Proposal

1. Name and Email

Emil Juzovitski - emiljuz@kth.se

2. Preliminary Thesis Title

Finding Clusters in a High-Dimensional Big Data Set

3. Background/Conditions

Soundtrack your brand offers a music streaming platform for companies . The songs you hear while shopping has a chance of being streamed from *Soundtrack your brand's* platform. The retail company picks characteristics e.g. Energy, Genre, Sound that fit their brand. This is then inputted into a machine learning model. The output is a playlist based on their selection.

Soundtrack your brand's streaming platform has *50 million* songs. Each song holds a multitude of attributes, not just metadata but also audio analysis.

Since obtaining training data is an expensive manual operation, the company is interested to see if it is possible to use exploratory clustering as an approximate labeling technique for the songs. Additionally, they are interested in using it to explore the music space and uncover interesting music themes.

With all these attributes and all these songs an additional task becomes to find a clustering technique that scales(*50 million songs*) and is reusable. Furthermore, an additional step is to determine an unknown k(amount of clusters).

Another thesis student will look at the UI components of the task and how to represent the themes to the end-user.

The project will be done at *Soundtrack your brand's* office in Stockholm.

4. Research Question

1. Question
    - How Can We Find Clusters in a Big Dataset With High Attribute-Dimensionality That Scales and is Reusable?
        - Traditional clustering algorithms assume k (the amount of clusters) is known beforehand. This is not the case for this thesis and so k approximation is necessary (that isn't too computationally heavy).
        - *K-means* and other well known algorithms taught in courses do not scale across multiple processing nodes in a network.

  - With *50 million songs k* could end up being too large too manage. We might want to use dimensionality reduction or ranking the clusters to prioritize “good” clusters over bad.
  - How do we rank clusters? We can look at properties of the cluster such as density and distance to other clusters with e.g. the silhouette coefficient.
  - Different samples of attributes might lead to different *music themes* so sampling attributes could be an idea.

2. Research Area
   Unsupervised Learning, specifically clustering analysis. This will be combined with *Scalability* research, which would mean the research area would be *Distributed Systems* and general *Computer science*.

3. Connection to Research/Development
   See Background for the interest of the company.

4. Examination method

   1. Researching different clustering techniques, how they perform while scaling.
   2. Researching k-approximation, how they scale.
   3. Researching cluster-ranking.
   4. Researching sampling methods suitable for the obtained clusters.
   5. Dimensionality Reduction?

5. Evaluation
   There are two ways in which the method will be examined.

   - One can look at the average e.g. silhouette coefficient of the clusters.
   - Songs don’t follow a strict mathematical function. *Soundtrack your brand* will therefore provide me with *Music Experts* that will manually given samples of a cluster determine whether a cluster is good or not, that is, the songs of the cluster have a high correlation between each other(a theme).
   - Scaling can be determined by looking at how the algorithm performs on different sizes of the dataset.
   - Re-usability, check how well the clustering works with another sample of attributes.

6. Hypothesis
   I believe that it is possible to achieve scalable clustering algorithm, however I do believe we’ll see that what the algorithm deems to be a great cluster with highly correlated songs doesn’t always prove to be so great by the experts, that is there is no common musical theme. But for the top clusters I think we will find some kind of a theme.

5. Background of the student

Given the courses below, I feel I am prepared to the task ahead.

- ID2223 Scalable Machine Learning and Deep Learning
- ID2222 Data Mining
- ID2221 Data-Intensive Computing
- DD2421 Machine Learning
- DD2434 Machine Learning, Advanced Course
- DD2352 Algorithms and Complexity
- DA2205 Introduction to the Philosophy of Science and Research Methodology

6. Supervisor(Company)

Supervisor will be *Omar Marzouk*, Senior Machine Learning Scientist with several years of experience doing Machine Learning research in the industry. His role revolves around applying Machine Learning to various areas within the company.

7. Limits/Resources

The company has had previous thesis students from KTH within Machine Learning. A team within the company including the supervisor work in the field of Machine Learning. Google's Cloud Platform is used by the company and computation is done through it, computational resources are available upon request.

The data is also provided by the company.

8. Eligibility

Approved by Ann Lantz