Preamble: A Brand New Jay

After an eventful season on season 8 of *A Brand New Jay*, the 3 remaining contestants were invited to Jay Stacksby's private island for the last three episodes. When the day of filming the finale came Mr. Stacksby was found with one of his Professional Series 8-inch Chef Knives plunged through his heart! After the initial investigation highlighted that the film crew all lived in a separate house on the other side of the island, it was concluded that only the three contestants were near enough to Stacksby in order to commit a crime. At the scene of the crime, a letter was left. Here are the contents of that letter:

> You may call me heartless, a killer, a monster, a murderer, but I'm still NOTHING compared to the villian that Jay was. This whole contest was a sham, an elaborate plot to shame the contestants and feed Jay's massive, massive ego. SURE you think you know him! You've seen him smiling for the cameras, laughing, joking, telling stories, waving his money around like a prop but off camera he was a sinister beast, a cruel cruel taskmaster, he treated all of us like slaves, like cattle, like animals! Do you remember Lindsay, she was the first to go, he called her such horrible things that she cried all night, keeping up all up, crying, crying, and more crying, he broke her with his words. I miss my former cast members, all of them very much. And we had to live with him, live in his home, live in his power, deal with his crazy demands. AND FOR WHAT! DID YOU KNOW THAT THE PRIZE ISN'T REAL? He never intended to marry one of us! The carrot on the stick was gone, all that was left was stick, he told us last night that we were all a terrible terrible disappointment and none of us would ever amount to anything, and that regardless of who won the contest he would never speak to any of us again! It's definitely the things like this you can feel in your gut how wrong he is! Well I showed him, he got what he deserved all right, I showed him, I showed him the person I am! I wasn't going to be pushed around any longer, and I wasn't going to let him go on pretending that he was some saint when all he was was a sick sick twisted man who deserved every bit of what he got. The fans need to know, Jay Stacksby is a vile amalgamation of all things evil and bad and the world is a better place without him.

Pretty sinister stuff! Luckily, in addition to this bold-faced admission, we have the introduction letters of the three contestants. Maybe there is a way to use this information to determine who the author of this murder letter is?

Myrtle Beech's introduction letter:

> Salutations. My name? Myrtle. Myrtle Beech. I am a woman of simple tastes. I enjoy reading, thinking, and doing my taxes. I entered this competition because I want a serious relationship. I want a commitment. The last man I dated was too whimsical. He wanted to go on dates that had no plan. No end goal. Sometimes we would just end up wandering the streets after dinner. He called it a "walk". A "walk" with no destination. Can you imagine? I like every action I take to have a measurable effect. When I see a movie, I like to walk away with insights that I did not have before. When I take a bike ride, there better be a worthy destination at the end of the bike path. Jay seems frivolous at times. This worries me. However, it is my staunch belief that one does not make and keep money without having a modicum of discipline. As such, I am hopeful. I will now list three things I cannot live without. Water. Emery boards. Dogs. Thank you for the opportunity to introduce myself. I look forward to the competition.

Lily Trebuchet's introduction letter:

> Hi, I'm Lily Trebuchet from East Egg, Long Island. I love cats, hiking, and curling up under a warm blanket with a book. So they gave this little questionnaire to use for our bios so lets get started. What are some of my least favorite household chores? Dishes, oh yes it's definitely the dishes, I just hate doing them, don't you? Who is your favorite actor and why? Hmm, that's a hard one, but I think recently I'll have to go with Michael B. Jordan, every bit of that man is handsome, HANDSOME! Do you remember seeing him shirtless? I can't believe what he does for the cameras! Okay okay next question, what is your perfect date? Well it starts with a nice dinner at a delicious but small restaurant, you know like one of those places where the owner is in the back and comes out to talk to you and ask you how your meal was. My favorite form of art? Another hard one, but I think I'll have to go with music, music you can feel in your whole body and it is electrifying and best of all, you can dance to it! Okay final question, let's see, What are three things you cannot live without? Well first off, my beautiful, beautiful cat Jerry, he is my heart and spirit animal. Second is pasta, definitely pasta, and the third I think is my family, I love all of them very much and they support me in everything I do. I know Jay Stacksby is a handsome man and all of us want to be the first to walk down the aisle with him, but I think he might truly be the one for me. Okay that's it for the bio, I hope you have fun watching the show!

Gregg T Fishy's introduction letter:

> A good day to you all, I am Gregg T Fishy, of the Fishy Enterprise fortune. I am 37 years young. An adventurous spirit and I've never lost my sense of childlike wonder. I do love to be in the backyard gardening and I have the most extraordinary time when I'm fishing. Fishing for what, you might ask? Why, fishing for compliments of course! I have a stunning pair of radiant blue eyes. They will pierce the soul of anyone who dare gaze upon my countenance. I quite enjoy going on long jaunts through garden paths and short walks through greenhouses. I hope that Jay will be as absolutely interesting as he appears on the television. I find that he has some of the most curious tastes in style and humor. When I'm out and about I quite enjoy hearing tales that instill in my heart of hearts the fascination that beguiles my every day life. Every fiber of my being scintillates and vascillates with extreme pleasure during one of these charming anecdotes and significantly pleases my beautiful personage. I cannot wait to enjoy being on A Brand New Jay. It certainly seems like a grand time to explore life and love.

Saving The Different Examples as Variables

First let's create variables to hold the text data in! Save the muder note as a string in a variable called `murder_note`. Save Lily Trebuchet's introduction into `lily_trebuchet_intro`. Save Myrtle Beech's introduction into `myrtle_beech_intro`.

In [257]:
```
# Add escape \ in front of special characters to return the entire note as one str
ing
murder_note="You may call me heartless, a killer, a monster, a murderer, but I'm s
till NOTHING compared to the villian that Jay was. This whole contest was a sham,
 an elaborate plot to shame the contestants and feed Jay's massive, massive ego. S
URE you think you know him! You've seen him smiling for the cameras, laughing, jok
ing, telling stories, waving his money around like a prop but off camera he was a
 sinister beast, a cruel cruel taskmaster, he treated all of us like slaves, like
 cattle, like animals! Do you remember Lindsay, she was the first to go, he called
her such horrible things that she cried all night, keeping up all up, crying, cryi
ng, and more crying, he broke her with his words. I miss my former cast members, a
ll of them very much. And we had to live with him, live in his home, live in his p
ower, deal with his crazy demands. AND FOR WHAT! DID YOU KNOW THAT THE PRIZE ISN'T
REAL? He never intended to marry one of us! The carrot on the stick was gone, all
 that was left was stick, he told us last night that we were all a terrible terrib
le disappointment and none of us would ever amount to anything, and that regardles
s of who won the contest he would never speak to any of us again! It's definitely
 the things like this you can feel in your gut how wrong he is! Well I showed him,
he got what he deserved all right, I showed him, I showed him the person I am! I w
asn't going to be pushed around any longer, and I wasn't going to let him go on pr
etending that he was some saint when all he was was a sick sick twisted man who de
served every bit of what he got. The fans need to know, Jay Stacksby is a vile ama
lgamation of all things evil and bad and the world is a better place without him."
lily_trebuchet_intro="Hi, I'm Lily Trebuchet from East Egg, Long Island. I love ca
ts, hiking, and curling up under a warm blanket with a book. So they gave this lit
tle questionnaire to use for our bios so lets get started. What are some of my lea
st favorite household chores? Dishes, oh yes it's definitely the dishes, I just ha
te doing them, don't you? Who is your favorite actor and why? Hmm, that's a hard o
ne, but I think recently I'll have to go with Michael B. Jordan, every bit of that
man is handsome, HANDSOME! Do you remember seeing him shirtless? I can't believe w
hat he does for the cameras! Okay okay next question, what is your perfect date? W
ell it starts with a nice dinner at a delicious but small restaurant, you know lik
e one of those places where the owner is in the back and comes out to talk to you
 and ask you how your meal was. My favorite form of art? Another hard one, but I t
hink I'll have to go with music, music you can feel in your whole body and it is e
lectrifying and best of all, you can dance to it! Okay final question, let's see,
 What are three things you cannot live without? Well first off, my beautiful, beau
tiful cat Jerry, he is my heart and spirit animal. Second is pasta, definitely pas
ta, and the third I think is my family, I love all of them very much and they supp
ort me in everything I do. I know Jay Stacksby is a handsome man and all of us wan
t to be the first to walk down the aisle with him, but I think he might truly be t
he one for me. Okay that's it for the bio, I hope you have fun watching the show!"
myrtle_beech_intro="Salutations. My name? Myrtle. Myrtle Beech. I am a woman of si
mple tastes. I enjoy reading, thinking, and doing my taxes. I entered this competi
tion because I want a serious relationship. I want a commitment. The last man I da
ted was too whimsical. He wanted to go on dates that had no plan. No end goal. Som
etimes we would just end up wandering the streets after dinner. He called it a \"w
alk\". A \"walk\" with no destination. Can you imagine? I like every action I take
to have a measurable effect. When I see a movie, I like to walk away with insights
that I did not have before. When I take a bike ride, there better be a worthy dest
ination at the end of the bike path. Jay seems frivolous at times. This worries m
e. However, it is my staunch belief that one does not make and keep money without
 having a modicum of discipline. As such, I am hopeful. I will now list three thin
gs I cannot live without. Water. Emery boards. Dogs. Thank you for the opportunity
to introduce myself. I look forward to the competition."
gregg_t_fishy_intro="A good day to you all, I am Gregg T Fishy, of the Fishy Enter
prise fortune. I am 37 years young. An adventurous spirit and I've never lost my s
ense of childlike wonder. I do love to be in the backyard gardening and I have the
most extraordinary time when I'm fishing. Fishing for what, you might ask? Why, fi
shing for compliments of course! I have a stunning pair of radiant blue eyes. They
will pierce the soul of anyone who dare gaze upon my countenance. I quite enjoy go
ing on long jaunts through garden paths and short walks through greenhouses. I hop
e that Jay will be as absolutely interesting as he appears on the television. I fi
```

```
nd that he has some of the most curious tastes in style and humor. When I\'m out a
nd about I quite enjoy hearing tales that instill in my heart of hearts the fascin
ation that beguiles my every day life. Every fiber of my being scintillates and va
scillates with extreme pleasure during one of these charming anecdotes and signifi
cantly pleases my beautiful personage. I cannot wait to enjoy being on A Brand New
Jay. It certainly seems like a grand time to explore life and love."
```

The First Indicator: Sentence Length

Perhaps some meaningful data can first be gleaned from these text examples if we measure how long the average sentence length is. Different authors have different patterns of written speech, so this could be very useful in tracking down the killer.

Write a function `get_average_sentence_length` that takes some `text` as an argument. This function should return the average length of a sentence in the text.

Hint (highlight this hint in order to reveal it): Use your knowledge of *string methods* to create a list of all of the sentences in a text, called **sentences_in_text**. Further break up each **sentences_in_text** into a list of words and save the *length* of that list of words to a new list that contains all the sentence lengths, called **sentence_lengths**. Take the average of all of the sentence lengths by adding them all together and dividing by the number of sentences (which should be the same as the length of the **sentence_lengths**).
Remember sentences can end with more than one kind of punctuation, you might find it easiest to use **.replace()** so you only have to split on one punctuation mark. Remember **.replace()** doesn't modify the string itself, it returns a new string!

In [258]:
```
def get_average_sentence_length(text):
    text_period=[text.replace("!",".").replace("?",".")]

# Create sentences_in_text list that breaks text into sentences
    sentences_in_text=[]
    for text in text_period:
       sentences_in_text.append(text.split('. '))

# Create words_in_sentence list that breaks sentences into words within text
    words_in_sentence=[]
    for text in sentences_in_text:
        for sentence in text:
            words_in_sentence.append(sentence.split())

# Create a list of length of each sentence for a note
    sentence_lengths=[]
    for i in range(len(words_in_sentence)):
        sentence_lengths.append(len(words_in_sentence[i]))

# Create a function that returns sum of length of each sentence/num of sentences
    sum_length=0
    for i in range(len(words_in_sentence)):
        sum_length+=len(words_in_sentence[i])
    return round(sum_length/(len(words_in_sentence)),2)
```

Creating The Definition for Our Model

Now that we have a metric we want to save and data that is coupled with that metric, it might be time to create our data type. Let's define a class called `TextSample` with a constructor. The constructor should take two arguments: `text` and `author`. `text` should be saved as `self.raw_text`. Call `get_average_sentence_length` with the raw text and save it to `self.average_sentence_length`. You should save the author of the text as `self.author`.

Additionally, define a string representation for the model. If you print a `TextSample` it should render:

- The author's name
- The average sentence length

This will be your main class for the problem at hand. All later instruction to update `TextSample` should be done in the code block below. After updating `TextSample`, click on the `Cell` option in the Jupyter Notebook main menu above, then click `Run All` to rerun the cells from top to bottom. If you need to restart your Jupyter Notebook either run the cells below first or move the `TextSample` class definition & instantiation cells to the bottom.

In [259]:
```
class TextSample:
    def __init__(self, raw_text, author, prepared_text,word_count_frequency,ngram_
frequency):
        self.raw_text=raw_text
        self.average_sentence_length=get_average_sentence_length(raw_text)
        self.author=author
        self.prepared_text=prepare_text(raw_text)
        self.word_count_frequency=build_frequency_table(prepared_text)
        self.ngram_frequency=build_frequency_table(ngram_creator(prepared_text))

# Create string representation
    def __repr__(self):
        return "The author's name is " + (self.author) +". The average sentence le
ngth is " + str(self.average_sentence_length) +"."

# Create a list of clean words
    def prepare_text(self, raw_text):
        prepared_text=raw_text.lower().replace("!"," ").replace("?"," ").replace(
"."," ").split()
        return prepared_text

# Create frequency table for the list of clean words
    def build_frequency_table(self, prepared_text):
        word_count_frequency={}
        for element in prepared_text:
            if element not in frequency_table:
                word_count_frequency[element]=0
            word_count_frequency[element]+=1
        return word_count_frequency

# Create a list of two-word n-grams
    def ngram_creator(self, prepared_text):
        ngram_list=[]
        for i in range(len(prepared_text)-1):
            ngram_list.append(prepared_text[i]+" "+prepared_text[i+1])
        return ngram_list

# Create frequency table for the list of two-word n-grams
    def frequency_comparison(self, table1,table2):
        appearances=0
        mutual_appearances=0
        for key in table1.keys():
            if key not in table2.keys():
                appearances+=table1[key]
            elif table1[key]<=table2[key]:
                appearances+=table2[key]
                mutual_appearances+=table1[key]
            else:
                appearances+=table1[key]
                mutual_appearances+=table2[key]
        for key in table2.keys():
            if key not in table1.keys():
                appearances+=table2[key]
        frequency_comparison_score=round(mutual_appearances/appearances,2)
        return frequency_comparison_score

# Create forumla to calculate percent of difference of sentence lenghs
    def percent_difference(self, value1, value2):
        sentence_length_difference=(abs((value1-value2)/((value1+value2)/2)))
        return sentence_length_difference

    #prepared_text
murder_prepared_text=prepare_text(murder_note)
```

```
lily_prepared_text=prepare_text(lily_trebuchet_intro)
myrtle_prepared_text=prepare_text(myrtle_beech_intro)
gregg_prepared_text=prepare_text(gregg_t_fishy_intro)

#word_count_frequency
murder_wc_fre=build_frequency_table(murder_prepared_text)
lily_wc_fre=build_frequency_table(lily_prepared_text)
myrtle_wc_fre=build_frequency_table(myrtle_prepared_text)
gregg_wc_fre=build_frequency_table(gregg_prepared_text)

#ngram_frequency
murder_ngram_fre=build_frequency_table(ngram_creator(murder_prepared_text))
lily_ngram_fre=build_frequency_table(ngram_creator(lily_prepared_text))
myrtle_ngram_fre=build_frequency_table(ngram_creator(myrtle_prepared_text))
gregg_ngram_fre=build_frequency_table(ngram_creator(gregg_prepared_text))
```

Creating our TextSample Instances

Now create a `TextSample` object for each of the samples of text that we have.

- `murderer_sample` for the murderer's note.
- `lily_sample` for Lily Trebuchet's note.
- `myrtle_sample` for Myrtle Beech's note.
- `gregg_sample` for Gregg T Fishy's note.

Print out each one after instantiating them.

In [260]:
```
murderer_sample=TextSample(murder_note,"Murderer",murder_prepared_text, murder_wc_
fre, murder_ngram_fre)
lily_sample=TextSample(lily_trebuchet_intro,"Lily", lily_prepared_text, lily_wc_fr
e, lily_ngram_fre)
myrtle_sample=TextSample(myrtle_beech_intro,"Myrtle", myrtle_prepared_text, murder
_wc_fre, myrtle_ngram_fre)
gregg_sample=TextSample(gregg_t_fishy_intro,"Gregg", gregg_prepared_text, gregg_wc
_fre, gregg_ngram_fre)

def find_text_similarity(sample1, sample2):
    sentence_length_similarity=abs(1-percent_difference(sample1.average_sentence_l
ength, sample2.average_sentence_length))
    word_count_similarity=frequency_comparison(sample1.word_count_frequency, sampl
e2.word_count_frequency)
    ngram_similarity=frequency_comparison(sample1.ngram_frequency,sample2.ngram_fr
equency)
    similarity_score=(sentence_length_similarity+word_count_similarity+ngram_simil
arity)/3
    return sample2.author+"'s similarity score is "+str(round(similarity_score,2))
+"."

print(murderer_sample)
print(lily_sample)
print(myrtle_sample)
print(gregg_sample)
print(find_text_similarity(murderer_sample,lily_sample))
print(find_text_similarity(murderer_sample,myrtle_sample))
print(find_text_similarity(murderer_sample,gregg_sample))
```

```
The author's name is Murderer. The average sentence length is 22.07.
The author's name is Lily. The average sentence length is 15.84.
The author's name is Myrtle. The average sentence length is 6.75.
The author's name is Gregg. The average sentence length is 13.47.
Lily's similarity score is 0.31.
Myrtle's similarity score is 0.08.
Gregg's similarity score is 0.23.
```

Cleaning Our Data

We want to compare the word choice and usage between the samples, but sentences make our text data fairly messy. In order to analyze the different messages fairly, we'll need to remove all the punctuation and uppercase letters from the samples.

Create a function called `prepare_text` that takes a single parameter `text`, makes the text entirely lowercase, removes all the punctuation and returns a list of the words in the text in order.

For example: `"Where did you go, friend? We nearly saw each other."` would become `['where', 'did', 'you', 'go', 'friend', 'we', 'nearly', 'saw', 'each', 'other']`.

In [261]:
```
# Create prepare_text
def prepare_text(text):
    return text.lower().replace("!"," ").replace("?"," ").replace("."," ").split()
```

Update the constructor for `TextSample` to save the prepared text as `self.prepared_text`.

Building A Frequency Table

Now we want to see which words were most frequently used in each of the samples. Create a function called `build_frequency_table`. It takes in a list called `corpus` and creates a dictionary called `frequency_table`. For every element in `corpus` the value `frequency_table[element]` should be equal to the number of times that element appears in `corpus`. For example the input `['do', 'you', 'see', 'what', 'i', 'see']` would create the frequency table `{'what': 1, 'you': 1, 'see' 2, 'i': 1}`.

In [262]:
```
# Build a frequency table
def build_frequency_table(corpus):
    frequency_table={}
    for element in corpus:
        if element not in frequency_table:
            frequency_table[element]=0
        frequency_table[element]+=1
    return frequency_table
```

The Second Indicator: Favorite Words

Use `build_frequency_table` with the prepared text to create a frequency table that counts how frequently all the words in each text sample appears. Call these functions in the constructor for `TextSample` and assign the word frequency table to a value called `self.word_count_frequency`.

The Third Indicator: N-Grams

An n-gram (https://en.wikipedia.org/wiki/N-gram) is a text analysis technique used for pattern recognition and applicable throughout lingusitics. We're going to use n-grams to find who uses similar word-pairs to the murderer, and we think it's going to make our evidence strong enough to conclusively find the killer.

Create a function called `ngram_creator` that takes a parameter `text_list`, a treated in-order list of the words in a text sample. `ngram_creator` should return a list of all adjacent pairs of words, styled as strings with a space in the center.

For instance, calling `ngram_creator` with the input `['what', 'in', 'the', 'world', 'is', 'going', 'on']` Should produce the output `['what in', 'in the', 'the world', 'world is', 'is going', 'going on']`.

These are two-word n-grams.

In [263]:
```
# reate ngram_creator
def ngram_creator(text_list):
    ngram_list=[]
    for i in range(len(text_list)-1):
        ngram_list.append(text_list[i]+" "+text_list[i+1])
    return ngram_list
```

Use `ngram_creator` along with the prepared text to create a list of all the two-word ngrams in each `TextSample`. Use `build_frequency_table` to tabulate the frequency of each ngram. In the constructor for `TextSample` save this frequency table as `self.ngram_frequency`.

Comparing Two Frequency Tables

We want to know how similar two frequency tables are, let's write a function that computes the comparison between two frequency tables and scores them based on similarity.

Write a function called `frequency_comparison` that takes two parameters, `table1` and `table2`. It should define two local variables, `appearances` and `mutual_appearances`.

Iterate through `table1`'s keys and check if `table2` has the same key defined. If it is, compare the two values for the key -- the smaller value should get added to `mutual_appearances` and the larger should get added to `appearances`. If the key doesn't exist in `table2` the value for the key in `table1` should be added to `appearances`.

Remember afterwards to iterate through all of `table2`'s keys that aren't in `table1` and add those to `appearances` as well.

Return a frequency comparison score equal to the mutual appearances divided by the total appearances.

In [264]:
```
#create frequency comparison score
def frequency_comparison(table1,table2):
    appearances=0
    mutual_appearances=0
    for key in table1.keys():
        if key not in table2.keys():
            appearances+=table1[key]
        elif table1[key]<=table2[key]:
            appearances+=table2[key]
            mutual_appearances+=table1[key]
        else:
            appearances+=table1[key]
            mutual_appearances+=table2[key]
    for key in table2.keys():
        if key not in table1.keys():
            appearances+=table2[key]
    frequency_comparison_score=round(mutual_appearances/appearances,2)
    return frequency_comparison_score
```

Comparing Average Sentence Length

In order to calculate the change between the average sentence lengths of two `TextSamples` we're going to use the formula for the percent difference.

Write a function called `percent_difference` that returns the percent difference as calculated from the following formula:

$$\frac{|\ value1 - value2\ |}{\frac{value1+value2}{2}}$$

In the numerator is the absolute value (use `abs()`) of the two values subtracted from each other. In the denominator is the average of the two values (value1 + value2 divided by two).

In [265]:
```
# Calculate Avg Sentence Length
def percent_difference(value1, value2):
    sentence_length_difference=(abs(value1-value2))/((value1+value2)/2)
    return sentence_length_difference
```

Scoring Similarity with All Three Indicators

We want to figure out who did it, so let's use all three of the indicators we built to score text similarity. Define a function `find_text_similarity` that takes two `TextSample` arguments and returns a float between 0 and 1 where 0 means completely different and 1 means the same exact sample. You can evaluate the similarity by the following criteria:

- Calculate the percent difference of their average sentence length using `percent_difference`. Save that into a variable called `sentence_length_difference`. Since we want to find how *similar* the two passages are calculate the inverse of `sentence_length_difference` by using the formula `abs(1 - sentence_length_difference)`. Save that into a variable called `sentence_length_similarity`.
- Calculate the difference between their word usage using `frequency_comparison` on both `TextSample`'s `word_count_frequency` attributes. Save that into a variable called `word_count_similarity`.
- Calculate the difference between their two-word ngram using `frequency_table` on both `TextSample`'s `ngram_frequency` attributes. Save that into a variable called `ngram_similarity`.
- Add all three similarities together and divide by 3.

In [266]:
```
#Scoring similarity
#def find_text_similarity(self, sample1, sample2):
    #sentence_length_similarity=abs(1-percent_difference(sample1, sample2))
    #word_count_similarity=frequency_comparison(sample1, sample2)
    #ngram_similarity=frequency_comparison(sample1,sample2)
    #similarity_score=(sentence_length_similarity+word_count_similarity+ngram_simi
larity)/3
    #return similarity_score
```

Rendering the Results

We want to print out the results in a way that we can read! For each contestant on *A Brand New Jay* print out the following:

- Their name
- Their similarity score to the murder letter

In [267]:
```
print(find_text_similarity(murderer_sample,lily_sample))
print(find_text_similarity(murderer_sample,myrtle_sample))
print(find_text_similarity(murderer_sample,gregg_sample))
```

```
Lily's similarity score is 0.31.
Myrtle's similarity score is 0.08.
Gregg's similarity score is 0.23.
```

Who Dunnit?

In the cell below, print the name of the person who killed Jay Stacksby.

In [268]:
```
print("Lily")
```

```
Lily
```