Lucas van Berkel, 2655014
Ruben Woortmann, 2657307

ASS1-PROPOSAL (max 1 page)

Due: April 26th midnight

Title of the proposal

Exploring the various opinions of the public as critics over movies and its directors and actors/actresses.

Name of the group

¯_(ツ)_/¯

Names of group members

Lucas van Berkel
Ruben Woortmann

Abstract describing of project (max 200 words)

The purpose of this project is to investigate influence of particular actors, genres, director or other factors may have on film ratings and reviews. We try to pursue this endeavour by abstracting the text and ratings of reviews from films on IMDB and connecting them with information about the films production. Adding information abstracted from established databases like wiki-data can provide extra information like genres being appreciated more with one group of reviewers. The final product will be able to present a complete overview of opinions or sentiment over various subjects like movies as actors and directors.

Datasets chosen (maximise variety of formats; one unstructured, one structured, one RDF)

Critic movie review abstracted from Rotten Tomatoes (text-format, semi-structured)
IMDB review of movies (CSV) → `https://www.imdb.com/interfaces/`
`https://www.kaggle.com/iarunava/imdb-movie-reviews-dataset`
Wiki-data, for general information over movies (RDF)

Overview of the approach

- *Conversion method: what needs to be converted and how.*
  The IMDB dataset consists of multiple CSV's with categorised details per film, which will be screened to determine if they may provide information of interest. The written reviews are from separate dataset to which DBpedia spotlight will be applied to the texts to extract named entities. The datasets will then be converted to RDF through the use of COW.

- *Ontology design, justifying choices: brief description of the domain to model, plus candidate ontologies/ontology design patterns to reuse; and expected inferences.*
  An ontology will be engineered where movies, actors/actresses and directors are connected with each other and the reviews they receive. Existing schemas will be used to make the ontology as generic as possible (like FOAF etc.)

- *Instance linking design, justifying choices: brief description of the strategy to link instances between the different datasets.*
  Entity resolution can be done by unsupervised graph clustering between the datasets. Similarly, the levenshtein distance between the labels of two given nodes can be used to connect them.

- *Querying, visualisation and app: your plans.*
  If successful, queries like the expected receival of movie when a certain actor plays a role in it when the movie has a certain genre. As well as a comparison in receival between critic audience.