Introduction to NLP

Scope of this lab

Not to bombard with NLP terminologies

Not to provide a NLP crash course in an hour!

But

to introduce you to NLP basics

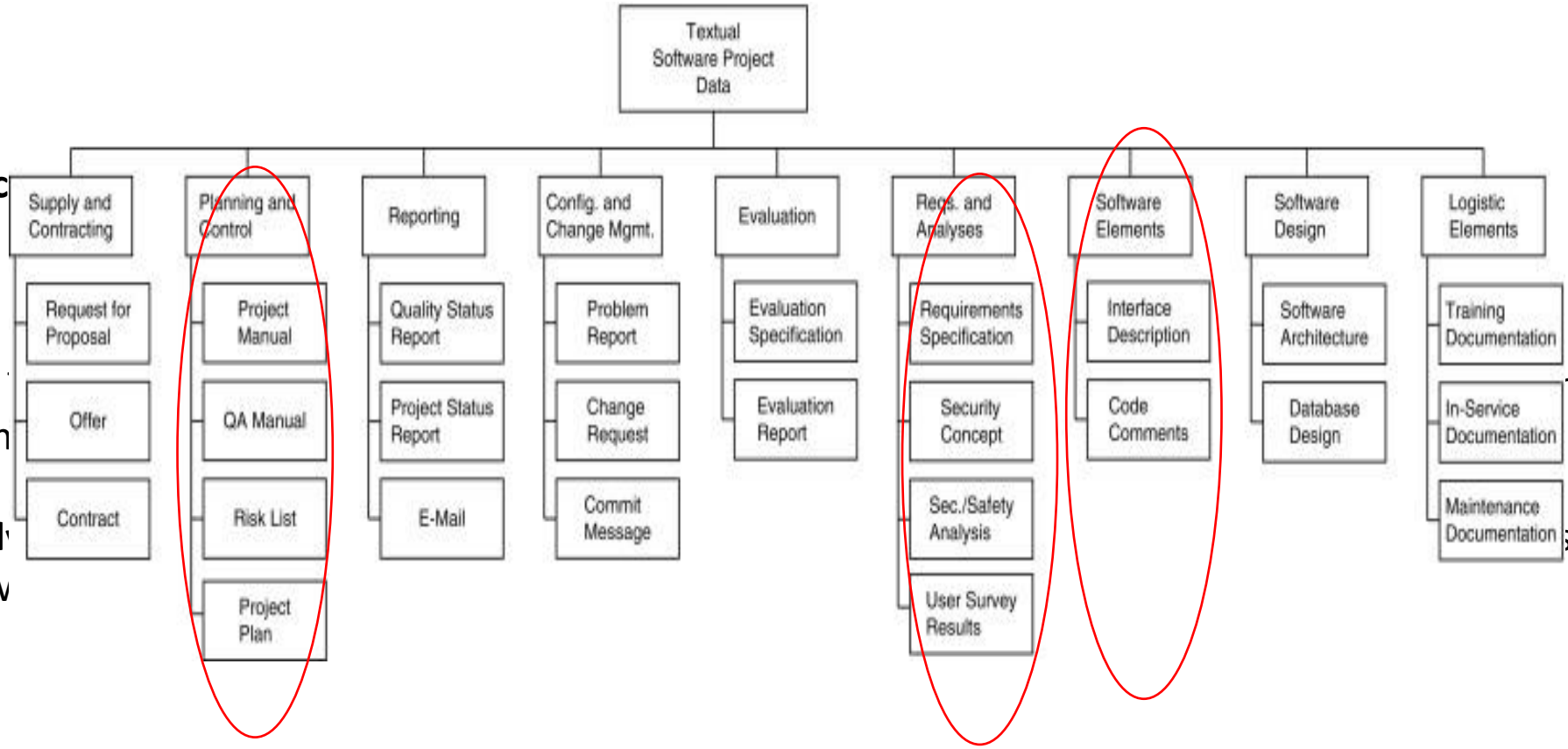
to explain the overall NLP platter that you have for your disposal

to provide a bit of hands-on to kindle your brain with what's in store and where to get started for your actual work

What are the various industry application of NLP that come to your mind?

- Search engines
- Advanced text editors: Such as Grammarly.com
- Computational advertising
- Fraud detection
- Sentiment analysis
- Opinion mining
- Text summarization
- Context analysis

Textual data in Software Project



Sourc

Also: nit messages, which

Finall s and product reviev

Bird, Christian, Tim Menzies, and Thomas Zimmermann, eds. The art and science of analyzing software data. Elsevier, 2015.

Source: Journal Gazett https://www.google.com/url?sa=i&url=https%3A%2F%2Fjournalgazette.net%2Fnews%2Flocal%2F20190614%2Fphotos-of-the-week&psig=AOvVaw1vy1qZ58aHaQ-P62fIIuU5&ust=1581277010885000&source=images&cd=vfe&ved=0CAMQjB1qFwoTCKjesofawucCFQAAAAAdAAAAABAD

Textual Software Project Data and Retrieval [1]

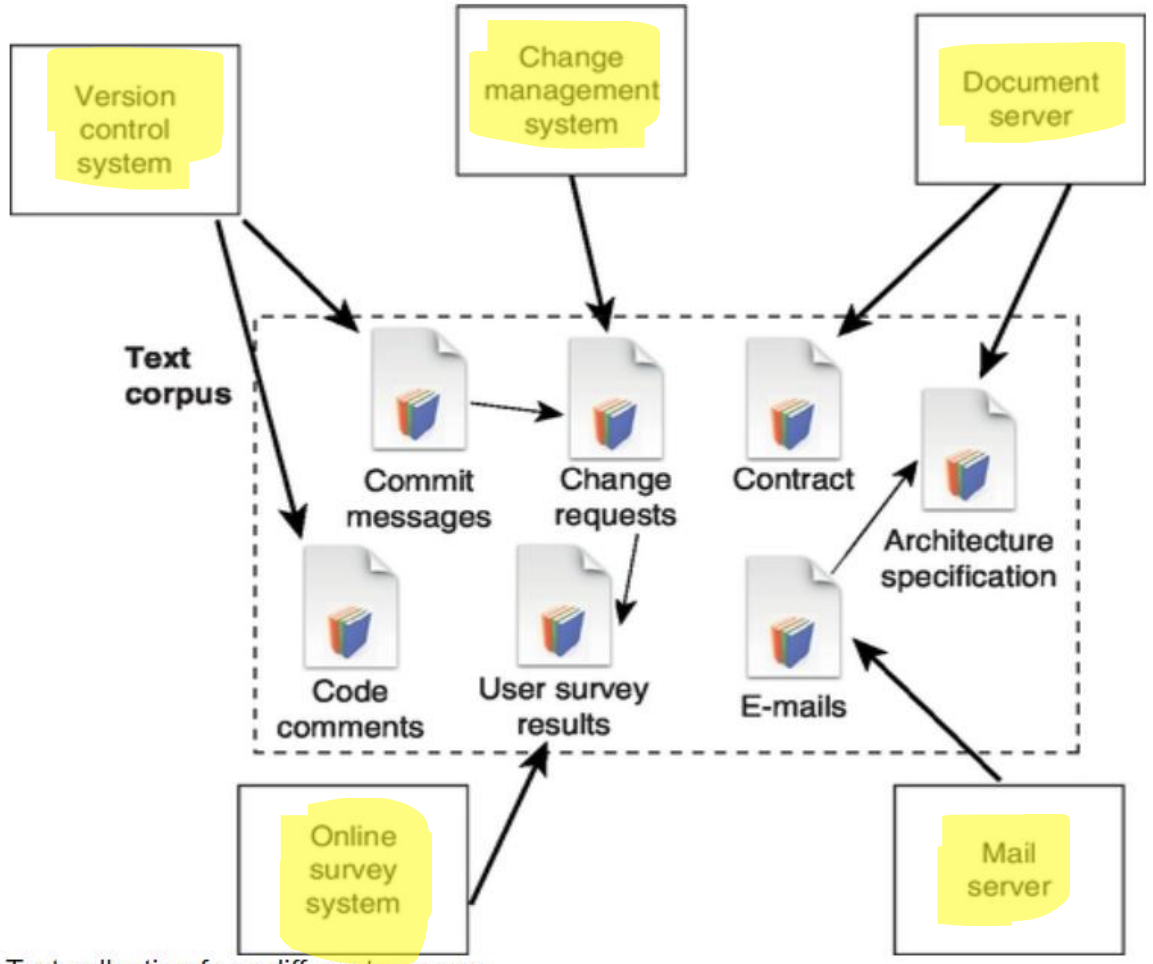
Custom extraction tool

API's

Formatted files

Anonymize

Web crawling

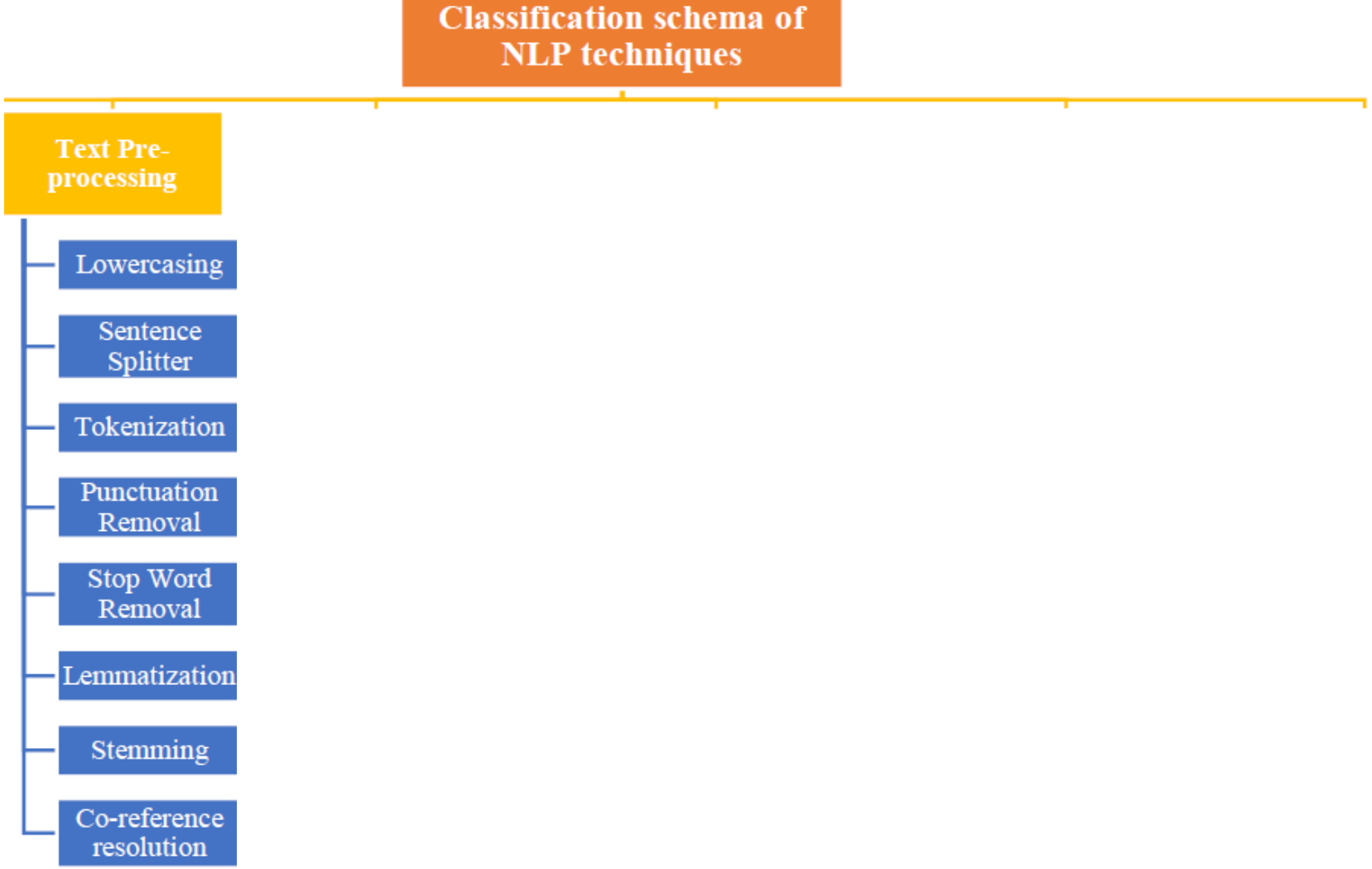


3.2 Text collection from different sources.

Arnaoudova, Venera, et al. "The use of text retrieval and natural language processing in software engineering." *Proceedings of the 37th International Conference on Software Engineering-Volume 2*. 2015.

Importance of NLP in Software Engineering

- Text Retrieval (TR) and NLP in software is one of the fastest growing areas of research in SE. [2]
- Exposing the SE community to these techniques and their applications in SE would help to fill a gap in their current background and allow them to immediately use TR and NLP to advance their research.
- In particular, for TR, approaches such as Vector Space Model, Latent Semantic Analysis, Latent Dirichlet Association, Language Models will be covered. NLP techniques covered will include part-of-speech tagging, stemming, stopword elimination, semantics analysis, sentiment analysis, etc.



Moises Gonzalez-Garcia, Speech recognition, NLP, and the use of ontologies to identify the problem-domain and solution requirements: A systematic mapping study, Information and Software Technology, 2019, In press

Part-of-Speech Tagging

The choice Tag Sets depends on the language and application
Example tag set sizes (for English)
– Brown corpus, 87 tags
– Penn treebank 45 tags
– BNC, 61 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	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PRP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PRP$	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*			

Source : Penn Tree Bank Tagset - http://coltekin.net/cagri/courses/snlp2017/slides/pos-tagging.pdf

Chunking

The color chooser requires **the presence** of **a hardware pointing device**

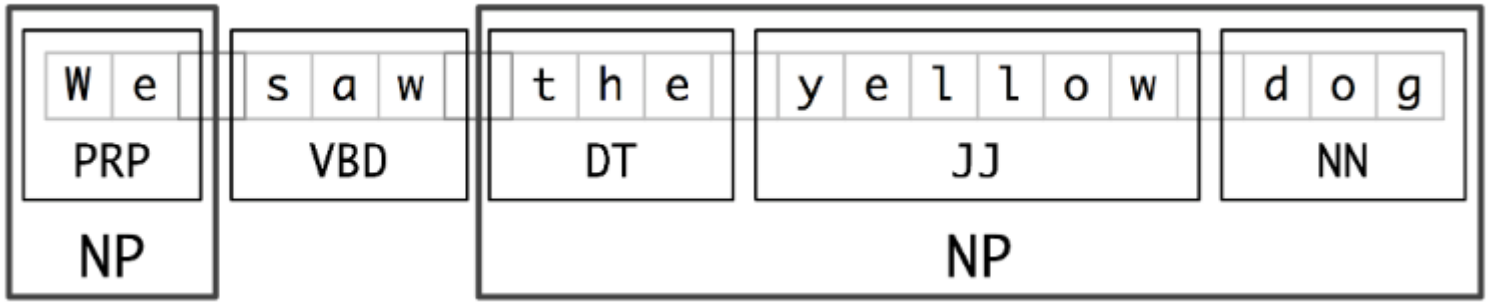
POS Tagging (`averaged_perceptron_tagger`)

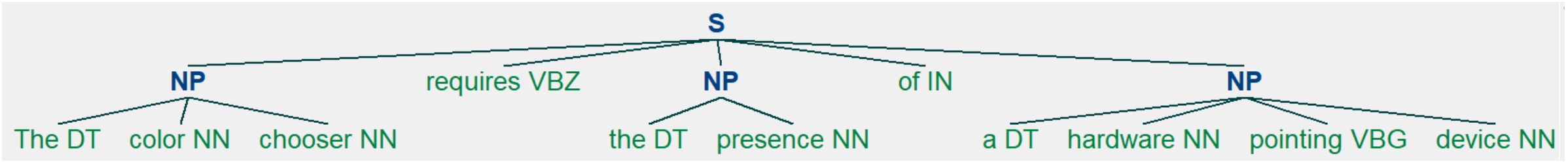
1

```
[('The', 'DT'), ('color', 'NN'),
                                  'VBZ'),
                                  of', 'IN'),

                                  )]
```

```
grammar = r"""NP:{<DT>?<NN><VBG><NN>}
```



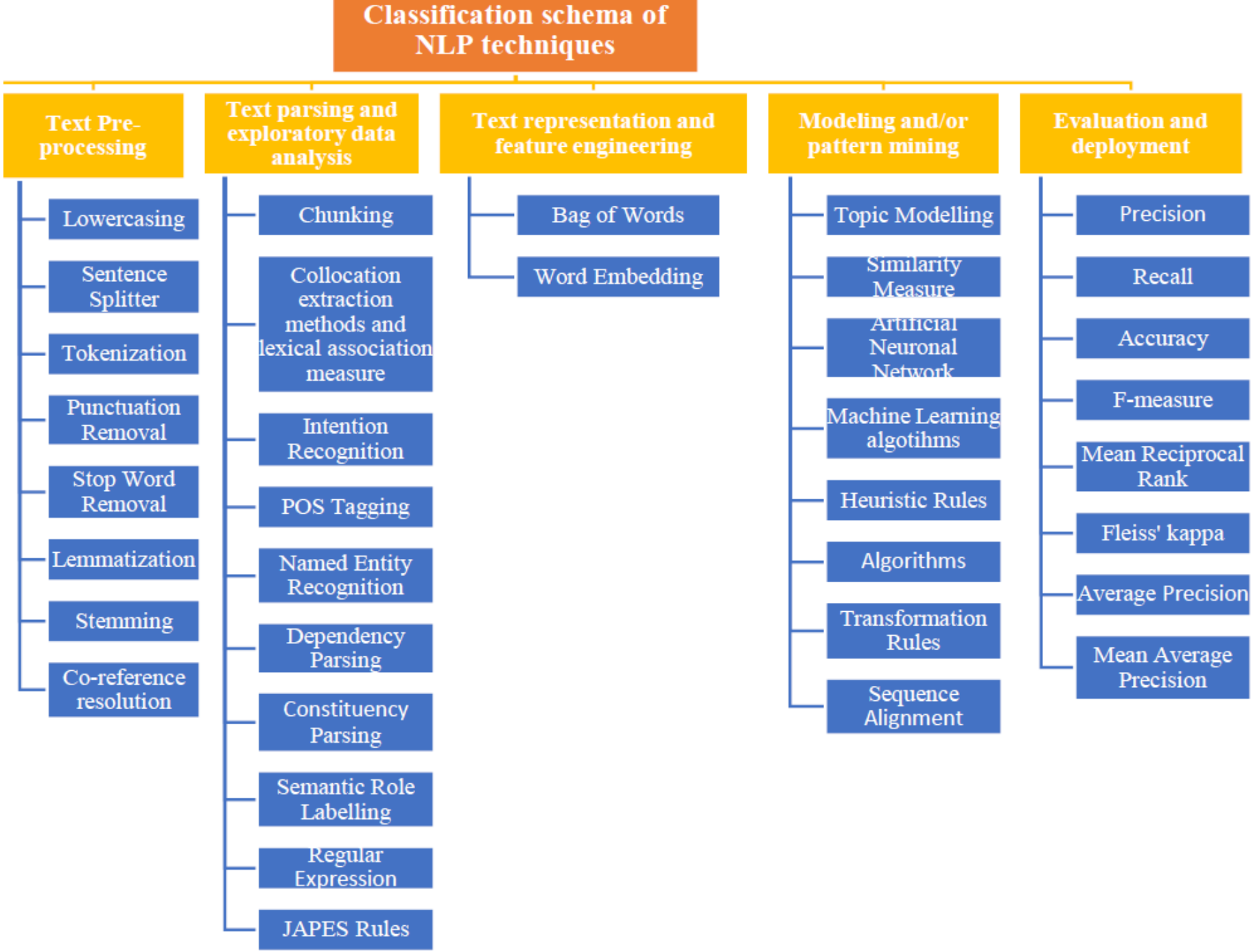


Named Entity Recognition

- Wikipedia: **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 entity mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

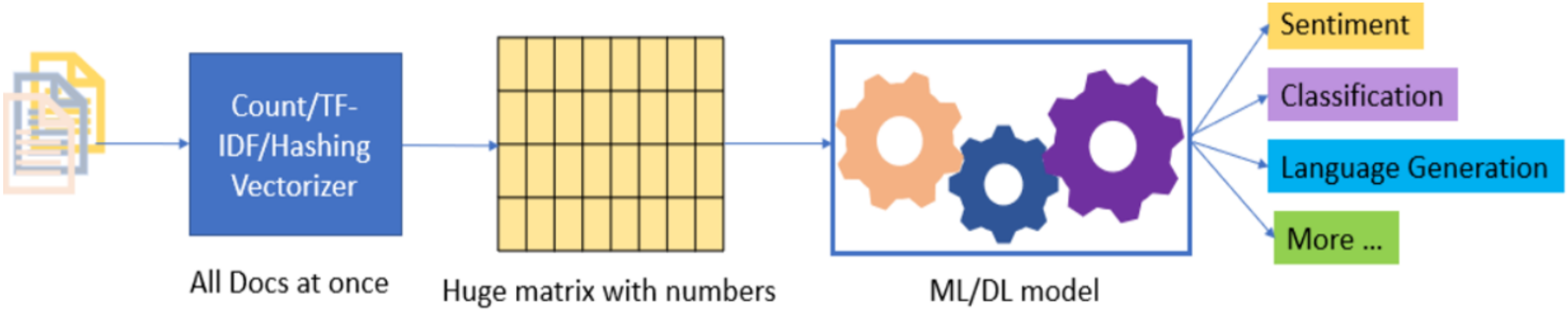
Input: "When Michael Jordan was at the peak of his powers as an NBA superstar, his Chicago Bulls team were moving down the completion, winning six National Basketball Association titles".

Output:
"Chicago Bulls"
"Michael Jordan"
"National Basketball Association"



Moises Gonzalez-Garcia, Speech recognition, NLP, and the use of ontologies to identify the problem-domain and solution requirements: A systematic mapping study, Information and Software Technology, 2019, In press

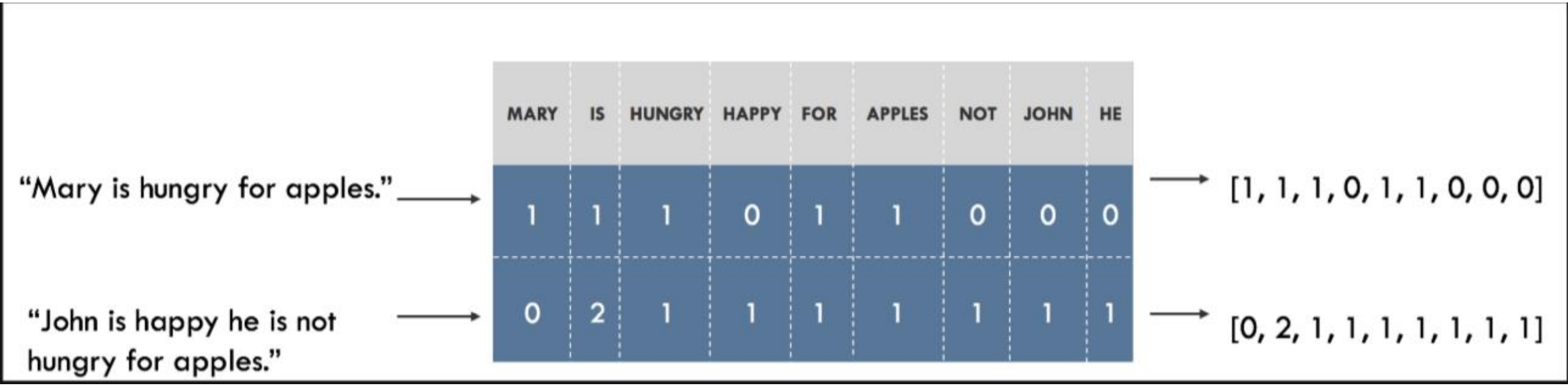
NLP and Machine learning



Source: https://www.datajango.com/deep-learning-and-nlp/

Bag of words

Methods that are used for natural language processing to represent documents where the order of words (grammar) is not important.



Source: https://blog.insightdatascience.com/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e

References

1. Bird, Christian, Tim Menzies, and Thomas Zimmermann, eds. The art and science of analyzing software data. Elsevier, 2015.
2. Arnaoudova, Venera, et al. "The use of text retrieval and natural language processing in software engineering." Proceedings of the 37th International Conference on Software Engineering-Volume 2. 2015.
3. Juergens, Elmar, et al. "Do code clones matter?." 2009 IEEE 31st International Conference on Software Engineering. IEEE, 2009.
4. Juergens, Elmar, et al. "Can clone detection support quality assessments of requirements specifications?." Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 2. ACM, 2010.
5. https://medium.com/square-corner-blog/topic-modeling-optimizing-for-human-interpretability-48a81f6ce0ed
6. Stevens, Keith, Kegelmeyer, Philip, Andrzejewski, David, and Buttler, David. Exploring topic coherence over many models and many topics. In EMNLP, 2012.