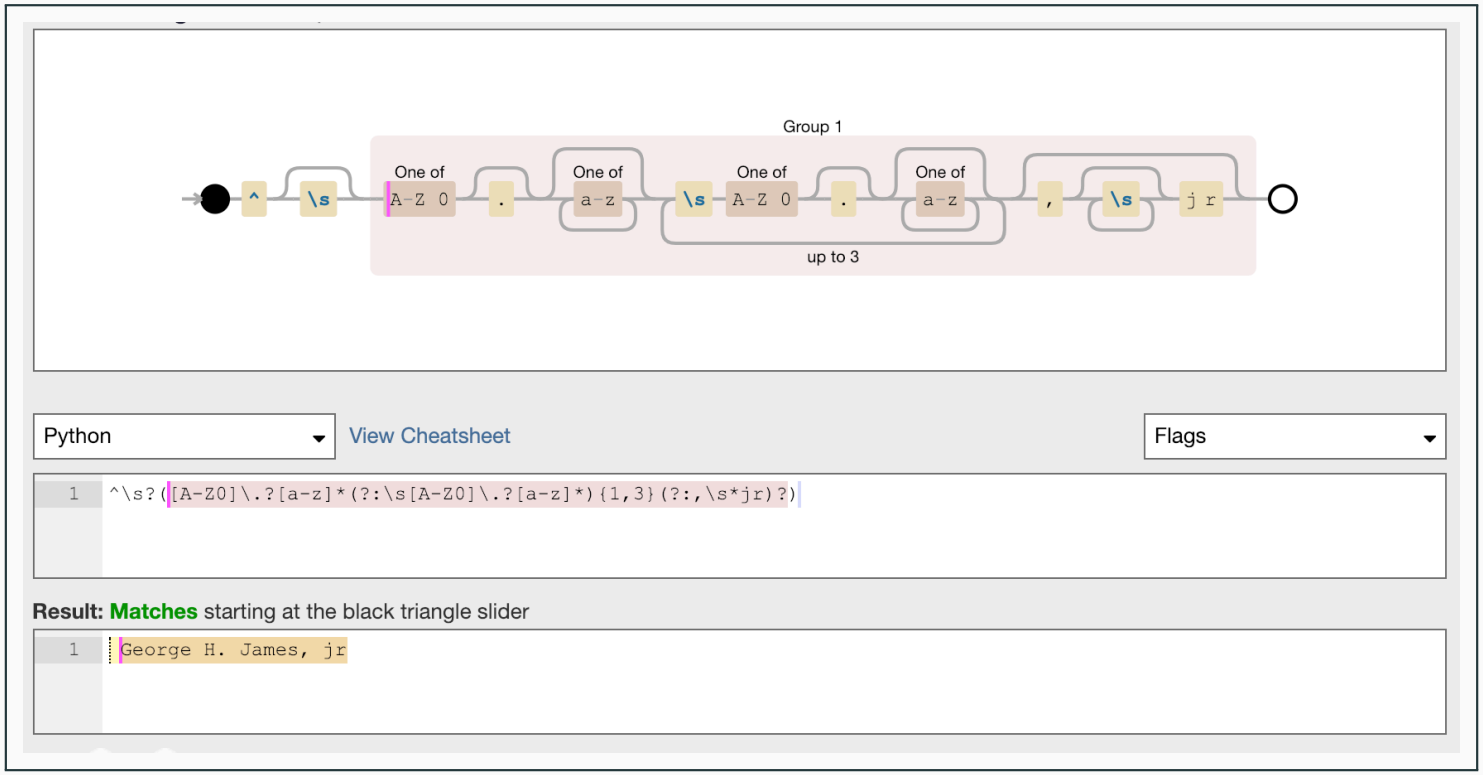
Regular Expressions

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



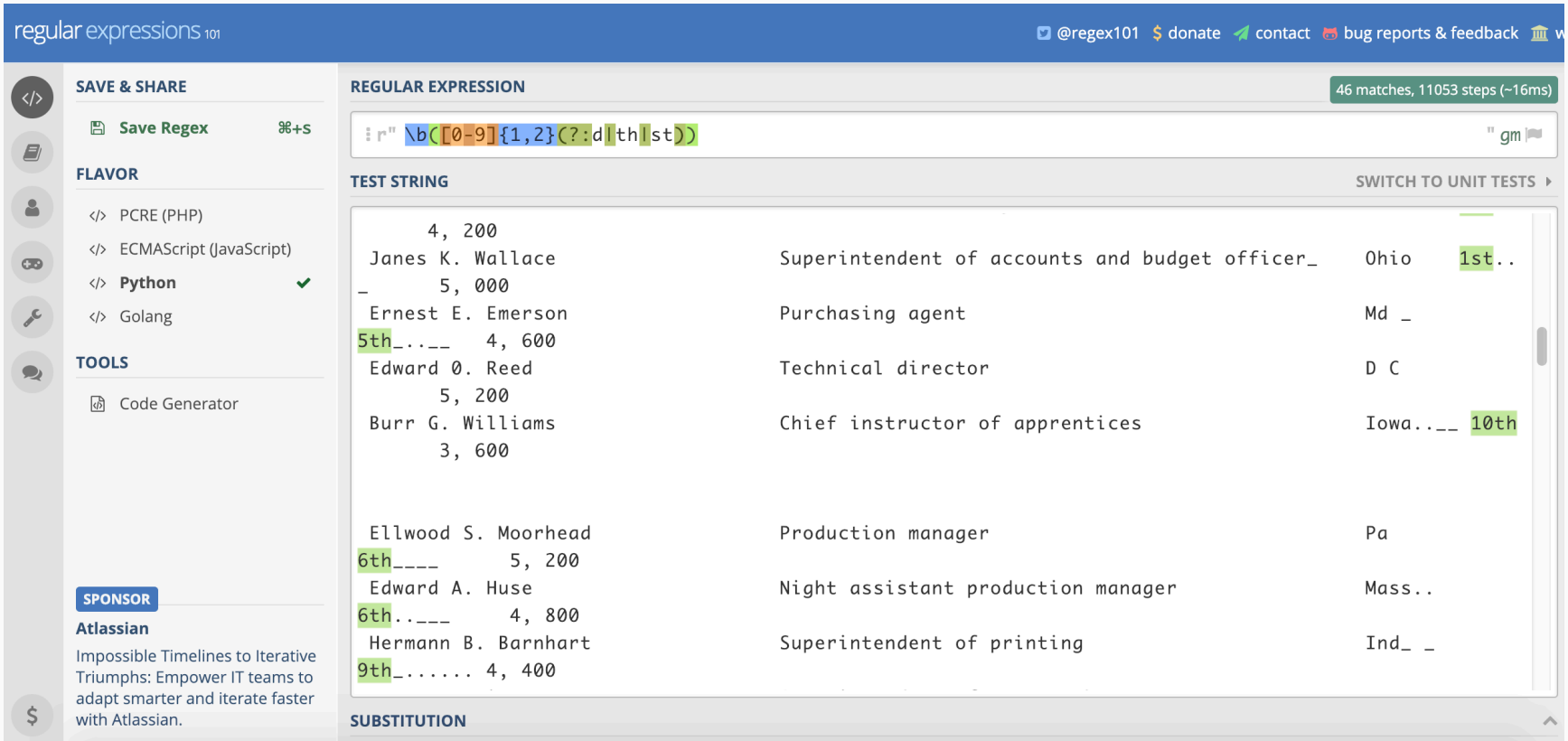
https://www.debuggex.com/

Chomsky Hierarchy

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

Regular Expressions $\equiv$ Regular Languages $\equiv$ Finite State Machines

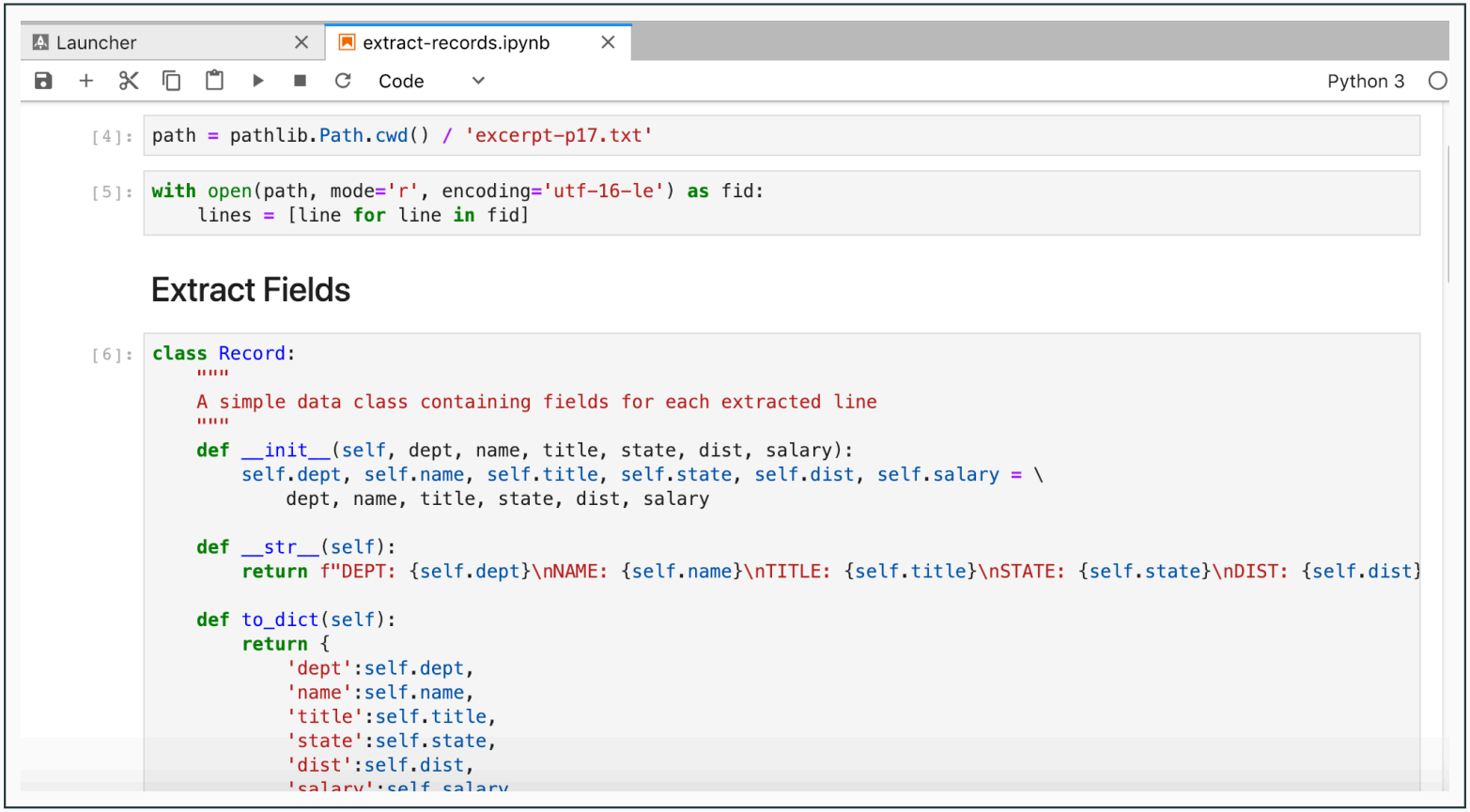
Demo: Developing Regular Expressions



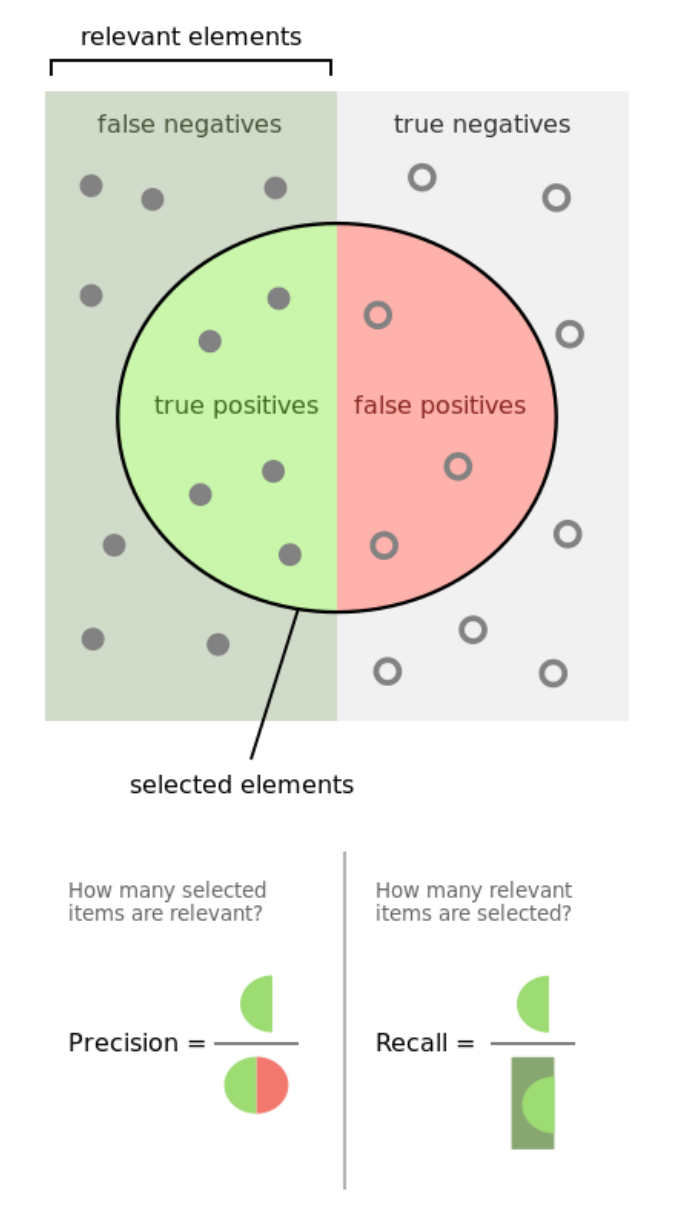
https://regex101.com/

(For fun: https://regexcrossword.com/)

Demo: Complete Working Example



Evaluating Performance



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

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

specificity or true negative rate (TNR)

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

precision or positive predictive value (PPV)

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

negative predictive value (NPV)

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

fall-out or false positive rate (FPR)

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

false discovery rate (FDR)

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

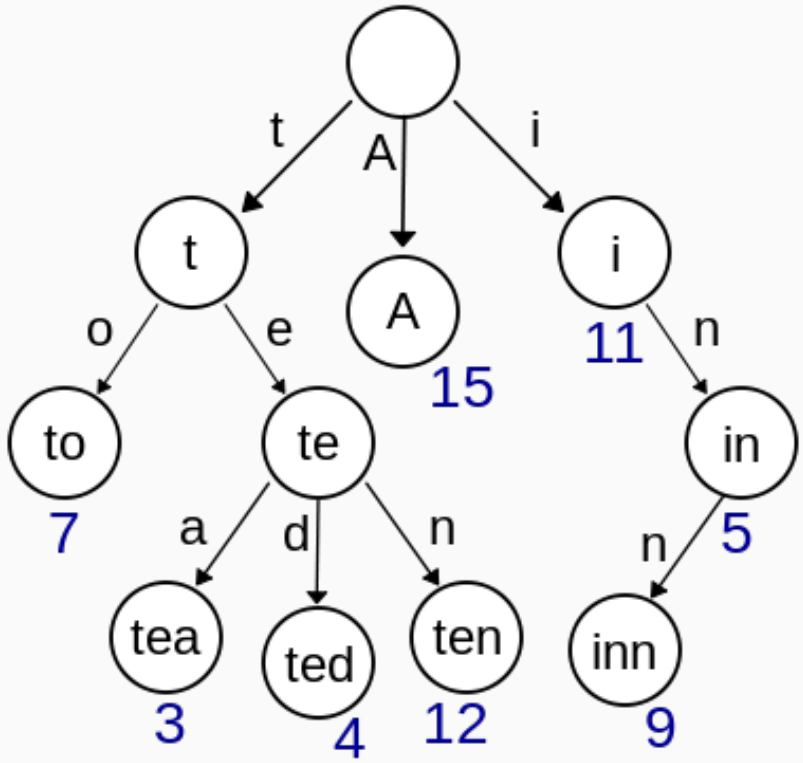
miss rate or false negative rate (FNR)

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

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

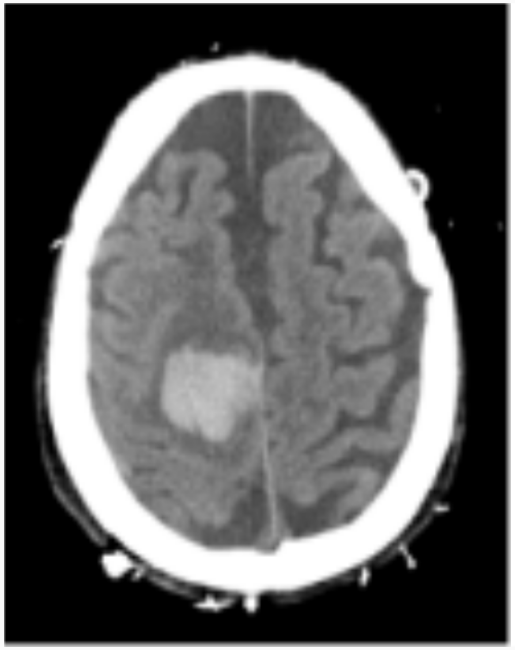
Advanced Pattern Matching

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.

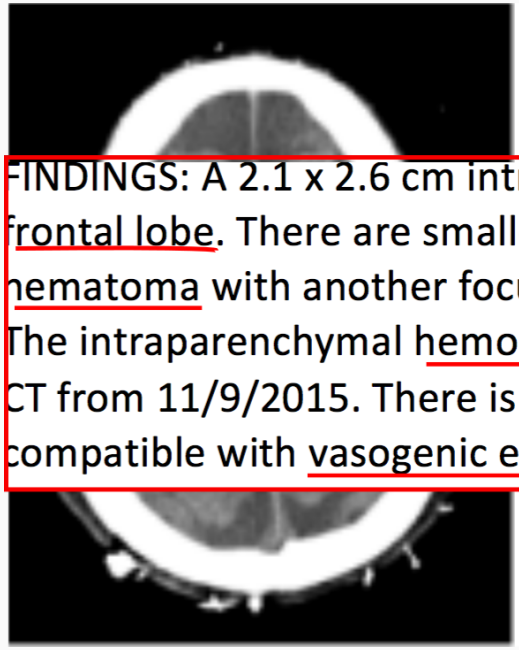
A more ventral region of low-attenuation is noted in the right frontal lobe (series 4, image 26), which is nonspecific and may represent an area of gliosis/encephalomalacia or small vessel ischemic disease. There is no midline shift and the basilar cisterns are preserved.

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 h 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 hematoma (Finding) is noted in the right high, posterior frontal lobe (AnatomicSite).

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

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

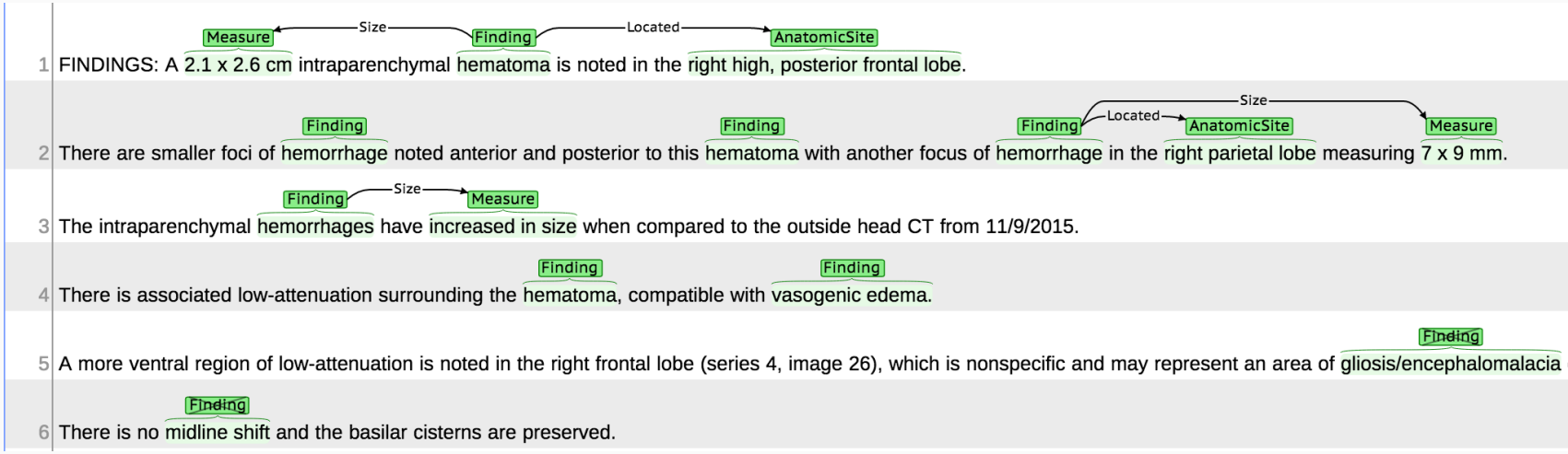
4 There is associated low-attenuation surrounding the hematoma (Finding), compatible with vasogenic edema (Finding).

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 (Finding) and the basilar cisterns are preserved.

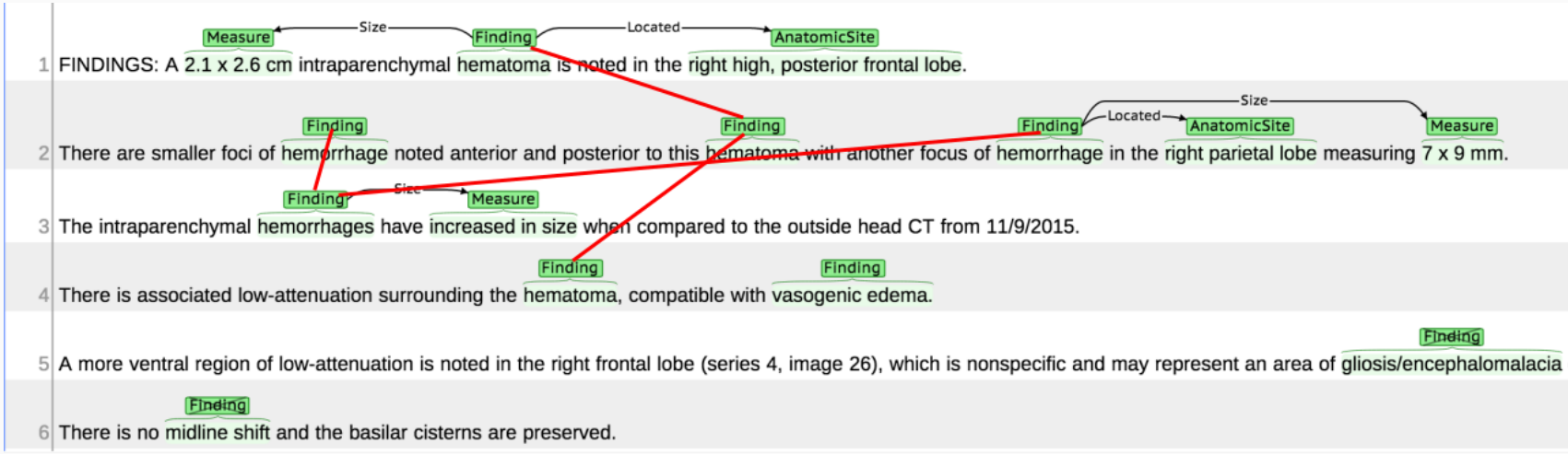
Concept	Type	Count
right high posterior frontal lobe	Site	1
right parietal lobe	Site	1
hematoma	Finding	3
hemorrhage	Finding	3
vasogenic edema	Finding	1
encephalomalacia	Finding	1
midline shift	Finding	1

Concept Relations



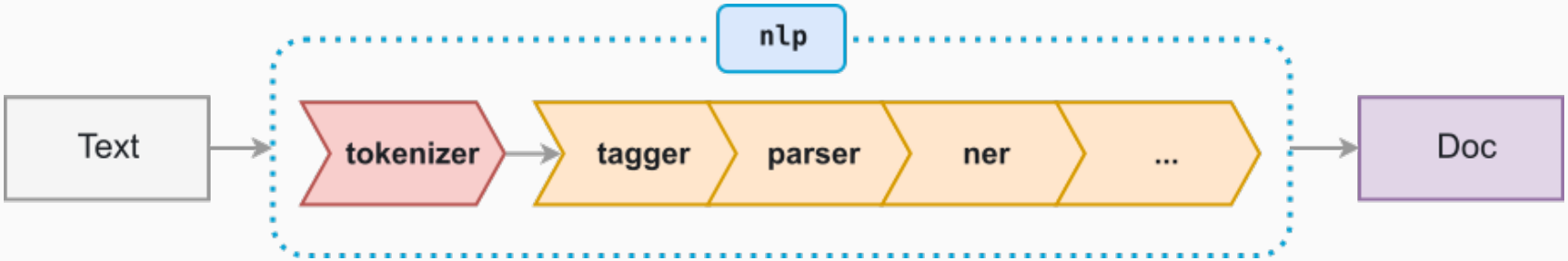
Concept	Type	Status	Site	Size
hematoma	Finding	Asserted	right high posteri	2.1 x 2.6 cm
hemorrhage	Finding	Asserted		
hematoma	Finding	Asserted		
hemorrhage	Finding	Asserted	right parietal lob	7 x 9 mm
hemorrhage	Finding	Asserted		increased
hematoma	Finding	Asserted		
encephalomalacia	Finding	Possible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

Coreference: A bridge too far?



Concept	Type	ID	Status	Site	Size
hematoma	Finding	F1	Asserted	right high poste	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