NLP Basics: Exploring the dataset

Read in text data

```
import pandas as pd

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

fullCorpus.head()
```

	label	body_text
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!!

Explore the dataset

```
# What is the shape of the dataset?

print("Input data has {} rows and {} columns".format(len(fullCorpus), len(fullCorpus.columns)))
```

```
     Input data has 5568 rows and 2 columns
```

```
# How many spam/ham are there?

print("Out of {} rows, {} are spam, {} are ham".format(len(fullCorpus),
                                                       len(fullCorpus[fullCorpus['label']=='spam']),
                                                       len(fullCorpus[fullCorpus['label']=='ham'])))
```

```
# How much missing data is there?

print("Number of null in label: {}".format(fullCorpus['label'].isnull().sum()))
print("Number of null in text: {}".format(fullCorpus['body_text'].isnull().sum()))
```

```
Number of null in label: 0
Number of null in text: 0
```

Colab paid products - Cancel contracts here