# CS 505 Homework 02: Data Wrangling and BOW

## Due Thursday 9/21 at midnight (1 minute after 11:59 pm) in Gradescope (with a grace period of 6 hours)

## You may submit the homework up to 24 hours late (with the same grace period) for a penalty of 10%.

All homeworks will be scored with a maximum of 100 points; point values are given for individual problems, and if parts of problems do not have point values given, they will be counted equally toward the total for that problem.

Note: I strongly recommend you work in **Google Colab** (the free version) to complete homeworks in this class; in addition to (probably) being faster than your laptop, all the necessary libraries will already be available to you, and you don't have to hassle with `conda`, `pip`, etc. and resolving problems when the install doesn't work. But it is up to you! You should go through the necessary tutorials listed on the web site concerning Colab and storing files on a Google Drive. And of course, Dr. Google is always ready to help you resolve your problems.

I will post a "walk-through" video ASAP on my Youtube Channel.

### Submission Instructions

You must complete the homework by editing **this notebook** and submitting the following two files in Gradescope by the due date and time:

- A file `HW02.ipynb` (be sure to select `Kernel -> Restart and Run All` before you submit, to make sure everything works); and
- A file `HW02.pdf` created from the previous.

  For best results obtaining a clean PDF file on the Mac, select `File -> Print Review` from the Jupyter window, then choose `File-> Print` in your browser and then `Save as PDF`. Something similar should be possible on a Windows machine -- just make sure it is readable and no cell contents have been cut off. Make it easy to grade!

The date and time of your submission is the last file you submitted, so if your IPYNB file is submitted on time, but your PDF is late, then your submission is late.

# Collaborators (5 pts)

Describe briefly but precisely

1. Any persons you discussed this homework with and the nature of the discussion;
2. Any online resources you consulted and what information you got from those resources; and
3. Any AI agents (such as chatGPT or CoPilot) or other applications you used to complete the homework, and the nature of the help you received.

A few brief sentences is all that I am looking for here.

```
<Your answer here>
1.
2. I used the two resources provided by the professor on the
assignment: https://docs.python.org/3/library/re.html and
https://docs.python.org/3/howto/regex.html.
I looked at the walkthrough video posted by the professor.
I searched up the 8 official different forms of of the verb
'to be'.
I searched up defaultdict in
https://www.geeksforgeeks.org/defaultdict-in-
python/#:~:text=Defaultdict%20is%20a%20sub%2Dclass,key%20that%20doe
 and https://stackoverflow.com/questions/31838823/create-a-
defaultdict-with-a-default-of-zero-0.

3. I searched chatGPT on using flags = re.MULTILINE for part
1.B and 1.C.
```

# Overview

We are going to practice converting raw (string form) text into a useful data set using the script of *Pirates of the Caribbean: The Curse of the Black Pearl* (2003), the first in a series of PotC movies starring Johnny Depp. The script is part of the `webtext` corpus in NLTK.

Under the assumption that we wish to perform an analysis of the words spoken by the characters in the movie, we will convert the text in a series of steps from a raw string of ASCII characters into a dictionary holding a sparse BOW model of the words spoken by two of the main characters, Elizabeth Swann and Jack Sparrow. These dictionaries could be the data set for a classification task, or for creating a vector-space model for each

character, which we will study later in the course. For this assignment, you will clean up, normalize and tokenize the text, create the dictionaries, and then simply print out the most common words spoken by the two characters.

Text normalization was covered in lecture on Tuesday 9/12 and the BOW model on Thursday 9/14. Before beginning the assignment, you should consult the following for information on using the Python regular expression library:

https://docs.python.org/3/library/re.html

Also useful is

https://docs.python.org/3/howto/regex.html

After reviewing the basic principles of regular expressions (which I will review in my walk-through video), read about the following useful functions:

```
result = re.split(...)

result = re.sub(...)
```

You may ONLY use standard functions from the `re` library for this homework, and you must perform your normalization starting with the string form of the script assigned below to the variable `pirates_txt`. You may NOT use indices of the string to perform your modifications (e.g., deleting the first line by counting how many characters to remove).

The point here is to use regular expressions to do the text wrangling. Don't worry, we shall use the `SpaCy` library later on the course to normalize text for a classification problem set.

In [196…
```
import numpy as np
import nltk
import re

# The first time you will need to download the corpus:

nltk.download('webtext')

pirates_txt = nltk.corpus.webtext.raw('pirates.txt')
```

```
[nltk_data] Downloading package webtext to
[nltk_data]     /Users/chrisyang/nltk_data...
[nltk_data]   Package webtext is already up-to-date!
```

In [197…
```
# raw string form
pirates_txt[:1000]
```

Out[197]:
```
"PIRATES OF THE CARRIBEAN: DEAD MAN'S CHEST, by Ted Elliott & Terry Rossio\
n[view looking straight down at rolling swells, sound of wind and thunder,
then a low heartbeat]\nScene: PORT ROYAL\n[teacups on a table in the rain]\
n[sheet music on music stands in the rain]\n[bouquet of white orchids, Eliz
abeth sitting in the rain holding the bouquet]\n[men rowing, men on horseba
ck, to the sound of thunder]\n[EITC logo on flag blowing in the wind]\n[man
y rowboats are entering the harbor]\n[Elizabeth sitting alone, at a distanc
e]\n[marines running, kick a door in] \n[a mule is seen on the left in the
barn where the marines enter]\n[Liz looking over her shoulder]\n[Elizabeth
drops her bouquet]\n[Will is in manacles, being escorted by red coats]\nELI
ZABETH SWANN: Will...!\n[Elizabeth runs to Will]\nELIZABETH SWANN: Why is t
his happening? \nWILL TURNER: I don't know. You look beautiful.\nELIZABETH
SWANN: I think it's bad luck for the groom to see the bride before the wedd
ing.\n[marines cross their long axes to bar Go"
```

In [198…
```
# printing it shows the formatting
print(pirates_txt[:1000])
```

```
PIRATES OF THE CARRIBEAN: DEAD MAN'S CHEST, by Ted Elliott & Terry Rossio
[view looking straight down at rolling swells, sound of wind and thunder, th
en a low heartbeat]
Scene: PORT ROYAL
[teacups on a table in the rain]
[sheet music on music stands in the rain]
[bouquet of white orchids, Elizabeth sitting in the rain holding the bouquet
]
[men rowing, men on horseback, to the sound of thunder]
[EITC logo on flag blowing in the wind]
[many rowboats are entering the harbor]
[Elizabeth sitting alone, at a distance]
[marines running, kick a door in]
[a mule is seen on the left in the barn where the marines enter]
[Liz looking over her shoulder]
[Elizabeth drops her bouquet]
[Will is in manacles, being escorted by red coats]
ELIZABETH SWANN: Will...!
[Elizabeth runs to Will]
ELIZABETH SWANN: Why is this happening?
WILL TURNER: I don't know. You look beautiful.
ELIZABETH SWANN: I think it's bad luck for the groom to see the bride before
the wedding.
[marines cross their long axes to bar Go
```

# Problem One (35 points): Cleaning up the lines

The first task is to clean up the text so that at the conclusion of this problem, you will have a text with punctuation and extraneous characters removed, and each line consisting of a character's name (human or otherwise), a colon, and a sequence of words, ending in a newline.
(You will keep this as a single string until Problem 3, but we will refer to the "lines" spoken by each character -- also, I hope it will not be confusing to speak of (ASCII) characters and characters played by the actors in the script!)

## Part 1.A (5 pts)

1. Convert the string into all lower-case letters.
2. Remove the first line which gives the title and authors. Print out the first 200 characters to show that you have done this.

Hint: Cut everything before the first '\n', using the 'beginning of string' special character `^` in the regular expression.

In [199…
```
# Your code here
# As the video from professor Snyder suggested, I did not convert the string
# since it is not important from parts 1.A to 1.C, I converted them in part
# that are part of the string.

# Therefore, the first part on converting the string into all lower-case let

pirates_without_title = re.sub('^[^\n]*\n', '', pirates_txt)

print(pirates_without_title[:200])
```

```
[view looking straight down at rolling swells, sound of wind and thunder, th
en a low heartbeat]
Scene: PORT ROYAL
[teacups on a table in the rain]
[sheet music on music stands in the rain]
[bouquet of
```

## Part 1.B (5 pts)

Cut out all the stage directions that are given in square brackets, including the newlines on those lines. Print out the first 200 characters as proof.

In [200…
```
# Your code here
pirates_without_stage_directions = re.sub('^\[.*\n', '', pirates_without_tit

print(pirates_without_stage_directions[:200])
```

```
Scene: PORT ROYAL
ELIZABETH SWANN: Will...!
ELIZABETH SWANN: Why is this happening?
WILL TURNER: I don't know. You look beautiful.
ELIZABETH SWANN: I think it's bad luck for the groom to see the brid
```

## Part 1.C (5 pts)

Cut out the lines where the 'scene' is specified. Again, print out the first 200 characters.

In [201…
```
# Your code here
pirates_without_scene = re.sub('^.*Scene:.*$\n*', '', pirates_without_stage_

print(pirates_without_scene[:200])
```

```
ELIZABETH SWANN: Will...!
ELIZABETH SWANN: Why is this happening?
WILL TURNER: I don't know. You look beautiful.
ELIZABETH SWANN: I think it's bad luck for the groom to see the bride before
the weddi
```

## Part 1.D (20 pts)

Now, we still have a lot of punctuation and some miscellaneous odd things occurring in this text, and we need to do further cleaning. But you will have to figure this out for yourself!

The main thing to do is to remove punctuation and anything that does not contribute to the goal of making a BOW model for our two characters.

But you can't just remove all non-word characters! Make sure you take account of the following:

1. You need to keep the character's names at the beginning of the line, so **do not remove the colon after the name** (note that these always occur at the beginning of a line, i.e., at the very beginning or immediately after the newline from the previous line).

2. In the next problem we will normalize the words, so **we DON'T want to change anything that might be a word,** such as,

```
    charges    don't    it's    'er   ah-ha      ha-ha-ha-ha-ha
stealin'
```

3. After observing the caveats above, **remove all punctuation.**

1. There are some places where apparently the transcriber was not sure what the word was and gave alternatives:

   ```
   weren't/wasn't
   ```

   and some places where it is not clear what is intended:

   ```
   oy /quick
   ```

Just treat `/` like ordinary punctuation and replace it by a blank.

1. Finally, there are miscellanous weird things in the text, such as

   ```
   ###         ?:
   ```

   (and possibly others) which **should be removed.**

**How to proceed:** To explore the data, print out the text after the modifications in Parts 1.A -- 1.D:

```
print(pirates_txt_01)
```

and think about what needs to be removed, paying careful attention to the comments above. (You could use the `Find` function in your browser to flip through various possibilities.)

After examining the text, **comment out the `print(pirates_txt_01)` so that we don't have to look at it!** This was just for exploration!

The result of your cleaning in this part should be assigned to `pirates_txt_01`.

Hint: At this stage, it might be better to **replace substrings with blanks** instead of deleting them (replacing with the empty string) to preserve the separation of words (just in case!).

## Part 1.D.1 (5 pts)

Write a short description here of what you removed, giving your reasoning. You must account for at least what is listed above, but you may find other things you want to change.

< your comments here > I first deleted all colons except the ones that come right after the character (except the first colon).

Then, I converted them into lower case letters, which was mentioned in section 1.A since the main reason I left them in capital form was to do the first task.

Then, changed the miscellanous weird things such as ?: in the text. Then, I changed the "weren't/wasn't" into "weren't wasn't".

I did these two first before setting up a pattern for all punctuations since they were weird punctuations that needed to be handled case by case.

Then, I set up a pattern to delete all puncutations except colon (:), hyphen (-), and apostrophe (') because these some of these punctuations are part of the word that should not be all deleted such as "don't", "ha-ha-ha", "elizabeth swann:".

After deleting all punctuations except the ones listed above, I deleted all hyphens (-) and apostrophes (') that were alone without a word before or after them, since if they are alone, they are not part of the word and act as punctuations, and therefore should be deleted.

After reading some texts, I realized that hyphens(-) and apostrophes(') existed at the end of the text before newline and therefore, I deleted them as well since they are not part of the word.

Also print out some portion of the text to show at least some of the changes you have made.

## Part 1.D.2 (15 pts)

Write your code in the following cell. The result at the end should be stored in `pirates_txt_01`. Print out the first 2000 characters.

In [202…

```
# Your code here
# print(pirates_txt_01)

# Delete all the colons right after the first one
pirates_clean = re.sub(r'([^A-Z]):', r'\1', pirates_without_scene)

lower_case_letters = pirates_clean.lower()

# change ?: to empty string
pirates_clean = re.sub('\?: ', '', lower_case_letters)

pirates_clean = re.sub("weren't/wasn't", "weren't wasn't", pirates_clean)

# change every punctuation except ' and : and -
pirates_clean = re.sub('[^\w\s:\'-]', '', pirates_clean)

#change every - that is not part of the word but acts as a punctuation
pirates_clean = re.sub(' - ', ' ', pirates_clean)

#change every ' that is not part of the word but acts as a punctuation
pirates_clean = re.sub(' \' ', ' ', pirates_clean)

#change every - that is at the end of the sentence that is not part of the w
pirates_clean = re.sub(' -\n', '\n', pirates_clean)

#change every ' that is at the end of the sentence that is not part of the w
pirates_clean = re.sub(' \'\n', '\n', pirates_clean)


pirates_txt_01 = pirates_clean
print(pirates_txt_01[:2000])
```

```
elizabeth swann: will
elizabeth swann: why is this happening
will turner: i don't know you look beautiful
elizabeth swann: i think it's bad luck for the groom to see the bride before
the wedding
lord cutler beckett: governor weatherby swann it's been too long
lord cutler beckett: his lord now actually
lord cutler beckett: in fact i do mister mercer the warrant for the arrest o
f one william turner
lord cutler beckett: oh is it that's annoying my mistake arrest her
elizabeth swann: on what charges
will turner: no
lord cutler beckett: ah-ha here's the one for william turner and i have anot
her one for a mister james norrington is he present
elizabeth swann: what are the charges
lord cutler beckett: i don't believe that's the answer to the question i ask
ed
will turner: lord beckett in the category of questions not answered
elizabeth swann: we are under the jurisdiction of the king's governor of por
t royal and you will tell us what we are charged with
lord cutler beckett: for which the punishment regrettably is also death perh
aps you remember a certain pirate named jack sparrow
elizabeth swann: captain jack sparrow
lord cutler beckett: captain jack sparrow yes i thought you might
gibbs: fifteen men on a dead man's chest yo ho ho and a bottle of rum drink
and the devil had done for the rest yo ho ho and a bottle of rum ha-ha-ha-ha
-ha
jack sparrow: sorry mate
jack sparrow: mind if we make a little side trip i didn't think so
gibbs: not quite according to plan
jack sparrow: complications arose ensued were overcome
gibbs: you got what you went in for then
jack sparrow: mm-hmm
gibbs: captain i think the crew meaning me as well were expecting something
a bit more shiny what with the isla de muerta going all pear shaped reclaime
d by the sea and the treasure with it
leech: and the royal navy chasing us all around the atlantic
marty: and the hurricane aye
crew: aye aye
gibbs: all in all it's seems some time since we did a speck of honest pirati
ng
jack sparrow: shiny
gibbs: ay
```

# Problem Two (30 points): Normalizing, Stemming, and Lemmatization

In this problem we are going to do **some** normalizing of the words, first of all to normalize certain words with apostrophes, and then performing stemming and lemmatization. We are not intended to be absolutely thorough here, just to try a few obvious possibilities.

## Part 2.A Normalizing (15 pts)

There are several ways that apostophes (single quotes) are used to compress two words into one (to give a better sense for how they are pronounced, *especially by pirates*):

```
    didn't  =  did not    we've  = we have       there'd  =
there would
```

**Your task: Find as many examples of these as you can, and replace the compressed word with the two-word phrase it represents.**

Note: **Do NOT process any words with `'s`,** as these will be done in the next part.

Do NOT simply compile a list of specific examples, but look for general patterns for substitution, for example:

```
          n't =>  _not       've =>  _have       # where _
represents a blank
```

Simply find as many examples which seem to have a general rule, and perform those substitutions, putting the result in `pirates_txt_02`.

Finally, print out the first 2000 characters.

In [203…

```
# Came up with as many patterns I can to deleted apostrophes that are not 's


pirates_clean = re.sub("n't", " not", pirates_txt_01)
pirates_clean = re.sub("\'re", " are", pirates_clean)

pirates_clean = re.sub("\'ve", " have", pirates_clean)
pirates_clean = re.sub("\'d", " would", pirates_clean)
pirates_clean = re.sub("i'm", "i am", pirates_clean)
pirates_clean = re.sub("\'ll", " will", pirates_clean)









pirates_txt_02 = pirates_clean



print(pirates_txt_02[:2000])
```

```
elizabeth swann: will
elizabeth swann: why is this happening
will turner: i do not know you look beautiful
elizabeth swann: i think it's bad luck for the groom to see the bride before
the wedding
lord cutler beckett: governor weatherby swann it's been too long
lord cutler beckett: his lord now actually
lord cutler beckett: in fact i do mister mercer the warrant for the arrest o
f one william turner
lord cutler beckett: oh is it that's annoying my mistake arrest her
elizabeth swann: on what charges
will turner: no
lord cutler beckett: ah-ha here's the one for william turner and i have anot
her one for a mister james norrington is he present
elizabeth swann: what are the charges
lord cutler beckett: i do not believe that's the answer to the question i as
ked
will turner: lord beckett in the category of questions not answered
elizabeth swann: we are under the jurisdiction of the king's governor of por
t royal and you will tell us what we are charged with
lord cutler beckett: for which the punishment regrettably is also death perh
aps you remember a certain pirate named jack sparrow
elizabeth swann: captain jack sparrow
lord cutler beckett: captain jack sparrow yes i thought you might
gibbs: fifteen men on a dead man's chest yo ho ho and a bottle of rum drink
and the devil had done for the rest yo ho ho and a bottle of rum ha-ha-ha-ha
-ha
jack sparrow: sorry mate
jack sparrow: mind if we make a little side trip i did not think so
gibbs: not quite according to plan
jack sparrow: complications arose ensued were overcome
gibbs: you got what you went in for then
jack sparrow: mm-hmm
gibbs: captain i think the crew meaning me as well were expecting something
a bit more shiny what with the isla de muerta going all pear shaped reclaime
d by the sea and the treasure with it
leech: and the royal navy chasing us all around the atlantic
marty: and the hurricane aye
crew: aye aye
gibbs: all in all it's seems some time since we did a speck of honest pirati
ng
jack sparrow: shiny
gibbs:
```

# Part 2.B Stemming and Lemmatization (15 pts)

## Stemming

There are multiple occurrence of the suffix `'s` in the text, some standing for a two word phrase:

```
   he's = he is       it's = it is       what's = what is
here's = here is
```

and some being possessives:

```
jack's     man's       hangman's
```

In the first case, the word `is` is very common, and would be removed later when we remove "stop words"; in the second, we will assume there is little difference in the BOW model between a noun and its possessive. So we will remove the `'s` from all words.

## Lemmatization

There are eight "official' different forms of the verb 'to be', all of which occur in the text. These must be replaced by the lemma 'be'. (These eight forms do not include modal expressions such as 'will be' or 'would be'.)

**Your tasks:**

1. Stem these words by removing all instances of `'s`.
2. Lemmatize all the 8 forms of the verb 'to be' by replacing them by their stem 'be'. Be sure to ONLY replace separate words, not substrings of other words, i.e., don't change 'mistake' to 'mbetake'!
3. Put the result in `pirates_txt_02` and print out the first 2000 characters.

In [204…

```
# Your code here

# first deleted all patterns that have the feature of "'s" becoming "is"
pirates_clean = pirates_txt_02
pirates_clean = re.sub("it's", "it is", pirates_clean)
pirates_clean = re.sub("that's", "that is", pirates_clean)
pirates_clean = re.sub("here's", "here is", pirates_clean)
pirates_clean = re.sub("there's", "there is", pirates_clean)
pirates_clean = re.sub("what's", "what is", pirates_clean)
pirates_clean = re.sub("he's", "he is", pirates_clean)
pirates_clean = re.sub("where's", "where is", pirates_clean)
pirates_clean = re.sub("eyesight's", "eyesight is", pirates_clean)
pirates_clean = re.sub("she's", "she is", pirates_clean)

# then deleted all patterns that possessive meanings

pirates_clean = re.sub(" \'s", "!!!", pirates_clean)
pirates_clean = re.sub("\'s", "", pirates_clean)
pirates_clean = re.sub("!!!", " \'s", pirates_clean)


# changed all 8 forms of "be" to "be": they were "be", "is", "am", "are", "w

pirates_clean = re.sub(" am ", " be ", pirates_clean)
pirates_clean = re.sub(" is ", " be ", pirates_clean)
pirates_clean = re.sub(" are ", " be ", pirates_clean)
pirates_clean = re.sub(" was ", " be ", pirates_clean)
pirates_clean = re.sub(" were ", " be ", pirates_clean)
pirates_clean = re.sub(" being ", " be ", pirates_clean)
pirates_clean = re.sub(" been ", " be ", pirates_clean)

pirates_txt_02 = pirates_clean


print(pirates_txt_02[:2000])
```

```
elizabeth swann: will
elizabeth swann: why be this happening
will turner: i do not know you look beautiful
elizabeth swann: i think it be bad luck for the groom to see the bride befor
e the wedding
lord cutler beckett: governor weatherby swann it be be too long
lord cutler beckett: his lord now actually
lord cutler beckett: in fact i do mister mercer the warrant for the arrest o
f one william turner
lord cutler beckett: oh be it that be annoying my mistake arrest her
elizabeth swann: on what charges
will turner: no
lord cutler beckett: ah-ha here be the one for william turner and i have ano
ther one for a mister james norrington be he present
elizabeth swann: what be the charges
lord cutler beckett: i do not believe that be the answer to the question i a
sked
will turner: lord beckett in the category of questions not answered
elizabeth swann: we be under the jurisdiction of the king governor of port r
oyal and you will tell us what we be charged with
lord cutler beckett: for which the punishment regrettably be also death perh
aps you remember a certain pirate named jack sparrow
elizabeth swann: captain jack sparrow
lord cutler beckett: captain jack sparrow yes i thought you might
gibbs: fifteen men on a dead man chest yo ho ho and a bottle of rum drink an
d the devil had done for the rest yo ho ho and a bottle of rum ha-ha-ha-ha-h
a
jack sparrow: sorry mate
jack sparrow: mind if we make a little side trip i did not think so
gibbs: not quite according to plan
jack sparrow: complications arose ensued be overcome
gibbs: you got what you went in for then
jack sparrow: mm-hmm
gibbs: captain i think the crew meaning me as well be expecting something a
bit more shiny what with the isla de muerta going all pear shaped reclaimed
by the sea and the treasure with it
leech: and the royal navy chasing us all around the atlantic
marty: and the hurricane aye
crew: aye aye
gibbs: all in all it be seems some time since we did a speck of honest pirat
ing
jack sparrow: shiny
gibbs: aye sh
```

# Problem Three (30 points): Removing Stop Words, Tokenizing, and Creating the BOW Models

## 3.A Removing Stop Words (10 pts)

"Stop words" are common words which do not give much information about a text, since they occur in almost all texts. There is a standard set of such words which can be accessed through NLTK (notice that these include some with apostrophes, which we will have already removed):

In [205…

```
import nltk
from nltk.corpus import stopwords

nltk.download('stopwords')
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're
", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', '
he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it'
, "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves',
'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had
', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', '
if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with
', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'af
ter', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off'
, 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'whe
n', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'mos
t', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so'
, 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 's
hould', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", '
hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma',
'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan'
t", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "wo
n't", 'wouldn', "wouldn't"]
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/chrisyang/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

For the first part of this problem, you must **remove all stop words from the text,** and store the result in `pirates_txt_03`. However, since `will` is the name of a character in the script, **do NOT remove the stopword `will`**! Make SURE that you only remove words, and not substrings of larger words, e.g., do not remove all occurrences of the character `i` from the text just because the word `i` is a stop word! Replace stop words with single blanks to preserve the word boundaries.

Put your code in the next cell and print out the first 2000 characters.

In [206…

```
# Your code here
pirates_clean = pirates_txt_02
for word in stopwords.words('english'):
    if word == 'will':
        continue
    add_space_word = " " + word + " "
    pirates_clean = re.sub(add_space_word, ' ', pirates_clean)
    end_word = " " + word + "\n"
    pirates_clean = re.sub(end_word, ' \n', pirates_clean)

pirates_txt_03 = pirates_clean

print(pirates_txt_03[:2000])
```

```
elizabeth swann: will
elizabeth swann: happening
will turner: know look beautiful
elizabeth swann: think bad luck groom see bride wedding
lord cutler beckett: governor weatherby swann be long
lord cutler beckett: lord actually
lord cutler beckett: fact mister mercer warrant arrest one william turner
lord cutler beckett: oh be annoying mistake arrest
elizabeth swann: charges
will turner:
lord cutler beckett: ah-ha one william turner another one mister james norri
ngton present
elizabeth swann: charges
lord cutler beckett: believe answer question asked
will turner: lord beckett category questions answered
elizabeth swann: jurisdiction king governor port royal will tell us charged
lord cutler beckett: punishment regrettably also death perhaps remember cert
ain pirate named jack sparrow
elizabeth swann: captain jack sparrow
lord cutler beckett: captain jack sparrow yes thought might
gibbs: fifteen men dead man chest yo ho ho bottle rum drink devil done rest
yo ho ho bottle rum ha-ha-ha-ha-ha
jack sparrow: sorry mate
jack sparrow: mind make little side trip think
gibbs: quite according plan
jack sparrow: complications arose ensued overcome
gibbs: got went
jack sparrow: mm-hmm
gibbs: captain think crew meaning well expecting something bit shiny isla de
muerta going pear shaped reclaimed sea treasure
leech: royal navy chasing us around atlantic
marty: hurricane aye
crew: aye aye
gibbs: all seems time since speck honest pirating
jack sparrow: shiny
gibbs: aye shiny
jack sparrow: feeling perhaps dear old jack serving best interests captain
cotton parrot: awk walk plank
jack sparrow: bird say
leech: blame bird show us piece cloth
jack sparrow: ohhh
gibbs: know good
jack sparrow:
marty: key
jack sparrow: much better drawing key
jack sparrow: gentlemen keys
leech: keys unlock things
gibbs: whatever key unlocks inside something valuable setting find whatever
key unlocks
jack sparrow: key ca open whatever unlocks purpose would served finding what
ever need unlocked w
```

## 3.B Tokenizing and Creating the BOW Dictionary (20 pts)

What we wish to do is to create a BOW model with a dictionary for two characters in the script, `elizabeth swann` and `jack sparrow`.

## Part 3.B.1 (2 pts)

Using `split(...)`, split the text on the newlines `\n` to get a list of each line as a string. Print out the first 10 lines.

In [207…
```
# Your code here
pirates_clean = pirates_txt_03
pirates_clean = re.split("\n", pirates_clean)
pirates_txt_03 = pirates_clean
print(pirates_txt_03[:10])
```

```
['elizabeth swann: will', 'elizabeth swann: happening ', 'will turner: know
look beautiful', 'elizabeth swann: think bad luck groom see bride wedding',
'lord cutler beckett: governor weatherby swann be long', 'lord cutler becket
t: lord actually', 'lord cutler beckett: fact mister mercer warrant arrest o
ne william turner', 'lord cutler beckett: oh be annoying mistake arrest ', '
elizabeth swann: charges', 'will turner: ']
```

## Part 3.B.2 (18 pts)

Create a dictionary to hold the BOW models for these two characters, each being a `defaultdict` with a default value of 0 (this is a representation of the sparce matrix representing the BOW for the character).

Then go through the lines and calculate the frequency of each word spoken by that character. Print out the 20 most common words spoken by each character and the number of times spoken.

Hint: Scan through the lines created in 3.B.1, and just check if the line contains that character's name. Hint: you can use `in` to check if a substring occurs in a string:

```
'wayne' in 'wayne snyder'   =>  True
```

Then split the line on blanks, and add all but the first two words (the name of the character) to the BOW for that character. If the empty word '' occurs, ignore it (do not add it to the BOW).

In [208…

```
# Elizabeth Swann's BOW (9 pts)


# Your code here

from collections import defaultdict

character_bow = {"elizabeth swann": defaultdict(lambda:0), "jack sparrow": d


words = []
for x in pirates_txt_03:
    words = x.split()
    if "swann:" in words:
        for w in range(len(words)):
            if w == 0 or w == 1:
                continue
            if words[w] == " ":
                continue
            character_bow["elizabeth swann"][words[w]] += 1


sorted_version_elizabeth = {word: count for word, count in sorted(character_
                                                    key=lambda item: item[1]

limit = 20
count = 0
print("The 20 most common words spoken by elizabeth swann are listed below w
for word, frequency in sorted_version_elizabeth.items():
    print(f"Word: {word}\t\t Number of times spoken: {frequency}")
    count += 1
    if count == limit:
        break
```

```
The 20 most common words spoken by elizabeth swann are listed below with the
corresponding frequencies:

Word: will              Number of times spoken: 22
Word: jack              Number of times spoken: 12
Word: find              Number of times spoken: 7
Word: know              Number of times spoken: 7
Word: oh                Number of times spoken: 7
Word: want              Number of times spoken: 6
Word: man               Number of times spoken: 6
Word: good              Number of times spoken: 5
Word: something         Number of times spoken: 5
Word: sparrow           Number of times spoken: 4
Word: would             Number of times spoken: 4
Word: chance            Number of times spoken: 4
Word: us                Number of times spoken: 3
Word: captain           Number of times spoken: 3
Word: compass           Number of times spoken: 3
Word: give              Number of times spoken: 3
Word: going             Number of times spoken: 3
Word: came              Number of times spoken: 3
Word: way               Number of times spoken: 3
Word: yes               Number of times spoken: 3
```

In [209…

```python
# Jack Sparrows's BOW   (9 pts)

# Your code here
words = []
for x in pirates_txt_03:
    words = x.split()
    if "sparrow:" in words:
        for w in range(len(words)):
            if w == 0 or w == 1:
                continue
            if words[w] == " ":
                continue
            character_bow["jack sparrow"][words[w]] += 1

sorted_version_jack = {word: count for word, count in sorted(character_bow["
                                                             key=lambda item: item[1]

count = 0
print("The 20 most common words spoken by jack sparrow are listed below with
for word, frequency in sorted_version_jack.items():
    print(f"Word: {word}\t\t Number of times spoken: {frequency}")
    count += 1
    if count == limit:
        break
```

```
The 20 most common words spoken by jack sparrow are listed below with the co
rresponding frequencies:

Word: want              Number of times spoken: 15
Word: come              Number of times spoken: 11
Word: know              Number of times spoken: 9
Word: oh                Number of times spoken: 9
Word: will              Number of times spoken: 9
Word: bugger            Number of times spoken: 8
Word: dirt              Number of times spoken: 8
Word: love              Number of times spoken: 8
Word: hey               Number of times spoken: 7
Word: one               Number of times spoken: 7
Word: mate              Number of times spoken: 6
Word: captain           Number of times spoken: 6
Word: key               Number of times spoken: 6
Word: would             Number of times spoken: 6
Word: jones             Number of times spoken: 6
Word: save              Number of times spoken: 6
Word: jar               Number of times spoken: 6
Word: chest             Number of times spoken: 6
Word: much              Number of times spoken: 5
Word: way               Number of times spoken: 5
```

## Optional:

Take a look at the most common words spoken by each; they include names. Who does each mention the most and what does this say about the characters?