NLP Libraries

① NLTK → Predefined function

↓

String processing library

Sentence tokenization

```
['Natural Language Processing (NLP) is a field that combines computer science, artificial intelligence and
language studies.', 'It helps computers understand, process and create human language in a way that
makes sense and is useful.', 'With the growing amount of text data from social media, websites and
other sources, NLP is becoming a key tool to gain insights and automate tasks like analyzing text or
translating languages.']
```

Word tokenization

```
['Natural', 'Language', 'Processing', '(', 'NLP', ')', 'is', 'a', 'field', 'that', 'combines', 'computer', 'science', ',',
'artificial', 'intelligence', 'and', 'language', 'studies', '.', 'It', 'helps', 'computers', 'understand', ',', 'process', 'and',
'create', 'human', 'language', 'in', 'a', 'way', 'that', 'makes', 'sense', 'and', 'is', 'useful', '.', 'With', 'the', 'growing',
'amount', 'of', 'text', 'data', 'from', 'social', 'media', ',', 'websites', 'and', 'other', 'sources', ',', 'NLP', 'is', 'becoming',
'a', 'key', 'tool', 'to', 'gain', 'insights', 'and', 'automate', 'tasks', 'like', 'analyzing', 'text', 'or', 'translating',
'languages', '.']
```

Normalise → lower case

filtered_tokens

```
['natural', 'language', 'processing', 'nlp', 'field', 'combines', 'computer', 'science', 'artificial', 'intelligence',
'language', 'studies', 'helps', 'computers', 'understand', 'process', 'create', 'human', 'language', 'way', 'makes',
'sense', 'useful', 'growing', 'amount', 'text', 'data', 'social', 'media', 'websites', 'sources', 'nlp', 'becoming', 'key',
'tool', 'gain', 'insights', 'automate', 'tasks', 'like', 'analyzing', 'text', 'translating', 'languages']
```

root words

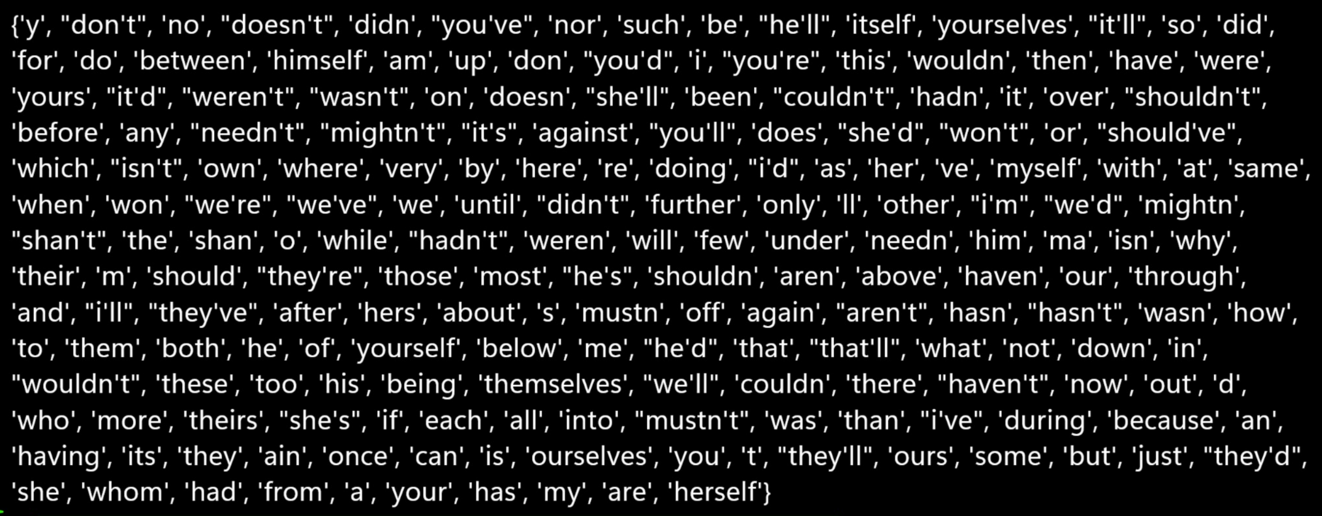
punkt_tab in NLTK refers to a specific data package used for sentence tokenization, a process of dividing text into individual sentences. It is the modern, pickle-free replacement for the older punkt package, which utilized unsafe pickles.

Stemming

```
['natur', 'languag', 'process', 'nlp', 'field', 'combin', 'comput', 'scienc', 'artifici', 'intellig', 'languag', 'studi', 'help',
'comput', 'understand', 'process', 'creat', 'human', 'languag', 'way', 'make', 'sens', 'use', 'grow', 'amount', 'text', 'data',
'social', 'media', 'websit', 'sourc', 'nlp', 'becom', 'key', 'tool', 'gain', 'insight', 'autom', 'task', 'like', 'analyz', 'text',
'translat', 'languag']
```

Lemmatization → meaningful output

```
['natural', 'language', 'processing', 'nlp', 'field', 'combine', 'computer', 'science', 'artificial', 'intelligence', 'language',
'study', 'help', 'computer', 'understand', 'process', 'create', 'human', 'language', 'way', 'make', 'sense', 'useful',
'growing', 'amount', 'text', 'data', 'social', 'medium', 'website', 'source', 'nlp', 'becoming', 'key', 'tool', 'gain',
'insight', 'automate', 'task', 'like', 'analyzing', 'text', 'translating', 'language']
```



```
{'y', "don't", 'no', "doesn't", 'didn', "you've", 'nor', 'such', 'be', "he'll", 'itself', 'yourselves', "it'll", 'so', 'did',
'for', 'do', 'between', 'himself', 'am', 'up', 'don', "you'd", 'i', "you're", 'this', 'wouldn', 'then', 'have', 'were',
'yours', "it'd", "weren't", "wasn't", 'on', 'doesn', "she'll", 'been', "couldn't", 'hadn', 'it', 'over', "shouldn't",
'before', 'any', "needn't", "mightn't", "it's", 'against', "you'll", 'does', "she'd", "won't", 'or', "should've",
'which', "isn't", 'own', 'where', 'very', 'by', 'here', 're', 'doing', "i'd", 'as', 'her', 've', 'myself', 'with', 'at', 'same',
'when', 'won', "we're", "we've", 'we', 'until', "didn't", 'further', 'only', 'll', 'other', "i'm", "we'd", 'mightn',
"shan't", 'the', 'shan', 'o', 'while', "hadn't", 'weren', 'will', 'few', 'under', 'needn', 'him', 'ma', 'isn', 'why',
'their', 'm', 'should', "they're", 'those', 'most', "he's", 'shouldn', 'aren', 'above', 'haven', 'our', 'through',
'and', "i'll", "they've", 'after', 'hers', 'about', 's', 'mustn', 'off', 'again', "aren't", 'hasn', "hasn't", 'wasn', 'how',
'to', 'them', 'both', 'he', 'of', 'yourself', 'below', 'me', "he'd", 'that', "that'll", 'what', 'not', 'down', 'in',
"wouldn't", 'these', 'too', 'his', 'being', 'themselves', "we'll", 'couldn', 'there', "haven't", 'now', 'out', 'd',
'who', 'more', 'theirs', "she's", 'if', 'each', 'all', 'into', "mustn't", 'was', 'than', "i've", 'during', 'because', 'an',
'having', 'its', 'they', 'ain', 'once', 'can', 'is', 'ourselves', 'you', 't', "they'll", 'ours', 'some', 'but', 'just', "they'd",
'she', 'whom', 'had', 'from', 'a', 'your', 'has', 'my', 'are', 'herself'}
```

stop words

doc

Natural Language Processing (NLP) is a field that combines computer science, artificial intelligence and language studies. It helps computers understand, process and create human language in a way that makes sense and is useful. With the growing amount of text data from social media, websites and other sources, NLP is becoming a key tool to gain insights and automate tasks like analyzing text or translating languages.

Natural Language Processing (NLP) is a field that combines computer science, artificial intelligence and language studies.

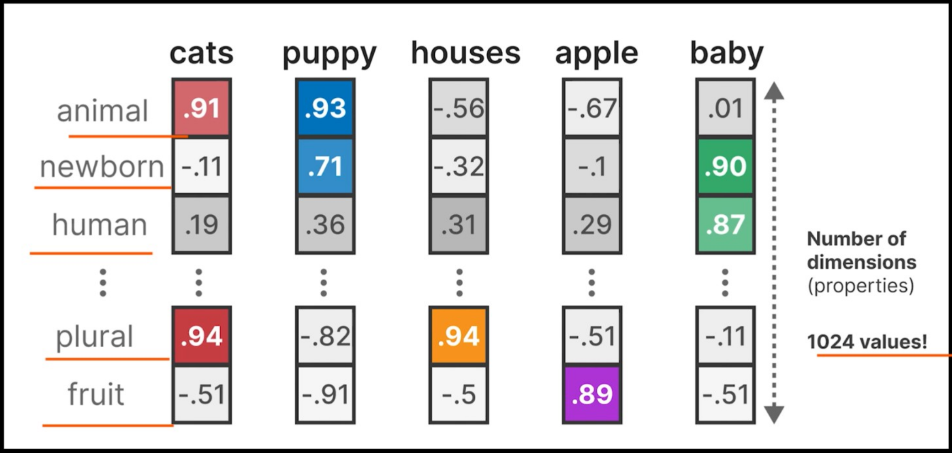
It helps computers understand, process and create human language in a way that makes sense and is useful.

With the growing amount of text data from social media, websites and other sources, NLP is becoming a key tool to gain insights and automate tasks like analyzing text or translating languages.

object oriented

③ gensim → word2vec

static embeddings



	cats	puppy	houses	apple	baby
animal	.91	.93	-.56	-.67	.01
newborn	-.11	.71	-.32	-.1	.90
human	.19	.36	.31	.29	.87
⋮	⋮	⋮	⋮	⋮	⋮
plural	.94	-.82	.94	-.51	-.11
fruit	-.51	-.91	-.5	.89	-.51

Number of dimensions (properties)

1024 values!

→ properties

→ static embeddings

Google

Attention is all you need → Transformers Architechure

English

Hindi

GPT

Encoder Decoder Architechure

GPT5

Input

output

BERT

NLP

Language Models

↓

to predict the next upcoming word

① Autocompletion task

I hope you are doing good.

Pretraining → $?(\hat{y} \mid x_1, x_2, x_3)$ → Conditional Probability & Bayes Theorem

DB

$\overset{x_1}{\text{I}}$ $\overset{x_2}{\text{love}}$ $\overset{x_3}{\text{reading}}$ → $\overset{\hat{y}}{\text{books}}$.

→ newspaper

finetuning

$$\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A) * P(A)}{P(B)}}$$

$$P(B \mid A) = \frac{P(A \overset{\checkmark}{\cap} B)}{P(A)}$$

Generative AI → generate the Content

GPT → LLM (Large Language Models)

Deeplearning.ai

3 Videos → Agentic AI, RAG → 1 Project, 2 RAG