Project

Intro

* Research question: What can quantitative text analysis reveil about the content of song lyrics across musical genres.

Motivation

* To kausale veje:
  + 1) musik påvirker os
  + 2) musik fortæller noget om befolkningsgrupper.

Ethics?

Data

* Data source
* Web scraping and data quality
  + Færdigt datasæt
* Data characteristics
  + Antal sange i genrer
  + Sange pr år (bias er ok – vil vil finde de sange folk hører i dag)
* Potential data issues
  + Multiple artists
  + Stadig mulighed for non-english
  + Folk på Genius er muligvis speciel gruppe

Exploratory analysis

* Language
  + Length of sentences/song
  + Words/unique words (boxplot)
  + Ordklasser og tegnsætninger (fx andel adjektiver og brug af !)
  + Characters per word
  + Brug af forkortelser
  + Slang
* Topics
  + Lav bag of words (count eller med vægt) – stemming og punctuation
  + Word cloud
  + Topic Modelling
    - Unsupervised
    - For hver enkelt genre
    - Kan vi finde clustre med andre genrer?
* Sentiment
  + Mood/tone
    - Positive/negative
    - Other feelings (fx sad songs, angry songs, happy songs)
    - Er nogle følelser mere i nogen chancer end andre?

Predictive power

* Den bedste model
  + Prøv at lave forskellige for vores egen skyld, men vis kun den bedste.
    - Random forest
    - Logistisk Regression
    - SVM
    - (obs, husk Snorres kommentar om transformering til forskelle – trække vektorerne fra hinanden. Er det relevant her?).

Diskussion

* Word2vec i forhold til CountVectorizer og ftidf
* Vi valgte de mest ppulære genre – kunne have predicted bedre med metal, christian.
* Al metode skal overvejes.

Concluding remarks