NLP Basics: Reading in text data & why do we need to clean the text?

Read in semi-structured text data

```
# Read in the raw text
rawData = open("SMSSpamCollection.tsv").read()

# Print the raw data
rawData[0:500]
```

```
"ham\tI've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and
will fulfil my promise. You have been wonderful and a blessing at all times.\nspam\tFree entry in 2 a wkly comp to win FA Cup
final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's\nham\tNah I
don't think he goes to usf, he lives around here though\nham\tEven my brother is not like to speak with me. They treat me like
aid"
```

```
parsedData = rawData.replace('\t', '\n').split('\n')
```

```
parsedData[0:5]
```

```
['ham',
 "I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will
fulfil my promise. You have been wonderful and a blessing at all times.",
 'spam',
 "Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt
rate)T&C's apply 08452810075over18's",
 'ham']
```

```
labelList = parsedData[0::2]
textList = parsedData[1::2]
```

```
print(labelList[0:5])
print(textList[0:5])
```

```
['ham', 'spam', 'ham', 'ham', 'ham']
["I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fu
```

```
import pandas as pd

fullCorpus = pd.DataFrame({
    'label': labelList,
    'body_list': textList
})

fullCorpus.head()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-27-25797b4f5cf0> in <module>()
      3 fullCorpus = pd.DataFrame({
      4     'label': labelList,
----> 5     'body_list': textList
      6 })
      7

--------------------------------------------------------------------------- 3 frames
~/anaconda3/lib/python3.6/site-packages/pandas/core/frame.py in extract_index(data)
   5542             lengths = list(set(raw_lengths))
   5543             if len(lengths) > 1:
-> 5544                 raise ValueError('arrays must all be same length')
   5545
   5546             if have_dicts:

ValueError: arrays must all be same length
```

SEARCH STACK OVERFLOW

```
print(len(labelList))
print(len(textList))
```

```
5571
5570
```

```
print(labelList[-5:])
```

```
['ham', 'ham', 'ham', 'ham', '']
```

```
fullCorpus = pd.DataFrame({
    'label': labelList[:-1],
    'body_list': textList
})

fullCorpus.head()
```

	body_list	label
0	I've been searching for the right words to tha...	ham
1	Free entry in 2 a wkly comp to win FA Cup fina...	spam
2	Nah I don't think he goes to usf, he lives aro...	ham
3	Even my brother is not like to speak with me. ...	ham
4	I HAVE A DATE ON SUNDAY WITH WILL!!	ham

```
dataset = pd.read_csv("SMSSpamCollection.tsv", sep="\t", header=None)
dataset.head()
```

	0	1
0	ham	I've been searching for the right words to tha...
1	spam	Free entry in 2 a wkly comp to win FA Cup fina...
2	ham	Nah I don't think he goes to usf, he lives aro...
3	ham	Even my brother is not like to speak with me. ...
4	ham	I HAVE A DATE ON SUNDAY WITH WILL!!

Colab paid products - Cancel contracts here