INFO411 Data Mining and Knowledge Discovery

Project 4

Instructions:

The following task is a real-world data mining problem. Part of the task is to retrieve a dataset for analysis. You are to create a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) a description of the task, (3) your proposed data mining approach and methodology; (4) the strengths and weaknesses of your proposed approach; (5) the performance measures that can evaluate your data mining results; (6) the results a brief discussion and a conclusion.

Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Mining Social Network Data

Background:

Social networking is a tool used by people all around the world. Its purpose is to promote and aid communication, and is widely regarded as a great way to connect across vast distanced. The mining of information posted on social networks has become a major area of interest to the data mining community in recent years.

Your task is to mine Twitter text data by:

- Retrieving text from Twitter. For this you may want to use the twitter API to request data from the twitter server. Create a sufficiently large corpus.
- Build a text corpus and stem the words in the corpus
- Build a term document matrix
- Find the co-occurrence of terms in the same tweet
- Show a network of tweets on the terms shared by them.

To keep the task manageable, you are allowed to follow the procedures described in the Book “R and Data Mining: Examples and Case Studies” (see chapter 11 and chapter 10 of the book). Keep in mind that you are not allowed to copy text or graphics from the book or any other source. Thus, you need to create, illustrate, and explain your own results.

Requirements:

1. Use R
2. Discuss the steps involved in text retrieval, feature extraction, feature presentation, data mining, and knowledge extraction.
3. Present a general description of the dataset and present a the general properties of the dataset. Explain the practicality of social network mining (why is it useful?).
4. Propose one suitable alternative (that differ from the approach in the text book) to mining data from social media. Note that you are only to propose an alternative. You are not required to deploy it.
5. Summarize: What new and interesting things did you discover while working on this project?