Project steps and report structure

Report Deadline: 1st of February

Project defense (15 minutes): 21th of January

Goal : Extract relevant terms of a domain from a corpus and classify or cluster them regarding core concepts, in order to build then an ontology of this domain. A cluster or class should be associated to a core concept and should include the terms that are specialisation of this core concept.

In this project, the domains considered are

- Music domain
- Computer science domain

You have to choose one of the three domains.

Step 1 : Analysis of the domain and Core ontology building (3pts).

- Description of the domain, description of its corpora (its size and kind)
- Core concepts
- Core relations
- Model of the core ontology
- Key terms of the domains

Step 3: Definition of your approach (3pts)

1. Natural language processing tool to use
2. Data mining/machine learning method to use : supervised or non-supervised
3. Kinds of lexical units/words/terms to extract and the method for relevant term extraction for example (Noun, NP, named entity, …)
4. Feature selection, for example
    - Verbs for which a term is a subject
    - Key words related to core concepts (core concepts and some ones that specialise it) that occur together with a term in windows of x words.
    - Words that occur with a term in windows of x words
    - others

Step 4 : Natural language processing of the corpus for NPs (or terms) extraction(3pts)

- Description of the step : how it is done ?
- Description of the results of this step (qualitative and quantitative descriptions)
- You can add a sub-step of NP filtring, to remove irrelevant ones
- File of the processing results, to attach to your report.

Step 4 : Parsing of the corpus for NP/feature space model definition (3pts)

Description of the result of this step:

- Naturel language processing of the corpus for feature extraction
- Matrix/vector space definition
- Dimension of the matrix and its % of sparsity
- Dimensionality reduction
- Filtration techniques and NP set to cluster or classify (You have to consider around 300 NPs whose 70% at least are relevant for the domain)

Step 5: Gold standard ontology building (3pts)

- To classify manually selected NPs according core concepts..

Step 6 : Clustering or classification (3pts)

- Description of the technique that will be used for clustering/classification and its parameters
- If the technique is supervised, first, you have to build a training data set (description of the training data set).
- Result of the clustering/classes: the list of terms of each cluster/class.

Step 7 : Analysis and Evaluation of the quality of the result (3pts)

Step6-1 : Visualisation of the results

Step6-2 Evaluation regarding the Gold Ontology (Precision and Recall)