Authorship Detection

Minutes for 4th meeting of 09/08/2010
Held in Innova21 inside office, 3:02 pm – 3.50 am
Participants: Clement, Joel, Jie Dong, Derek, Brian, Matthew, Maryam,
Absent: None

Minutes

1. **<u>Project status/progress during last week</u>**
   a) Done on comparison of different data conversion and classification.
   b) WBS is roughly done.
   c) Matthew:
      a. Get java and matlab working for SVM
      b. Use default test files
      c. Try reproduce Talis data
      d. Get train data, test data, check if is correct, test on known author
   d) The classification in SVM is similar to LDA
   e) SVM can handle non-linearity.
   f) Put in train data, it comes out as math function (not sure about this)
   g) Block diagram of flow of data in SVM
   h) 3 methods use: WRI, word frequency, and Trigram Markov (featured vector)
   i) Applications:
      a. New search engine
         i. not by keyword, but put entire document and find similar document
         ii. any document written by same author
      b. plagiarism in music
      c. comparing software version-control
   j) need 1 slides on risk, budget and work hazard
   k) can test by chop half, train first half, then feed 2nd half of known author
   l) TALIS DID NOT USE SVM, he uses WRI, Markov and MDA (multidimentional Discriminant analysis)

2. **<u>Project goals for upcoming week</u>**
   a) Need slide on work hazard and risk
   b) Get flow chart on data flow
   c) Do a brief introduction slide on the 3 technique use

3. **<u>Individual reports</u>**
   a) Joel:
      a. Do milestone and budget for power point
      b. Do work hazard and risk slides
      c. Study on controversial things (Federalist Paper, Shakespeare, bible, and book of Mormon)
   b) Dong Jie:
      a. Do flow chart of data
      b. Study on SVM
      c. Later on focus on matlab of SVM
   c) Clement:

a. Do the slides on 3 technique use (WRI, trigram Markov, and word frequency)
b. Study SVM for java