In [1]:
```
import nltk
#nltk.download('stopwords')
```

In [2]:
```
import os
import re
import emoji
import pandas as pd
import numpy as np


from collections import Counter, defaultdict
from nltk.corpus import stopwords
from string import punctuation

sw = stopwords.words("english")
```

In [3]:
```
# change `data_location` to the location of the folder on your machine.
data_location = "/Users/Abana/Downloads"
# These subfolders should still work if you correctly stored the
# data from the Module 1 assignment
twitter_folder = "/Users/Abana/Downloads/twitter"
lyrics_folder = "/Users/Abana/Downloads/lyrics"
```

In [4]:
```
def descriptive_stats(tokens, num_tokens = 5, verbose=True) :
    """
        Given a list of tokens, print number of tokens, number of unique tokens,
        number of characters, lexical diversity (https://en.wikipedia.org/wiki/Lexical_
        and num_tokens most common tokens. Return a list with the number of tokens, num
        of unique tokens, lexical diversity, and number of characters.

    """

    # Fill in the correct values here.
    #num_tokens = 0
    #num_unique_tokens = 0
    #lexical_diversity = 0.0
    #num_characters = 0
    num_tokens = len(tokens)
    num_unique_tokens = len(set(tokens))
    num_characters = sum(len(token) for token in tokens)
    lexical_diversity = num_unique_tokens / num_tokens

    if verbose :
        print(f"There are {num_tokens} tokens in the data.")
        print(f"There are {num_unique_tokens} unique tokens in the data.")
        print(f"There are {num_characters} characters in the data.")
        print(f"The lexical diversity is {lexical_diversity:.3f} in the data.")

        # print the five most common tokens

    return([num_tokens, num_unique_tokens,
            lexical_diversity,
            num_characters])
```

In [5]:
```
text = """here is some example text with other example text here in this text""".split(
assert(descriptive_stats(text, verbose=True)[0] == 13)
assert(descriptive_stats(text, verbose=False)[1] == 9)
assert(abs(descriptive_stats(text, verbose=False)[2] - 0.69) < 0.02)
assert(descriptive_stats(text, verbose=False)[3] == 55)
```

```
There are 13 tokens in the data.
There are 9 unique tokens in the data.
There are 55 characters in the data.
The lexical diversity is 0.692 in the data.
```

Q: Why is it beneficial to use assertion statements in your code?

A: One of the reasons why it's beneficial to use assertion statements within code is because it will check errors if there is an error code will stop. If that is the case then we can fix it.

Data Input

Now read in each of the corpora. For the lyrics data, it may be convenient to store the entire contents of the file to make it easier to inspect the titles individually, as you'll do in the last part of the assignment. In the solution, I stored the lyrics data in a dictionary with two dimensions of keys: artist and song. The value was the file contents. A data frame would work equally well.

For the Twitter data, we only need the description field for this assignment. Feel free all the descriptions read it into a data structure. In the solution, I stored the descriptions as a dictionary of lists, with the key being the artist.

In [6]:
```
df_lyrics = {
    'artist': [],
    'song_name': [],
    'contents': []
}

for artist_folder in os.listdir(lyrics_folder):
    for song_file in os.listdir(os.path.join(lyrics_folder, artist_folder)):
        with open(os.path.join(lyrics_folder, artist_folder, song_file), 'r') as f:
            song_lyrics = f.read()
        df_lyrics['artist'].append(artist_folder)
        df_lyrics['song_name'].append(song_file.replace('www_azlyrics_comkcijojo_', ''))
        df_lyrics['contents'].append(song_lyrics)
```

In [7]:
```
df_lyrics = pd.DataFrame(df_lyrics)
df_lyrics
```

Out[7]:

	artist	song_name	contents
0	realkcijojo	allmylife	"All My Life"\n\nBaby, baby, baby, baby, baby,...
1	realkcijojo	babycomeback	"Baby Come Back"\n\n[Verse 1].I was a fool to ...
2	realkcijojo	dontrushtakeloveslowly	"Don't Rush (Take Love Slowly)"\n\nThe look wi...
3	realkcijojo	feefiefoefum	"Fee Fie Foe Fum"\n\nOhh baby.You been leaving...
4	realkcijojo	girl	"Girl"\n\nBaby I was born to give you all of m...

	artist	song_name	contents
5	realkcijojo	hbi	"HBI"\n\nI really love you.Girl I really love ...
6	realkcijojo	hellodarlin	"Hello Darlin'"\n\nWish that I could have you ...
7	realkcijojo	howcouldyou	"How Could You"\n\nAll I can do.Is sit alone.I...
8	realkcijojo	howlongmusticry	"How Long Must I Cry"\n\nBaby, listen.I never ...
9	realkcijojo	howmanytimes	"How Many Times"\n\nHow many times you're gonn...
10	realkcijojo	intro	"Intro"\n\nOhh wee.My darlin.Can I make love t...
11	realkcijojo	iwannagettoknowyou	"I Wanna Get To Know You"\n\nHey pretty lady, ...
12	realkcijojo	iwannamakelovetoyou	"I Wanna Make Love To You"\n\nEither you're wi...
13	realkcijojo	justforyourlove	"Just For Your Love"\n\n[* = speaking].For you...
14	realkcijojo	lastnightsletter	"Last Night's Letter"\n\n[Verse 1].I was sitti...
15	realkcijojo	life	"Life"\n\nJust like a birdie.I just wanna fly ...
16	realkcijojo	loveballad	"Love Ballad"\n\nI, have never been so much.In...
17	realkcijojo	makinmesaygoodbye	"Makin' Me Say Goodbye"\n\nIt's tree o'clock.A...
18	realkcijojo	nowandforever	"Now And Forever"\n\nThey're always running ar...
19	realkcijojo	stillwaiting	"Still Waiting"\n\nCheck this out.It's Devante...
20	realkcijojo	tellmeitsreal	"Tell Me It's Real"\n\n[Chorus:].Tell me it's ...
21	realkcijojo	youbringmeup	"You Bring Me Up"\n\nIsn't it funny.The things...
22	SammHenshaw	816	"8.16"\n\n(Run to me, girl, run to me).Hey, lo...
23	SammHenshaw	autonomyslave	"Autonomy (Slave)"\n\nNeed laws of my own.No f...
24	SammHenshaw	better	"Better"\n\nSaid I, I need something to ease m...
25	SammHenshaw	chances	"Chances"\n\nI know you know.About the thrill ...
26	SammHenshaw	chickenwings	"Chicken Wings"\n\nCos the heart wants what it...
27	SammHenshaw	easy	"Easy"\n\nI'm a broken man.Yes I am.But I won'...
28	SammHenshaw	everything	"Everything"\n\nPower-hungry politicians make ...
29	SammHenshaw	grow	"Grow"\n\nI just need you near me.With you, it...
30	SammHenshaw	lovedbyyou	"Loved By You"\n\nI remember this thing that s...
31	SammHenshaw	mrintrovert	"Mr Introvert"\n\nShe said she likes it when I...
32	SammHenshaw	mrintrovertreprise	"Mr Introvert (Reprise)"\n\nAh Bruv.No, Mate.N...
33	SammHenshaw	nightcalls	"Night Calls"\n\nWe slept.Under the safety of ...
34	SammHenshaw	onlywannabewithyouunplugged	"Only Wanna Be With You (Unplugged)"\n\nSaid, ...
35	SammHenshaw	ourlove	"Our Love"\n\nI know.I know its been hard on y...
36	SammHenshaw	redemption	"Redemption"\n\nOh no... oho....If I die today...
37	SammHenshaw	stillnoalbumintro	"Still No Album (Intro)"\n\nYeah bro?.Ah, stop...

	artist	song_name	contents
38	SammHenshaw	temptationintro	"Temptation (Intro)"\n\nMy mom told me, "Stop"...
39	SammHenshaw	thesehands	"These Hands"\n\nI've been procrastinating for...
40	SammHenshaw	thoughtsandprayers	"Thoughts And Prayers"\n\nHello stranger.The g...

In [8]:
```
#Reading in data
df_twitter = {
    'artist': [],
    'description': []
}
for filename in os.listdir(twitter_folder):
    if 'data' in filename:
        artist = filename.split('_')[0]
        with open(os.path.join(twitter_folder, filename), 'r') as f:
            for line in f:
                fields = [t.strip() for t in line.split('    ') if t.strip()]
                description = fields[-2]
                if description=='description':
                    continue
                df_twitter['artist'].append(artist)
                df_twitter['description'].append(description)
```

In [9]:
```
df_twitter = pd.DataFrame(df_twitter)
df_twitter
```

Out[9]:

	artist	description
0	realkcijojo	Thanking God n loving life \U0001f600.
1	realkcijojo	\U0001F497 Young, Black, hard working & humble...
2	realkcijojo	Sudan
3	realkcijojo	I live in the 3rd pyramid on the left
4	realkcijojo	God Fearing,Daughter,Sister, Auntie!!! Im a st...
...	...	...
189	SammHenshaw	Friendly, fun loving, Jesus Freak, charismati...
190	SammHenshaw	Life full of mysteries
191	SammHenshaw	London
192	SammHenshaw	None
193	SammHenshaw	Outer Space

194 rows × 2 columns

Data Cleaning

Now clean and tokenize your data. Remove punctuation chacters (available in the punctuation object in the string library), split on whitespace, fold to lowercase, and remove stopwords. Store

your cleaned data, which must be accessible as an interable for descriptive_stats, in new objects or in new columns in your data frame.

In [10]:
```
punctuation = set(punctuation) # speeds up comparison
```

In [11]:
```
def clean_data(s, punctuation):
    s = ''.join(ch for ch in s if ch not in punctuation)
    word_list = s.replace('\n', ' ').lower().split(' ')
    filtered_words = [word for word in word_list if (word not in sw) and word !='']
    return filtered_words
```

In [12]:
```
# create your clean twitter data here
df_twitter['re_punc'] = df_twitter['description'].apply(lambda s: clean_data(s, punctua
df_twitter
```

Out[12]:

	artist	description	re_punc
0	realkcijojo	Thanking God n loving life \U0001f600.	[thanking, god, n, loving, life, u0001f600]
1	realkcijojo	\U0001F497 Young, Black, hard working & humble...	[u0001f497, young, black, hard, working, humbl...
2	realkcijojo	Sudan	[sudan]
3	realkcijojo	I live in the 3rd pyramid on the left	[live, 3rd, pyramid, left]
4	realkcijojo	God Fearing,Daughter,Sister, Auntie!!! Im a st...	[god, fearingdaughtersister, auntie, im, stron...
...	...	...	...
189	SammHenshaw	Friendly, fun loving, Jesus Freak, charismati...	[friendly, fun, loving, jesus, freak, charisma...
190	SammHenshaw	Life full of mysteries	[life, full, mysteries]
191	SammHenshaw	London	[london]
192	SammHenshaw	None	[none]
193	SammHenshaw	Outer Space	[outer, space]

194 rows × 3 columns

In [13]:
```
# create your clean lyrics data here
df_lyrics['re_punc'] = df_lyrics['contents'].apply(lambda s: clean_data(s, punctuation)
df_lyrics
```

Out[13]:

	artist	song_name	contents	re_punc
0	realkcijojo	allmylife	"All My Life"\n\nBaby, baby, baby, baby, baby,...	[life, baby, baby, baby, baby, baby, baby, bab...
1	realkcijojo	babycomeback	"Baby Come Back"\n\n[Verse 1].I was a fool to ...	[baby, come, back, verse, 1i, fool, let, golet...

	artist	song_name	contents	re_punc
2	realkcijojo	dontrushtakeloveslowly	"Don't Rush (Take Love Slowly)"\n\nThe look wi...	[dont, rush, take, love, slowly, look, within,...
3	realkcijojo	feefiefoefum	"Fee Fie Foe Fum"\n\nOhh baby.You been leaving...	[fee, fie, foe, fum, ohh, babyyou, leaving, ev...
4	realkcijojo	girl	"Girl"\n\nBaby I was born to give you all of m...	[girl, baby, born, give, lovebaby, born, give,...
5	realkcijojo	hbi	"HBI"\n\nI really love you.Girl I really love ...	[hbi, really, love, yougirl, really, love, you...
6	realkcijojo	hellodarlin	"Hello Darlin'"\n\nWish that I could have you ...	[hello, darlin, wish, could, spacewish, could,...
7	realkcijojo	howcouldyou	"How Could You"\n\nAll I can do.Is sit alone.I...	[could, dois, sit, alonein, roomthinking, youh...
8	realkcijojo	howlongmusticry	"How Long Must I Cry"\n\nBaby, listen.I never ...	[long, must, cry, baby, listeni, never, meant,...
9	realkcijojo	howmanytimes	"How Many Times"\n\nHow many times you're gonn...	[many, times, many, times, youre, gonna, let, ...
10	realkcijojo	intro	"Intro"\n\nOhh wee.My darlin.Can I make love t...	[intro, ohh, weemy, darlincan, make, love, ton...
11	realkcijojo	iwannagettoknowyou	"I Wanna Get To Know You"\n\nHey pretty lady, ...	[wanna, get, know, hey, pretty, lady, look, fi...
12	realkcijojo	iwannamakelovetoyou	"I Wanna Make Love To You"\n\nEither you're wi...	[wanna, make, love, either, youre, metell, wan...
13	realkcijojo	justforyourlove	"Just For Your Love"\n\n[* = speaking].For you...	[love, speakingfor, lovebaby, im, like, sad, m...
14	realkcijojo	lastnightsletter	"Last Night's Letter"\n\n[Verse 1].I was sitti...	[last, nights, letter, verse, 1i, sittin, home...
15	realkcijojo	life	"Life"\n\nJust like a birdie.I just wanna fly ...	[life, like, birdiei, wanna, fly, freeand, pie...
16	realkcijojo	loveballad	"Love Ballad"\n\nI, have never been so much.In...	[love, ballad, never, muchin, love, beforewhat...
17	realkcijojo	makinmesaygoodbye	"Makin' Me Say Goodbye"\n\nIt's tree o'clock.A...	[makin, say, goodbye, tree, oclockand, youre, ...
18	realkcijojo	nowandforever	"Now And Forever"\n\nThey're always running ar...	[forever, theyre, always, running, aroundtelli...
19	realkcijojo	stillwaiting	"Still Waiting"\n\nCheck this	[still, waiting, check,

	artist	song_name	contents	re_punc
			out.It's Devante...	outits, devante, haile...
20	realkcijojo	tellmeitsreal	"Tell Me It's Real"\n\n[Chorus:].Tell me it's ...	[tell, real, chorustell, realthe, feeling, fee...
21	realkcijojo	youbringmeup	"You Bring Me Up"\n\nIsn't it funny.The things...	[bring, isnt, funnythe, things, said, done, me...
22	SammHenshaw	816	"8.16"\n\n(Run to me, girl, run to me).Hey, lo...	[816, run, girl, run, mehey, lovesay, lovehey,...
23	SammHenshaw	autonomyslave	"Autonomy (Slave)"\n\nNeed laws of my own.No f...	[autonomy, slave, need, laws, ownno, foreign, ...
24	SammHenshaw	better	"Better"\n\nSaid I, I need something to ease m...	[better, said, need, something, ease, soul, li...
25	SammHenshaw	chances	"Chances"\n\nI know you know.About the thrill ...	[chances, know, knowabout, thrill, play, fired...
26	SammHenshaw	chickenwings	"Chicken Wings"\n\nCos the heart wants what it...	[chicken, wings, cos, heart, wants, wantsand, ...
27	SammHenshaw	easy	"Easy"\n\nI'm a broken man.Yes I am.But I won'...	[easy, im, broken, manyes, ambut, wont, let, s...
28	SammHenshaw	everything	"Everything"\n\nPower-hungry politicians make ...	[everything, powerhungry, politicians, make, s...
29	SammHenshaw	grow	"Grow"\n\nI just need you near me.With you, it...	[grow, need, near, mewith, easier, get, bylike...
30	SammHenshaw	lovedbyyou	"Loved By You"\n\nI remember this thing that s...	[loved, remember, thing, used, doshe, would, w...
31	SammHenshaw	mrintrovert	"Mr Introvert"\n\nShe said she likes it when I...	[mr, introvert, said, likes, im, socialbut, wo...
32	SammHenshaw	mrintrovertreprise	"Mr Introvert (Reprise)"\n\nAh Bruv.No, Mate.N...	[mr, introvert, reprise, ah, bruvno, mateno, t...
33	SammHenshaw	nightcalls	"Night Calls"\n\nWe slept.Under the safety of ...	[night, calls, sleptunder, safety, darknesswe,...
34	SammHenshaw	onlywannabewithyouunplugged	"Only Wanna Be With You (Unplugged)"\n\nSaid, ...	[wanna, unplugged, said, weve, together, minut...
35	SammHenshaw	ourlove	"Our Love"\n\nI know.I know its been hard on y...	[love, knowi, know, hard, youi, know, getting,...

	artist	song_name	contents	re_punc
36	SammHenshaw	redemption	"Redemption"\n\nOh no... oho....If I die today...	[redemption, oh, ohoif, die, today, would, way...
37	SammHenshaw	stillnoalbumintro	"Still No Album (Intro)"\n\nYeah bro?.Ah, stop...	[still, album, intro, yeah, broah, stop, thatn...
38	SammHenshaw	temptationintro	"Temptation (Intro)"\n\nMy mom told me, "Stop"...	[temptation, intro, mom, told, stopshes, like,...
39	SammHenshaw	thesehands	"These Hands"\n\nI've been procrastinating for...	[hands, ive, procrastinating, far, longi, dont...
40	SammHenshaw	thoughtsandprayers	"Thoughts And Prayers"\n\nHello stranger.The g...	[thoughts, prayers, hello, strangerthe, girls,...

Basic Descriptive Statistics

Call your descriptive_stats function on both your lyrics data and your twitter data and for both artists (four total calls).

In [14]:
```
# calls to descriptive_stats here
print("Twitter")
for artist in df_twitter['artist'].unique():
    print("Artist:", artist)
    df_twitter_artist = df_twitter[df_twitter['artist']==artist]
    tokens = ' '.join([' '.join(c) for c in df_twitter_artist['re_punc']])
    descriptive_stats(tokens, verbose=True)

print('\n')
print("Lyrics")
for artist in df_lyrics['artist'].unique():
    print("Artist:", artist)
    df_lyrics_artist = df_lyrics[df_lyrics['artist']==artist]
    tokens = ' '.join([' '.join(c) for c in df_lyrics_artist['re_punc']])
    descriptive_stats(tokens, verbose=True)
```

```
Twitter
Artist: realkcijojo
There are 4419 tokens in the data.
There are 44 unique tokens in the data.
There are 4419 characters in the data.
The lexical diversity is 0.010 in the data.
Artist: SammHenshaw
There are 4390 tokens in the data.
There are 43 unique tokens in the data.
There are 4390 characters in the data.
The lexical diversity is 0.010 in the data.


Lyrics
Artist: realkcijojo
There are 17213 tokens in the data.
There are 38 unique tokens in the data.
There are 17213 characters in the data.
```

```
The lexical diversity is 0.002 in the data.
Artist: SammHenshaw
There are 14445 tokens in the data.
There are 36 unique tokens in the data.
There are 14445 characters in the data.
The lexical diversity is 0.002 in the data.
```

Q: How do you think the "top 5 words" would be different if we left stopwords in the data?

A: Well if stopwords were to be kept there would more words that are commonly used like "and" "the" also" within the top 5 words. Which does not bring many insights when we're looking for the top 5 words.

Q: What were your prior beliefs about the lexical diversity between the artists? Does the difference (or lack thereof) in lexical diversity between the artists conform to your prior beliefs?

A: Well, my prior belief was that there could be some difference between the two artists in terms of lexical diversity. But it appears that they're quite similar. Artist: realkcijojo has lexical diversity of 0.010 and artist: SammHenshaw has lexical diversity of 0.010 as far as Twitter information goes. But it appears that the same thing follows for the lyrics information which both artists come with a lexical diversity of 0.002. In conclusion, this shows that there is a small amount of range of vocabulary being used. In this case, it makes sense because the song lyrics are short words as well as the tweet which has a small number of characters allowed.

Specialty Statistics

The descriptive statistics we have calculated are quite generic. You will now calculate a handful of statistics tailored to these data.

Ten most common emojis by artist in the twitter descriptions. Ten most common hashtags by artist in the twitter descriptions. Five most common words in song titles by artist. For each artist, a histogram of song lengths (in terms of number of tokens) We can use the emoji library to help us identify emojis and you have been given a function to help you.

In [15]:
```
assert(emoji.is_emoji("❤️"))
assert(not emoji.is_emoji(":-)"))
```

In [16]:
```
df_twitter
```

Out[16]:

	artist	description	re_punc
0	realkcijojo	Thanking God n loving life \U0001f600.	[thanking, god, n, loving, life, u0001f600]
1	realkcijojo	\U0001F497 Young, Black, hard working & humble...	[u0001f497, young, black, hard, working, humbl...
2	realkcijojo	Sudan	[sudan]
3	realkcijojo	I live in the 3rd pyramid on the left	[live, 3rd, pyramid, left]
4	realkcijojo	God Fearing,Daughter,Sister, Auntie!!! Im a st...	[god, fearingdaughtersister, auntie, im, stron...
...	...	...	...

	artist	description	re_punc
189	SammHenshaw	Friendly, fun loving, Jesus Freak, charismati...	[friendly, fun, loving, jesus, freak, charisma...
190	SammHenshaw	Life full of mysteries	[life, full, mysteries]
191	SammHenshaw	London	[london]
192	SammHenshaw	None	[none]
193	SammHenshaw	Outer Space	[outer, space]

194 rows × 3 columns

Emojis 😁 What are the ten most common emojis by artist in the twitter descriptions?

In [17]:
```
# Your code here
twitter_emoji = {}
for artist in df_twitter['artist'].unique():
    twitter_emoji[artist] = {}
    df_twitter_artist = df_twitter[df_twitter['artist']==artist]
    for idx, row in df_twitter_artist.iterrows():
        for word in row['description'].split(' '):
            word = word.replace('.', '').encode().decode('unicode_escape')
            if emoji.is_emoji(word):
                if word not in twitter_emoji[artist].keys():
                    twitter_emoji[artist][word] = 1
                else:
                    twitter_emoji[artist][word] +=1
```

```
<ipython-input-17-ed3e6d952586>:8: DeprecationWarning: invalid escape sequence '\,'
  word = word.replace('.', '').encode().decode('unicode_escape')
```

In [18]:
```
for artist in twitter_emoji.keys():
    twitter_emoji[artist] = {k: v for k, v in sorted(twitter_emoji[artist].items(), key
```

In [19]:
```
twitter_emoji
```

Out[19]:
```
{'realkcijojo': {'🤩': 3, '💗': 1, '🤗': 1, '🍸': 1, '🍀': 1, '😵': 1},
 'SammHenshaw': {'💗': 4, '😁': 2, '🥲': 1, '😇': 1, '😅': 1, '😗': 1}}
```

Hashtags

What are the ten most common hashtags by artist in the twitter descriptions?

In [20]:
```
# Your code here
twitter_hashtags = {}
for artist in df_twitter['artist'].unique():
    twitter_hashtags[artist] = {}
    df_twitter_artist = df_twitter[df_twitter['artist']==artist]
    for idx, row in df_twitter_artist.iterrows():
        for word in row['description'].split(' '):
            if '#' in word:
                if word not in twitter_hashtags[artist].keys():
                    twitter_hashtags[artist][word] = 1
```

```
                else:
                    twitter_hashtags[artist][word] +=1
```

In [21]:
```
for artist in twitter_hashtags.keys():
    twitter_hashtags[artist] = {k: v for k, v in sorted(twitter_hashtags[artist].items(
```

In [22]:
```
twitter_hashtags
```

Out[22]:
```
{'realkcijojo': {'#GodBless': 1,
  '#StayPrayedUp!!': 1,
  '#BLM': 1,
  '#TeamLeo': 1,
  '#Bayareabornandraised': 1,
  '#teamIfollowback': 1,
  '#4everBrandy': 1,
  '#teamfollowback': 1,
  '#TeamTaureanDream': 1,
  '#lovey4life': 1},
 'SammHenshaw': {'#MUFC.': 1,
  '#MUFC': 1,
  '#Federer': 1,
  '#TVD': 1,
  '#TheSecretCircle': 1,
  '#TEENWOLF': 1,
  '#PLL': 1,
  '#spaceshost': 1,
  '#BlackInTheWorkspace': 1,
  '#teamEsRo': 1}}
```

Song Titles

What are the five most common words in song titles by artist? The song titles should be on the first line of the lyrics pages, so if you have kept the raw file contents around, you will not need to re-read the data.

In [23]:
```
song_name = {}
for artist in df_lyrics['artist'].unique():
    song_name[artist] = {}
    df_lyrics_artist = df_lyrics[df_lyrics['artist']==artist]
    for idx, row in df_lyrics_artist.iterrows():
        title = row['contents'].split('"')[1]
        for word in title.split(' '):
            if word not in song_name[artist].keys():
                song_name[artist][word] = 1
            else:
                song_name[artist][word] +=1
```

In [24]:
```
for artist in song_name.keys():
    song_name[artist] = {k: v for k, v in sorted(song_name[artist].items(), key=lambda
```

In [25]:
```
song_name
```

Out[25]:
```
{'realkcijojo': {'Love': 4, 'You': 4, 'How': 3, 'I': 3, 'Me': 3},
 'SammHenshaw': {'You': 2, 'Mr': 2, 'Introvert': 2, '(Intro)': 2, '8.16': 1}}
```

Song Lengths

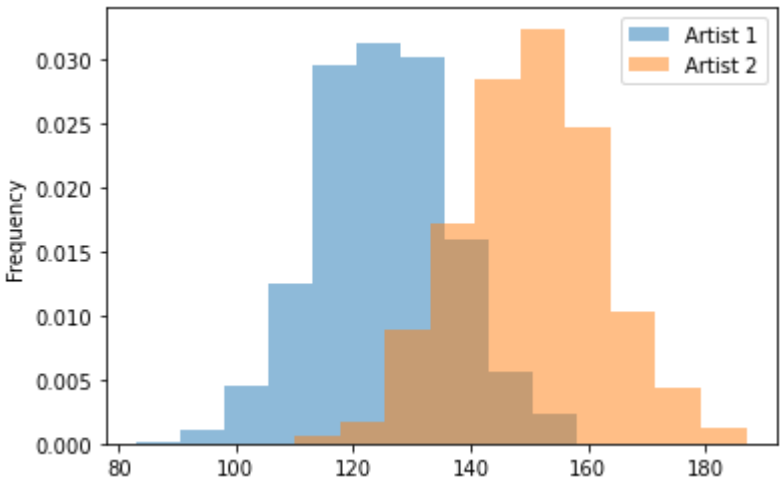
For each artist, a histogram of song lengths (in terms of number of tokens). If you put the song lengths in a data frame with an artist column, matplotlib will make the plotting quite easy. An example is given to help you out.

In [26]:
```
num_replicates = 1000

df = pd.DataFrame({
    "artist" : ['Artist 1'] * num_replicates + ['Artist 2']*num_replicates,
    "length" : np.concatenate((np.random.poisson(125,num_replicates),np.random.poisson(
})

df.groupby('artist')['length'].plot(kind="hist",density=True,alpha=0.5,legend=True)
```

Out[26]:
```
artist
Artist 1    AxesSubplot(0.125,0.125;0.775x0.755)
Artist 2    AxesSubplot(0.125,0.125;0.775x0.755)
Name: length, dtype: object
```



Q: What does the regular expression '\s+' match on?

A: It matches on 1 or more space character

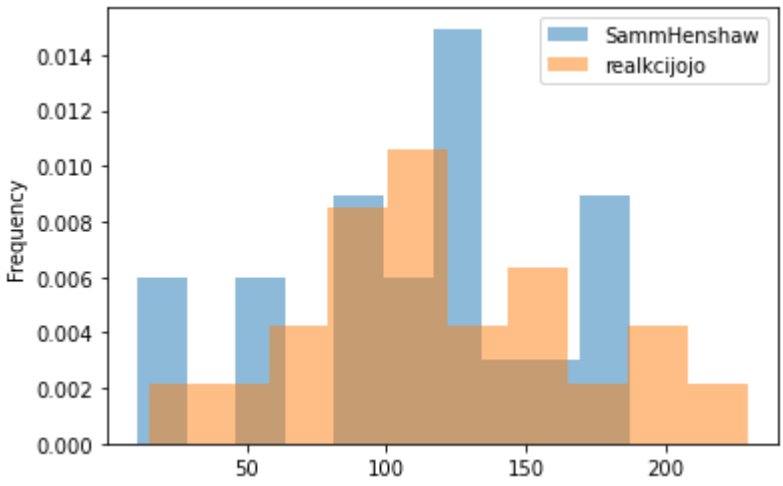
In [27]:
```
collapse_whitespace = re.compile(r'\s+')

def tokenize_lyrics(lyric) :
    """strip and split on whitespace"""
    return([item.lower() for item in collapse_whitespace.split(lyric)])
```

In [28]:
```
# Your lyric length comparison chart here.
df_lyrics['length'] = df_lyrics['re_punc'].apply(lambda x: len(tokenize_lyrics(' '.join
```

In [29]:
```
df_lyrics.groupby('artist')['length'].plot(kind="hist",density=True,alpha=0.5,legend=Tr
```

Out[29]:
```
artist
SammHenshaw    AxesSubplot(0.125,0.125;0.775x0.755)
realkcijojo    AxesSubplot(0.125,0.125;0.775x0.755)
Name: length, dtype: object
```


SammHenshaw
realkcijojo
Frequency
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
50
100
150
200