Supplemental Data Cleaning: Using Stemming

Test out Porter stemmer

```
import nltk

ps = nltk.PorterStemmer()
```

```
dir(ps)
```

```
['MARTIN_EXTENSIONS',
 'NLTK_EXTENSIONS',
 'ORIGINAL_ALGORITHM',
 '__abstractmethods__',
 '__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__',
 '_abc_cache',
 '_abc_negative_cache',
 '_abc_negative_cache_version',
 '_abc_registry',
 '_apply_rule_list',
 '_contains_vowel',
 '_ends_cvc',
 '_ends_double_consonant',
 '_has_positive_measure',
 '_is_consonant',
 '_measure',
 '_replace_suffix',
 '_step1a',
 '_step1b',
 '_step1c',
 '_step2',
 '_step3',
 '_step4',
 '_step5a',
 '_step5b',
 'mode',
 'pool',
 'stem',
 'unicode_repr',
 'vowels']
```

```
print(ps.stem('grows'))
print(ps.stem('growing'))
print(ps.stem('grow'))
```

```
     grow
     grow
     grow
```

```
print(ps.stem('run'))
print(ps.stem('running'))
print(ps.stem('runner'))
```

```
     run
     run
     runner
```

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]

Stem text

```
def stemming(tokenized_text):
    text = [ps.stem(word) for word in tokenized_text]
    return text

data['body_text_stemmed'] = data['body_text_nostop'].apply(lambda x: stemming(x))

data.head()
```

	label	body_text	body_text_nostop	body_text_stemmed
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, entri, 2, wkli, comp, win, fa, cup, final, tkt, 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, goe, usf, live, 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]

Colab paid products - Cancel contracts here