**Project proposal**

*Answer these questions concisely below each other according to the given numbering. Use the Cambria font, no Bold.*

**Date: 23 March 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.**

What is the problem that the visualization will try to explain or solve? Why is this problem important? Why is this problem complex?

Problem: What are songs about and how are they connected

Why the problem: the song lyrics is a key component, what can we discover when ignoring melodies and examining solely lyrics

Why is this problem complex:

* big dataset
* word frequency count is on word stems
* lyrics haven’t been analyzed in detail

1. **Insight.**

What data-driven insight(s) do you expect your visualization to be able to reveal?

(Be aware that the visualization will have to be developed primarily for this objective)  
  
Insight 1: the change of the frequency of the words in relation to the popularity of songs

Insight 2: How the frequency of words changes by genre and period

Insight 3: Relation of artist/song hotness to other parameters

Insight 4: Connected bubble chart of keywords (e.g. keyword love is most connected to words x,y,z)

1. **Data.**

Which datasets will you use? What kind of data attributes do they contain? What kind of licenses are connected to these datasets? How big are they? How recent are they? How often are they refreshed? Etc.

(Fill in the table below, or make your own table)

Title dataset 1: MusixMatch

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

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

License: CCA

Title dataset 2: Million song dataset

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 3: TagTraum

Source: http://www.tagtraum.com/msd\_genre\_datasets.html

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

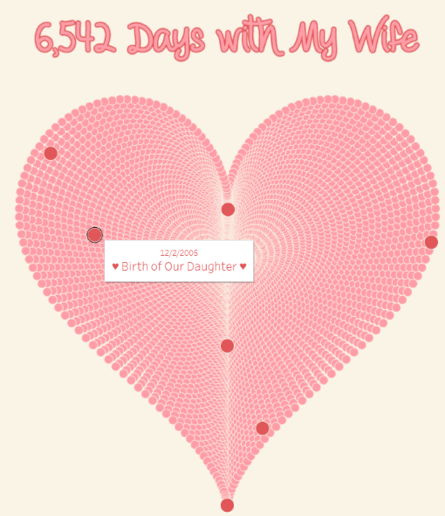
License: Non-commercial only - Hendrik Schreiber. Improving Genre Annotations for the Million Song Dataset. In Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR), pages 241-247, Málaga, Spain, Oct. 2015.

1. **Inspiration.**

Which existing visualizations will inspire you? Which aspects specifically do you find interesting? Why?

(Fill in the table below, or create your own table; Give at least 3 different examples)

Title example 1: Connected bubble chart

Short description:

Source: https://www.conceptdraw.com/How-To-Guide/picture/Bubble-chart-sample.png

Inspiration aspect + argumentation: natural way to show connections. Show importance among connections by bubble size.

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

Title example 2: Charles Minard's map of Napoleon's disastrous Russian campaign of 1812.

Short description:

Source: https://en.wikipedia.org/wiki/Charles\_Joseph\_Minard#/media/File:Minard.png

Inspiration aspect + argumentation: dense but accessible. Benchmark for good visualization.

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

Title example 3: Hearts

Short description:

Source: https://www.flerlagetwins.com/2017/02/creating-heart-in-tableau\_43.html

Inspiration aspect + argumentation: create visualizations that people can remember / make figurative impact

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

1. **Provisional sketches**

What will your visualization potentially look like? What are the interactive possibilities? What will be the narrative with which the user can use the visualization and understand the insights?

(Sketch - by hand, Photoshop or through other simple means - some strategies that illustrate how the user will come to the insights of point 2 via the visual decoding, interactive functions and narrative storytelling of the visualization) (if there are multiple ideas, display them below)



Idea 1

Screenshots (copy / paste image (s)):

Short description: Voyage of Words

Name visualization technique: Connected bubble chart

Interaction: Select keyword and see first-generation relations

Narrative: Walk through keywords to see which ones are most heavily connected

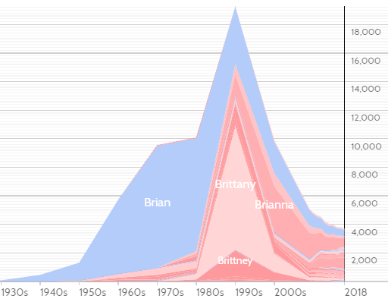
Idea 2

Screenshots (copy / paste image (s)):

Short description: Compose your own hit song

Name visualization technique: Tableau set actions

Interaction: Track user-selected words (chosen from displayed branches)

Narrative: Choose the most popular to write your hit

Idea 3

Screenshots (copy / paste image (s)):

Short description: Word hot or not

Name visualization technique: Stacked line graph

Interaction: Select keyword and see popularity over time

Narrative: If you see a song using the word “scrubs” then it was probably made in the 1990s

1. **Questions / problems**

* Data cleaning: null values
* The categories in two genre datasets are either too general or too specific
* Different languages of the songs cannot be treated separately
* Biased in the datasets in relation to the period the songs are collected
* Word stems are very useful statistically but less helpful for visualization. Users want to see regular full words instead of just stems.