Department of Computing and Information Systems
The University of Melbourne
COMP30018/90049 Knowledge Technologies, Semester 1 2016

Project 1: Which films are good?

Executive Summary

- Get your data from the CIS servers

- *Optional: Read accompanying paper: Maas et al. (2011)*

- *Try exact matching the reviews against the film titles, e.g. using* `grep`

- Choose:

    - COMP30018: 1 or more approximate matching methods
    - COMP90049: 2 or more approximate matching methods

- *Find package(s) which implement chosen methods*

- Brainstorm mechanism for dealing with multi-word titles (subject to method)

- For each review text:

    - Find most similar title according to matching method
    - *Reject review if worse than threshold*

- Look at some reviews with approximate (not only exact) match to title:

    - Evaluate (by hand) correctness
    - Calculate evaluation metric(s)
    - Choose suitable examples demonstrating approximate matching behaviour

- **Submit Part A to CIS Servers** (by 5pm 20 Apr)

- Brainstorm and apply "good"ness strategy

- *Can't think of any suitable strategy*:

    - *Choose one or more positive/negative words (e.g.* "good" *vs.* "bad"*)*
    - *For each review with identified title:*
        * *Apply approximate matching method(s) from above to pos/neg words*
        * *Assess overall pos vs. neg of review*
    - *For each title identified in reviews:*
        * *Count whether more reviews are pos or neg*

- Consider list of "good" films to determine suitability of method(s)

- Write report:

  - Summarise problem and data set
  - Describe how approximate matching method(s) were applied
  - Present evaluation of methods on approximate matching, use examples to strengthen analysis
  - Describe how "good"ness was determined; discuss reasons for some "good" films
  - Make conclusions about suitability of methods for finding "good" films
  - Cite relevant materials (**Don't forget Maas et al. (2011)!**)

- **Submit Part B to Turnitin on LMS** (by 5pm 27 Apr)

Items *in italics* are optional. Items **in bold** are important.