Natural Language Processing (NLP)

Lecture 2

NLP Pipeline/Tools

Before we go.. let's have a look on Recent NLP Libraries

- **NLTK:** Released 2001Latest 3.5 in April 2020
- **Spacy** **:** Released 2015
- **RE**
- **Genism**
- **Fasttext**
- **Pandas**
- **etc ..**

Remember

How to:

- Create/read/write/append for text/csv/pdf(PyPDF lib) files
- Use Pandas & Anakonda libraries
- Use RE library for searching text patterns in text context

NLP traditional PipeLine

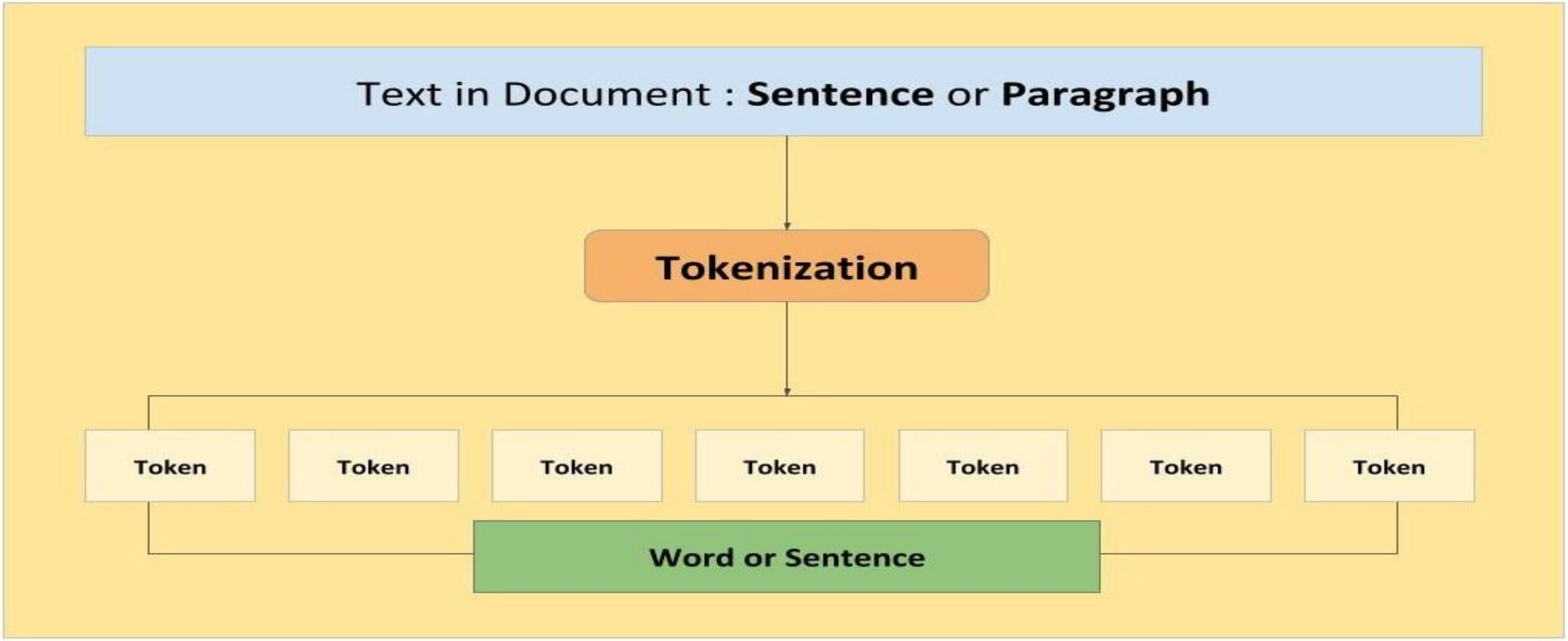
1. Tokenization
2. Sentence Segmentation
3. POS tagging
4. Stemming
5. NER
6. Stopwords
7. Matchers
8. Syntactic structure
9. T Visualization

1.Tokenization

- Dividing the sentence into a set of tokens/words
- Different from splitting as it considers the word meaning.

for Example: I’m from New York

2 tokens 1 token



Text in Document : **Sentence** or **Paragraph**

Tokenization

Token Token Token Token Token Token Token

Word or Sentence

1.Tokenization

Challenges:

1. Noun compound that are not segmented

   Such as: German & Turkish languages

2. No spaces between words such as Japanese and Chinese languages

Solution

Mix match: looking for the max length of letters to form understood meaning)

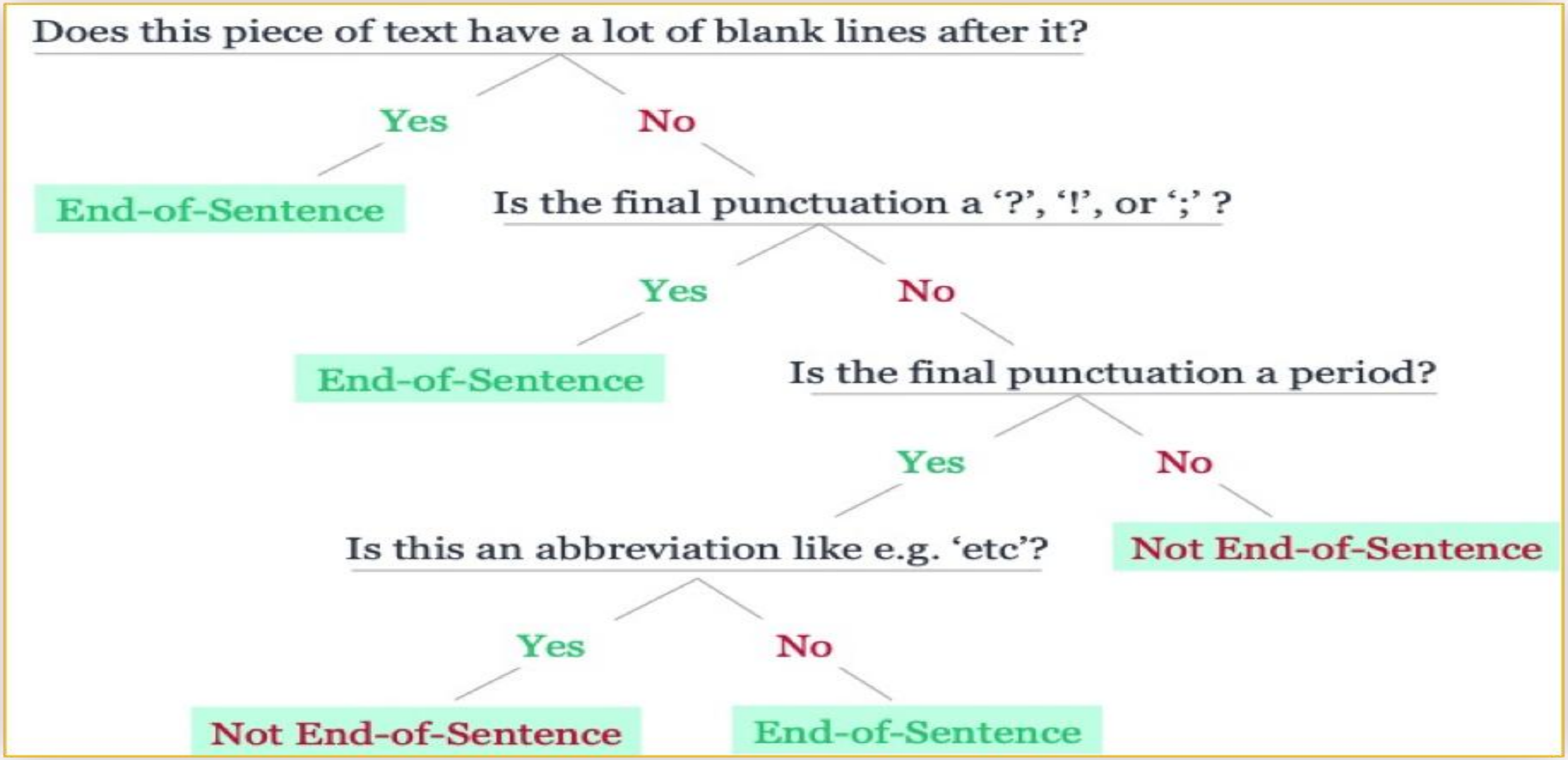
Mix/match negatives

Thecatinthehat → the cat in the hat

Thetabledownthere → theta bled own there

2. Sentence segmentation

- Dividing the context into a set of sentences
- Using ML algorithms to find End Of Statements (EOS)
- Example using Decision Tree (DR):



Does this piece of text have a lot of blank lines after it?

Yes

End-of-Sentence

No

Is the final punctuation a '?', '!', or ';' ?

Yes

End-of-Sentence

No

Is the final punctuation a period?

Yes

Is this an abbreviation like e.g. 'etc'?

Yes

Not End-of-Sentence

No

End-of-Sentence

No

Not End-of-Sentence

3. POS Tagging

- Part Of Speech ➔ POS
- Idea of POS started by Aristotle (384-322)BC
- Determine lexical category of the word based on its meaning in the context.
- Thrax (100 BC): had proposed 8 POS

  (noun, verb, article, adverb, proposition, conjunction, participle, pronoun)
- Today in our schools

  (noun, verb, adjective, adverb, preposition, conjunction, pronoun, interjection)

Closed vs. Open POS

- **Open Classes**

✓Nouns

(Proper: Egypt, KSA, Mansoura,…)

(Common : cat, dog, sky…)

✓Adjectives(new, old, long, taller, shorter)

✓Adverbs(slowly, firstly, tightly,.)

- **Closed Classes**

✓ Pronouns: I, He, she,they,his..

✓ Determiners: The, a, an

✓ Conjunctions : and, or

✓ Prepositions: on, over, under, by, in,..

✓ Particles: up, off,

✓ Interjections: oh, hey, yes, no, ..

- **Common**

- Numbers:

One, two,..1,2,3,.

- Verbs

(play, eat, run, ..)

- Numbers

…more

- Verbs

Modals (can, may, have ..)

POS Tagging challenge

- Word meaning varies according to the context
- Current models don't exceed 97% accurate
- About 11% of the word types are ambiguous regarding POS

- Example 1:

The back door ➔ adjective JJ

on my back ➔ noun NN

please, back the receipt ➔ verb VB

Information sources for POS Tagging

1. **Knowledge of neighboring words in the context**

   I saw Dena yesterday➔ verb VB⇧

   I have used my saw to cut the tree ➔ noun NN⇧

2. **Knowledge of word probabilities**

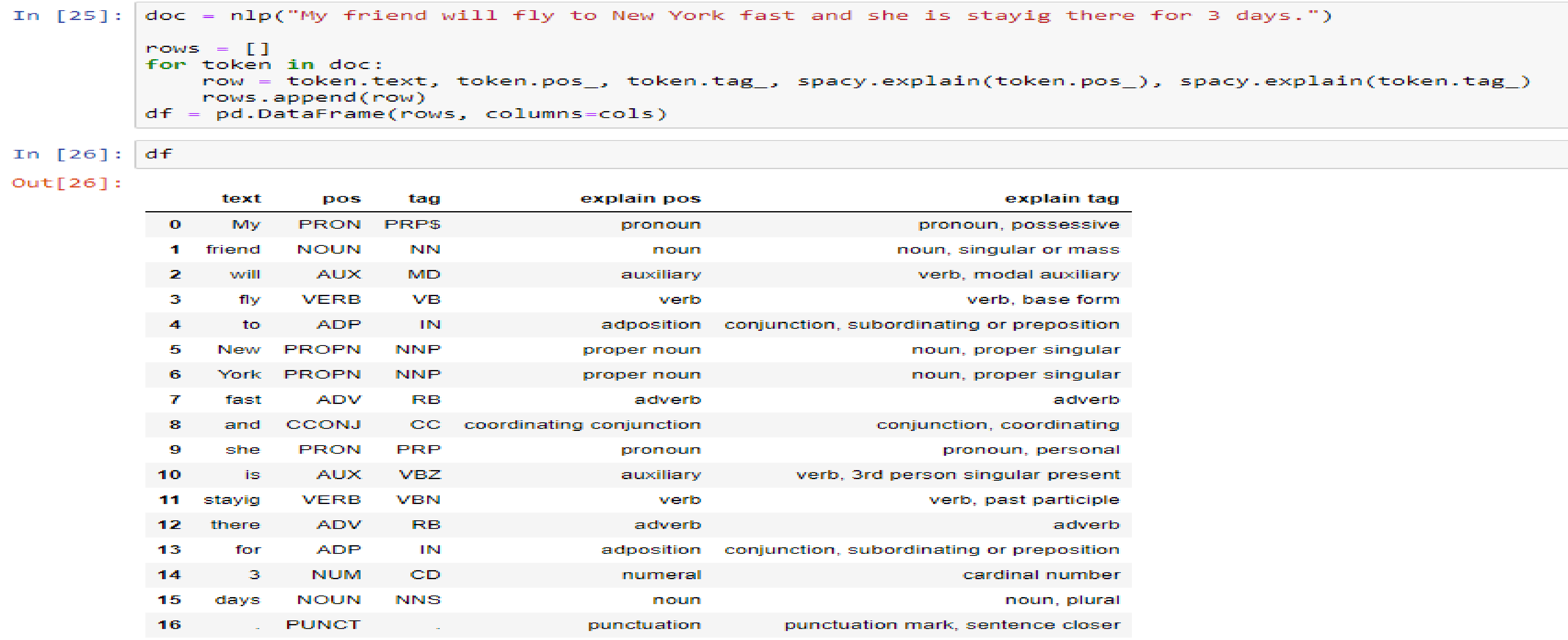
   I saw Mona yesterday➔ verb VB ⇧

   I saw this piece of *wood* ➔ another verb VB⇧

3. **Information about the word itself**

   - **Capitalization:** Egypt, .. ➔ noun NN ⇧
   - **Prefixes:** Uncomfortable, misunderstanding… ➔ adjective JJ⇧
   - **Suffixes:** importantly, .. ➔ adverb RB⇧
   - **Word shape :** 2-years old boy➔ adjective JJ⇧

POS tagging in Spacy



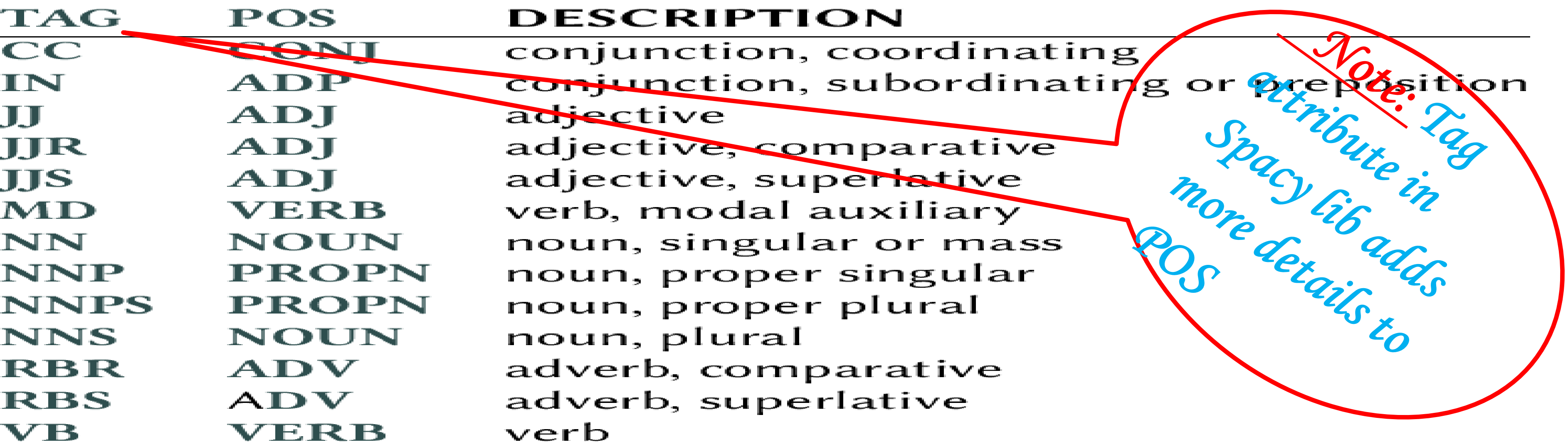
In [25]:
```
doc = nlp("My friend will fly to New York fast and she is stayig there for 3 days.")

rows = []
for token in doc:
    row = token.text, token.pos_, token.tag_, spacy.explain(token.pos_), spacy.explain(token.tag_)
    rows.append(row)
df = pd.DataFrame(rows, columns=cols)
```

In [26]:
```
df
```

Out[26]:

	text	pos	tag	explain pos	explain tag
0	My	PRON	PRP$	pronoun	pronoun, possessive
1	friend	NOUN	NN	noun	noun, singular or mass
2	will	AUX	MD	auxiliary	verb, modal auxiliary
3	fly	VERB	VB	verb	verb, base form
4	to	ADP	IN	adposition	conjunction, subordinating or preposition
5	New	PROPN	NNP	proper noun	noun, proper singular
6	York	PROPN	NNP	proper noun	noun, proper singular
7	fast	ADV	RB	adverb	adverb
8	and	CCONJ	CC	coordinating conjunction	conjunction, coordinating
9	she	PRON	PRP	pronoun	pronoun, personal
10	is	AUX	VBZ	auxiliary	verb, 3rd person singular present
11	stayig	VERB	VBN	verb	verb, past participle
12	there	ADV	RB	adverb	adverb
13	for	ADP	IN	adposition	conjunction, subordinating or preposition
14	3	NUM	CD	numeral	cardinal number
15	days	NOUN	NNS	noun	noun, plural
16	.	PUNCT	.	punctuation	punctuation mark, sentence closer



TAG	POS	DESCRIPTION
CC	CONJ	conjunction, coordinating
IN	ADP	conjunction, subordinating or preposition
JJ	ADJ	adjective
JJR	ADJ	adjective, comparative
JJS	ADJ	adjective, superlative
MD	VERB	verb, modal auxiliary
NN	NOUN	noun, singular or mass
NNP	PROPN	noun, proper singular
NNPS	PROPN	noun, proper plural
NNS	NOUN	noun, plural
RBR	ADV	adverb, comparative
RBS	ADV	adverb, superlative
VB	VERB	verb

Note: Tag attribute in Spacy lib adds more details to POS

4. Stemming & Lemmatization

- Stemming reduces the word to its stem by removing all affixes
- Plays, played, playing, player ➔ stem: play
- NlTK supports stemming
- Spacy doesn't support stemming, instead supports Lemmatization
- Lemmatization additionally reduces the word to its root
- Am, are, is, was, been ➔ be
- Lemmatization is useful in word disambiguation

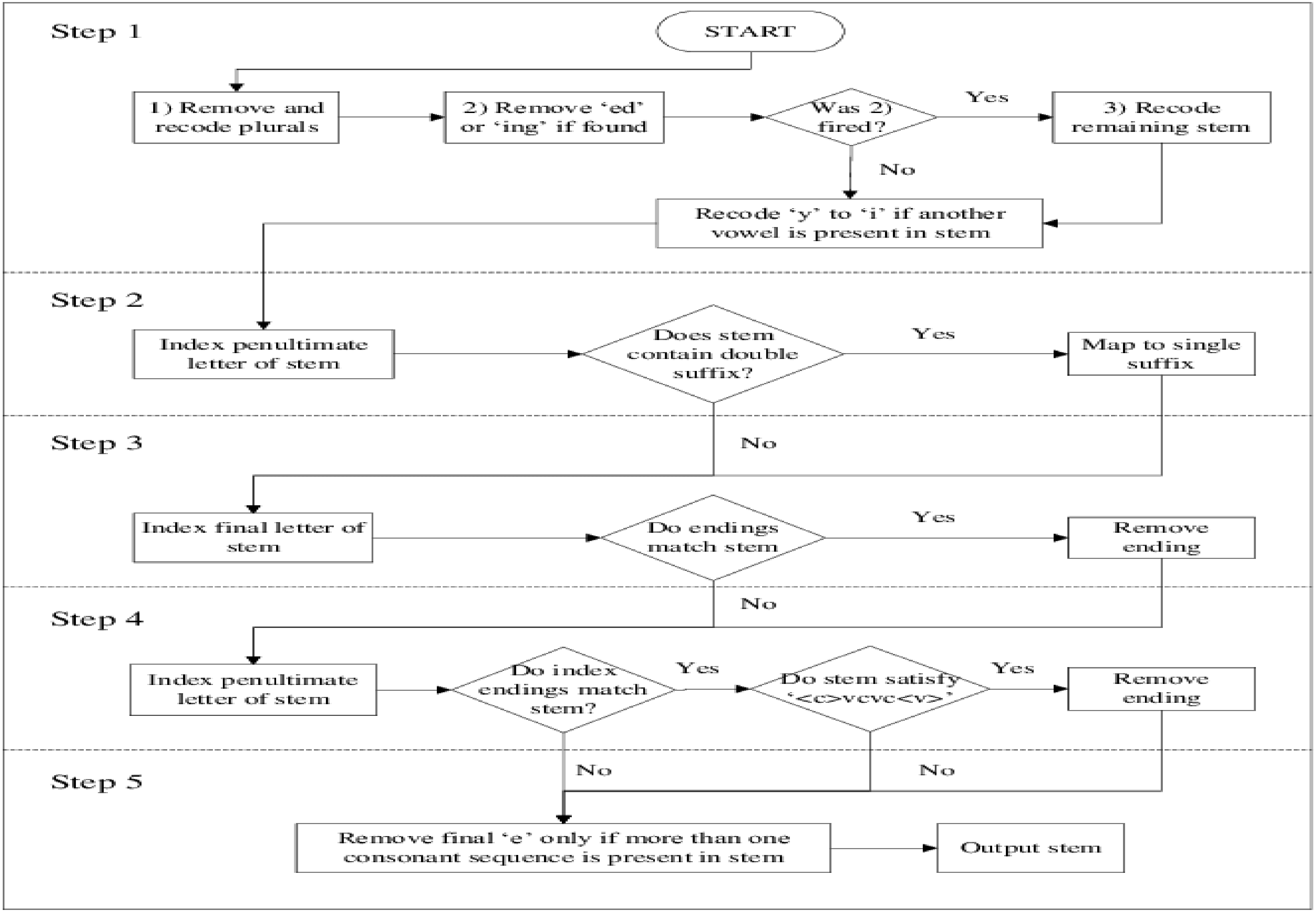
In Stemming & Lemmatization

- Normalization :
  - Remove punctuations : U.S.A ➔ USA
  - Remove plural S:plays➔ play
- Case folding : capital initials cause determination according to context

  US Vs. us ➔ Unites States or us

  Fed Vs. fed ➔ Federal Reserve System or PP of Feed
- Word reduction to its stem or root due to the training of Stemming /Lemmatization model

Porter Stemming algorithm (NLTK)



Step 1

START

1) Remove and recode plurals

2) Remove 'ed' or 'ing' if found

Was 2) fired?

Yes

3) Recode remaining stem

No

Recode 'y' to 'i' if another vowel is present in stem

Step 2

Index penultimate letter of stem

Does stem contain double suffix?

Yes

Map to single suffix

No

Step 3

Index final letter of stem

Do endings match stem

Yes

Remove ending

No

Step 4

Index penultimate letter of stem

Do index endings match stem?

Yes

Do stem satisfy '<c>vcvc<v>'

Yes

Remove ending

No

No

Step 5

Remove final 'e' only if more than one consonant sequence is present in stem

Output stem

Assignment 1

Check for the quality of stemming of NLTK Vs Lemmatization of Spacy by coding on simple text document.

5. Name Entity Recognition (NER)

Find and classify important names in text such as person names, organizations names, cities, countries, Dates, currencies, ..etc.

Example:

The decision by the independent MP **Andrew Wikie** to withdraw his support for the minority **Labor** government sounded dramatic but it should not further threaten its stability. When, after the **2010** election, **Wikie**, **Rob Oakeshott**, **TonyWindsor** and the **Greens** agreed to support **Labor**, they gave just two guarantee: confidence and supply.

Andrew → **PER**

Wikie → **PER**

Labor → **ORG**

Decision → **O …(Other)**

For successful NER model

- Huge data collection for entities names
- Manual detection for large amount of entities:

Egypt ➔ country, IBM ➔ Organization, …

- Efficient features detection (may be : pos ,current token, last token, etc)
- Good training for NER model

Remember to review entity attribute in Spacy which refers to NER

6. Stopwords in NLP

- Frequently repeated words along the context
- Its removal doesn't affect the meaning of the context
- Such as: the, a, was, and, or,..
- Some applications are affected by the Stopwords removal such as Chatbot
- Stopwords lists vary among NLP libraries
- You can edit Stopwords list by removing or appending to the open source libraries such as Spacy.
- NLTK supports Stopwords list for Arabic language.

7. Matchers

- A tool that admits the connection between different words for referring to the same meaning.

- Such as different typing for words:
  - ➔ Youssef, Yossef, Yossuf, Yusuf ,Yossof ,Yusf ,…
  - ➔ colour, color,..
  - ➔ solar power, solar-power, solarpower
  - ➔ cupboard, cupbord

- Or different words with one meaning such as :
  - ➔ put on, wear,…
  - ➔ wardrobe, closet, cupboard

- By coding, you create a set of different pattern objects, then add them to one matcher.

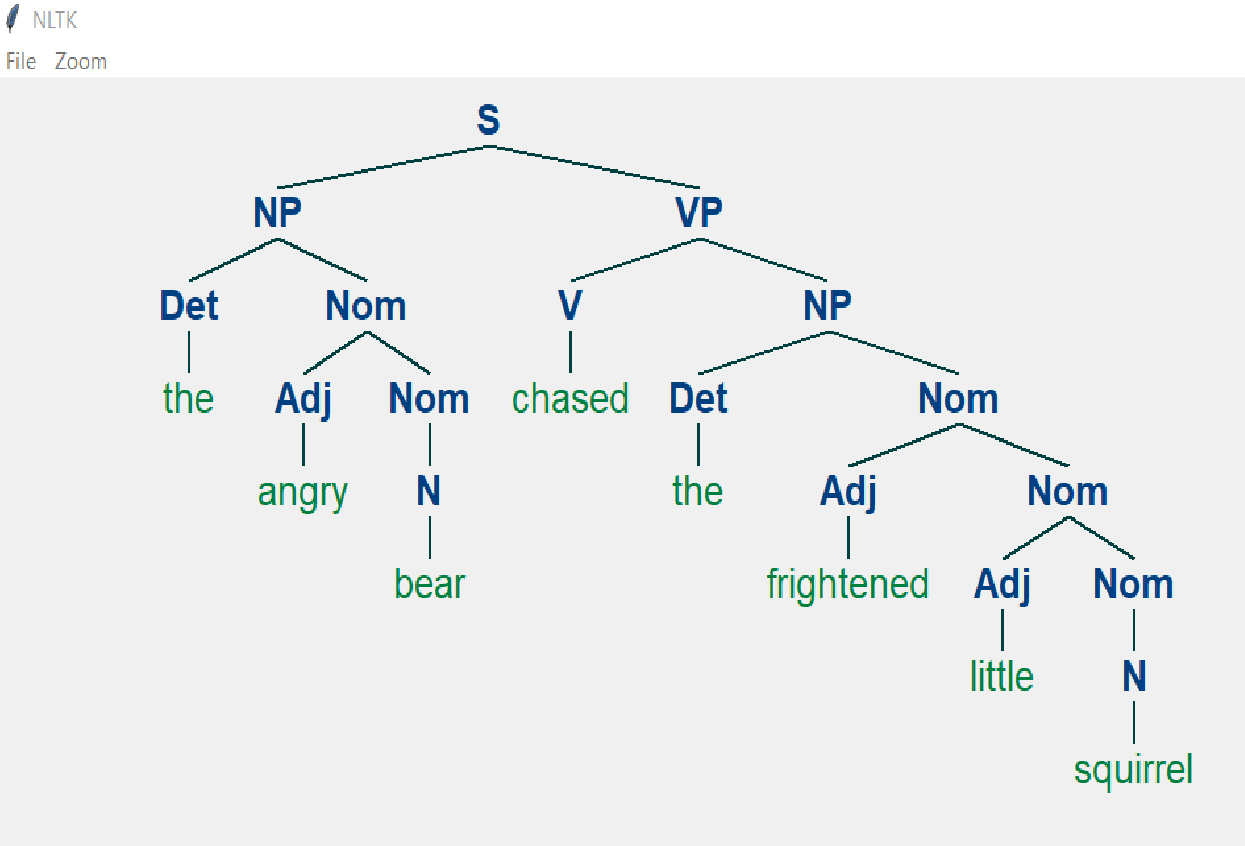
8. Syntactic structure

- Structuring the words in sentence based on its grammar type and its dependency on other words(over regular POS)

- No standard syntactic structure for every sentence especially ambiguous sentences

Example: **S : the angry bear chased the frightened little squirrel**

- S: Sentence
- NP: Noun Phrase
- VP: Verb Phrase
- Det: Determiner
- PP : Prepositional Phrase
- ADJP : Adjective Phrase
- ADVP : Adverb Phrase
- N : Noun
- V : Verb
- P : Preposition



Syntactic Structure Models

1. Constituency model

Divide the sentence into small pieces with collecting the pieces that refer to a complete meaning.

Ex: FED raises interest rate

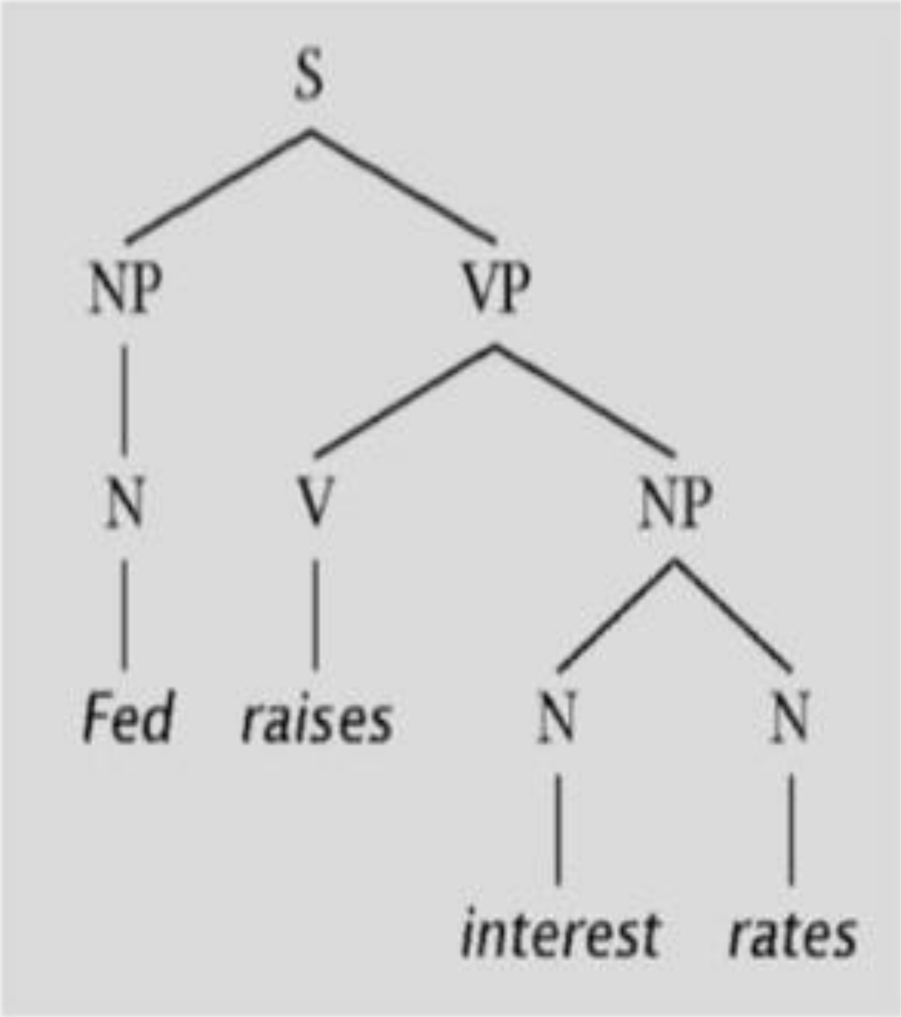
FED .. No meaning N

FED raises .. No meaning X

raises interest .. No meaning X

interest rate .. Has a meaning ADJP

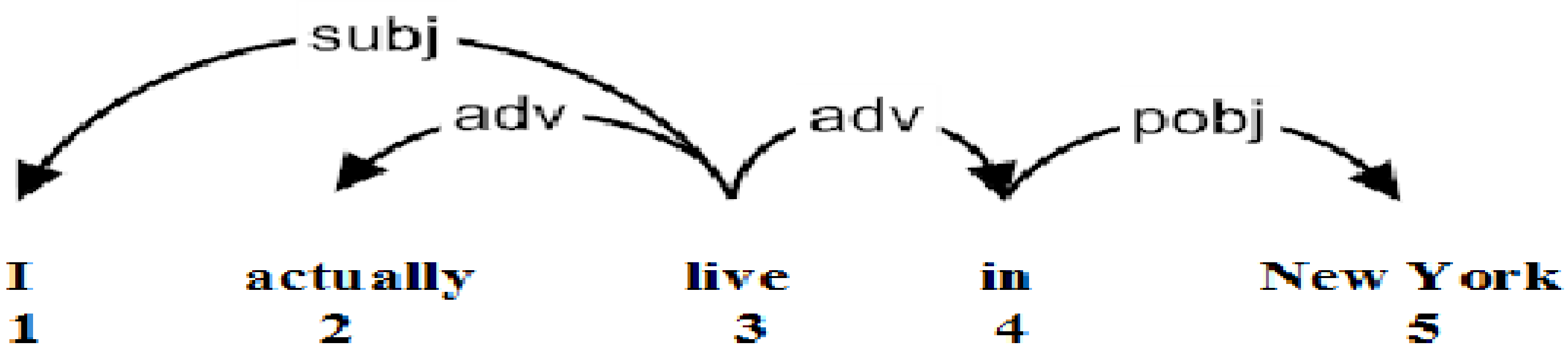
Raises interest rate .. Has a meaning VP



Syntactic Structure Models

2. Dependency Model

- Starts with the most important word in the sentence
- Append other words that have relations with this word



9. T-Visualization

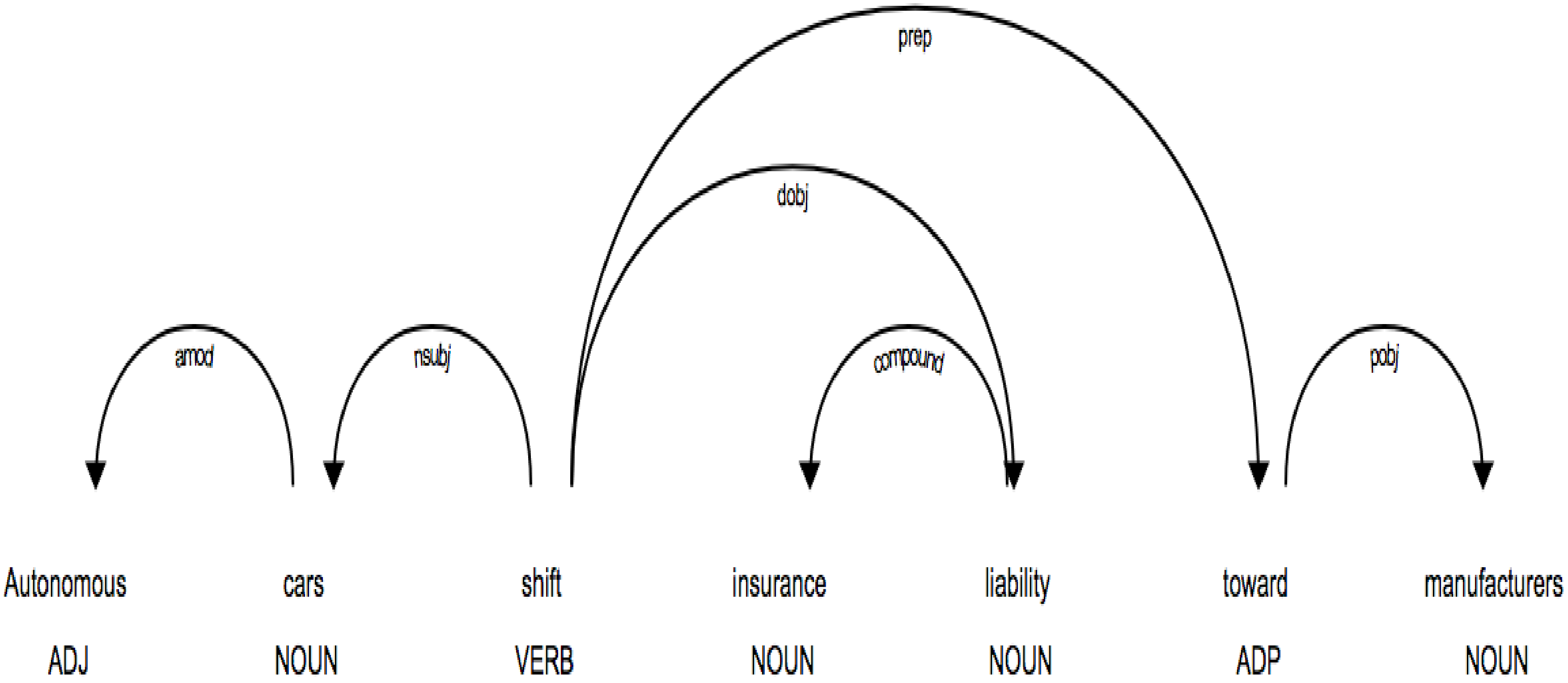
- A tool to display the relations (arrows and graphs) between the words visually and clearly

- By displaCy tool from Spacy library

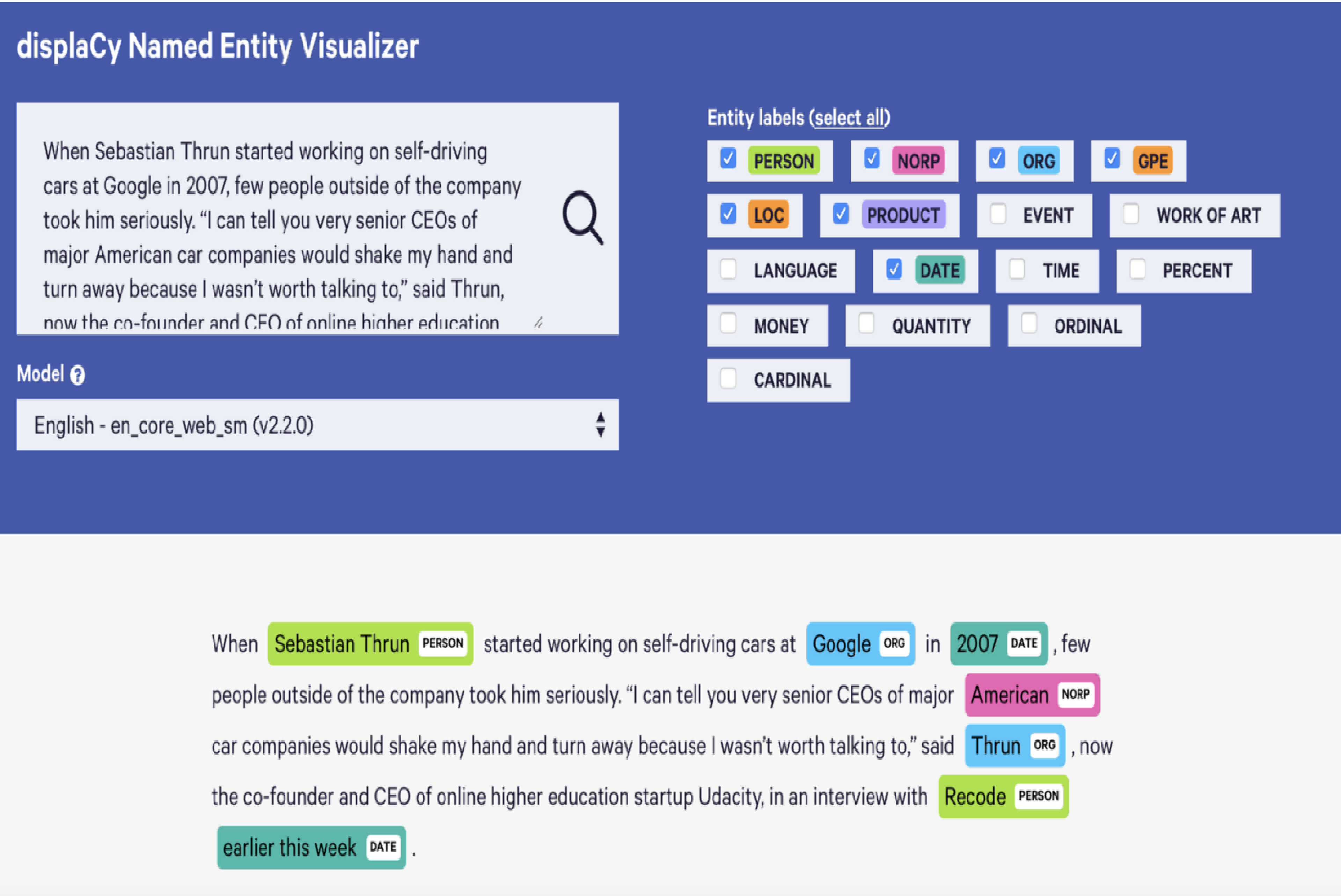
- **Two styles** : 1. Dependencies

  2. Entities

Dependency T visualization



Entity T visualization



displaCy Named Entity Visualizer

When Sebastian Thrun started working on self-driving cars at Google in 2007, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun, now the co-founder and CEO of online higher education

Entity labels (select all)

☑ PERSON ☑ NORP ☑ ORG ☑ GPE
☑ LOC ☑ PRODUCT □ EVENT □ WORK OF ART
□ LANGUAGE ☑ DATE □ TIME □ PERCENT
□ MONEY □ QUANTITY □ ORDINAL
□ CARDINAL

Model

English - en_core_web_sm (v2.2.0)

When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE , few people outside of the company took him seriously. "I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun ORG , now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode PERSON earlier this week DATE .

THANK YOU … ☺