Supplemental Data Cleaning: Using a Lemmatizer

Test out WordNet lemmatizer (read more about WordNet here)

```
import nltk

wn = nltk.WordNetLemmatizer()
ps = nltk.PorterStemmer()


dir(wn)
```

```
['__class__',
 '__delattr__',
 '__dict__',
 '__dir__',
 '__doc__',
 '__eq__',
 '__format__',
 '__ge__',
 '__getattribute__',
 '__gt__',
 '__hash__',
 '__init__',
 '__init_subclass__',
 '__le__',
 '__lt__',
 '__module__',
 '__ne__',
 '__new__',
 '__reduce__',
 '__reduce_ex__',
 '__repr__',
 '__setattr__',
 '__sizeof__',
 '__str__',
 '__subclasshook__',
 '__unicode__',
 '__weakref__',
```

```
      'lemmatize',
      'unicode_repr']
```

```
print(ps.stem('meanness'))
print(ps.stem('meaning'))
```

```
     mean
     mean
```

```
print(wn.lemmatize('meanness'))
print(wn.lemmatize('meaning'))
```

```
     meanness
     meaning
```

```
print(ps.stem('goose'))
print(ps.stem('geese'))
```

```
     goos
     gees
```

```
print(wn.lemmatize('goose'))
print(wn.lemmatize('geese'))
```

```
     goose
     goose
```

Read in raw text

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

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

data = pd.read_csv("SMSSpamCollection.tsv", sep='\t')
data.columns = ['label', 'body_text']

data.head()
```

	label	body_text
0	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive ...
1	ham	Nah I don't think he goes to usf, he lives around here though
2	ham	Even my brother is not like to speak with me. They treat me like aids patent.
3	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
4	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your call...

Clean up text

```
def clean_text(text):
    text = "".join([word for word in text if word not in string.punctuation])
    tokens = re.split('\W+', text)
    text = [word for word in tokens if word not in stopwords]
    return text

data['body_text_nostop'] = data['body_text'].apply(lambda x: clean_text(x.lower()))

data.head()
```

	label	body_text	body_text_nostop
0	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, text, fa, 87121, receiv...
1	ham	Nah I don't think he goes to usf, he lives around here though	[nah, dont, think, goes, usf, lives, around, though]
2	ham	Even my brother is not like to speak with me. They treat me like aids patent.	[even, brother, like, speak, treat, like, aids, patent]
3	ham	I HAVE A DATE ON SUNDAY WITH WILL!!	[date, sunday]

Lemmatize text

```
def lemmatizing(tokenized_text):
```

```
    text = [wn.lemmatize(word) for word in tokenized_text]
    return text

data['body_text_lemmatized'] = data['body_text_nostop'].apply(lambda x: lemmatizing(x))

data.head(10)
```

	label	body_text	body_text_nostop	body_text_lemmatized
0	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, text, fa, 87121, receiv...	[free, entry, 2, wkly, comp, win, fa, cup, final, tkts, 21st, may, 2005, text, fa, 87121, receiv...
1	ham	Nah I don't think he goes to usf, he lives around here though	[nah, dont, think, goes, usf, lives, around, though]	[nah, dont, think, go, usf, life, around, though]
2	ham	Even my brother is not like to speak with me. They treat me like aids patent.	[even, brother, like, speak, treat, like, aids, patent]	[even, brother, like, speak, treat, like, aid, patent]
3	ham	I HAVE A DATE ON SUNDAY WITH WILL!!	[date, sunday]	[date, sunday]
4	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your call...	[per, request, melle, melle, oru, minnaminunginte, nurungu, vettam, set, callertune, callers, pr...	[per, request, melle, melle, oru, minnaminunginte, nurungu, vettam, set, callertune, caller, pre...
5	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To c...	[winner, valued, network, customer, selected, receivea, 900, prize, reward, claim, call, 0906170...	[winner, valued, network, customer, selected, receivea, 900, prize, reward, claim, call, 0906170...
6	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with came...	[mobile, 11, months, u, r, entitled, update, latest, colour, mobiles, camera, free, call, mobile...	[mobile, 11, month, u, r, entitled, update, latest, colour, mobile, camera, free, call, mobile, ...

Colab paid products - Cancel contracts here