University of Bolton

UNIVERSITY OF LEEDS

Combining Teaching and Research in Text-Mining from Social and Cultural Data

Claire Brierley and Eric Atwell

School of Games Computing and Creative Technologies, University of Bolton

and School of Computing, University of Leeds

INTRODUCTION: 2 research uses of Computing students

University of Bolton

UNIVERSITY OF LEEDS

- A rich resource for e-Social Science text mining research: Computing students, working on coursework projects
- Computing students can apply text-mining tools to eSS data, and/or provide social text-data at micro-level
- We will present 2 research uses of Computing students:

A) supply e-Social Science text data for students to mine; coordinated “intelligent agents” generate research results: software, text-mining outputs, research papers

B) Computing student project logs are a source of social interaction data, for the Projects Coordinator to text-mine

A) INTRODUCTION: controversial assumptions?

Q: What's the greatest commercial success on the Internet?

A) INTRODUCTION: controversial assumptions?

University of Bolton

UNIVERSITY OF LEEDS

Q: What's the greatest commercial success on the Internet?

A: not PORN ... but ADVERTISING!

SPAM is a particularly successful innovation: generating large numbers of adverts and sending to potential customers

Spam WORKS: generate LOTS of outputs, only a fraction are successful, but this amounts to many successes!

A) INTRODUCTION: controversial assumptions?

University of Bolton

UNIVERSITY OF LEEDS

Q: What is the aim of academic research?

A) INTRODUCTION: controversial assumptions?

University of Bolton

UNIVERSITY OF LEEDS

A: The aim of academic research is to generate journal papers (for RAE, for publicity, for promotion, ?)

RAE: Researchers must produce 4 journal papers in 6 years

A hybrid of student and machine intelligence can produce 60 draft journal papers in 6 weeks

-a BIG advance in Machine AND Human Intelligence?

- AND great publicity for AI !?

A) INTRODUCTION:
Students as intelligent agents

University of Bolton

UNIVERSITY OF LEEDS

Bio-Inspired Computing researchers aim to develop software which behaves like ants, bees, etc to achieve complex results

Why not use students as “super-intelligent agents”??

Prof David Cliff: this is “cheating” – his goal is software agents

BUT our goal is to generate research journal papers,

not to build bio-inspired computing software!

A) METHOD: how to generate a journal paper on eSS text mining

University of Bolton

UNIVERSITY OF LEEDS

Provide students with research journal paper generic structure: Introduction, Methods, Results, Conclusions.

DEMO at BCS Machine Intelligence Contest (AI'2007):

… a volunteer from the audience demonstrated how student + AI software, with help from an eSS text-mining researcher (me!), can generate a draft journal paper

I am the QB "queen bee": I guide the hive (students+MI)

We had 10-15 minutes, not 6 weeks, so key steps only…

A) METHOD:
How to create a journal paper

University of Bolton

UNIVERSITY OF LEEDS

QB) Design the overall HI-MI hybrid: coursework specification

http://www.comp.leeds.ac.uk/db32/assessment.htm

QB) Select a domain + research question for text-mining

Social and Cultural studies for a region; specifically: Do British or American influences dominate the Web in this region?

1) Use AI search tool to choose a region and journal for this question; and find related research to cite, in the Introduction of your paper.

2) Choose 3+ countries in this region, use AI search tool to harvest a Web-Corpus for each country

QB) harvest 10 UK and 10 US Web-corpus data-samples

A) How to create a journal paper (continued…)

University of Bolton

UNIVERSITY OF LEEDS

QB) Use AI tool to find significant differences: candidate Text-Mining features characteristic of UK v. US English

3) Choose a small set of features, encode in uk-us ARFF file

4) Chosen region: encode features from (3) in test ARFF file

5) Use AI ML toolkit (WEKA) to build text-mining evidence of uk-us decision; copy-and-paste into journal paper

6) Decision-tree predictions for region samples: UK or US? (Test options: Supplied test set); copy into journal paper

7) Finish paper: Introduction, Methods, Results (ML evidence: novel to this research journal readership), Conclusions

8) Submit paper via intranet Knowledge Management tool

QB) assess course-works, aka review and improve

A) RESULTS

Student: learning through practical experience of text-mining; outline paper as coursework assessment towards Degree

QB) 60 draft research papers to polish and submit to journals!

(also: research papers on combining teaching and research…)

B) Student projects as data source

1. Exploit the opportunity afforded by student projects to undertake e-Social Science text-mining research within limited resources and time

2. Use recorded computer-mediated social interactions that arise naturally from collaborative learning situations to gain empirical insights into the learning process itself

B) More about the data source (1)

University of Bolton

UNIVERSITY OF LEEDS

- Games Design Team Project
- Project generates a lot of data: documentation; presentations; game-play artefacts
- Data of interest to this study: team and individual online project journals
- Strong first cohort of final year students on GAD in 2007-2008 who did a lot of blogging

blogger.com **blogspot.com** **MSN**

Wikispaces **Google Docs**

B) More about the data source (2)

1. Dynamics: **collaborative tie strength** (Cummings & Kiesler, 2007)
2. Mechanics: **norms** in online communities (Arms et al, 2006)

TEAM DYNAMICS

Strong ties ← frequent communication and emotional closeness

Observation: on the whole, positive dynamics within teams on this project

TEAM MECHANICS

Team contexts upheld by different styles of leadership

Observation: emergence of norms (Arms et al, 2006) for joint team effort, and compliance with these norms, was bottom-up and aided by online social interactions

B) Elements of emerging study

University of Bolton

UNIVERSITY OF LEEDS

- Access to a self-organising online social network of students influencing one another, helped along by frequent face-to-face contact
- Data is digital records of learning and team-working from 4 different student cohorts over 4 semesters
- Could compare groups that differed in reliance on outside moderation
- More inclined to look at **lived experience** and **hopes and fears** (Ahmad et al, 2005) and **digital documents of life** (Crabtree & Rouncefield, 2005) of individuals and groups over a particular period of the project

PROJECT START → FIRST MILESTONE OUTPUT

(DESIGN DOCUMENT, PITCH and PLAN)

- Apply text-mining techniques of corpus linguistics and information extraction to these spontaneous, expressive texts to explore **values** held by students - values associated with ***meaningful learning gained through team-working***

B) What's involved?

- Use of keyword filters to track salience and sentiment in student texts
- Determine whether there is a **special language*** (Ahmad et al, 2005) in these texts expressing values associated with *meaningful learning gained through team-working*
- Achieve this by computing and contrasting the frequency of keyword filters in student texts relative to a general language corpus such as the BNC (British National Corpus)
- Choice of keyword filters may be subjective but a starting point may be the module specification for GAD Team Project which is a concentrated statement of values in itself
- I'm interested to see what students make of these values
- Principal software will be latest version of NLTK or Natural Language ToolKit (Bird et al, 2008) which conveniently has a probability module with a set of Classes applicable to experiments planned

*** i.e. frequency of certain keywords**

CONCLUSIONS

University of Bolton

UNIVERSITY OF LEEDS

A) hybrid of human and machine intelligence:

AI architecture applied to students + smart choice of journals and instructions + use of AI tools by AI students

… can produce 60 draft journal papers in 6 weeks

B) Computing student project logs provide rich data about student social interaction, for Text Mining and collaborative research with Social Scientists