Computational Content Analysis: Memo 2

Xingyun Wu

1/17/2018

1. Summarize Results

In this assignment, I scraped Wikipedia webpages for the following types of music: blues music, house music, gospel music, hip-hop music, country music, and popular music. After extracting the text description of these types of music, I tokenized, normalized and stored them in a Pandas dataframe. In the first 3 exercises, I treated them as a corpus to look for aggregated patterns. But in Exercise 4, I used each of them as an individual corpus, and then analyzed the distances among four types of music: blues music, house music, gospel music, and hip-hop music. Although these corpora are from Wikipedia, they are individual texts that are used to describe different concepts. So it makes sense to treat them as different corpora in Exercise 4.

1.1 Exercise 1

In this part, I downloaded the corpus of House Music and explored all the required features. Compared to the example, I found very similar distribution of raw counts, as well as log-log counts.

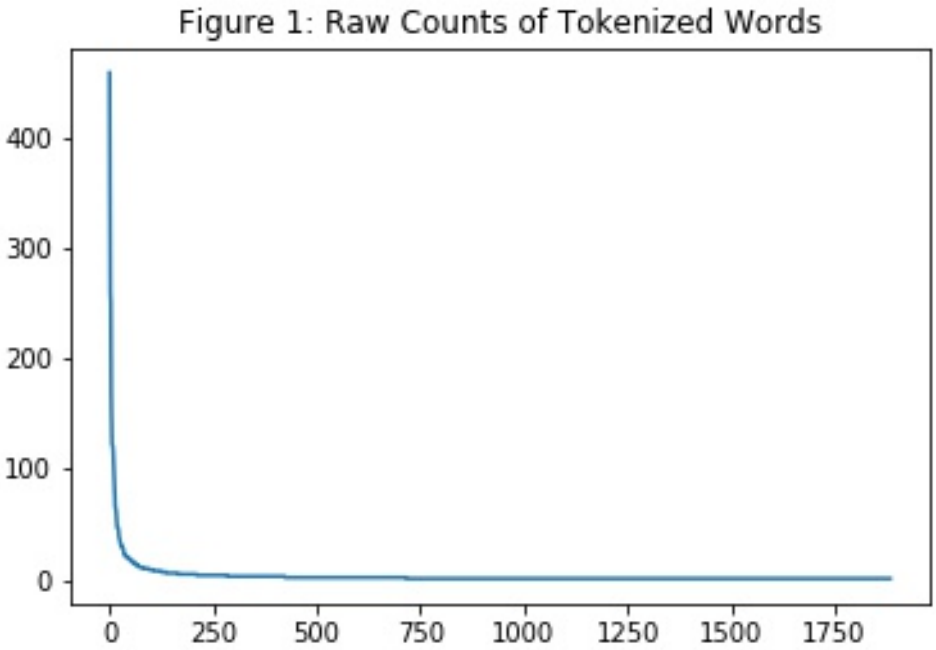
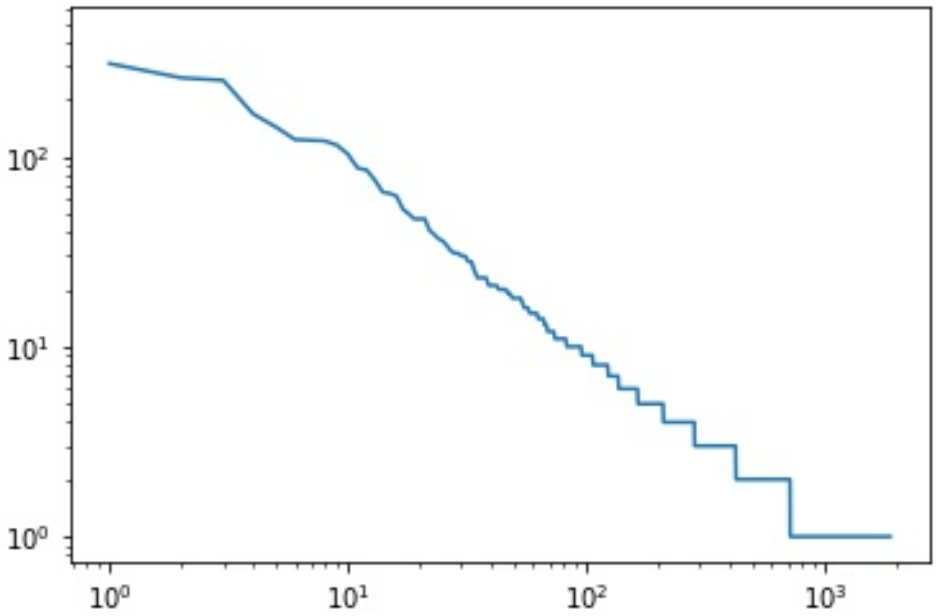


Figure 2: Log-log Counts of Tokenized Words



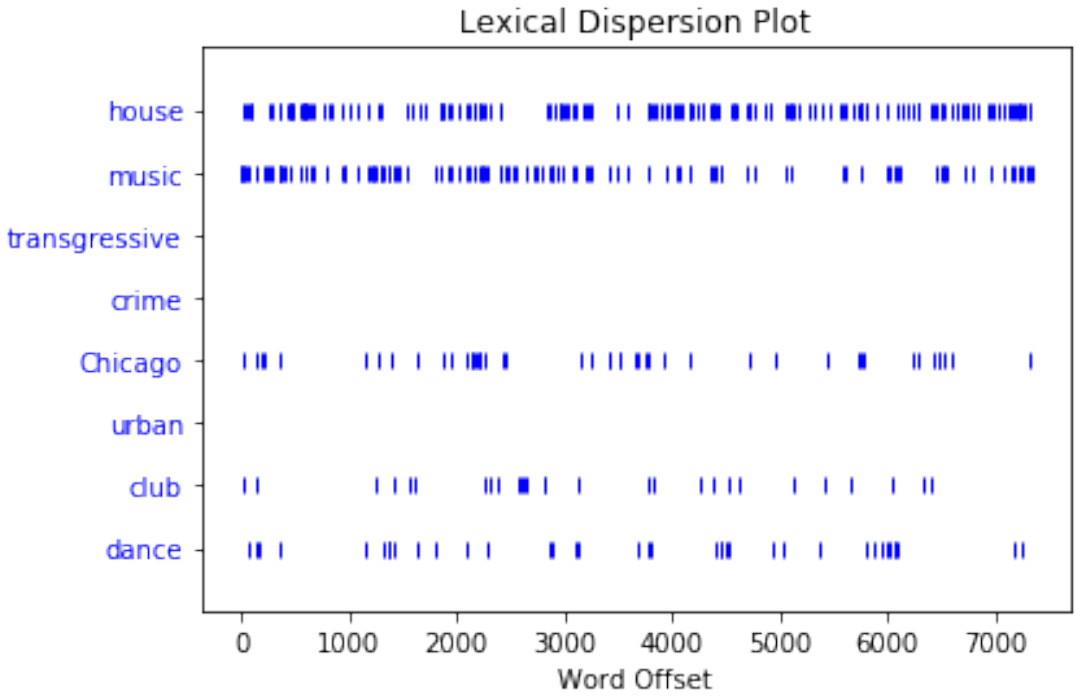

And I got collocations, which are basically phrases, people's names, and cities' names. As is shown in the following words, the machine could identify common collocations very nicely.

```
house music; New York; Frankie Knuckles; citation needed; Ron Hardy;
Marshall Jefferson; drum machine; Derrick May; acid house; Kevin
Saunderson; Magic Orchestra; Yellow Magic; dance music; drum machines;
Juan Atkins; Music Box; Calvin Harris; Dada Nada; Daft Punk; Los
Angeles
```

I also got some common contexts of the word "house":

```
*START*_music early_music while_displayed ,_was of_was many_songs
._music ,_music the_'' chicago_music deep_to euro_, tech_, electro_and
jump_. acid_, progressive_( ghetto_, deep_, future_and
```

My lexical dispersion plot shows that some of my input words do not show up in the Wikipedia description of the house music. These words are *transgressive*, *crime*, and *urban*. Although these words should, by my qualitative observation, have associated with the House music, they still do not show up in the description. The words, *club* and *dance*, seems to occur in nearby places, which indicates that these two words are associated in the text. And the word, *Chicago*, appears very frequently. This is consistent with the fact that the house music originates in Chicago, which is a reasonable explanation for the frequent occurrence of this word.




At the end of this exercise, I also downloaded texts for blues music, gospel music, hip-pop music, country music, and popular music. I tokenized these texts, and stored all the available information in a Pandas dataframe. The types and corresponding word counts of texts are listed as below.

| Music Type | Word Counts |
|---|---|
| Blues Music | 9513 |
| House Music | 7366 |
| Gospel Music | 2537 |
| Hip-hop Music | 14782 |
| Country Music | 13191 |
| Popular Music | 3634 |

1.2 Exercise 2

In this exercise, I treat the texts of the 6 types of music as a corpus. I normalized the tokens with built-in stop words and Porter Stemmer. By now, the dataframe look like this:

| | music_type | text | tokenized_text | word_counts | normalized_tokens | normalized_tokens_count |
|---|---|---|---|---|---|---|
| 0 | blues_music | Blues is a music genre and musical form origin... | [Blues, is, a, music, genre, and, musical, for... | 9513 | [blue, music, genr, music, form, origin, afric... | 4990 |
| 1 | house_music | House music is a genre of electronic music cre... | [House, music, is, a, genre, of, electronic, m... | 7366 | [hous, music, genr, electron, music, creat, cl... | 3832 |
| 2 | gospel_music | Gospel music is a genre of Christian music. Th... | [Gospel, music, is, a, genre, of, Christian, m... | 2537 | [gospel, music, genr, christian, music, creati... | 1323 |
| 3 | hip_hop_music | Hip hop music, also called hip-hop or rap musi... | [Hip, hop, music, ,, also, called, hip-hop, or... | 14782 | [hip, hop, music, also, call, rap, music, musi... | 7717 |
| 4 | country_music | Country (or country and western) is a musical ... | [Country, (, or, country, and, western, ), is,... | 13191 | [countri, countri, western, music, genr, origi... | 6882 |
| 5 | popular_music | Popular music is music with wide appeal that i... | [Popular, music, is, music, with, wide, appeal... | 3634 | [popular, music, music, wide, appeal, typic, d... | 1942 |

The frequency distribution of the whole corpus and the frequency distribution of the house music corpus look very similar.

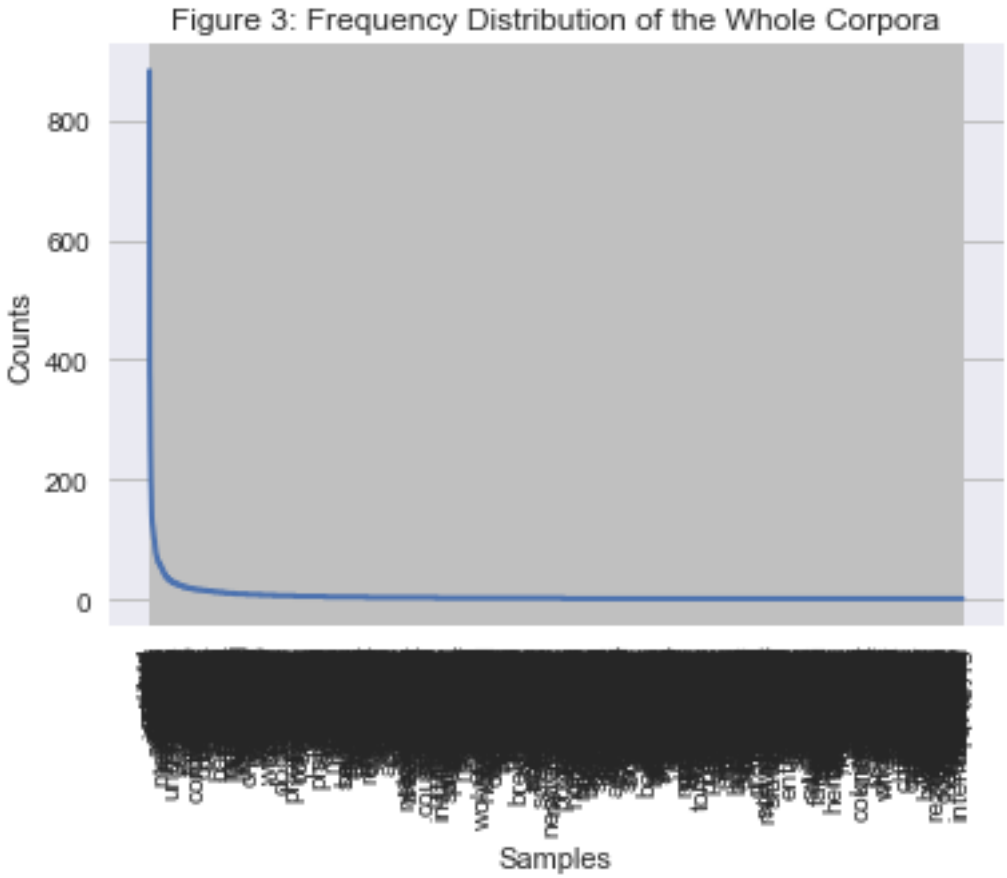
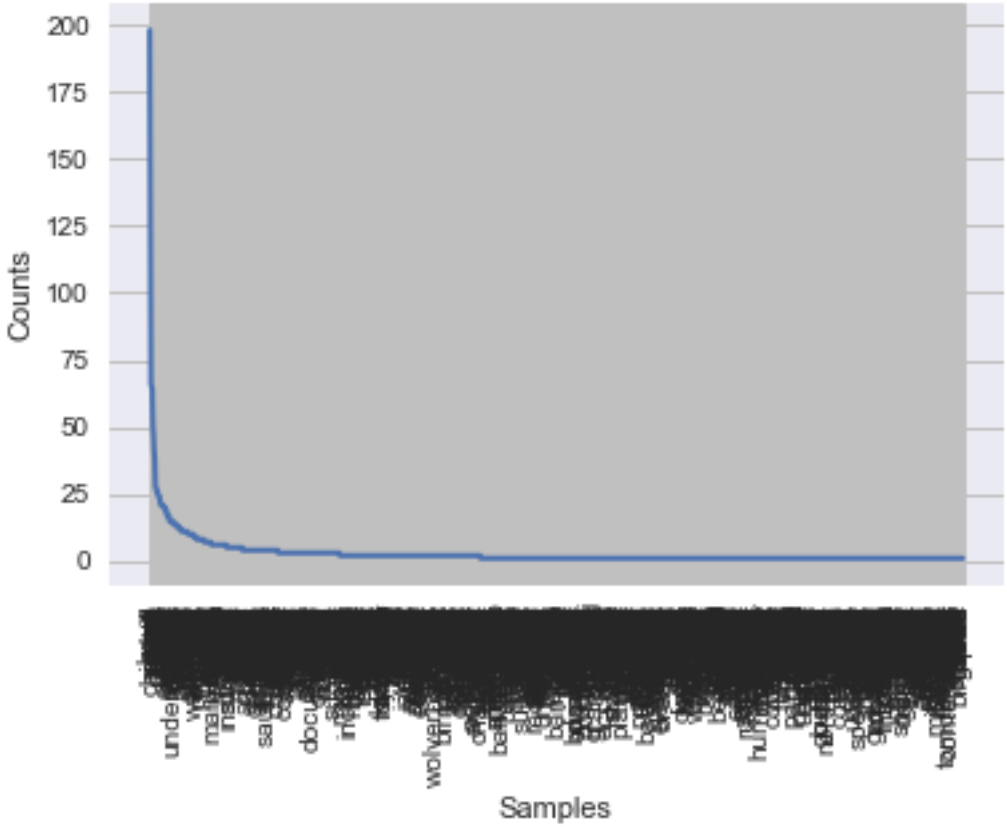
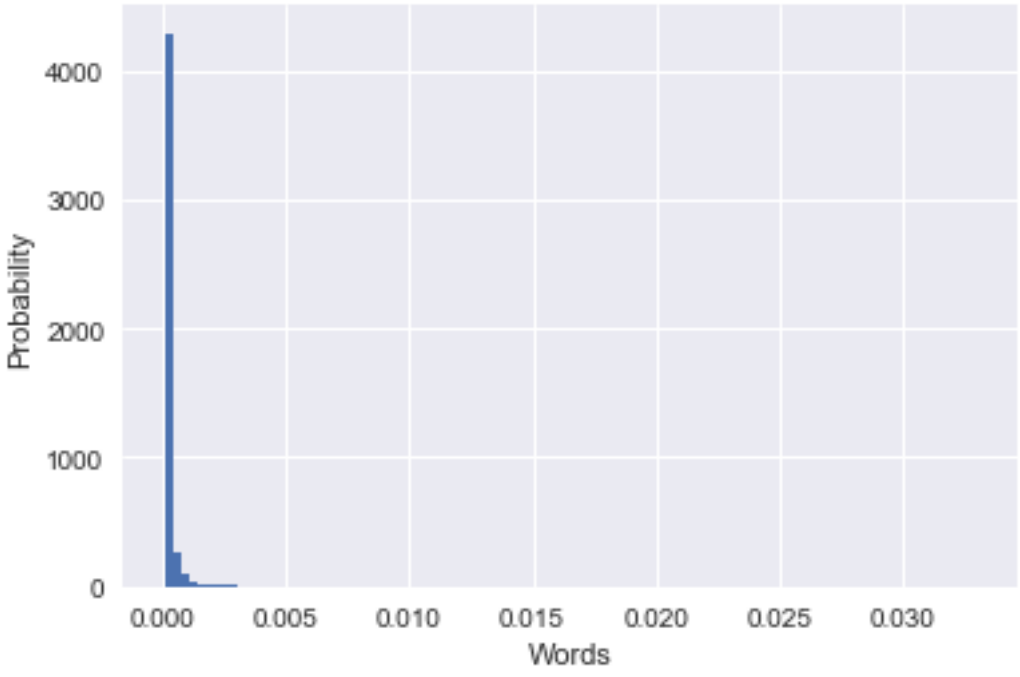


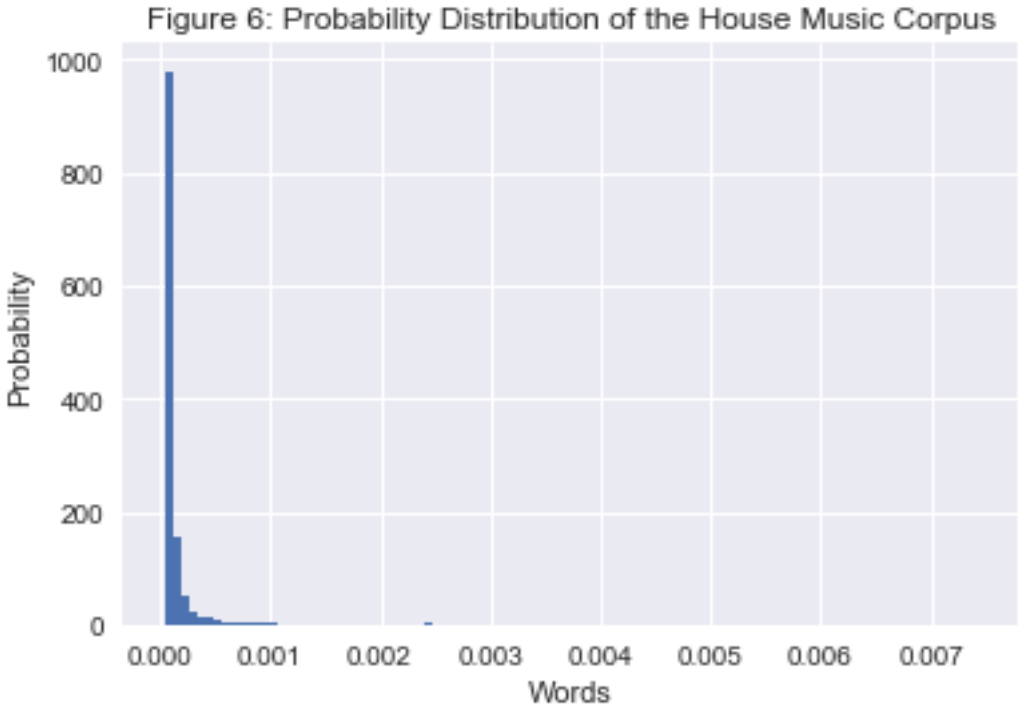
Figure 4: Frequency Distribution of the House Music Corpus




Their probability distributions also look similar.

Figure 5: Probability Distribution of the Whole Corpora






Then I plot two conditional frequency distributions, which indicates three interesting points with social meaning. One is of the words whose length are 11, and the other one is of the words whose length are 12. They follow similar distribution of the previous ones, indicating that this distribution is very stable, no matter what are the conditions. They also suggest that the most frequently used words are words that are common. The special words directly associated with the described concepts are less frequently used. In addition, the number of words with length 12 is smaller than the number of words with length 11.

I also generate word clouds for all the texts as whole, and for each type of music. They indicate that different types of music have different frequently used key words. Some types of music, the blues music, house music, and hip-hop music, are having very classic words in description. And the other types of music do not have such strong key words. The former ones might have more stable styles and features, compared to the latter ones, and this pattern influences their description in texts.

Figure: Word Cloud for all_type music


Figure: Word Cloud for blues music


Figure: Word Cloud for house music


Figure: Word Cloud for gospel music

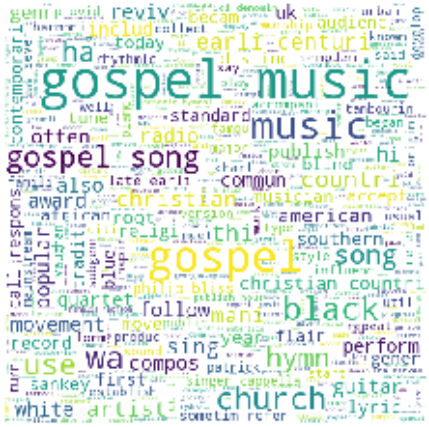



Figure: Word Cloud for hip_pop music


Figure: Word Cloud for country music

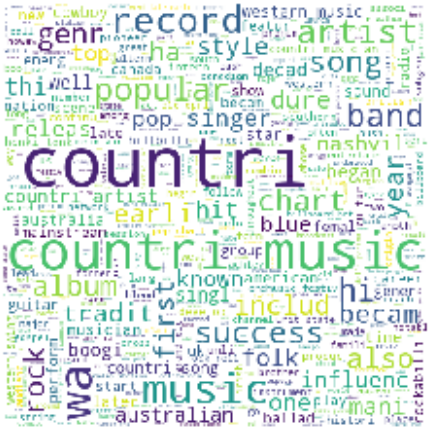


Figure: Word Cloud for popular music


1.3 Exercise 3

In this part, I check the bigrams, trigrams, quadgrams, 5-grams, and 3 types of skipgrams. I use both student-t and likelihood ratio to look for statistically significant grams. And I find different methods could provide different n-grams. For example, the top 20 student-t based trigrams are different from the top 20 likelihood-ratio-based trigrams. We need to be careful when we select the methods.

In [108]:
```
# Check statistically significant trigrams with student_t: the first 20
trigram_measures = nltk.collocations.TrigramAssocMeasures()
musicTrigrams = nltk.collocations.TrigramCollocationFinder.from_words(music_df['normalized_toker
musicTrigrams.score_ngrams(trigram_measures.student_t)[:20]
```

Out[108]:
```
[(('hip', 'hop', 'music'), 6.916077641931419),
 (('new', 'york', 'citi'), 4.898885259000975),
 (('world', 'war', 'ii'), 3.316621218669452),
 (('coast', 'hip', 'hop'), 3.3157632417855947),
 (('hip', 'hop', 'wa'), 3.306228770946415),
 (('altern', 'hip', 'hop'), 3.1615849003499714),
 (('hip', 'hop', 'cultur'), 3.1604102206578895),
 (('dj', 'kool', 'herc'), 2.8284131989219774),
 (('first', 'hip', 'hop'), 2.824251403056866),
 (('hip', 'hop', 'artist'), 2.82196149116272),
 (('countri', 'western', 'music'), 2.8188506556547295),
 (('grand', 'ole', 'opri'), 2.6457510138486873),
 (('east', 'coast', 'hip'), 2.6456660472523565),
 (('hip', 'hop', 'movement'), 2.6445273016711917),
 (('hip', 'hop', 'ha'), 2.6419352817792885),
 (('countri', 'music', 'festiv'), 2.641201232802333),
 (('hip', 'hop', 'record'), 2.635635233430913),
 (('hop', 'music', 'wa'), 2.6006755769052003),
 (('john', 'lee', 'hooker'), 2.4494880573763784),
 (('late', 'earli', 'centuri'), 2.4492678308878273)]
```

```
In [117]: musicTrigrams.score_ngrams(trigram_measures.student_t)[-20:]
```

```
Out[117]: [(('record', 'music', 'record'), 0.9022055380605263),
           (('popular', 'music', 'wa'), 0.9000195573191074),
           (('countri', 'music', 'song'), 0.8996802039348872),
           (('blue', 'popular', 'music'), 0.8956005322282393),
           (('music', 'gospel', 'music'), 0.8896705176163904),
           (('countri', 'music', 'popular'), 0.8881434273873993),
           (('rap', 'music', 'music'), 0.8656383531367923),
           (('countri', 'music', 'record'), 0.8590506865285166),
           (('record', 'countri', 'music'), 0.8590506865285166),
           (('music', 'music', 'genr'), 0.8448833019953211),
           (('music', 'blue', 'wa'), 0.8305264245140028),
           (('music', 'wa', 'blue'), 0.8305264245140028),
           (('blue', 'music', 'blue'), 0.823035879741141),
           (('countri', 'music', 'blue'), 0.8103955854369369),
           (('music', 'blue', 'countri'), 0.8103955854369369),
           (('style', 'music', 'music'), 0.80228082859967),
           (('hous', 'music', 'music'), 0.7695096952183997),
           (('music', 'popular', 'music'), 0.7564012418658916),
           (('popular', 'music', 'music'), 0.7564012418658916),
           (('music', 'countri', 'music'), 0.5575896993528526)]
```

```
In [109]: # Check statistically significant trigrams with likelihood ratio: the first 20
          musicTrigrams.score_ngrams(trigram_measures.likelihood_ratio)[:20]
```

## 1.4 Exercise 4

In this part, I use four separated corpora: blues music corpus, house music corpus, gospel music corpus, and hip-hop music corpus. I compute the distances with KL, Chi-square, KS, and Wasserstein distance. After creating heatmaps with every required method, I draw multidimensional scaling for each of these methods. I also construct the Jensen-Shannon Divergence, compute corresponding matrix, and draw heatmaps and multidimensional scaling for it.

From these plots, based on the tests, I find that the distances among these four types of music are unstable. With different methods, I could classify these four types of music differently. So it is hard to say explicitly which types of music are similar. And based on the matrices, heatmaps, and multidimensional scaling alone, it is very hard to decide which method I should apply.

Figure: Heatmap with KL

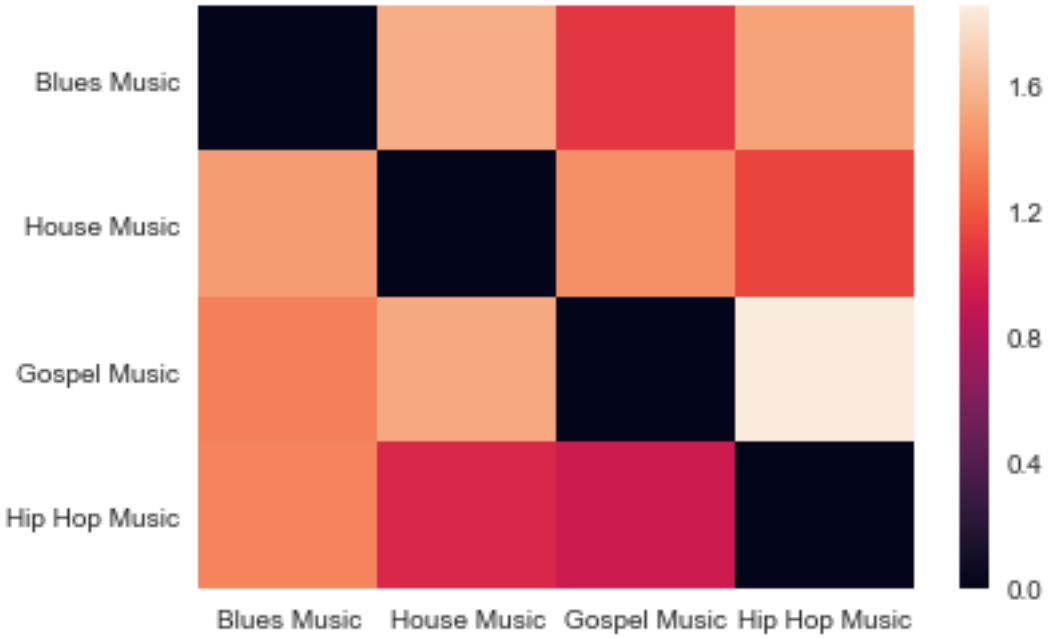



Figure: Multidimensional Scaling with KL

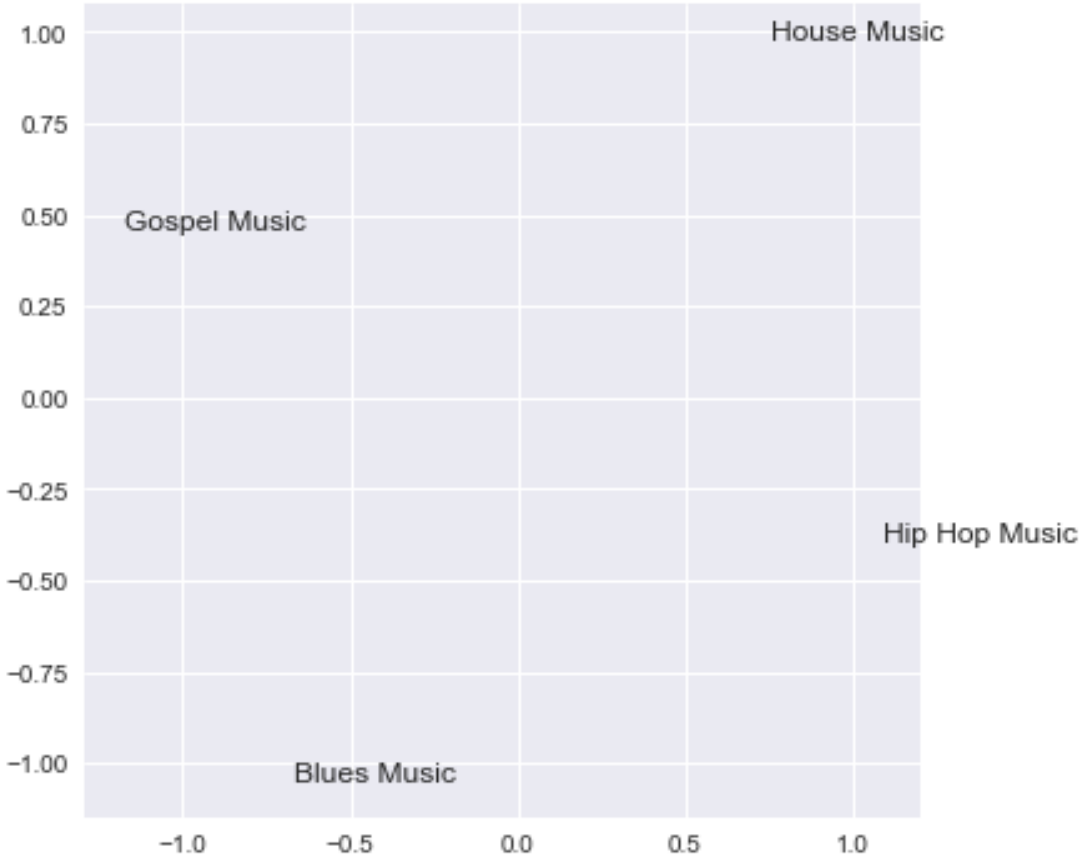


Figure: Heatmap with Chi-square

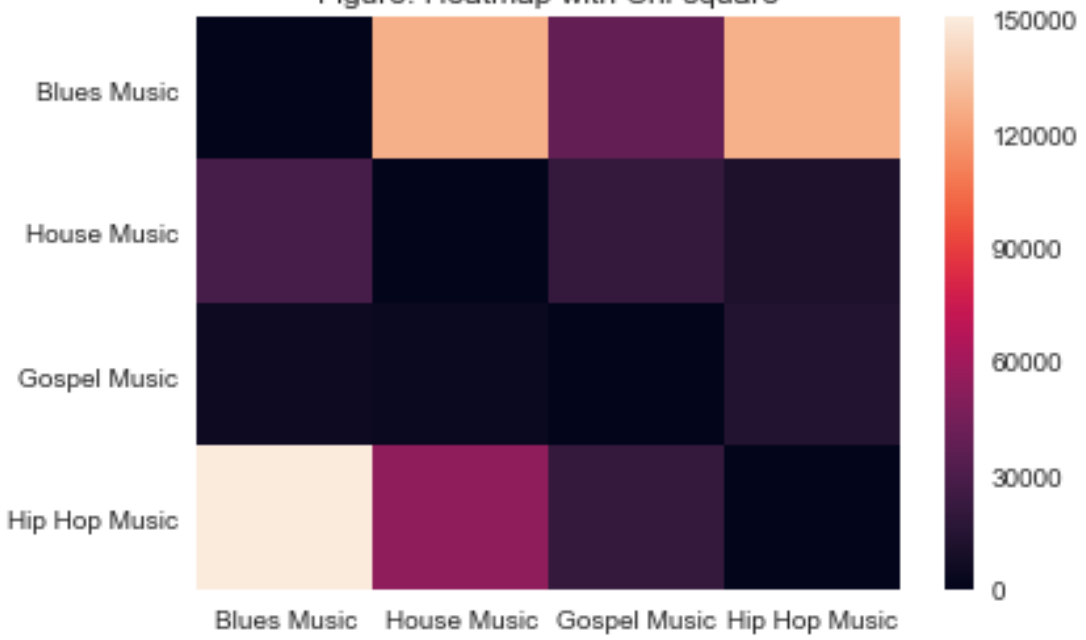



Figure: Multidimensional Scaling with Chi-square

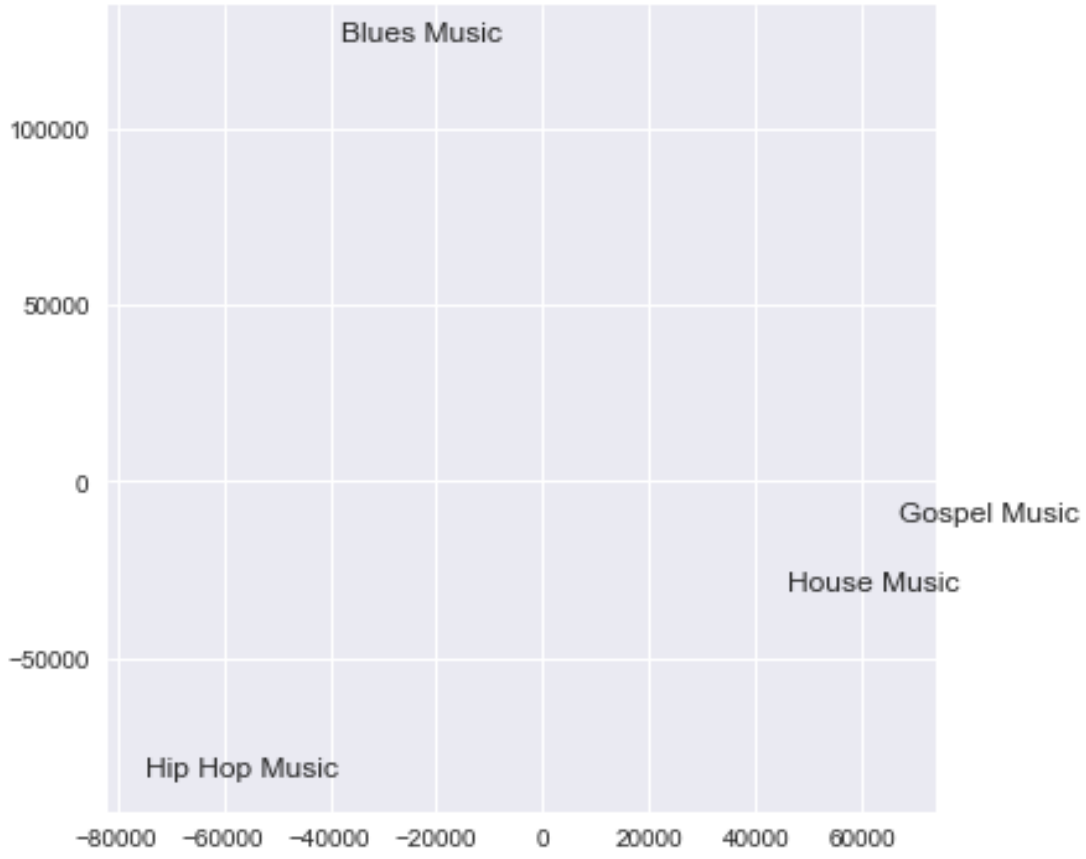


Figure: Heatmap with KS


Figure: Multidimensional Scaling with KS

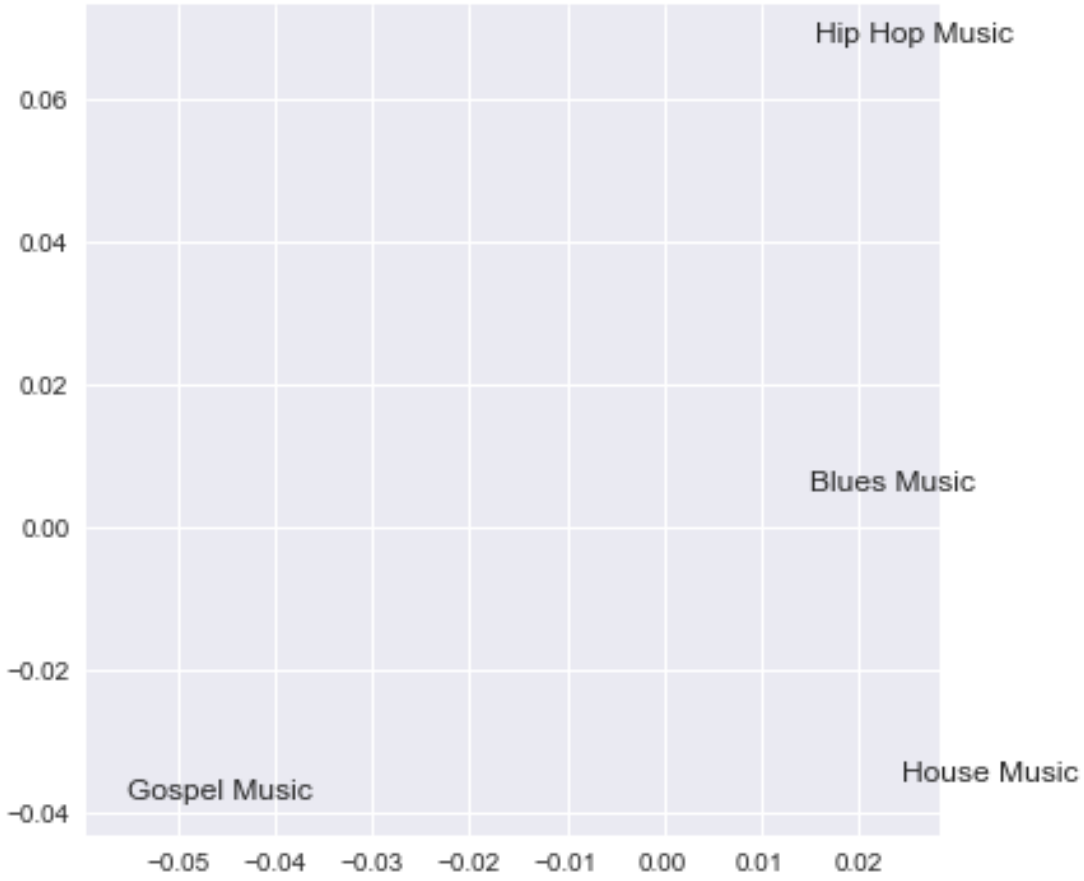


Figure: Heatmap with Wasserstein Distance

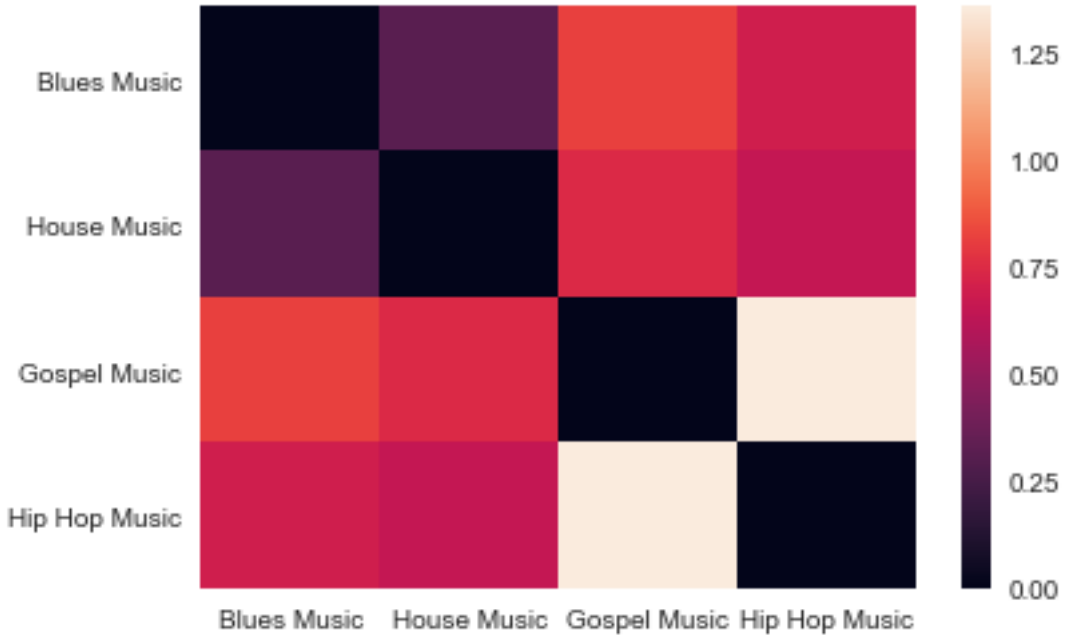



Figure: Multidimensional Scaling with Wasserstein Distance

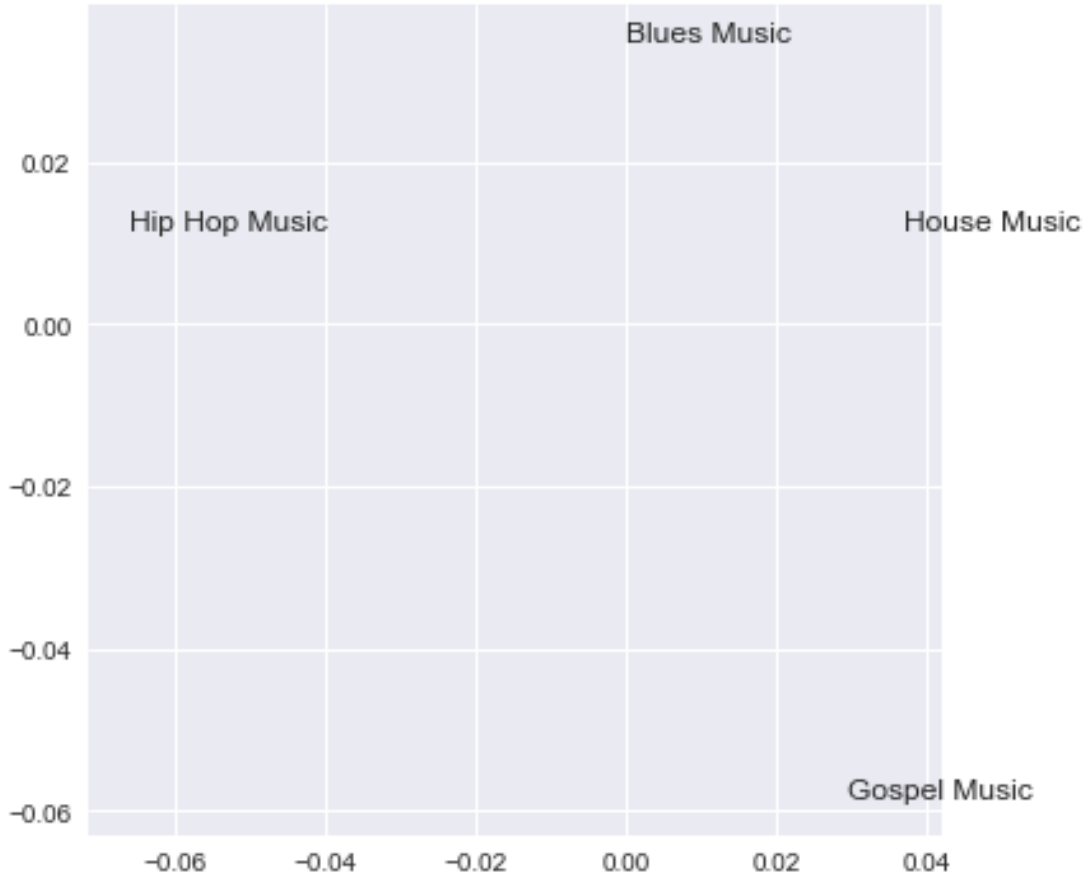


Figure: Heatmap with Jensen-Shannon Divergence

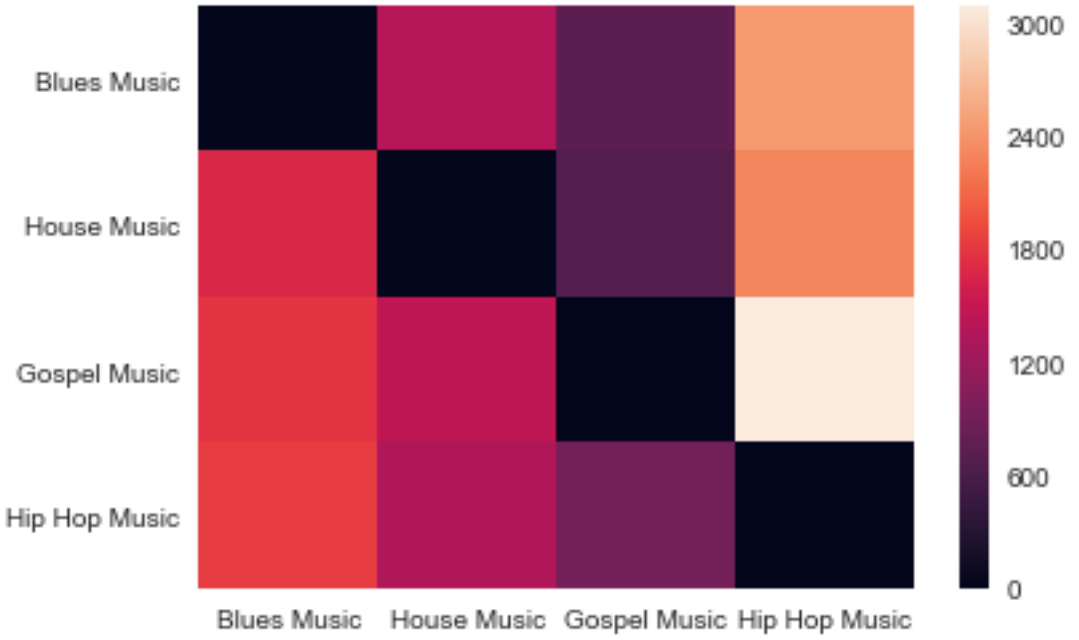
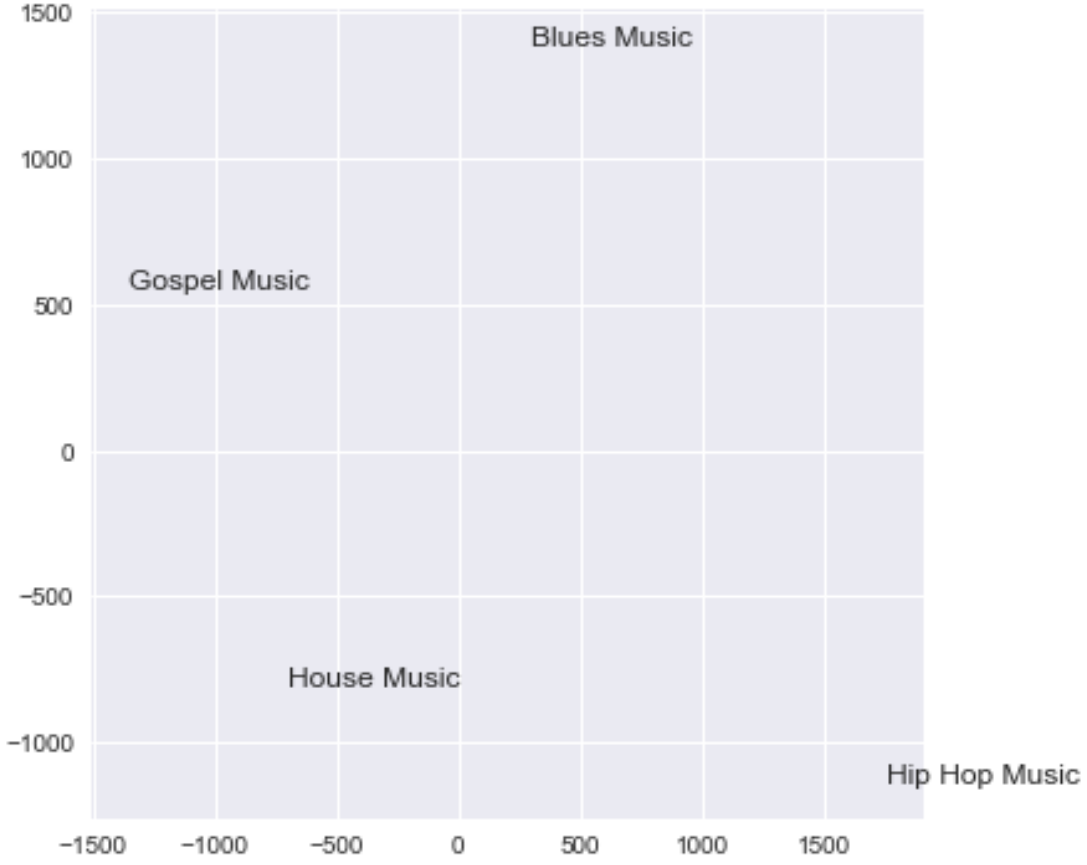



Figure: Multidimensional Scaling with Jensen–Shannon Divergence




2. Identification and Interpretation for of Textual Examples that Facilitate Qualitative Validation of Patterns Summarized

   This part is done as above.

3. Methods' Drawback(s) and Scope Conditions for its Beneficial Deployment

The drawback of the methods is obvious: their results could be very unstable. So it would be hard to find robust results. The trickiest part is the distributional distances. I tried to classify the selected four types of music into fewer categories, by looking at their distances. However, with KL, gospel music is very similar to hip-hop music, which does not make sense. And with Chi-square, blues music is very similar to house music and hip-hop music, but gospel music is not similar to hip-hop music. Maybe the sample size in this case is too small (only 4), but the robustness of results needs to be checked even with larger sample size.

But the benefit is that we could apply different methods to find the optimal one, which best fit our projects and their interpretation. And the word clouds are very useful to look for some qualitative explanation, as a good tool for data visualization.