In [27]:
```
import nltk
import re
```

In [39]:
```
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/ahmedhanif/nltk_data...
```

Out[39]: True

In [28]:
```
nltk.download('stopwords')
from nltk.corpus import stopwords
english_stopwords = stopwords.words("English")
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/ahmedhanif/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [29]:
```
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
wn = nltk.WordNetLemmatizer()
ps = nltk.PorterStemmer()
```

In [30]:
```
text = '''
I am a fourth-year bachelor's student in Data Science and AI, an exciting
cutting-edge technology with the power of data to solve real-world proble
'''
```

In [31]:
```
import string

def remove_punctuation(text):
    return "".join([char for char in text if char not in string.punctuati
```

In [32]:
```
def performStemming(text):
    return" ".join([ps.stem(word) for word in re.split('\W+', text)])
```

In [33]:
```
def performLemmatization(text):
    return" ".join([wn.lemmatize(word,'v') for word in re.split('\W+', te
```

In [34]:
```
def tokenize(text):
    return nltk.word_tokenize(text)
```

In [35]:
```
def clean_tokens(tokenized_words):
    filtered_words = []
    for word in tokenized_words:
        if word not in english_stopwords:
            filtered_words.append(word)

    return filtered_words
```

In [61]:
```
def preprocess(text):
    text = remove_punctuation(text)
    text_tokenized = tokenize(text)
    tokenized_clean = clean_tokens(text_tokenized)
    print(tokenized_clean)
    tokenized_clean_text = " ".join(tokenized_clean)
```

```
    stemed= performStemming(tokenized_clean_text)
    lemmatized= performLemmatization(tokenized_clean_text)
    return stemed, lemmatized
```

In [62]:
```
stemed, lemmatized = preprocess(text)
```

```
['I', 'fourthyear', 'bachelors', 'student', 'Data', 'Science', 'AI', 'exci
ting', 'field', 'combines', 'cuttingedge', 'technology', 'power', 'data',
'solve', 'realworld', 'problems']
```

In [63]:
```
lemmatized
```

Out[63]:
```
'I fourthyear bachelor student Data Science AI excite field combine cutt
ingedge technology power data solve realworld problems'
```

In [65]:
```
stemed
```

Out[65]:
```
'i fourthyear bachelor student data scienc ai excit field combin cutting
edg technolog power data solv realworld problem'
```

Interpretation

Stemming tries to find the root of the word... example "excit" is the root of many words like excited, excitement etc

Lemma tries to do the same but it ensures that the word actually also exists in the language itself.