Project report

Date: 31.05.2020

Name of students:

Tsz Kin Chau	tszkin.chau@student.kuleuven.be
Fengan Li	fengan.li@student.kuleuven.be
Hugo Enrique Montaño Castillo	hugoenrique.montanocastillo@student.kuleuven.be
Lars Magne Tungland	larsmagne.tungland@student.kuleuven.be
Martina Verna	martina.verna@student.kuleuven.be
Xin Wang	xin.wang@student.kuleuven.be

1. Motivation.

The aim of our project is to visualize songs. We show interesting patterns and encourage the user to explore the data. Music represents a broad cultural memory and reveals how music is produced and consumed.

One of the main challenges with our dataset was to avoid trivial connections. On the one hand, we had to tackle words that are used in almost all songs (including words like *love*, *baby*) and all kinds of stop words (pronouns, for instance). And on the other hand, we also had words that appear in too few songs, such as symbols, names, onomatopoeias, etc. In the end, we managed to present words with a relatively good balance between popularity and uniqueness.

The complexity of this problem arose from creating an interactive retrieval of the co-occurrences. We allow the user to enter a word of its interest and discover some relevant connections.

2. Data.

Title dataset 1: Million song dataset (MSD)

Source: http://millionsongdataset.com/

Data attributes: 46 fields: song name, artist, duration, hotness

License: CCA - Proceedings of the 12th International Conference on Music Information Retrieval {ISMIR} 2011

Title dataset 2: MusixMatch

Source: http://millionsongdataset.com/musixmatch/

Data attributes: ~20 fields: word stem frequency of lyrics

License: CCA

Title dataset 3: TagTraum

Source: http://www.tagtraum.com/msd_genre_datasets.html

Data attributes: compilation of tabs from last.fm, musiXmatch, MSD

License: Non-commercial only - Hendrik Schreiber.

We worked with a subset of these data sets: a lot of data processing was required. We sliced the data using certain constraints: 1) using songs with lyrics with genre (given that MSD has also instrumental songs); 2) only English lyrics; 3) songs from 1980 to 2010.

3. Inspiration.

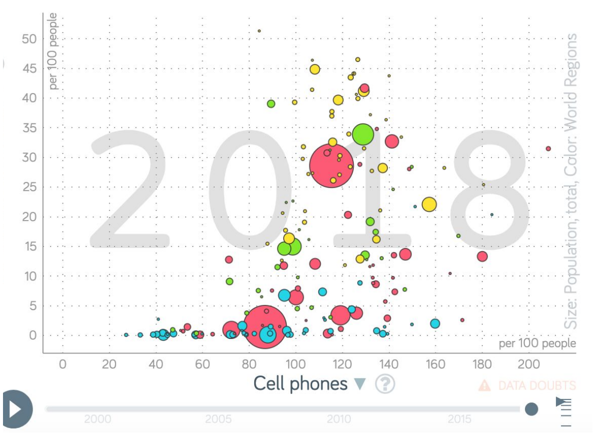
Example 1: Charles Minard's Map of Napoleon's Russian Campaign
Description: Famous visualization showing the route, temperature and losses of Napoleon's attack on Russia in 1812.
Source: Charles-Joseph Minard's map of Napoleon's flawed Russian campaign: An ever-current classic
Inspiration aspect: Minard's map is an absolute classic, and in many ways the benchmark of a good visualization. The way it condenses a lot of information and tells a clear narrative was a clear inspiration to our group.

Title example 2: Gapminder bubble chart
Short description: Iconic bubble chart showing the relation between two metrics of your choice against the population of a country. The animation shows the changes over time.
Source: https://www.gapminder.org/tools/
Inspiration aspect + argumentation: By limiting to only four dimensions: x-position, y-position, size and colour, this refined viz finds a balance between cognitive cost and level of details. Besides, it is also pleasing to the eyes.

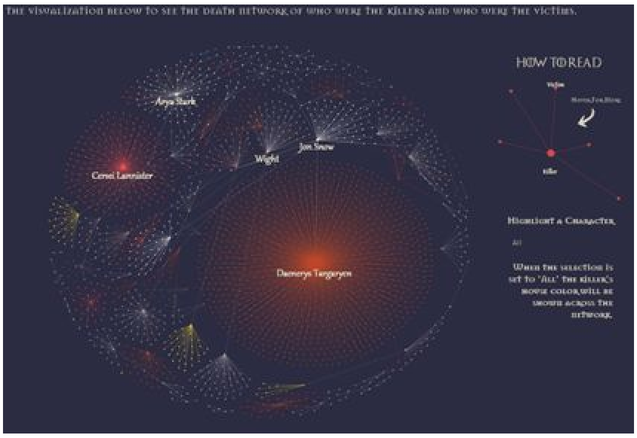
Title example 3: GoT Network Diagram

Short description: Network diagram of who kills who in the tv-series Game of Thrones made by importing a Gephi chart into Tableau.
Source: https://thedatasurfers.com/2019/08/27/how-to-use-gephi-to-create-network-visualizations-for-tableau/
Inspiration aspect: Shows how Gephi can be used together with Tableau.**n**

Relevant screenshots (copy / paste image (s)):



Gapminder bubble chart


The visualization below to see the death network of who were the killers and who were the victims.
Arya Stark
Jon Snow
Wight
Cersei Lannister
Daenerys Targaryen
HOW TO READ
Victim
Killer
Highlight a Character
All
When the selection is set to 'All' the killer's node color will be shown across the network.

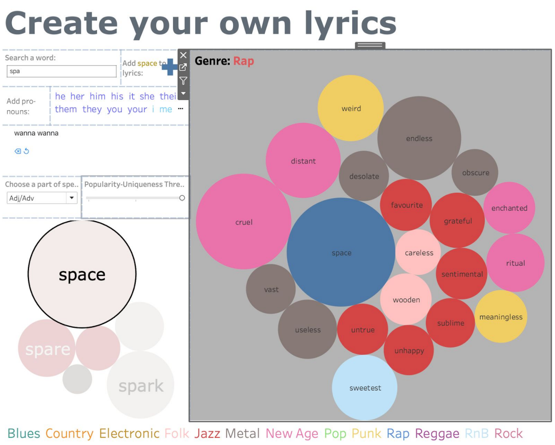
GoT Network Diagram

4. Visualisation design and interaction.

For our project we used different kinds of visualizations. The main ones are Gephi network visualizations, bubble charts and chord diagrams.

The main narrative behind these visualizations is the discovery of word co-occurrences at different levels. Starting with the gephi visualizations, the users can get a broad overview of the data. Then, the bubbles charts provide nuanced insights on words connections at the song and musical genre levels. Finally, the chord plots summarise both levels by displaying word connections in song titles.

Create your own lyric:



Create your own lyrics

Search a word: spa

Add space to lyrics:

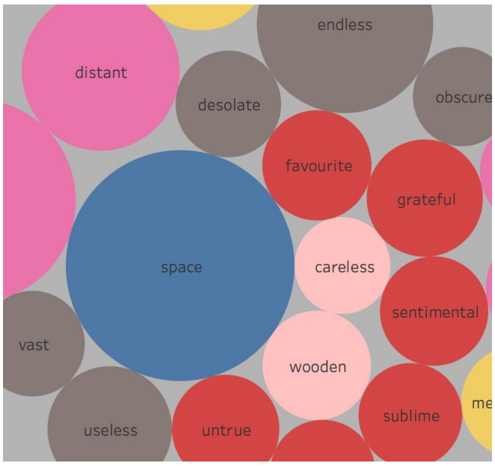
Add pronouns: he her him his it she their them they you your i me

wanna wanna

Choose a part of spe... Adj/Adv

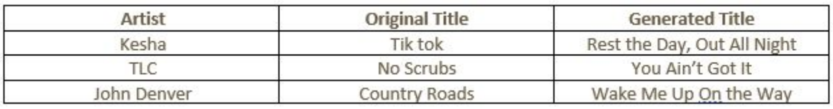
Popularity-Uniqueness Thre...

Genre: Rap



Short description: Users can explore the words connecting to a chosen word via the packed bubble chart and create their own lyric based on the insight obtained in the exploration.
Interaction: Primary interaction: clicking a word bubble triggers the rendering of word connection to the chosen word which potentially forms a chain of endless connection. During the process, the chosen word is always positioned at the centre of the re-rendered bubble chart.
Other interaction includes word search, switching parts of speech, adjusting 'popularity-uniqueness threshold', add a word to the lyric composer, lyric composer backspace and reset, switching song genre.
Narrative: Odyssey in the lyrical space.
Insight: Gain personalized insight as the user explores according to his/her interest. Users can see the relative popularity-uniqueness of words via the bubble size. The user-generated lyric could be further analyzed by ego-position, genre correlation and nearest neighbor song ratings. (not implemented)

Chord Diagram: Hot titles use uncommon word pairs to create descriptive and memorable titles. Poorly-ranked songs use more common words. We can also use this to create new titles for existing songs.



Artist	Original Title	Generated Title
Kesha	Tik tok	Rest the Day, Out All Night
TLC	No Scrubs	You Ain't Got It
John Denver	Country Roads	Wake Me Up On the Way

Some common words are not here, most noticeably being the word "Love." "Love" is one of the most used, or overused words, in every bottom-ranked song. Great love songs describe the theme, they don't need to put it directly in the title.

5. Effort.

It required substantial effort and collaboration to deliver this project. We spent significant effort cleaning the data and cutting it into usable slices, which was repeated as we found which preliminary visualizations did or did not work. Additionally, we were required to learn new tools to execute the visualizations, all of which through many rounds of revision.

Xin - Proposed this MSD/MM dataset and initial avenues of exploration. Data cleanup and prep on song title data as the rest of the group used song lyric data. Learned D3 wrapper for chord (ref: https://shahinrostami.com/ and http://holoviews.org/). Implemented chord diagram.

Lars - Organized meetings and wrote up the agenda and tasks. Made the Gephi networks charts. The biggest challenge related to this was the extraction of the network edges from the dataset. This required a fair amount of data engineering in Python. Because of the size of the dataset computing the edges necessitated a significant amount of code optimization.

Fengan - Cleaning the word list, tagging stopwords and part of speech; integrating dashboards. Main challenge was to find a suitable natural language processing library.

Tsz Kin - Explore workable dataset addons and subsets. Data transformation (tfidf). Realized the 'Create your own lyric' dashboard. One challenge is to optimize speed for retrieving insightful co-occurrence from a very large dataset. Learned how to use Tableau .hyper to enhance speed and reduce file size. Another challenge is that it is tricky to blend results from multiple aspects (different POS, different ranking position) on a single bubble chart in Tableau.

Hugo - Explore the dataset. Analysis for the assignment of the words to a musical genre. Work in a way to retrieve relevant word co-occurrences for individual words. Main development of the Ego Position pronoun-analysis and charts. The main effort was in the retrieval of the words co-occurrences, in constant collaboration with Tsz Kin we analyzed the dataset and tried different implementations in Tableau.

Martina - Encountered many difficulties in dealing with the initial data processing. Explored the data set; supported design decisions, contributed to design implementation and prepared the final Tableau Story.

6. Reflection.

In our project we implemented two of the tools that we learned during the Exercises sessions (Tableau and Gephi), plus a third method using D3 visualization (Chord Diagram).

The major challenge in presenting our data is its high-dimensionality. In total, we had over 15 dimensions (e.g. title, artist, hotness) from the base data. Then we added dimensions through POS tagging, colors, links between words, etc. We strove to keep the balance between preserving insights and simplifying the visualizations.

We learnt that the same data can be visualized in different manners, but that not all of them are significant in the same way. In fact, we learnt that the most important result of a visualization (and the most challenging) is not visualizing the data itself, but showing insights and correct interpretations.

If we were to start over, we would start with a small subset. Large sets are generally difficult to visualize. Additionally, it allows us to develop test visualizations earlier on and from there decide on the best paths to pursue.