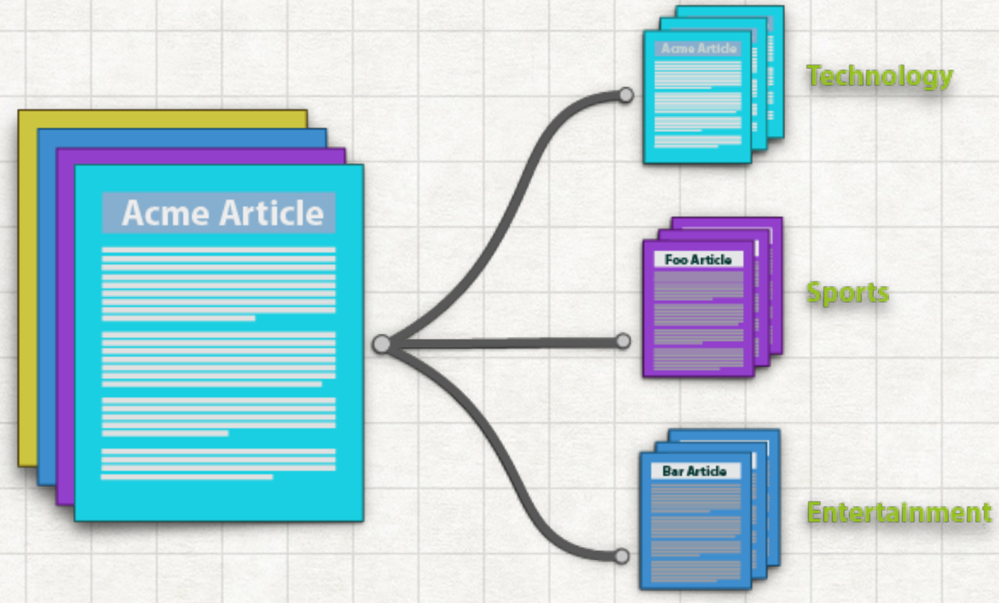
Pulsar AI

NLP APPLICATIONS

- TEXT CLUSTERING
- TEXT CLASSIFICATION
- TEXT SUMMARISATION
- MACHINE TRANSLATION
- SEMANTIC SEARCH
- SENTIMENT ANALYSIS
- QUESTION ANSWERING
- INFORMATION EXTRACTION

Pulsar AI

NLP. TEXT CLASSIFICATION



Document classification or document categorization is a problem in library science, information science and computer science. The task is to assign a document to one or more classes or categories. This may be done "manually" or algorithmically.

Popular algorithms:

1. Multinomial Naive Bayes
2. SVM
3. Neural Networks

Pulsar AI

NLP. TEXT CLUSTERING

Document **clustering** (or **text clustering**) is the application of **cluster** analysis to textual documents. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering.

Popular algorithms:

1. k-Means
2. DBSCAN
3. Deep Learning

Pulsar AI

NLP. TEXT SUMMARISATION

Automatic **summarization** is the process of shortening a **text** document with software, in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax.

Popular algorithms:

1. LDA
2. Deep Learning

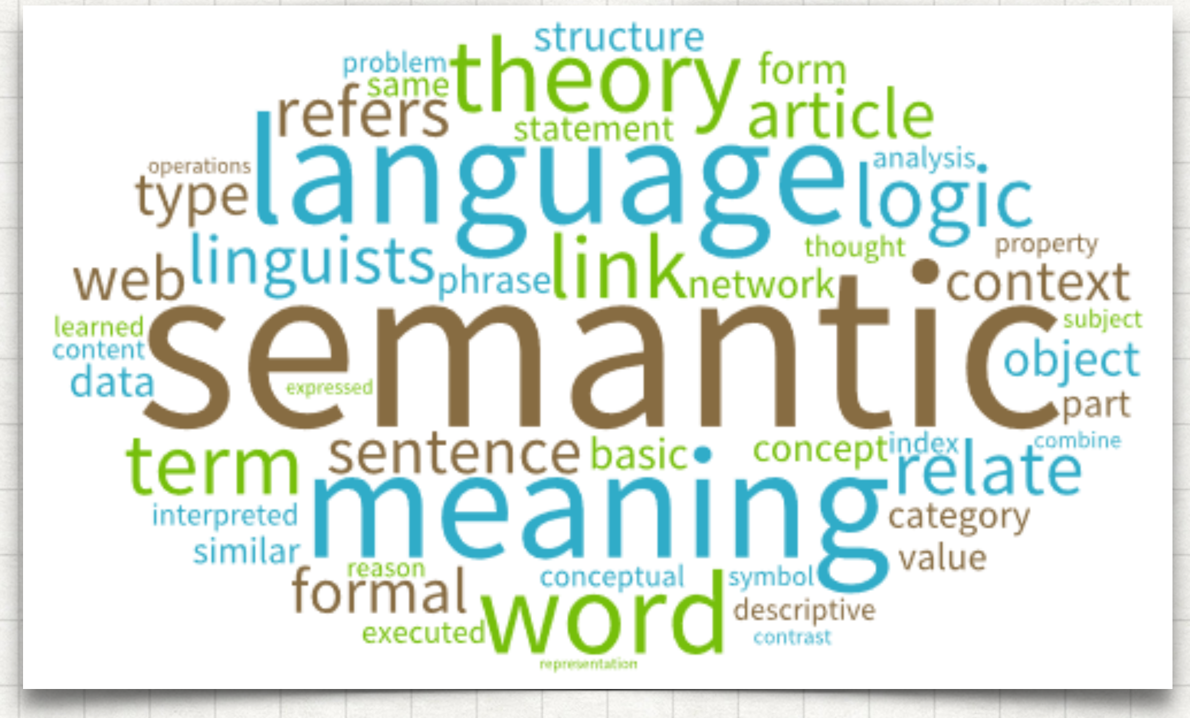
NLP. MACHINE TRANSLATION

MT performs simple substitution of words in one language for words in another, but that alone usually cannot produce a good translation of a text because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with corpus statistical, and neural techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies

Algorithms:

1. Rule based
2. Statistical methods
3. Encoder-Decoder

NLP. SEMANTIC SEARCH



Semantic search seeks to improve **search** accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable dataspace, whether on the Web or within a closed system, to generate more relevant results.

Approaches:

1. Entity Recognition
2. User context

Pulsar AI

NLP. SENTIMENT ANALYSIS

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker.

Algorithms:

1. Lexicon-based
2. Machine Learning (SVM)
3. Deep Learning (RNN, LSTM)

Pulsar AI

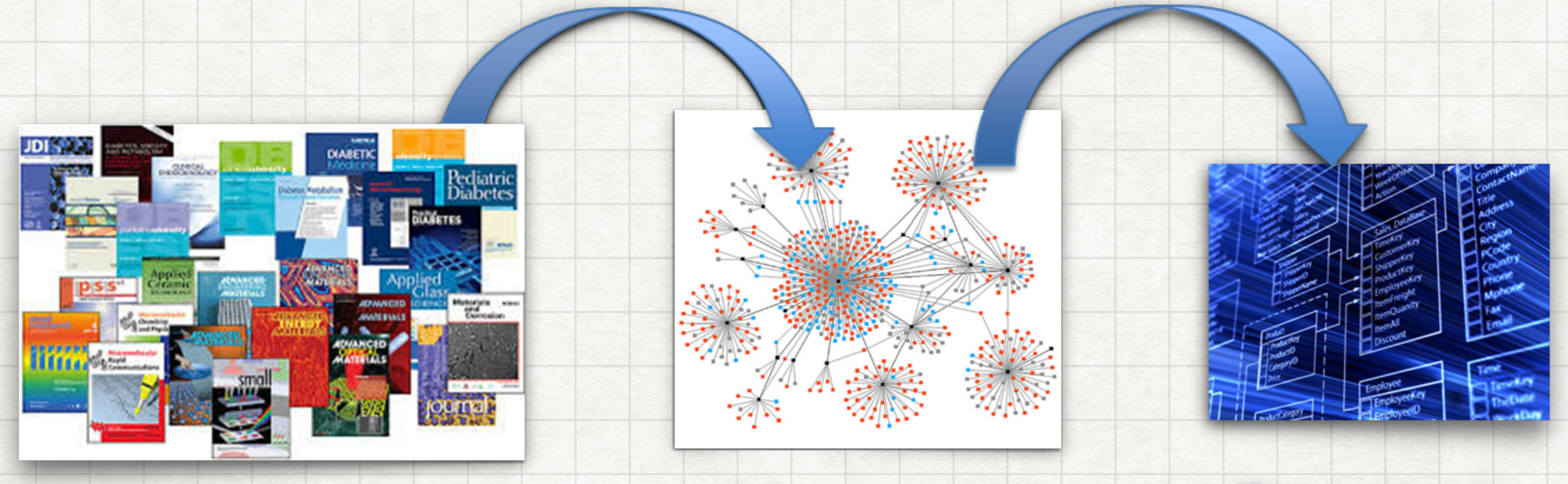
NLP. QUESTION ANSWERING

Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically **answer questions** posed by humans in a natural language.

Algorithms:

1. Rule based
2. Machine Learning
3. Deep Learning

NLP. INFORMATION EXTRACTION



Publications

Relationships

Structured Information

Information extraction is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents.

NLP TOOLS

1. MORPHOLOGICAL ANALYZER
2. POS TAGGER
3. STEMMER
4. PARSERS
5. NAMED ENTITY RECOGNIZER

Pulsar AI

NLP. STEMMER

Stemmers remove morphological affixes from words, leaving only the word stem.

bananas -> banana

cats -> cat

dogs -> dog

flies -> fli

How about “flies” -> fly?

NLP. MORPHOLOGICAL ANALYZER

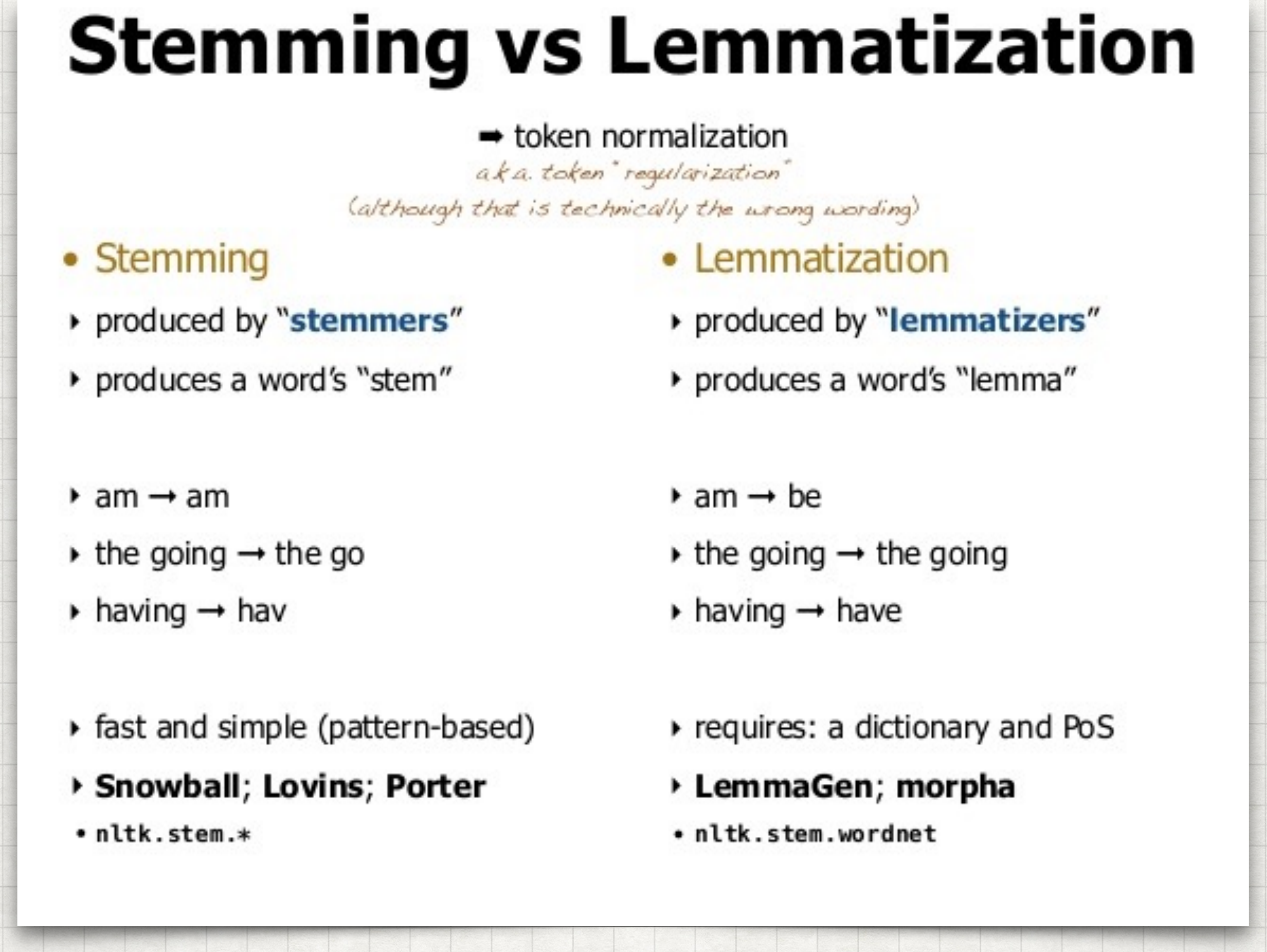
Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma .

flies -> fly

went -> go

am, are, is -> be

NLP. MORPHOLOGICAL ANALYZER



Stemming vs Lemmatization

➡ token normalization

a.k.a. token "regularization"

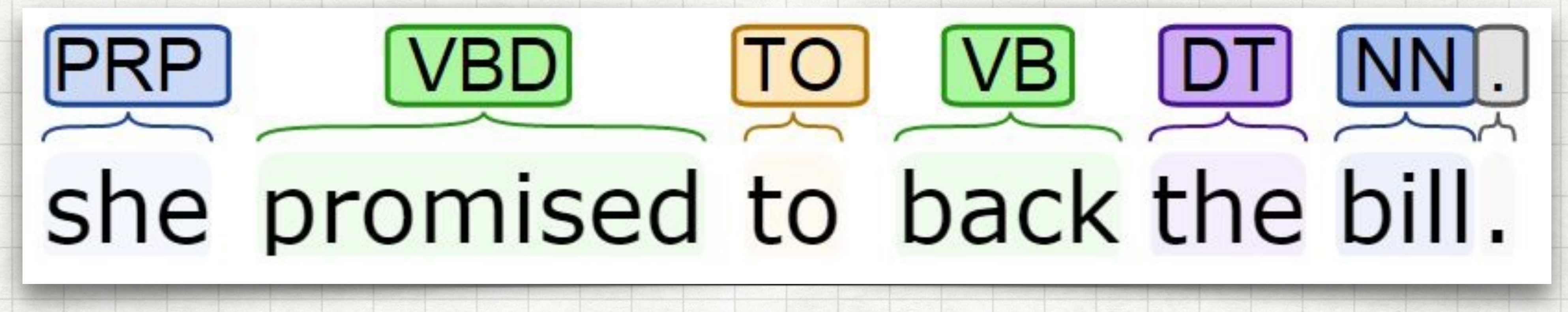
(although that is technically the wrong wording)

- Stemming
  - produced by "**stemmers**"
  - produces a word's "stem"
  - am → am
  - the going → the go
  - having → hav
  - fast and simple (pattern-based)
  - **Snowball**; **Lovins**; **Porter**
  - `nltk.stem.*`

- Lemmatization
  - produced by "**lemmatizers**"
  - produces a word's "lemma"
  - am → be
  - the going → the going
  - having → have
  - requires: a dictionary and PoS
  - **LemmaGen**; **morpha**
  - `nltk.stem.wordnet`

Pulsar AI

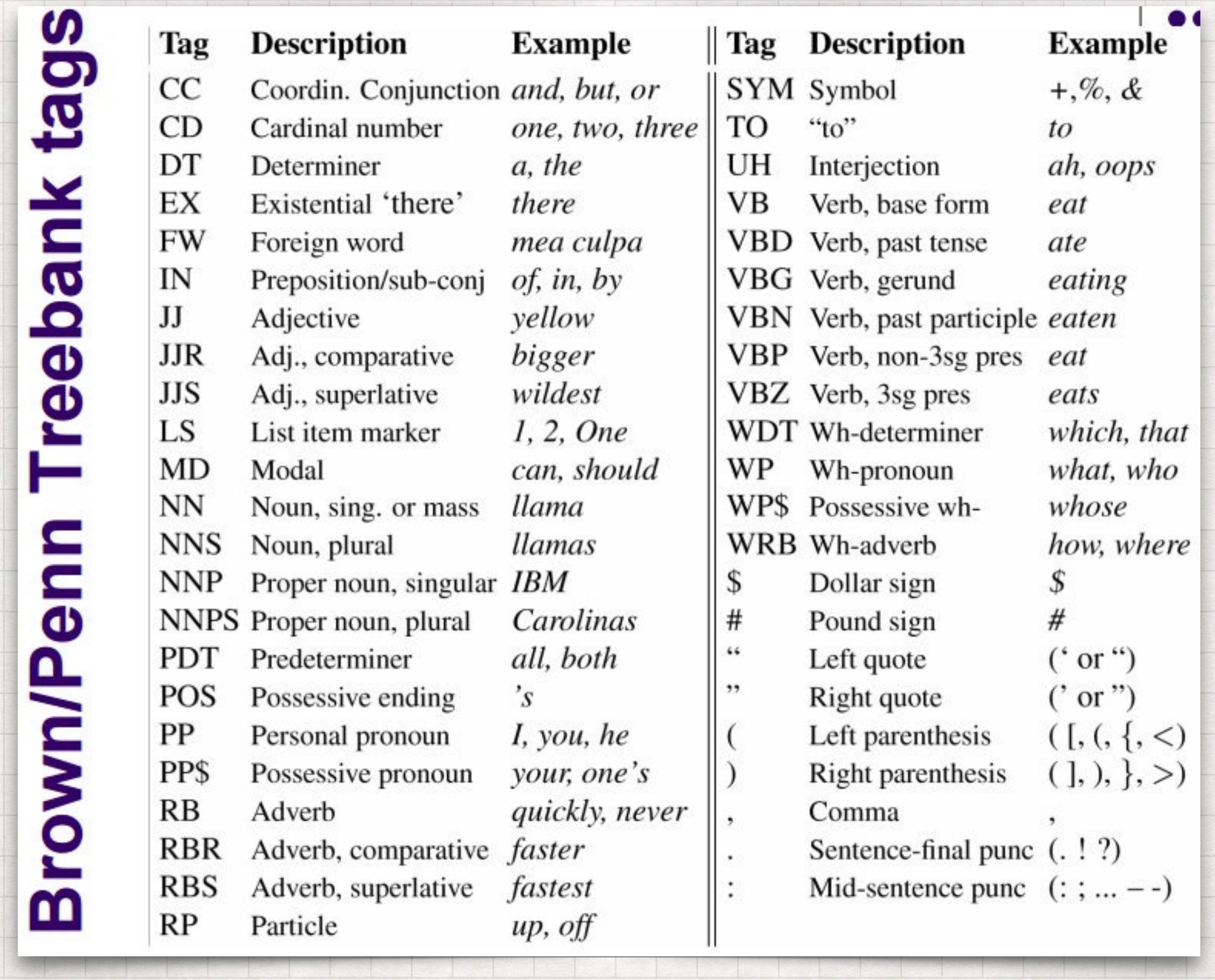
NLP. POS TAGGER



A **Part-Of-Speech Tagger** (**POS Tagger**) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained **POS** tags like 'noun-plural'.

Pulsar AI

NLP. POS TAGGER



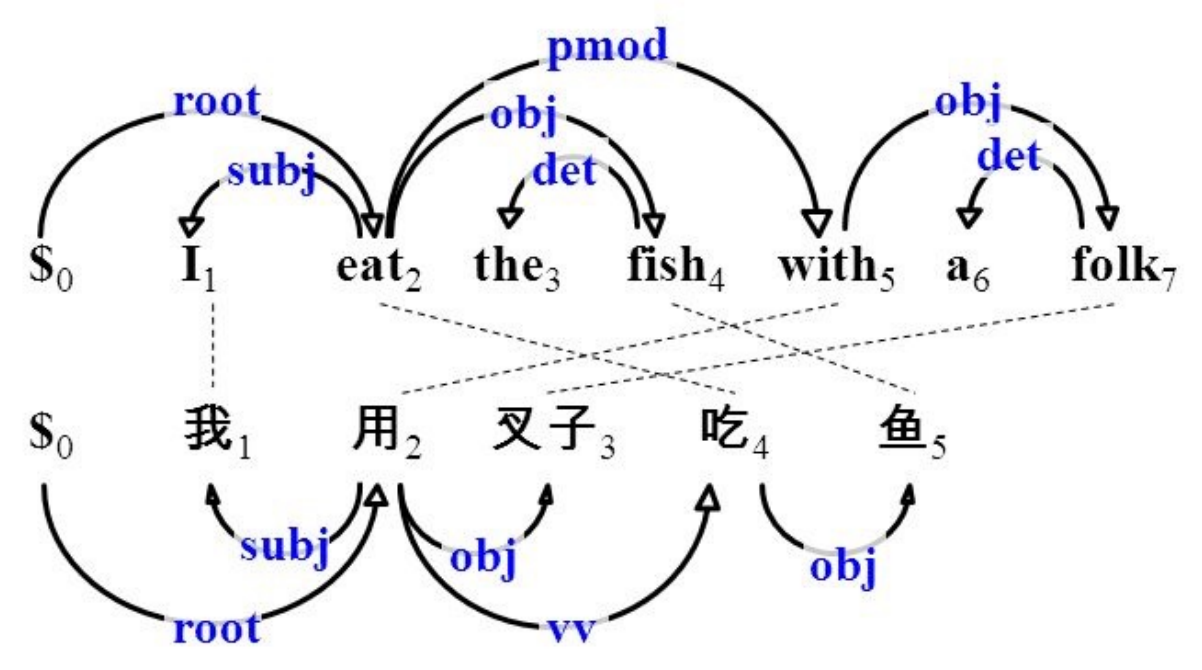
Brown/Penn Treebank tags

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	“to”	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential ‘there’	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*’s*	”	Right quote	(’ or ”)
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one’s*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

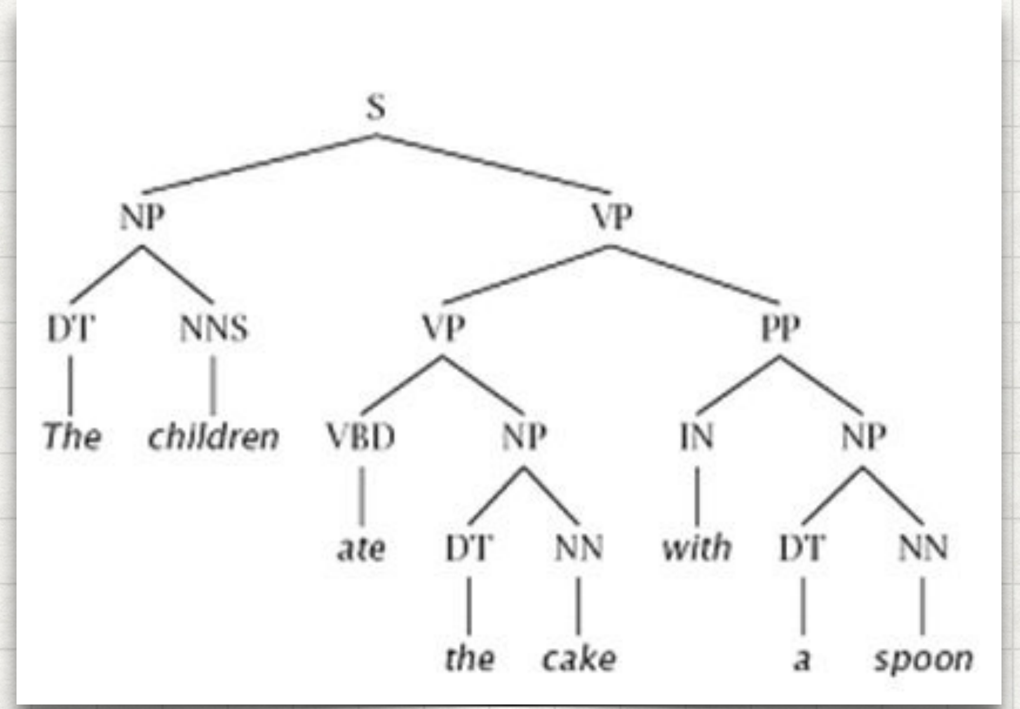
NLP. PARSER

A natural language parser is a program that works out the grammatical **structure of sentences**, for instance, which groups of words go together (as "phrases") and which words are the **subject** or **object** of a verb.

Dependency tree

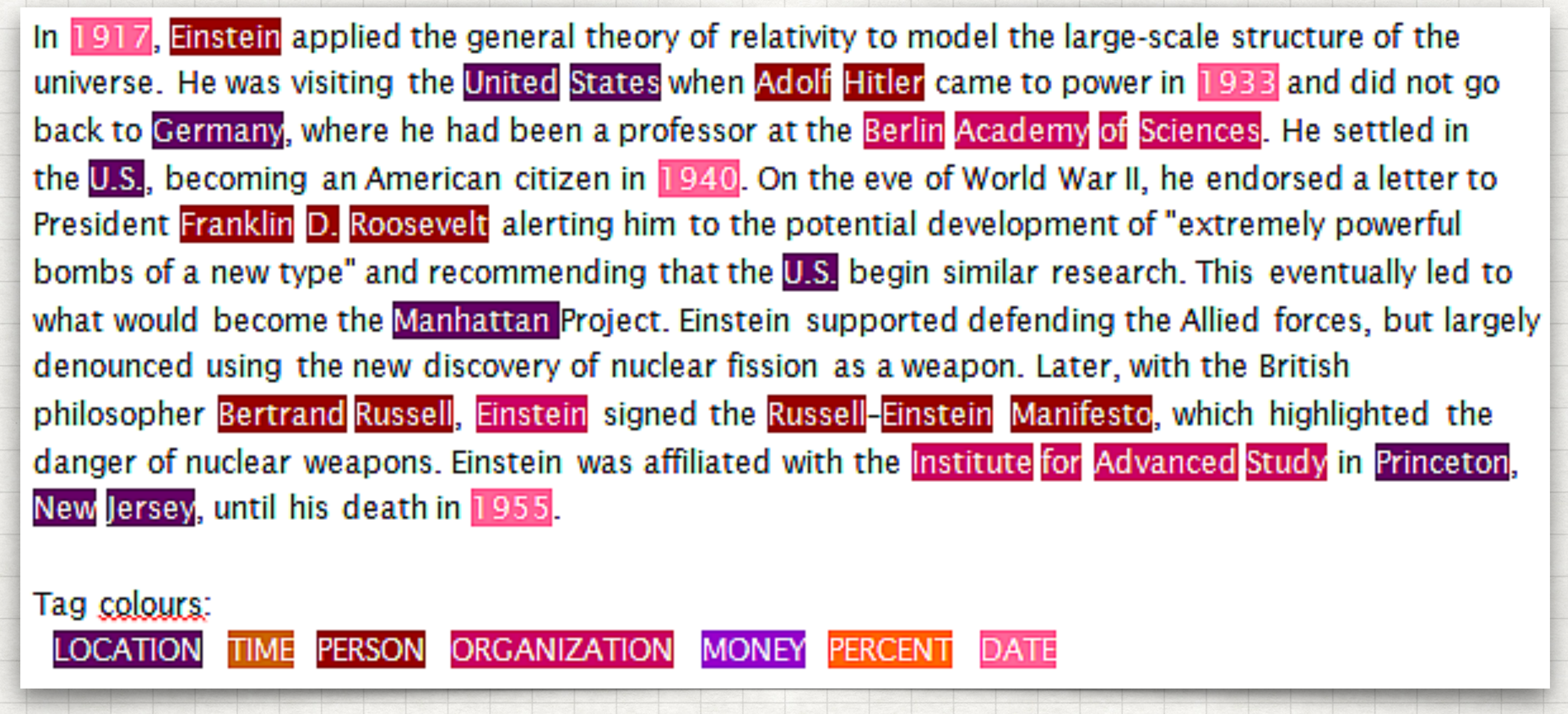


Constituency tree



NLP. NAMED ENTITY RECOGNIZER

Named-entity recognition (**NER**) (also known as **entity identification**, **entity chunking** and **entity extraction**) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.



In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:

LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE