School of Computing and Information Technology

ISIT219

Knowledge and Information Engineering

Assignment 2

Group members: minimum 3, maximum 5
Total mark: 40
Contribution to the final mark: 40%

Submissions: soft copy via Moodle

- report in MS Word or Pdf format (maximum 2500 words)
- submission time: 28 May at 9:00 am
- source code files (such as the RapidMiner process or any other preferred programming languages)

Business Case

YouTube is one of the largest video-sharing websites worldwide, with an estimated monthly viewership of 1 billion and serves as an important source for analyzing online user activity. In this assignment, we are taking YouTube as the main resource. There is a great potential of using YouTube data in a wide range of real-life applications. As a group of knowledge engineers, your team is required to use knowledge creation and representation techniques to analysis available YouTube data, for gaining an in-depth knowledge of user online activity. You will need to decide one topic that is of your interest, and clearly state that in your report. The data structure from YouTube is shown as follows:

Table. 1 Data structure for harvested YouTube content

Columns/Attributes	Description	Columns/Attributes	Description
video_id	ID for a video	***channel_title***	Name of video channels
category_id	Type of the video	***trending_date***	Date of video trending
tags	Tags for the comments/videos	***views***	How many views of the video
likes	The accumulated number of likes	***dislikes***	The accumulated number of dislikes
comment_count	The accumulated number of comments until the ***publish_time***	***description***	Comments content

Description of category_id:

1 - Film & Animation
2 - Autos & Vehicles
10 - Music
15 - Pets & Animals
17 - Sports
18 - Short Movies
19 - Travel & Events
20 - Gaming
21 - Videoblogging
22 - People & Blogs
23 - Comedy
24 - Entertainment
25 - News & Politics
26 - Howto & Style
27 - Education
28 - Science & Technology
29 - Nonprofits & Activism
30 - Movies
31 - Anime/Animation
32 - Action/Adventure
33 - Classics
34 - Comedy
35 - Documentary
36 - Drama
37 - Family
38 - Foreign
39 - Horror
40 - Sci-Fi/Fantasy
41 - Thriller
42 - Shorts
43 - Shows
44 - Trailers

Your tasks:

1. Some related topics include, but not limited to:
   - the influence analysis from video channels (tips: identify popular video channels and explore their influence in relation to type of video, likes/dislikes and received comments, etc., over the time span)
   - sentiment analysis of comments (tips: find out the relationship between "likes" ("dislikes") and "description")
   - NLG (nature language generator) (tips: find out the relationship between "tags" and "description")
   - categorising videos based on comments (tips: find out the relationship between "category_id" and "description")
   - prediction of video popularity (tips: find out the relationship between "views" and "description, comment_count, category_id", etc)

You need to choose a YouTube-related topic, and state it explicitly in your report.

2. Apart from the available datasets, it is expected that you collect other necessary information and/or existing case studies from academic resources (such as journal papers and books) to facilitate your research. This will be presented as the knowledge acquisition part in your project.

3. Various knowledge creation techniques can be employed including, but not limited to:

   - Classification (such as DT or ANN)
   - Clustering (such as SOM)
   - Association analysis (such as rule mining)

4. Finally, you need to write a report (maximum 2500 words) to elaborate on the following item:

   - Knowledge Acquisition or elicitation process
   - The techniques that you have employed for knowledge creation
     - You need to justify the choice of techniques
     - You need to provide at least 2 techniques to achieve full mark of knowledge creation section
   - Results and Discussions
     - The information resource that you have gathered to assess the generated knowledge
     - You can compare and contrast each knowledge category that is generated in the previous section with the existing documents or case studies from existing academic papers
     - Minimum 2 pieces for each knowledge category are expected to achieve full mark
   - Explain and justify the possible inconsistencies in the gathered knowledge

Marking Criteria:

		Very good	Good	Satisfactory	Marginal	poor
Acquiring knowledge	Through literature and the previous methods that have been applied	6	5	4	3-2.5	1.5
Knowledge creation	Justification of the methods chosen	6	5	4	3-2.5	1.5
	Software development –RapidMiner or other programming tools (marked online in lab)	10	8	6	5-4	3
	Presentation of the work in the report with explanation	6	5	4	3-2.5	1.5
Discussions and conclusion	Compare and contrast each knowledge category that is generated in the previous section with the existing documents or case studies from existing academic papers	8	6	5	4-3	2
Report writing (presentation, quality of writing, writing style, spelling grammar and use of resources		4	3	2.5	2-1.75	1