CSE 398/498 Text Mining

Project 2

Sihong Xie

September 17, 2016

1. Project Description

In this project, you are asked to use a set of relevant techniques you have learned in this course (you are welcome to go beyond those) to implement a real world text mining program that accomplishes something significant and useful. This hand-off project is open-ended in terms of project choice and research depth. You can discuss with the instructor about the techniques, solutions and other resources, but your creativity and originality are key to the success of this project. This project can be accomplished by a team with a maximal of three members.

2. Deliverables

2.1 Proposal

The team will submit a two-page document specifying which project is chosen, what resources (data and softwares) and techniques (algorithms and models) are needed to accomplish it. A high-level flow-chart is required to convince me that the team has thought about the path leading to the project implementation.

2.2 Milestone check

The team will submit a refined version of the proposal by adding more details to the flow-chart. Details about how to implement data pre-processing step or a key algorithm shall be described. The team shall have downloaded the data and implemented data preprocessing. Any issues that the team is facing or can foresee shall also be reported here and discussed with the instructor.

2.3 Report, codes and results

The team will submit a report consisting of the following parts:

1. Project description: what's background of the project and what goals the project accomplishes.
2. Dataset description: give details of the dataset, such as the source (how you obtained the data), format (free texts or structured database or semi-structured data), size (number of records), etc.
3. Algorithm description: give details of your algorithm(s) to achieve the goal you mentioned above. Importantly, you need to justify the techniques you choose. For example, if you mine frequent phrases from a text data, then you shall justify why you use frequent item mining algorithms instead of statistical word collocation algorithms.

The codes and the output of your programs shall be submitted during the rehearsal of the presentation (see below).

2.4 Presentation

The team is required to arrange a meeting in the instructor's office to have a rehearsal of the formal presentation **before** the last lecture. During the office visit, the team will submit their codes and results in person.

3. Options for your project

The following items are ideal for project 2 in terms of usefulness, relevance and complexity. You can pick one from the following options, or you can propose your own projects that are similar to these options (with the consent of the instructor).

1. Extraction of Drug-Drug Interactions from BioMedical Texts

   https://www.cs.york.ac.uk/semeval-2013/task9.html

   An exciting application of text mining to the bio-medicine area.

2. Twitter sentiment analysis

   http://alt.qcri.org/semeval2014/task9/

   Tweets are short texts, so the challenge is different from the aspect-based sentiment analysis project below.

3. Aspect based sentiment analysis

   http://alt.qcri.org/semeval2014/task4/

   This is close to the sentiment analysis we will teach in class.

4. Analysis of Clinical Text

   http://alt.qcri.org/semeval2015/task14/

   Since this dataset is a big sensitive, you are required to go through a series of training (free) to be certificated.

5. Hypernym-hyponym relation extraction

   http://alt.qcri.org/semeval2015/task17/

4. Grading

The grades depend on the following factors:

1. Completeness: the team shall submit all deliverables (proposal 10 points, milestone report 10 points, codes and results 60 points, presentation 20 points).

   The final codes and results will be evaluated in the following three aspects.

2. Complexity: deep neural network is considered to be more complicated than simple parsing. A project with multiple components is more complicated than one has a single component.

3. Usefulness: counting the words is not very useful, while finding medical terminologies is of practical value.

4. Novelty: whether the team think of something less standard in the text mining area. Examples include but not limited to new research problems or new techniques. Given the data or problems in the above project options, you can still come up with something missing from their project descriptions.

5. The presentation will be evaluated by the instructor and other teams in class (a team will submit a confidential report of the scores of other teams' presentations).

6. Individuals in a team may not receive the same grades. Your contributions to the team project will be evaluated by your teammates (each team member will include a confidential evaluation each of his/her teammates). Self-evaluation (describe your contribution in the project).