Socially intelligent computing for coding of qualitative data*

Kevin Crowston[1] & Nancy McCracken[2], PIs

Syracuse University School of Information Studies Grant 09–68470

with: Nora Misiolek PhD Students: Joshua Seymour & Jasy Liew Suet Yan Developers: Steven Rowe & Yatish Hegde

Research Problem and Goals

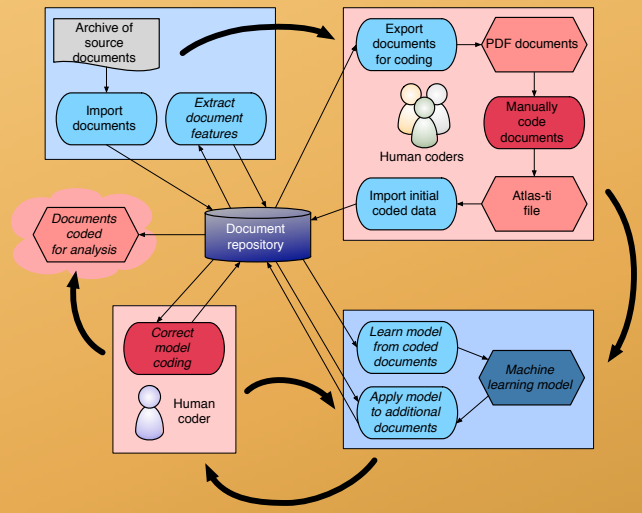
- Social science researchers often study texts to analyze practices of research populations
- But analysis of textual data is very labor-intensive and does not scale to the increasingly large amounts of data available

The goal of the research is to develop and test an innovative Natural Language Processing (NLP) and Machine Learning (ML)-based research tool that supports a computer-human partnership for content analysis for qualitative social science.

Proposed approach

- A human-computer partnership integrating manual coding with NLP information extraction and active ML
- System will learn model to extract coded text from sample hand coded by researchers
- Then apply model to code additional documents
- Human coders will correct machine coding to create corpus for further ML and final analysis
- 100% ML accuracy is not attainable, but if NLP can identify text containing phenomena of interest with high recall and reasonable precision, it will greatly speed manual coding

Planned workflow



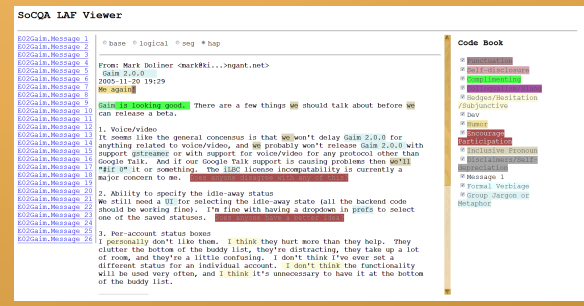
System components

- Document repository (Fedora) to store raw documents annotated with features and human and machine codes
- Import/export of source documents and import of Atlas-ti file with human coding
- Feature extraction and machine learning; application of a model to a document corpus
- Web user interface to control operations and to correct machine applied codes
- Experimenter's interface to explore different feature sets, learning parameters, etc.

Project status and plans

Year 1 (just finishing)

- System design and development (about 50%)
- Hand coding sample dataset on leadership behaviours in open source software projects



Year 2

- Experiment with system to determine useful features and ML settings for extraction
- Continued development of system (e.g., import additional document types, new NLP features, UI improvements)
- Complete leadership and other studies using system on large sample of text

Year 3

- Recruit research projects to further test system
- Continued system development and tuning to adapt system to additional users
- Study what codes system can and can't handle

Please let us know if you'd like to be a beta tester!

* http://socqa.org/

[1] crowston@syr.edu, http://crowston.syr.edu

[2] njmccrac@syr.edu

Poster presented at the SOCS PI Meeting, Ann Arbor, MI 17–19 June 2012