NLP Basics: Implementing a pipeline to clean text

Pre-processing text data

Cleaning up the text data is necessary to highlight attributes that you're going to want your machine learning system to pick up on. Cleaning (or pre-processing) the data typically consists of a number of steps:

1. **Remove punctuation**
2. **Tokenization**
3. **Remove stopwords**
4. Lemmatize/Stem

The first three steps are covered in this chapter as they're implemented in pretty much any text cleaning pipeline. Lemmatizing and stemming are covered in the next chapter as they're helpful but not critical.

```
import pandas as pd
pd.set_option('display.max_colwidth', 100)

data = pd.read_csv("SMSSpamCollection.tsv", sep='\t', header=None)
data.columns = ['label', 'body_text']

data.head()
```

	label	body_text
0	ham	I've been searching for the right words to thank you for this breather. I promise i wont take yo...
1	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive ...
2	ham	Nah I don't think he goes to usf, he lives around here though
3	ham	Even my brother is not like to speak with me. They treat me like aids patent.
4	ham	I HAVE A DATE ON SUNDAY WITH WILL!!

```
# What does the cleaned version look like?
```

```
data_cleaned = pd.read_csv("SMSSpamCollection_cleaned.tsv", sep='\t')
data_cleaned.head()
```

	label	body_text	body_text_nostop
0	ham	I've been searching for the right words to thank you for this breather. I promise i wont take yo...	['ive', 'searching', 'right', 'words', 'thank', 'breather', 'promise', 'wont', 'take', 'help', '...
1	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive ...	['free', 'entry', '2', 'wkly', 'comp', 'win', 'fa', 'cup', 'final', 'tkts', '21st', 'may', '2005...
2	ham	Nah I don't think he goes to usf, he lives around here though	['nah', 'dont', 'think', 'goes', 'usf', 'lives', 'around', 'though']
3	.	Even my brother is not like to speak with me. They treat me	['even', 'brother', 'like', 'speak', 'treat', 'like', 'aids',

Remove punctuation

```
import string
string.punctuation
```

```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

```
"I like NLP." == "I like NLP"
```

```
False
```

```
def remove_punct(text):
    text_nopunct = "".join([char for char in text if char not in string.punctuation])
    return text_nopunct

data['body_text_clean'] = data['body_text'].apply(lambda x: remove_punct(x))

data.head()
```

	label	body_text	body_text_clean
0	ham	I've been searching for the right words to thank you for this breather. I promise i wont take yo...	Ive been searching for the right words to thank you for this breather I promise i wont take your...

Tokenization

```
import re

def tokenize(text):
    tokens = re.split('\W+', text)
    return tokens

data['body_text_tokenized'] = data['body_text_clean'].apply(lambda x: tokenize(x.lower()))

data.head()
```

	label	body_text	body_text_clean	body_text_tokenized
0	ham	I've been searching for the right words to thank you for this breather. I promise i wont take yo...	Ive been searching for the right words to thank you for this breather I promise i wont take your...	[ive, been, searching, for, the, right, words, to, thank, you, for, this, breather, i, promise, ...
1	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive ...	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005 Text FA to 87121 to receive e...	[free, entry, in, 2, a, wkly, comp, to, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, to...
2	ham	Nah I don't think he goes to usf, he lives around here though	Nah I dont think he goes to usf he lives around here though	[nah, i, dont, think, he, goes, to, usf, he, lives, around, here, though]
3		Even my brother is not like to speak	Even my brother is not like to speak	[even, my, brother, is, not, like, to,

```
'NLP' == 'nlp'
```

```
False
```

Remove stopwords

```
import nltk
```

```
stopword = nltk.corpus.stopwords.words('english')

def remove_stopwords(tokenized_list):
    text = [word for word in tokenized_list if word not in stopword]
    return text

data['body_text_nostop'] = data['body_text_tokenized'].apply(lambda x: remove_stopwords(x))

data.head()
```

	label	body_text	body_text_clean	body_text_tokenized	body_text_nostop
0	ham	I've been searching for the right words to thank you for this breather. I promise i wont take yo...	Ive been searching for the right words to thank you for this breather I promise i wont take your...	[ive, been, searching, for, the, right, words, to, thank, you, for, this, breather, i, promise, ...	[ive, searching, right, words, thank, breather, promise, wont, take, help, granted, fulfil, prom...
1	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive ...	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005 Text FA to 87121 to receive e...	[free, entry, in, 2, a, wkly, comp, to, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, to...	[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receiv...
2	ham	Nah I don't think he goes to usf, he lives around here though	Nah I dont think he goes to usf he lives around here though	[nah, i, dont, think, he, goes, to, usf, he, lives, around, here, though]	[nah, dont, think, goes, usf, lives, around, though]
3		Even my brother is not like	Even my brother is not like	[even, my, brother, is,	[even, brother, like

Colab paid products - Cancel contracts here