DECODING
DATA SCIENCE

NATURA
LANGUAG
PROCES IN

NLP

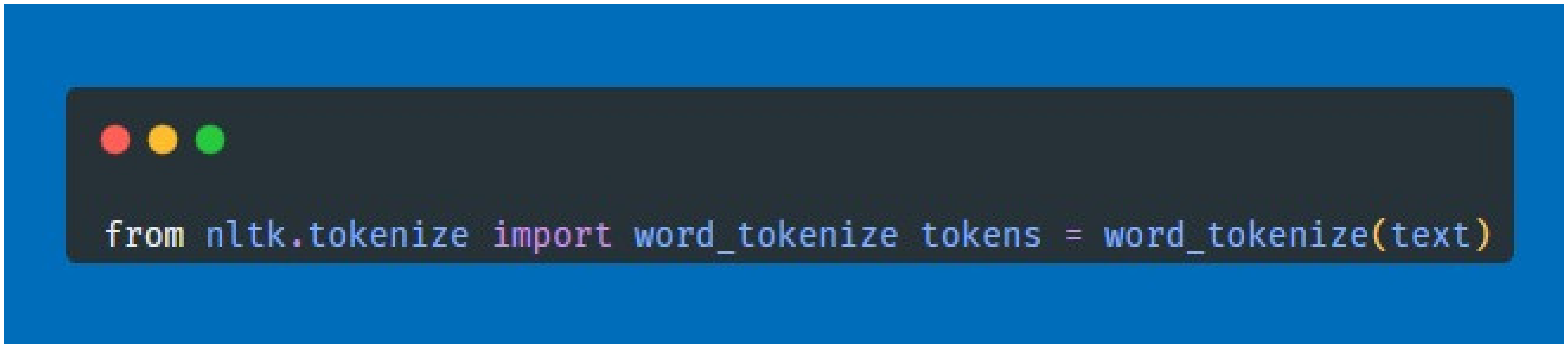
Cheat Sheet

NAS.IO/ARTIFICIALINTELLIGENCE

1. Tokenization

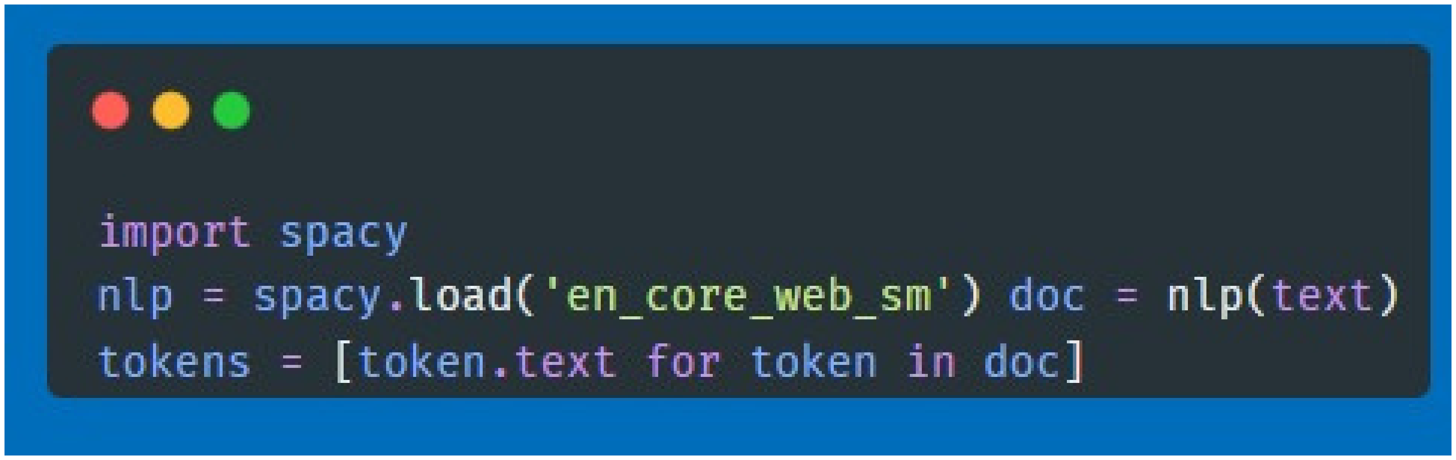
Tokenization is the process of breaking up text into words, phrases, symbols, or other meaningful elements, which are called tokens.

- NLTK Word Tokenization:



```
from nltk.tokenize import word_tokenize tokens = word_tokenize(text)
```

- Spacy Word Tokenization:

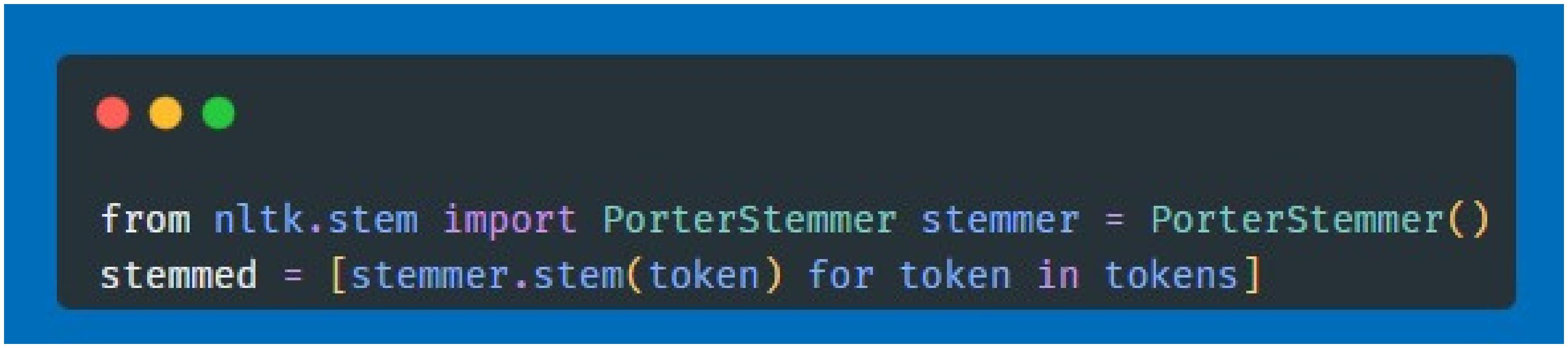


```
import spacy
nlp = spacy.load('en_core_web_sm') doc = nlp(text)
tokens = [token.text for token in doc]
```

2. Stemming and Lemmatization

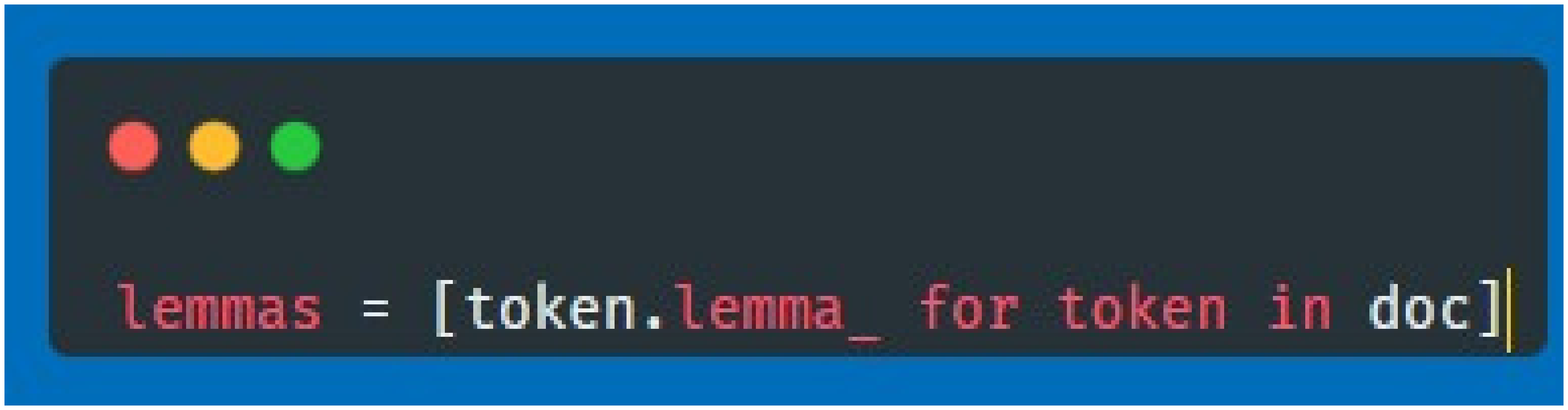
Stemming and Lemmatization are techniques used to extract the base form of the words by removing its inflection.

- NLTK Stemming:



```
from nltk.stem import PorterStemmer stemmer = PorterStemmer()
stemmed = [stemmer.stem(token) for token in tokens]
```

- Spacy Lemmatization:

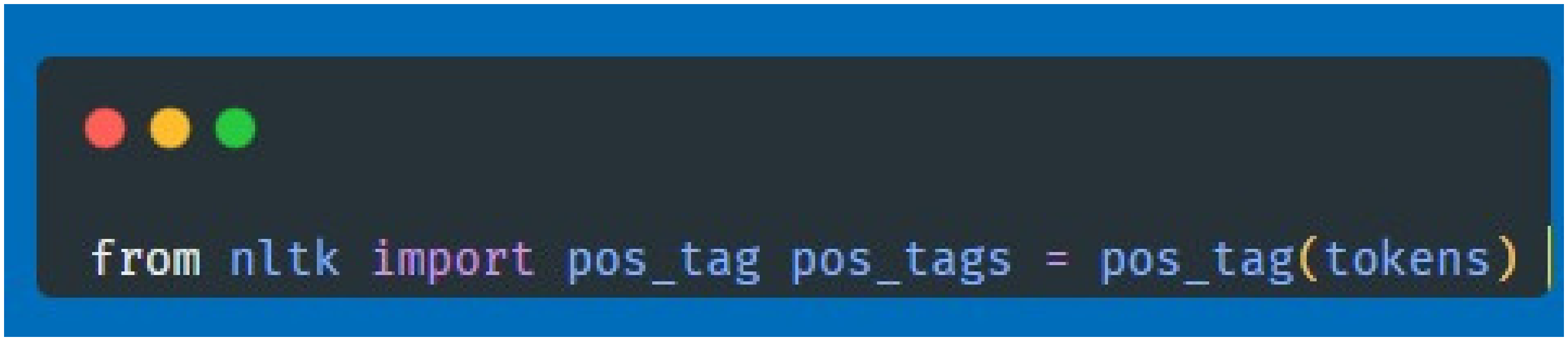


```
lemmas = [token.lemma_ for token in doc]
```

3. Part-of-Speech (POS) Tagging

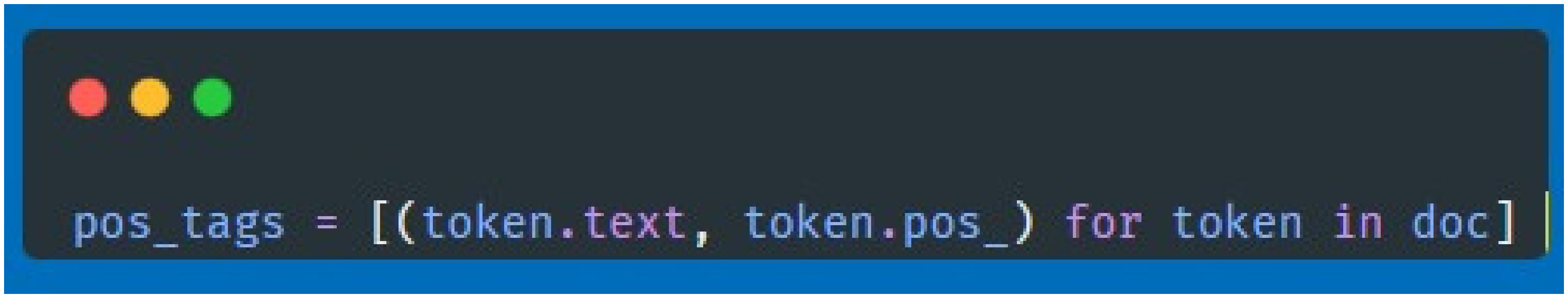
POS tagging is the task of labeling the words in a sentence with their appropriate part of speech.

- NLTK POS Tagging:



```
from nltk import pos_tag pos_tags = pos_tag(tokens)
```

- Spacy POS Tagging:

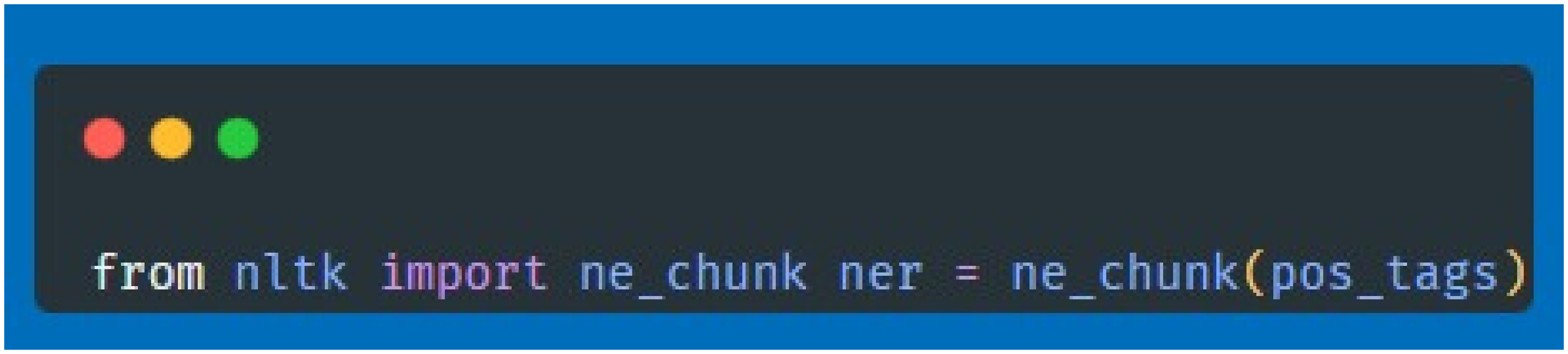


```
pos_tags = [(token.text, token.pos_) for token in doc]
```

4. Named Entity Recognition (NER)

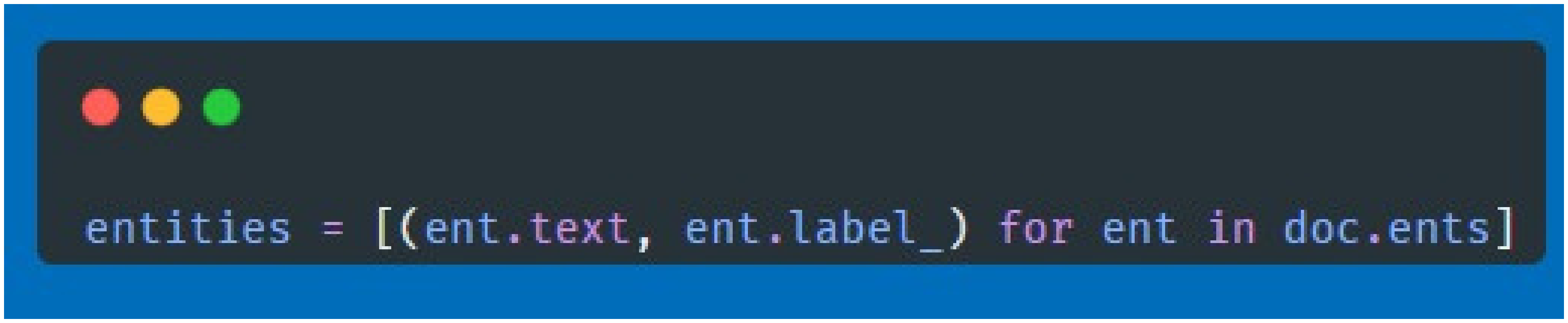
NER is the process of locating named entities in text and classifying them into predefined categories.

- NLTK NER:



```
from nltk import ne_chunk ner = ne_chunk(pos_tags)
```

- Spacy NER:

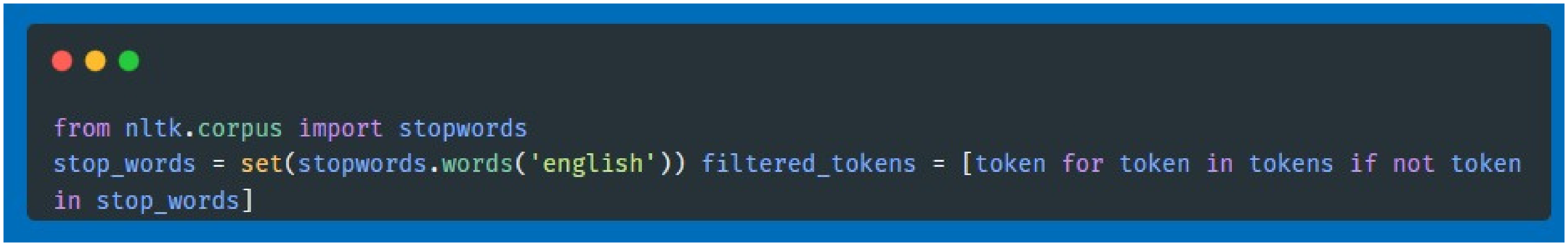


```
entities = [(ent.text, ent.label_) for ent in doc.ents]
```

5. Stopword Removal

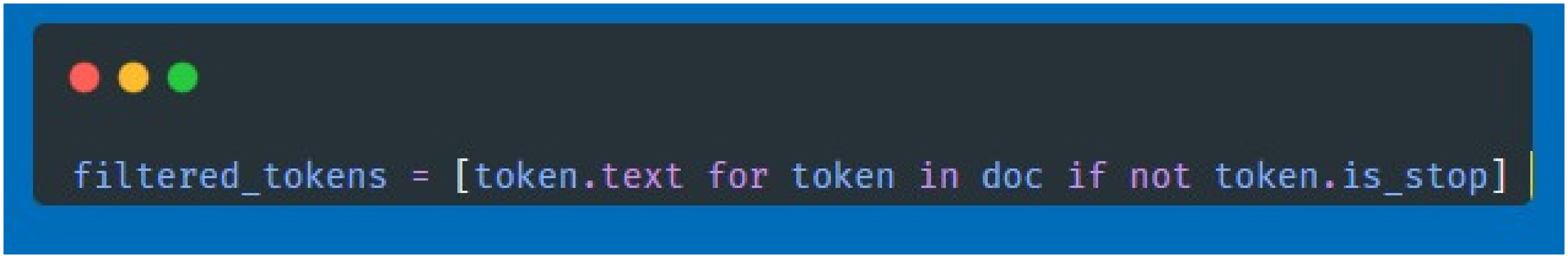
Stopwords are the most common words in a language that are to be filtered out before processing the text data.

- NLTK Stopword Removal:



```
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english')) filtered_tokens = [token for token in tokens if not token
in stop_words]
```

- Spacy Stopword Removal:

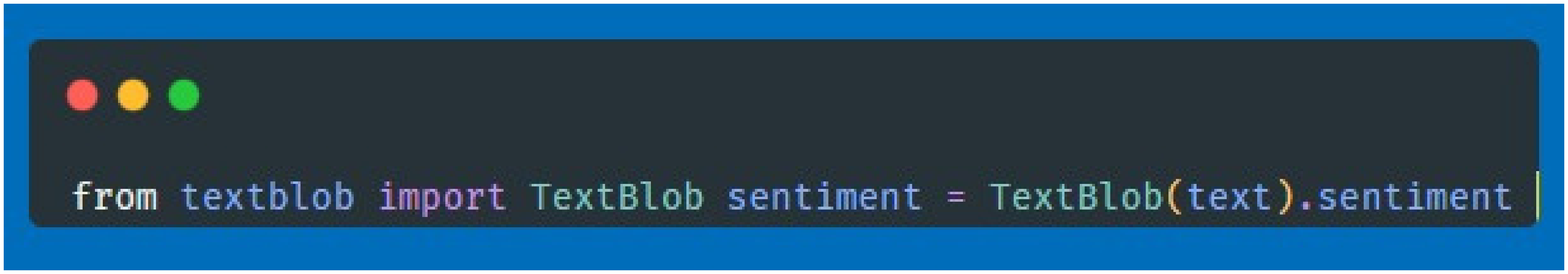


```
filtered_tokens = [token.text for token in doc if not token.is_stop]
```

6. Sentiment Analysis

Sentiment Analysis is the process of determining the sentiment or emotion of a piece of text.

- TextBlob Sentiment Analysis:

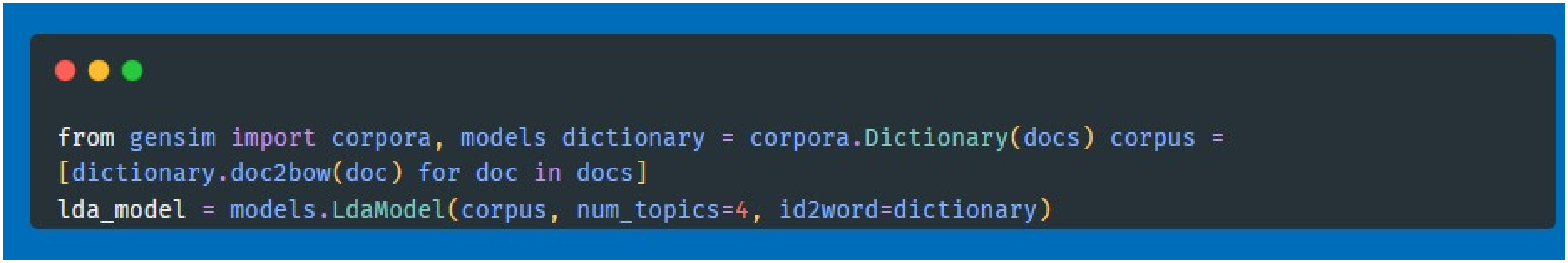


```
from textblob import TextBlob sentiment = TextBlob(text).sentiment
```

7. Topic Modeling

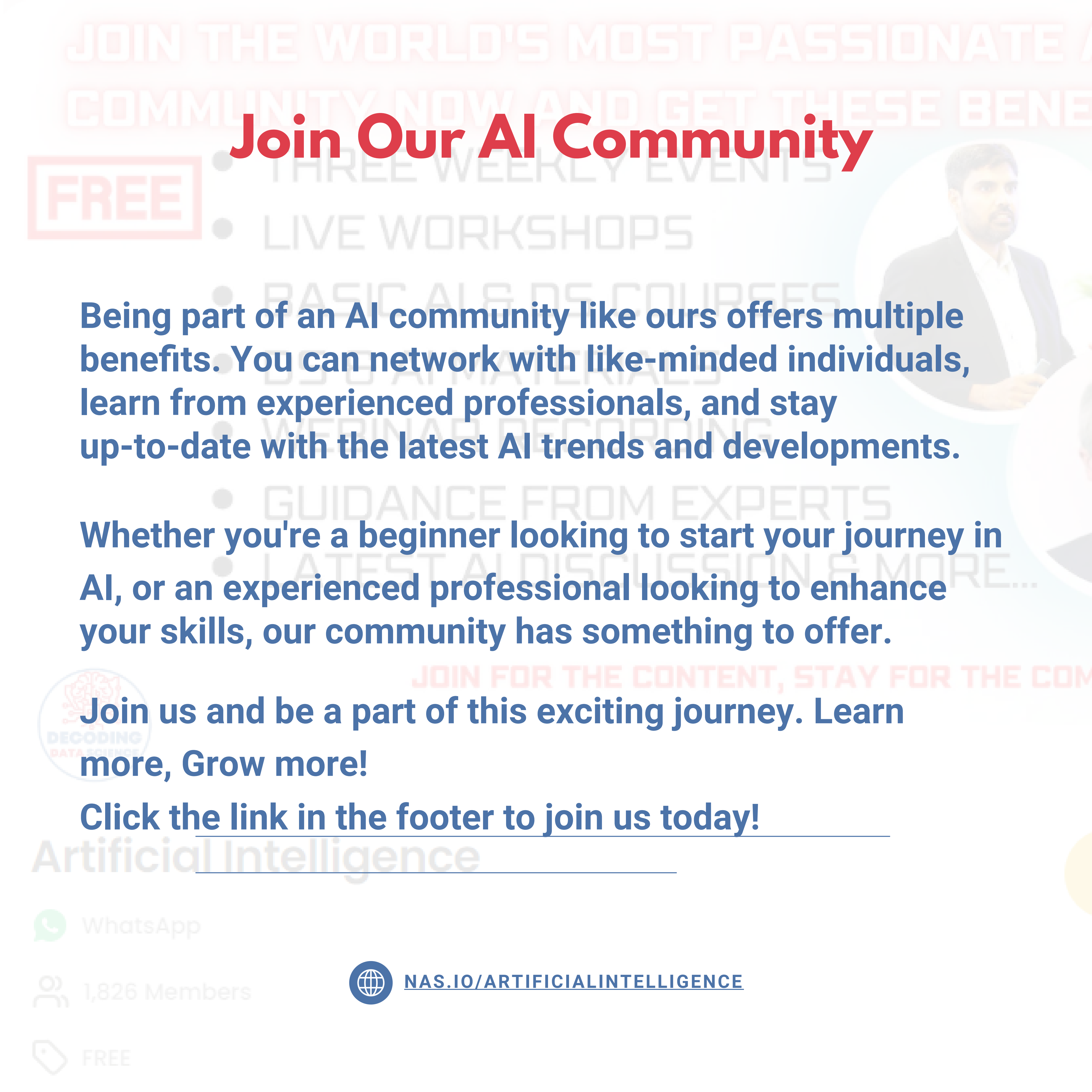
Topic Modeling is the process of identifying topics in a set of documents.

- Gensim LDA Topic Modeling:



```
from gensim import corpora, models dictionary = corpora.Dictionary(docs) corpus =
[dictionary.doc2bow(doc) for doc in docs]
lda_model = models.LdaModel(corpus, num_topics=4, id2word=dictionary)
```

Join Our AI Community



Being part of an AI community like ours offers multiple benefits. You can network with like-minded individuals, learn from experienced professionals, and stay up-to-date with the latest AI trends and developments.

Whether you're a beginner looking to start your journey in AI, or an experienced professional looking to enhance your skills, our community has something to offer.

Join us and be a part of this exciting journey. Learn more, Grow more!

Click the link in the footer to join us today!

NAS.IO/ARTIFICIALINTELLIGENCE