

Information Retrieval and Text Mining

Emotion Mining and Topic Extraction on Lyrics

Christoph Emunds (i6146758)
Richard Polzin (???)

May 29, 2017

Contents

1	Introduction	1
2	Dataset	1
3	Preprocessing	2
4	Visualization	2
5	Emotion Mining	2
5.1	Results	3
6	Topic Extraction	6
6.1	Results	6
6.2	Evaluation	7
7	Other techniques	8
8	Conclusion	8
	References	9

1 Introduction

Leaving melody aside, there is a lot of information that can be extracted from a song's lyrics. In this work, we aim to measure the similarity of artists with respect to emotions and topics of their songs based on their lyrics.

In Section 2, we give an overview over our dataset and explain how it was gathered. Section 3 describes the preprocessing that has been done. We mention the tools we used for the visualization in Section 4, followed by the actual emotion mining in Section 5 and the topic extraction in Section 6 and the visualization of their results. All of the processing is done in Python 3.6.

2 Dataset

The *pylyrics3* package is able to download the lyrics for all songs of a given artist if it exists on *lyrics.wikia.com*. The lyrics are returned in a dictionary, where the keys are the song names and the values are the songs' lyrics. The usage of this package is straightforward and removes the

necessity to deal with the raw HTML data. The retrieved songs are stored in JSON format, each song having a title, the artist's name and the actual lyrics.

Some (how many?) of the top 100 Billboard artists could not be resolved when querying pylyrics3. This could be due to spelling variations. Moreover, we scraped the names of 30 metal bands from got-djent.com and requested their lyrics via pylyrics3. Some (how many?) of them could also not be resolved, either due to spelling variations or because they are not famous enough to have an entry in die LyricsWiki. Altogether, we ended up with 11.??? songs from 120 artists.

In addition to the dataset we gathered ourselves we were given a database of more than 16.000 artists and 228.000 originally built by Sibirtsev and Schuster (???). In the remainder of this document, we call this dataset ???

Some of the lyrics are not very clean. For example, they contain markers for the chorus and the verse part...

3 Preprocessing

For tokenization and stemming, the *nltk* Python package has been used. It provides easy to use function for splitting a document into sentences and splitting the sentences in words. Furthermore it implements several stemming algorithms. In our implementations, we use the Porter algorithm. We downloaded a list of stopwords that are more appropriate for lyrics. Moreover, in the process of topic extraction, we iteratively added words to this stopwords list whenever we found...

The tf-idf features are built with the help of the *sklearn* package. It offers the *TfidfVectorizer* class, which...

For the topic extraction, all songs should be in the same language, since otherwise the used algorithm will create an extra topic for every language.

In our dataset, all songs, except for one, are in English. For the ??? dataset, we implemented a rudimentary language detection using the *langdetect* Python package (Danilak, 2014). This package is a port of Google's language detection library.

4 Visualization

To visualize our results, we use the open-source platform *Gephi*, which is able to display a plethora of graphs and networks that can also be explored interactively. Gephi provides a rich set of features. Some of the features we make use of include filtering out edges with low weight, calculating modularity classes for the nodes in a graph, ...

To be able to import our results to Gephi, we made use of the *pygraphml* Python package. It provides a couple of methods to dump the extracted information into a *GraphML* file, that can be imported by Gephi. However, the *pygraphml* package is very rudimentary. An issue was that the weight attribute that has been added to edges is stored as a string, rather than a number, which makes it impossible for Gephi to recognize the edge weight properly. We circumvented this problem by simply exchanging a single line in the beginning of the GraphML file.

5 Emotion Mining

The emotion mining is done by first tokenizing every song's text. Afterwards, every token is looked up in an emotion lexicon for its emotional value. For our experiments, we used the *NRC Emotion Lexicon* (EmoLex) by... This lexicon specifies the emotional value for the eight emotions *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*.

The emotion mining we apply to the lyrics is only rudimentary, as it does not handle negations. We were not able to find an emotion lexicon that specifies the opposite emotional value for every common emotion word.

After we acquired the eight dimensional emotion vector for every song, the vectors for all songs of an artist have been summed up and normalized. The resulting vectors contain information about how strong each of the eight emotions is for a certain artist. Furthermore, when calculating the distance from one artist's emotion vector to the emotion vector of every other artist, we are able to see the emotional similarity between the artists. For our experiments, we used the cosine distance as it...

5.1 Results

Figure 1 shows a graph of all 120 artists, where each artist is connected to every emotion. The edges themselves have weights depending on the strength of the emotion in an artist's song. The weights are normalized between 0 and 1.

The full graph is of course quite confusing and does not yield any useful insight. Therefore, Figure ?? shows a graph that contains only those edges with weights higher than 0.3.

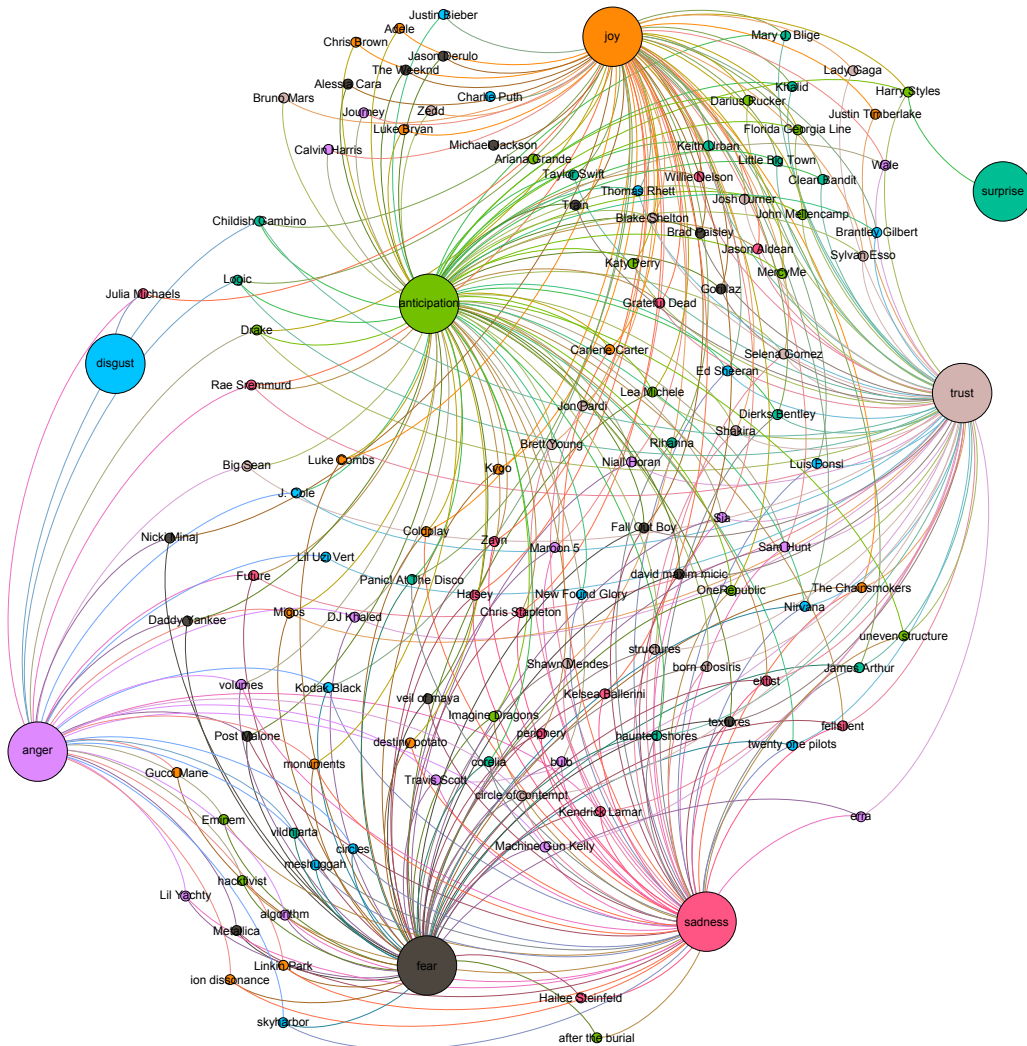


Figure 1: Emotion graph with edge weight greater than 0.37

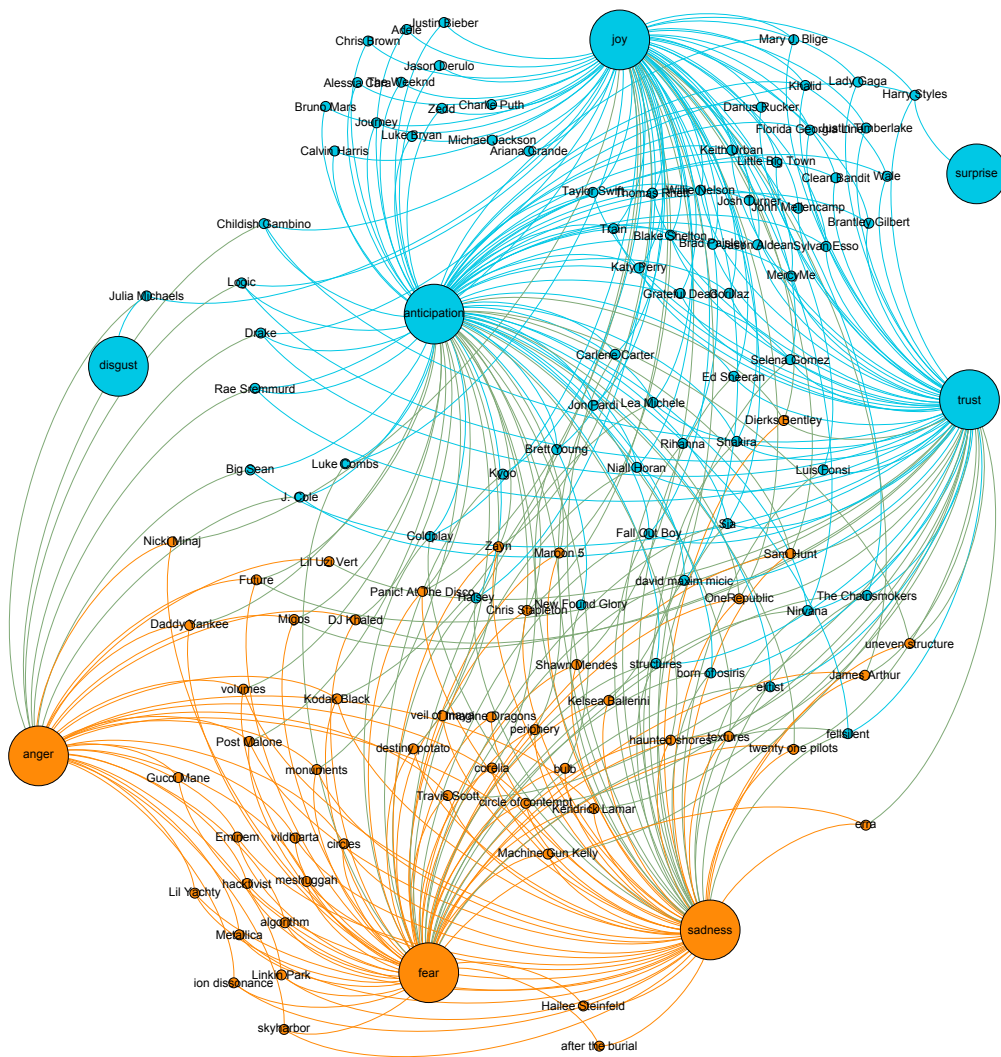


Figure 2: Emotion graph with recalculated modularity

Figure 3 shows the emotion similarity between individual artists. The modularity score separates the graph into three main regions, which...

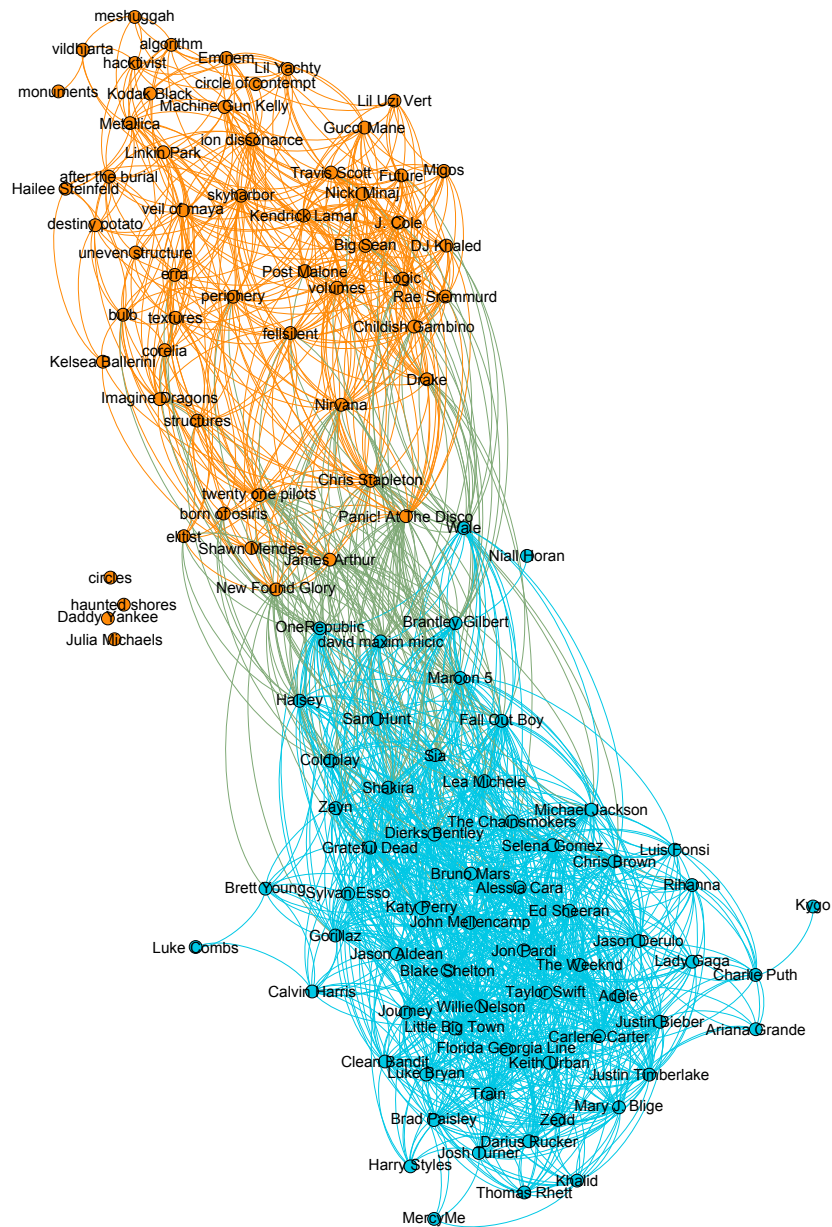


Figure 3: Emotion similarity between artists greater than 0.68

6 Topic Extraction

The Non-negative Matrix Factorization (NMF) algorithm has been used to extract potential topics from the lyrics. The *sklearn* Python package provides everything necessary to apply NMF to the data. First, a tf-idf feature vector for each song text is calculated. Afterwards, these are fed into the NMF algorithm, which returns two matrices that...

We found that a number of ten topics is appropriate for our dataset. When choosing more topics, they start to overlap too much. When choosing fewer topics, too much information is put into every single topic. This is of course subject to the experimentalist's taste.

6.1 Results

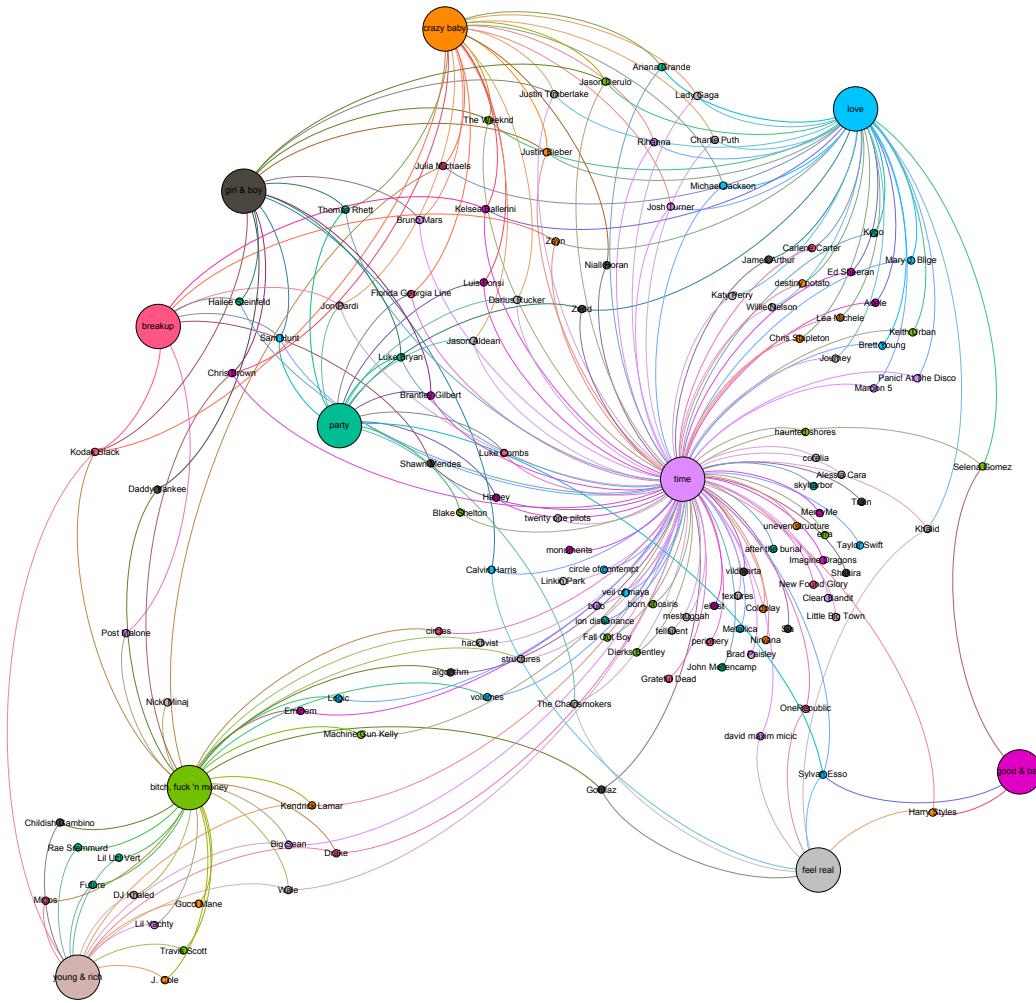


Figure 4: Topic graph with edge weights greater than 0.32

Figure 6 shows the similarity of topics between artists. It is interesting to see that a lot of the rappers form their own subgraph, that is completely isolated from the rest of the graph.

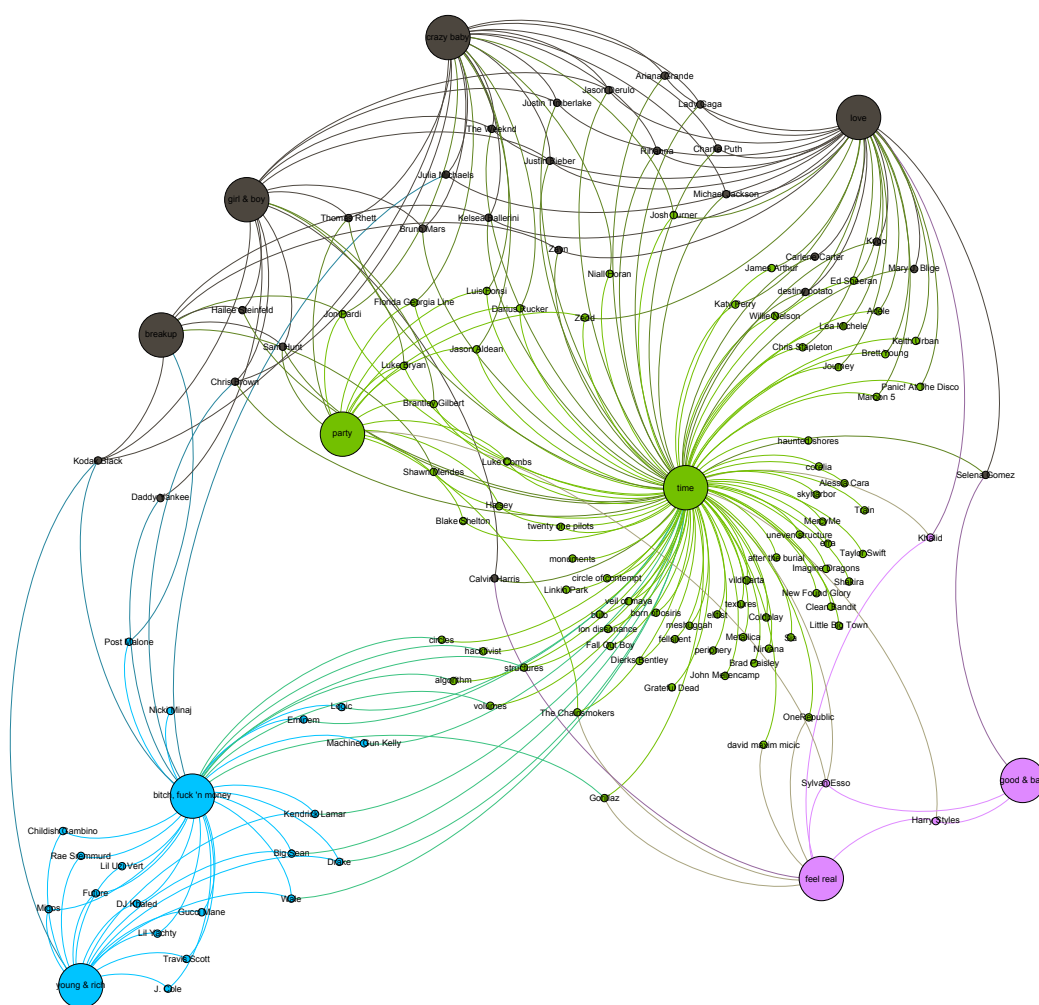


Figure 5: Topic graph with recalculated modularity

6.2 Evaluation

We tried to evaluate the resulting topics via held-out tests.

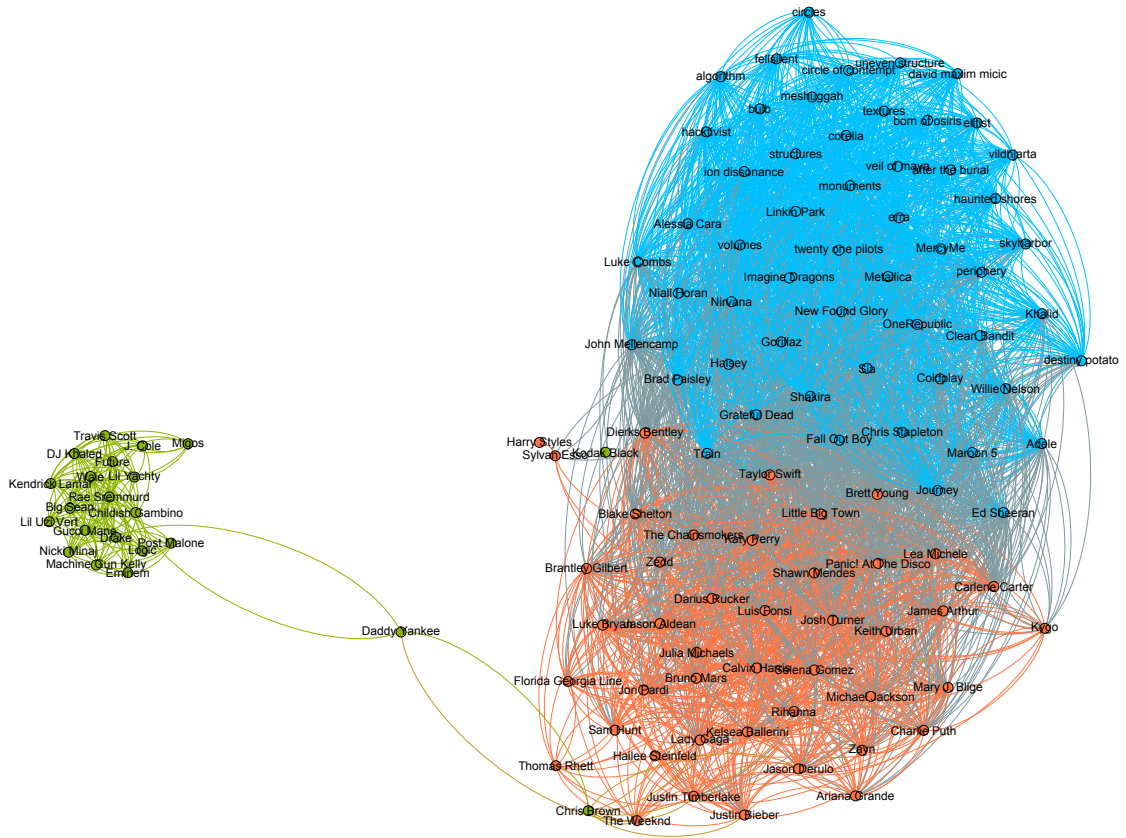


Figure 6: Topic similarity between artists greater than 0.82

7 Other techniques

In this section we want to mention some of the techniques we used that were neglected as we settled for the approach we took in the end. POS Tagging Named Entity Extraction RAKE Summarization Text statistics (Flesch Score, etc.)

8 Conclusion

References

Danilak, M. (2014). *Language detection library*.