John Fields
Dr. Norma Palomino
IST664 - Homework #1
Comparison of the song lyrics of Hank Williams Sr. and Holly Williams
10/23/19

PHOTO: WIKIPEDIA
Hank Williams Sr. (1923-1953)


PHOTO: DAVID MCCLISTER
Holly Williams (1981-)

Introduction

Ken Burns 2019 documentary, *Country Music*, traces the roots of this popular music form from the 1920's to today. Country music originated in the southeast United States and Hank Williams, Senior, was a major influence who is highlighted in the film. Rolling Stone magazine ranks Mr. Williams #2 on the list of the 100 Greatest Country Artists of all time.[i] The family tradition continued with Hank Williams, Junior, #50 on the Rolling Stones list, and his daughter Holly Williams, who has recorded over 40 country songs.

The purpose of this paper is to utilize Natural Language Processing (NLP) techniques to compare and contrast the song lyrics of Hank Williams Sr. and his grand-daughter Holly Williams. Some of the questions that will be explored:

- Are there differences in the number of words used Holly and Hank in their songs?
- Are similar themes or words used in the song lyrics?
- What are the differences or similarities in the bigrams scored by frequency and pointwise mutual information (PMI)?

Analysis

About the Text

The lyrics for this analysis were obtained from the website Genius.com which claims to have the "largest collection of song lyrics and music knowledge".[ii] Genius.com provides an application programming interface (API) which allows the song lyrics to be imported into programming software such as Python. Prior to downloading the song lyrics, it was decided to download an equal number of songs from each artist for the song comparisons. Since Holly Williams only has 40 songs on Genius.com, all of her songs were downloaded. Hank Williams has over 200 songs on Genius.com and only the first 40 were downloaded to provide a sample for comparison with his granddaughter's songs. Figures 1 and 2 below show examples of the lyrics from both artists prior to any processing.

```
"Did your lover leave you stranded
Did your heart endure the truth
Will you fight to find the healing or will time heal all your wounds
Do you drown in your desire for a time you used to know
Memories can't touch you they are just a picture show
Dont let go, oh honey can't you see
Everythings a circle in itself
All as it should be

Have you stood inside a nightmare
Did it slap you in the face
Were you left to lose control
Did you ever find a way
Was it somebody that loved you or a face you've never seen
So final, so quick are the hands of destiny
Don't let go, honey can't you see
Everythings a circle in itself
All as it should be
```

Figure 1- Sample of Holly Williams song lyric

```
I love you baby, but you gotta understand
When the Lord made me, he made a Ramblin' Man

Some folks might sa-ay that I'm no good
That I wouldn't settle down if I could
But when that open ro-oad starts to callin' me
There's somethin' o'er the hill that I gotta see

Sometimes it's har-rd but you gotta understand
When the Lord made me, He made a Ra-amblin' Man

I love to see the towns a-passin' by
And to ride these rails 'neath God's blue sky
Let me travel this land from the mountains to the sea
Cause that's the life I believe, He meant for me
And when I'm gone and at my grave, you stand
Just say God's called home your Ramblin' Man"
"On that resurrection morning
When all dead in Christ shall rise
I'll have a new body
Praise the Lord, I'll have a new life
```

Figure 2 - Sample of Hank Williams song lyric

Processing the Text

Below is a summary of the steps followed to download the files and preliminary processing of the text with Python's NLTK package:

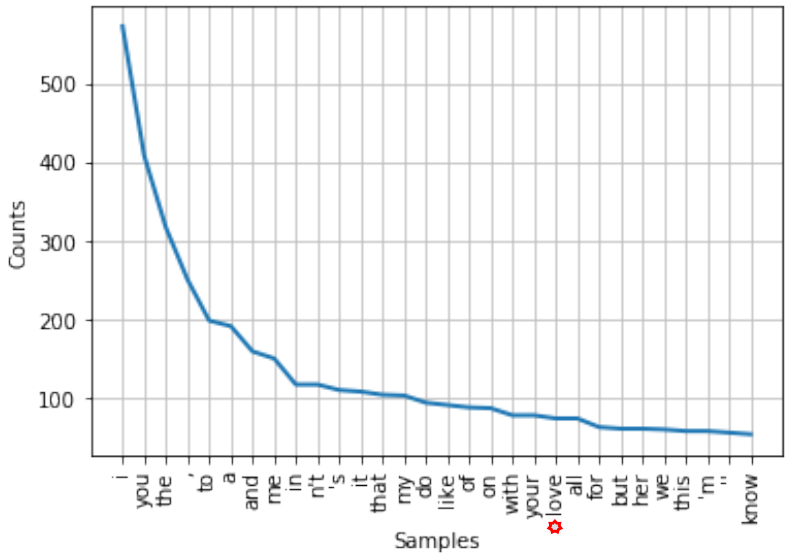
1. Sign up for an API key through Genius.com
2. Write the Python code to import the lyrics for each artist in JSON (java script object notation) files.
3. Merge the JSON files.
4. Create a function to extract only the song lyrics from the JSON file.
5. Read in separate lyric files for Holly Williams and Hank Williams using NLTK. The Hank text contained 8,269 words and the Holly text contained 28,584 words. Holly Williams did write longer songs than her grandfather, but this didn't explain a 3.5X difference. After a review of the Python code, it was discovered that the JSON files from Holly Williams were duplicated. After this was corrected, the Holly word count was 10,133.
6. During a review of the text content, it was also discovered that some songs have recording artist information and instrument (eg Gwyneth Paltrow – background vocals) that needed to be removed.
7. After consultation with Professor Palomino, manual removal of the artist/instrument information was chosen as the preferred method since there was not an automated way to accomplish this task. After these edits, the Holly word count was 9588 and Hank word count was 8050.
8. The final processing step were accomplished by:
   a. Removing stop words
      - The standard NLTK stop word list of 179 words was used.
      - The following additional custom stop words were added: ['hank','holly','williams','could','would','might','must','need','sha','wo','y',"'s","'d","'ll","'t","'m","'re","'ve", "n't", 'repeat','chorus','verse']
   b. Remove punctuation, numbers and separators.
      - The song lyrics contain more word contractions with an apostrophe (') and Hank uses a hyphen (-) as shown in Figure 2. These were removed with other punctuation, numbers and separators to provide more standardized text for the analysis.
   c. Change upper case to lower case.
      - Hank also uses upper case to highlight some words like "Ramblin' Man" in Figure 2. However, the decision was made to make everything lower case for the comparison between the two artists.

d. Lemmatization was also tested on the Holly lyrics and there were some changes to words like "was" to "wa". There did not appear to be an advantage to using lemmatization so it was not used for this project.
e. Tokenize using the NLTK tokenizer that uses WordNet.
   - The Porter tokenizer was also evaluated but it was more aggressive as shown in this example from Holly's lyrics:
     - (Porter)`'A', 'memori', 'fall', 'down', 'from', 'where', 'it', 'wa', 'with', 'thi', 'confess'`
     - (NLTK) `"'a', 'memory', 'falls', 'down', 'from', 'where', 'it', 'was', 'with', 'this', 'confession'`

After the steps above, the Hank text contained 3,435 words and the Holly text contained 3,934 words.

Analyzing the Text

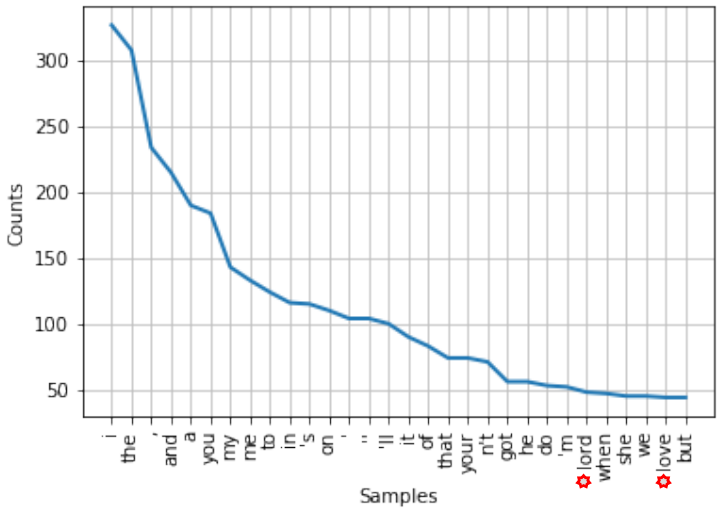
Prior to removing the stop words and non-alphabetical characters, a word frequency for each artist was summarized as shown in Figures 3 and 4 below. Stop words and pronouns are the most common, however, words like "love" appear in both artist's lyrics and the word "lord" is frequently used by Hank Williams.



i	572
you	408
the	317
,	251
to	199
a	192
and	160
me	151
in	118
n't	118
's	111
it	109
that	105
my	104
do	95
like	92
of	89
on	88
with	79
your	79
love	75
all	75
for	64
but	62
her	62

we	61
this	59
'm	59
''	57
know	55
could	55
is	55
never	46
have	46
if	45
when	44
let	44
are	44
there	43
've	43
be	43
was	41
go	41
're	39
he	39
up	38
why	38
heart	37
out	37
'll	36

Figure 3 - Holly Williams word count frequency (top 30 above, top 50 right)



i	327
the	308
,	234
and	215
a	190
you	184
my	143
me	133
to	124
in	116
's	115
on	110
'	104
''	104
'll	100
it	90
of	83
that	74
your	74
n't	71
got	56
he	56
do	53
'm	52
lord	48

when	47
she	45
we	45
love	44
but	44
no	44
have	44
be	43
all	43
for	42
over	42
so	40
is	40
heart	36
was	36
with	35
there	32
oh	30
like	30
-	29
just	28
down	28
will	27
if	26
new	26

Figure 4 - Hank Williams word count frequency (top 30 above, top 50 right)

Searching for words with the letter "z" reveals other differences in the types of words used by each artist. The word "crazy" is the only "z" word that is common to both artists.

```
Holly Williams
['memorized', 'realize', 'crazy', 'crazy', 'cuz', 'criticize',
'philosophize']

Hank Williams
['lazarus', "buzzin'", 'dozen', 'magazines', 'magazines', 'crazy']
```

Additional word frequency charts were created in Figure 5 and Figure 6 to show the changes in word frequency for each artist after removing the stop words and non-alphabetical characters. This reveals that Holly's favorite word is "like" and Hank's favorite word is "got". Both artists also use the word "heart" frequently with it occurring ~35 times in both artists' lyrics.

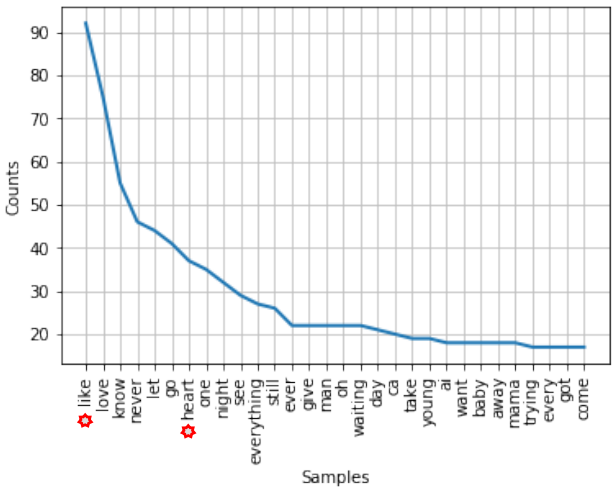


Figure 5 - Holly Williams word frequency after removing stop words and non-alphabetical characters (top 30 above, top 50 right)

```
('like', 92)
('love', 75)
('know', 55)
('never', 46)
('let', 44)
('go', 41)
('heart', 37)
('one', 35)
('night', 32)
('see', 29)
('everything', 27)
('still', 26)
('ever', 22)
('give', 22)
('man', 22)
('oh', 22)
('waiting', 22)
('day', 21)
('ca', 20)
('take', 19)
('young', 19)
('ai', 18)
('want', 18)
('baby', 18)
('away', 18)
```

```
('mama', 18)
('trying', 17)
('every', 17)
('got', 17)
('come', 17)
('well', 17)
('home', 17)
('without', 17)
('june', 17)
('used', 16)
('end', 16)
('believe', 16)
('yeah', 16)
('blue', 16)
('say', 16)
('nothing', 15)
('maybe', 15)
('hands', 15)
('made', 15)
('lie', 14)
('knew', 14)
('way', 14)
('always', 14)
('boy', 14)
('little', 14)
```

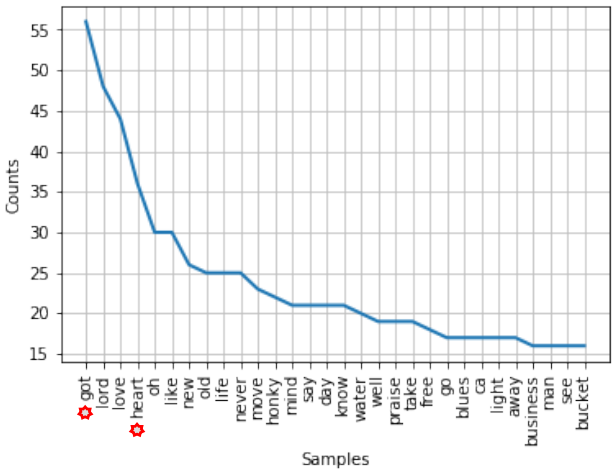


Figure 6 - Hank Williams word frequency after removing stop words and non-alphabetical characters (top 30 above, top 50 right)

```
('got', 56)
('lord', 48)
('love', 44)
('heart', 36)
('oh', 30)
('like', 30)
('new', 26)
('old', 25)
('life', 25)
('never', 25)
('move', 23)
('honky', 22)
('mind', 21)
('say', 21)
('day', 21)
('know', 21)
('water', 20)
('well', 19)
('praise', 19)
('take', 19)
('free', 18)
('go', 17)
('blues', 17)
('ca', 17)
('light', 17)
```

```
('away', 17)
('business', 16)
('man', 16)
('see', 16)
('bucket', 16)
('soul', 16)
('dog', 16)
('let', 15)
('lost', 15)
('yea', 15)
('hole', 15)
('home', 14)
('made', 14)
('long', 14)
('cry', 14)
('cold', 14)
('tonk', 13)
('god', 13)
('gone', 13)
('come', 13)
('used', 13)
('lonesome', 13)
('one', 13)
('kaw-liga', 13)
('mine', 12)
```

Next, the bi-grams of words will be analyzed. Bi-grams are two-word sequences that can be analyzed with the NLTK package to reveal insights into words that occur together in a text.[iii] Figure 7 shows the bi-grams for Holly Williams with the percentage frequency.

```
(('gon', 'na'), 0.0007300792657488527)
(('runs', 'dry'), 0.0006257822277847309)
(('woman', 'like'), 0.0006257822277847309)
(('go', 'let'), 0.0005214851898206091)
(('good', 'man'), 0.0005214851898206091)
(('heart', 'blue'), 0.0005214851898206091)
(('life', 'like'), 0.0005214851898206091)
(('rodeo', 'road'), 0.0005214851898206091)
(('beat', 'til'), 0.00041718815185648727)
(('break', 'free'), 0.00041718815185648727)
(('day', 'without'), 0.00041718815185648727)
(('every', 'night'), 0.00041718815185648727)
(('finally', 'found'), 0.00041718815185648727)
(('gone', 'away'), 0.00041718815185648727)
(('let', 'go'), 0.00041718815185648727)
(('missing', 'home'), 0.00041718815185648727)
(('still', 'trying'), 0.00041718815185648727)
(('wan', 'na'), 0.00041718815185648727)
(('without', 'jesus'), 0.00041718815185648727)
(('alright', 'mama'), 0.00031289111389236547)
(('always', 'remember'), 0.00031289111389236547)
(('believe', 'everything'), 0.00031289111389236547)
(('blood', 'brothers'), 0.00031289111389236547)
(('brand', 'new'), 0.00031289111389236547)
(('even', 'though'), 0.00031289111389236547)
(('ever', 'make'), 0.00031289111389236547)
(('ever', 'slip'), 0.00031289111389236547)
(('every', 'time'), 0.00031289111389236547)
(('every', 'wave'), 0.00031289111389236547)
(('free', 'cause'), 0.00031289111389236547)
(('hands', 'let'), 0.00031289111389236547)
(('hardest', 'part'), 0.00031289111389236547)
(('heart', 'taught'), 0.00031289111389236547)
(('honey', 'ca'), 0.00031289111389236547)
(('like', 'mine'), 0.00031289111389236547)
(('narrow', 'line'), 0.00031289111389236547)
(('never', 'come'), 0.00031289111389236547)
(('never', 'knew'), 0.00031289111389236547)
(('never', 'see'), 0.00031289111389236547)
(('never', 'wore'), 0.00031289111389236547)
(('night', 'mama'), 0.00031289111389236547)
(('old', 'railroad'), 0.00031289111389236547)
(('one', 'day'), 0.00031289111389236547)
(('ones', 'dying'), 0.00031289111389236547)
(('pony', 'free'), 0.00031289111389236547)
(('railroad', 'delivers'), 0.00031289111389236547)
(('refuse', 'jesus'), 0.00031289111389236547)
(('road', 'set'), 0.00031289111389236547)
(('searching', 'anymore'), 0.00031289111389236547)
(('see', 'everythings'), 0.00031289111389236547)
```

Figure 7 - Holly Williams 50 most frequent bi-grams (without stop words)

Figure 8 shows the bi-grams for Hank with the percentage frequency.

```
(('tonk', 'blues'), 0.0014906832298136647)
(('body', 'praise'), 0.0011180124223602484)
(('gon', 'na'), 0.0011180124223602484)
(('honky', 'tonk'), 0.0011180124223602484)
(('honky', 'tonkin'), 0.0011180124223602484)
(('new', 'body'), 0.0011180124223602484)
(('new', 'life'), 0.0011180124223602484)
(('clear', 'water'), 0.0009937888198757764)
(('big', 'fun'), 0.0008695652173913044)
(('blues', 'well'), 0.0008695652173913044)
(('dog', 'cause'), 0.0008695652173913044)
(('precious', 'lord'), 0.0008695652173913044)
(('country', 'church'), 0.0007453416149068323)
(('go', 'honky'), 0.0007453416149068323)
(('oh', 'lord'), 0.0007453416149068323)
(('old', 'country'), 0.0007453416149068323)
(('poor', 'old'), 0.0007453416149068323)
(('fly', 'away'), 0.0006211180124223603)
(('got', 'ta'), 0.0006211180124223603)
(('never', 'let'), 0.0006211180124223603)
(('old', 'kaw-liga'), 0.0006211180124223603)
(('poor', 'soul'), 0.0006211180124223603)
(('wedding', 'bells'), 0.0006211180124223603)
(('bayou', 'jambalaya'), 0.0004968944099378882)
(('beer', 'well'), 0.0004968944099378882)
(('blues', 'yeah'), 0.0004968944099378882)
(('cold', 'heart'), 0.0004968944099378882)
(('crawfish', 'pie'), 0.0004968944099378882)
(('doubtful', 'mind'), 0.0004968944099378882)
(('god', 'dips'), 0.0004968944099378882)
(('hand', 'precious'), 0.0004968944099378882)
(('ho-on-ky', 'tonk'), 0.0004968944099378882)
(('honey', 'baby'), 0.0004968944099378882)
(('lonesome', 'blue'), 0.0004968944099378882)
(('long', 'gone'), 0.0004968944099378882)
(('lost', 'highway'), 0.0004968944099378882)
(('said', 'goodbye'), 0.0004968944099378882)
(('saviour', 'call'), 0.0004968944099378882)
(('sight', 'praise'), 0.0004968944099378882)
(('well', 'lord'), 0.0004968944099378882)
(("'cause", 'tonight'), 0.00037267080745341616)
(('amio', 'pick'), 0.00037267080745341616)
(('bad', 'news'), 0.00037267080745341616)
(('carry', 'bad'), 0.00037267080745341616)
(('cher', 'amio'), 0.00037267080745341616)
(('die', 'like'), 0.00037267080745341616)
(('dying', 'breath'), 0.00037267080745341616)
(('fall', 'praise'), 0.00037267080745341616)
(('filled', 'fruit'), 0.00037267080745341616)
(('filé', 'gumbo'), 0.00037267080745341616)
```

Figure 8 - Hank Williams 50 most frequent bi-grams (without stop words)

Comparison and contrast of the artists bi-grams will be discussed in the Questions to Answer section below.

Pointwise Mutual Information (PMI) is another NLP technique to "measure of how often two events x and y occur, compared with what we would expect if they were independent".[iv] The word frequency variable was set to 5 so that pairs of words must occur at least 5 times to be included. Since the text of the lyrics had only ~8-10K words in the raw form, the PMI results are more limited than with a corpus of more words. However, the PMI results provided excellent insights into the types of phrases used by both artists.

```
(('runs', 'dry'), 10.642051692927978)
(('gon', 'na'), 9.767582575011836)
(('rodeo', 'road'), 9.767582575011836)
(('good', 'man'), 7.389070951758107)
(('woman', 'like'), 6.481059816255673)
(('heart', 'blue'), 6.339488922907547)
(('life', 'like'), 5.440417831758328)
(('go', 'let'), 4.731958665281116)
```

Figure 9 - Holly Williams PMI scores without stop words

```
(('wedding', 'bells'), 10.389810567168187)
(('country', 'church'), 9.974773067889341)
(('gon', 'na'), 9.65284497300198)
(('big', 'fun'), 9.290274893617273)
(('tonk', 'blues'), 8.771833009219067)
(('clear', 'water'), 8.652844973001981)
(('fly', 'away'), 8.62427582080521)
(('body', 'praise'), 8.574842461000706)
(('honky', 'tonkin'), 8.515341449252045)
(('dog', 'cause'), 8.32269637130965)
(('new', 'body'), 8.122330256303199)
(('honky', 'tonk'), 7.984826732553264)
(('old', 'country'), 7.915879378835775)
(('poor', 'soul'), 7.837269544139407)
(('poor', 'old'), 7.4564477601984755)
(('blues', 'well'), 7.446737635253022)
(('got', 'ta'), 7.167418145831739)
(('precious', 'lord'), 7.027240487783477)
(('go', 'honky'), 7.012841108722862)
(('old', 'kaw-liga'), 6.952405254860889)
(('new', 'life'), 6.800402161415835)
(('never', 'let'), 6.745954377393462)
(('oh', 'lord'), 5.067882472280823)
```

Figure 10 - Hank Williams PMI scores without stop words

Questions to Answer

1. Are there differences in the number of words used by Holly and Hank in their songs?

The overall raw word count of 40 songs from each artist was 10,133 for Holly Williams and 8,269 for Hank Williams. The difference in the word counts could be due to stylistic variations between grandfather and granddaughter. However, a more likely explanation is technology limitations for audio recordings during the 1940's and 1950's. Hank Williams music was played on vinyl LP's with song limits of 2-3 minutes for a 10 song album. Holly Williams' music was recorded in the age of CD's and streaming music which had fewer limitations on song length. Figure 11 illustrates the change in song duration from 1944-2008. This supports the theory that shorter songs were more common in Hank Williams' time compared to Holly Williams', who started her recording career in 2003.

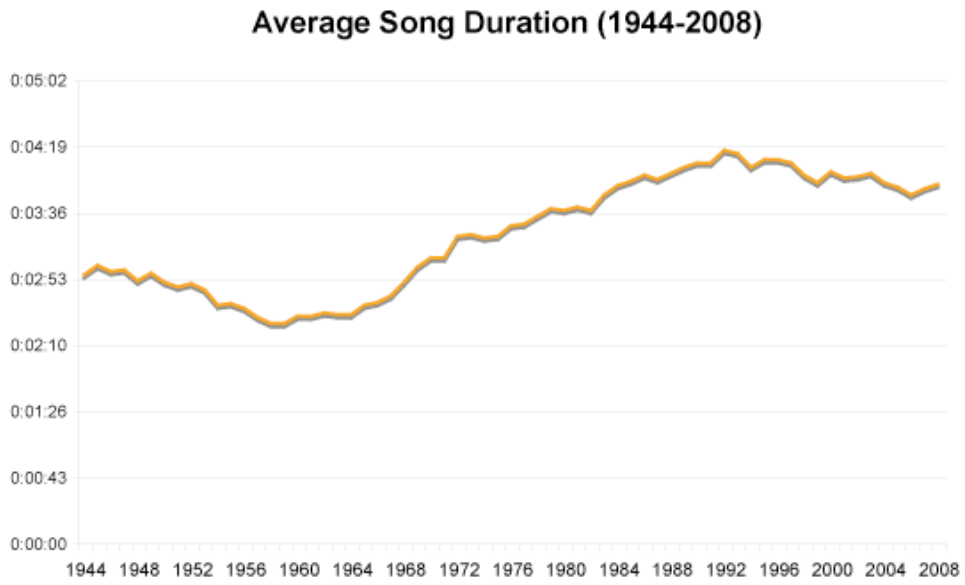


Figure 11 - Average Song Duration (1944-2008) - https://waxy.org/2008/05/the_whitburn_project/

2. Are similar themes or words used in the song lyrics?

Both artists utilize lyrics from the country genre of music which are a mixture of words that are more common in the South and West part of the U.S. The PMI list for Hank shown in Figure 10, uses words such as "country church", "big fun", and "tonk blues". These phrases provide insights into the spiritual/emotional/cultural influences during the 1940's and 1950's in the US

after the Great Depression and World War II. Holly Williams uses the phrases like "runs dry" which is likely a reference to the struggles of the depression era droughts that occurred before she was born (Figure 9). She also has more relational phrases like "good man" and "women like" which touches on the common topic of relationships in the country music genre. The only phrase that is common to both artists is the word "gon'na". This is a contraction of "going to" which may have been borrowed by Holly Williams from her grandfather, but we don't know for certain.

3. What are the differences or similarities in the bigrams scored by frequency and pointwise mutual information (PMI)?

There are similarities between the bi-grams scored by frequency and PMI as shown in Figure 12 and 13. However, for Hank Williams, the top two bi-grams for PMI (wedding bells and country church), do not show up in the bi-grams scored by frequency. It can be argued that Holly Williams tends to use bigrams that are more likely to occur together and also have a high frequency. Whereas her grandfather uses more varied bi-grams as measured by PMI and frequency.

Holly Williams
Frequency
(('gon', 'na'), 0.0007300792657488527)
(('runs', 'dry'), 0.0006257822277847309)
(('woman', 'like'), 0.0006257822277847309)
(('go', 'let'), 0.0005214851898206091)
(('good', 'man'), 0.0005214851898206091)

PMI
(('runs', 'dry'), 10.642051692927978)
(('gon', 'na'), 9.767582575011836)
(('rodeo', 'road'), 9.767582575011836)
(('good', 'man'), 7.389070951758107)

Figure 12 - Holly Williams comparison of bi-gram frequency and PMI

Hank Williams
Frequency
(('**tonk**', '**blues**'), 0.0014906832298136647)
(('body', 'praise'), 0.0011180124223602484)
(('**gon**', '**na**'), 0.0011180124223602484)
(('honky', 'tonk'), 0.0011180124223602484)
(('honky', 'tonkin'), 0.0011180124223602484)
(('new', 'body'), 0.0011180124223602484)
(('new', 'life'), 0.0011180124223602484)
(('clear', 'water'), 0.0009937888198757764)

PMI
(('wedding', 'bells'), 10.389810567168187)
(('country', 'church'), 9.974773067889341)
(('**gon**', '**na**'), 9.65284497300198)
(('big', 'fun'), 9.290274893617273)
(('**tonk**', '**blues**'), 8.771833009219067)
(('**clear**', '**water**'), 8.652844973001981)

Figure 13 - Hank Williams comparison of bi-gram frequency and PMI

Conclusion

Since Hank Williams lived a short life that only lasted 29 years, his grand-daughter Holly Williams never had a chance to meet him. However, this analysis of the song lyrics of both artists shows a common bond through the language and style of country music with some unique differences based on the different time periods in which they lived. Holly Williams probably could have "re-used" more of famous grandfather's lyrics and phrases, but this analysis shows that she created her own distinct style and words for her songs.

This lyric from Holly Williams song "Sometimes" summarizes how she wished that she could be in the back of the blue Cadillac in 1952 when her grandfather died from a heart attack that was caused by drugs and alcohol.[v]

I wish I were an angel in 52'
In a blue Cadillac on the eve of the new year
And there I would have saved him, the man who sang the blues
But maybe he is listening right now

[i] Browne, David, et al. "100 Greatest Country Artists of All Time." *Rolling Stone*, 15 Sept. 2019, www.rollingstone.com/music/music-lists/100-greatest-country-artists-of-all-time-195775/.

[ii] "Song Lyrics & Knowledge." *Genius*, 2019, genius.com/.

[iii] Jurafsky, Dan, and James H. Martin. *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Dorling Kindersley Pvt, Ltd., 2014.

[iv] Jurafsky, Dan, and James H. Martin. *Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Dorling Kindersley Pvt, Ltd., 2014.

[v] "Death of Hank Williams." *Wikipedia*, Wikimedia Foundation, 22 Sept. 2019, en.wikipedia.org/wiki/Death_of_Hank_Williams.