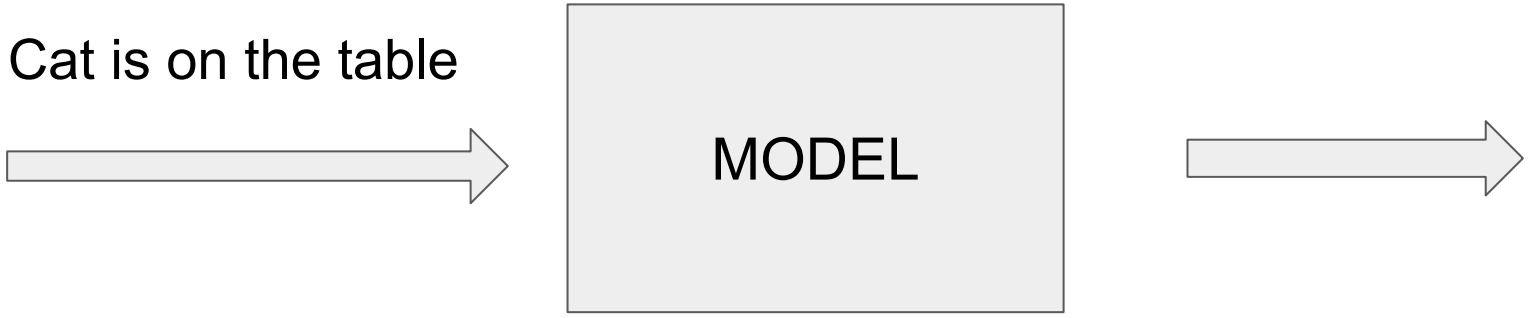
Text representation

Problem

The Cat is on the table

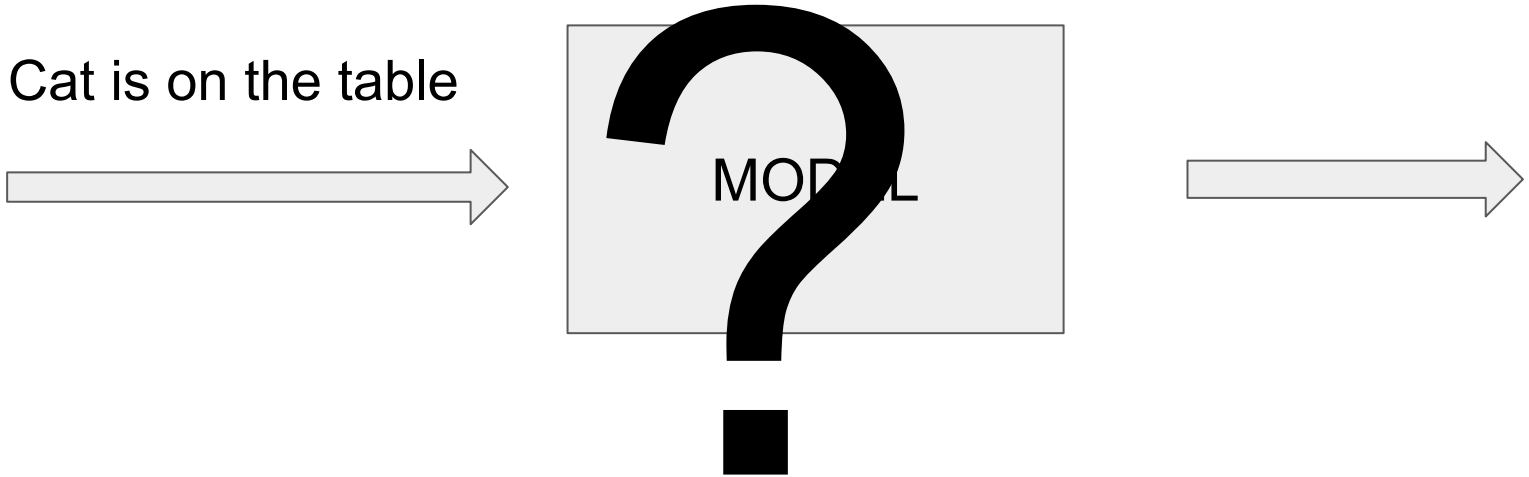
Problem

The Cat is on the table



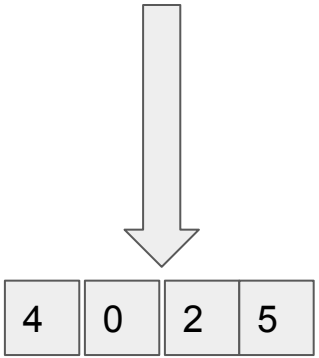
Problem

The Cat is on the table



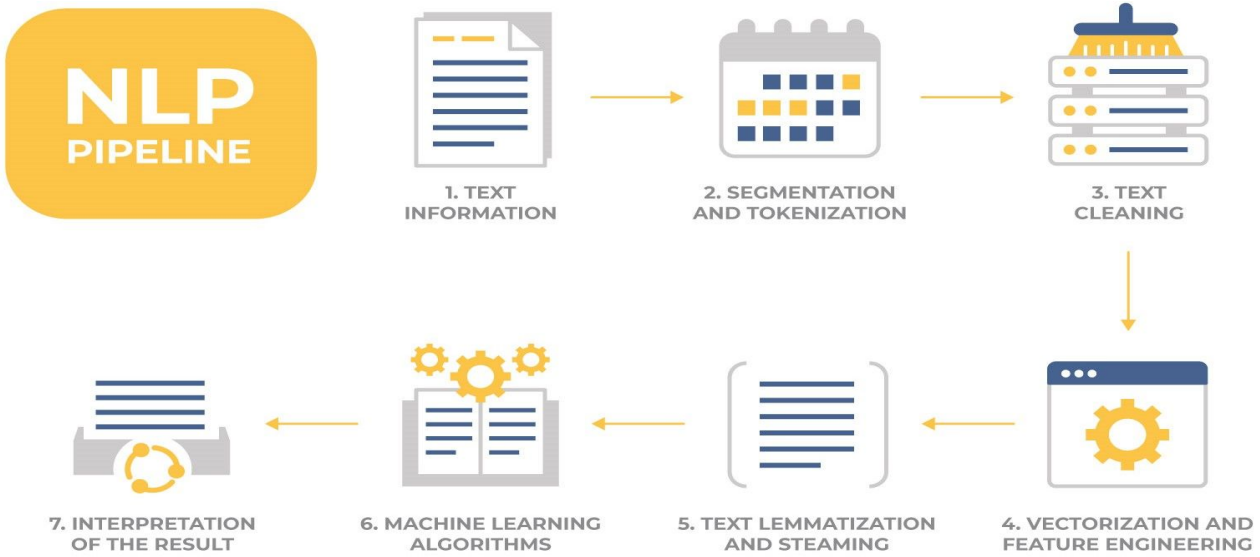
Problem

The Cat is on the table

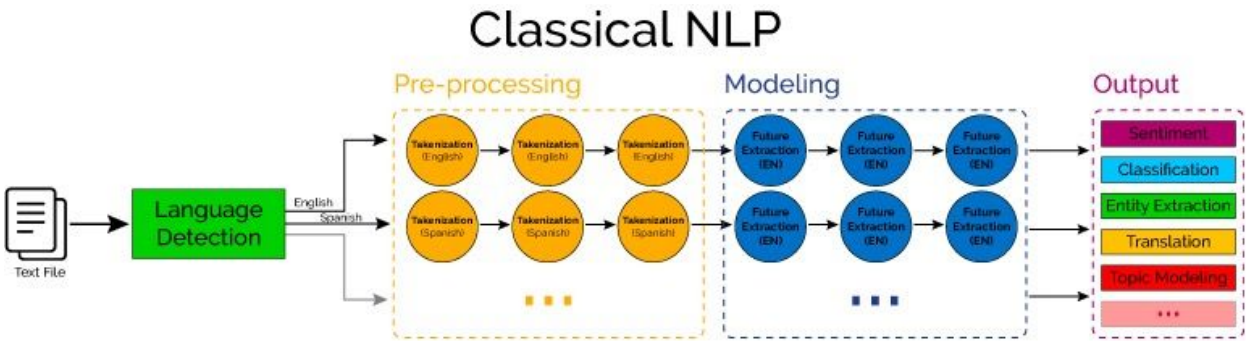


Traditional NLP

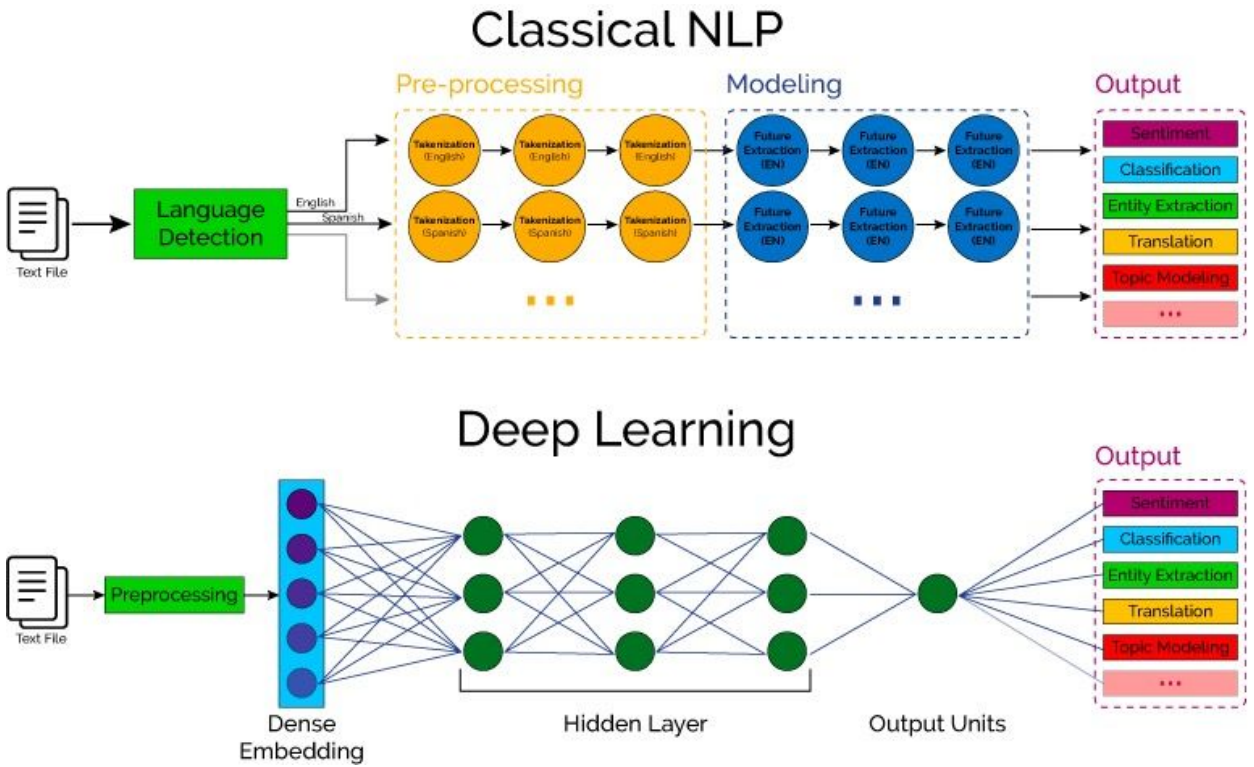
Traditional NLP



Traditional NLP



Traditional NLP vs Deep Learning



NLTK

Old but gold

Most features and algorithms

No embedding

Academic

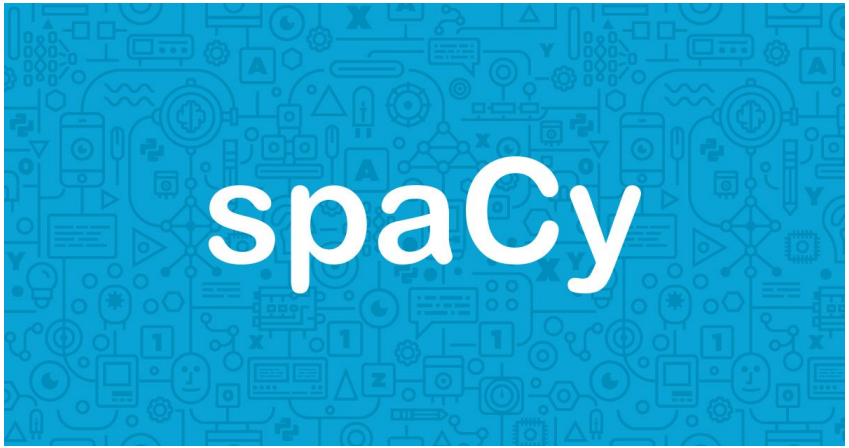
spaCy

Practicality over completeness

Modern

Great doc

Word vectors



Gensim

First embedding

Modern

Great doc

Word vectors

GloVe

Embedding

Modern

Great doc

Word vectors

Huggingface

State-of the-art models

Transformers

Ultra-Fast tokenization

Word vectors

Tokenizer

Definition

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens , often removing punctuation. Here is an example of tokenization:

Example

Input: "This is the best movie of all time."

Output: ["This" "is" "the" "best" "movie" "of" "all" "time"]

NLTK

```
import nltk
nltk.download('punkt')

text_tokens = word_tokenize(sample_text)
text_tokens
```

Stemming

Stemming and Lemmatization

reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

Reduce the dimensionality

Example

am, are, is ⇒ be

car, cars, car's, cars' ⇒ car

Spacymoji

Transform emoji in text

Stopwords

What are Stopwords?

Very common words

Don’t add signal

After tokenization

What are Stopwords?

The pizza is on the table.

The pizza is on the table.

Methods

List

Frequency

List

the
is
in
where
when

Frequency

If word has high frequency

Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', ' the', 'lazy', 'dog']

	The	quick	brown	fox	jumps	over	lazy	dog
Data	2	1	1	1	1	1	1	1

Frequency - How to Count

- Countvectorizer
- HashingVectorizer
- TF-IDF

When to remove them

- Text Classification
- Spam Filtering
- Language Classification
- Genre Classification
- Caption Generation
- Auto-Tag Generation

When to keep them

- Machine Translation
- Language Modeling
- Text Summarization
- Question-Answering problems

Part-Of-Speech

Problem

“Help” can be used as a name or a noun

Depends on the context

When to keep them

Part-Of-Speech (POS) tagging consist on adding some extra information to the word to specify its role in the sentence.

Examples

I eat pizza.

becomes:

I (PRP) eat (VB) pizza (NN)

Where PRP stands for personal pronoun, VB for verb and NN for noun.

Methods

- Machine Translation
- Language Modeling
- Text Summarization
- Question-Answering problems

Encoders

One-host encoder

- How do I represent words that don’t have a non linear

One, two, three, ... - > 1, 2, 3, …
Paris, London, NY, … -> ???

One-hot encoder

	o_ENE	o_ESE	o_East	o_NE	o_NNE	o_NNW	o_NW	o_North	o_SE	o_SSE	o_SSW	o_SW	o_South	o_Variable	o_WSW
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Classifiers

Classifiers

- Naive Bayes
- EM algorithms
- Support vector machines
- Random Forest
- ….

THANK YOU!

You can contact me at

www.ideai.io info@ideai.io