Who wrote the letter to Hebrews?

--Data mining for detection of text authorship

Honours Project Proposal Seminar

Date:
12/August/2011

Presented by:
Kai He, Yan Xie, Zhaokun Wang

Supervisor:
Derek Abbott

Co-Supervisor:
Brian Ng

Background information for Data Mining

What is Data Mining?

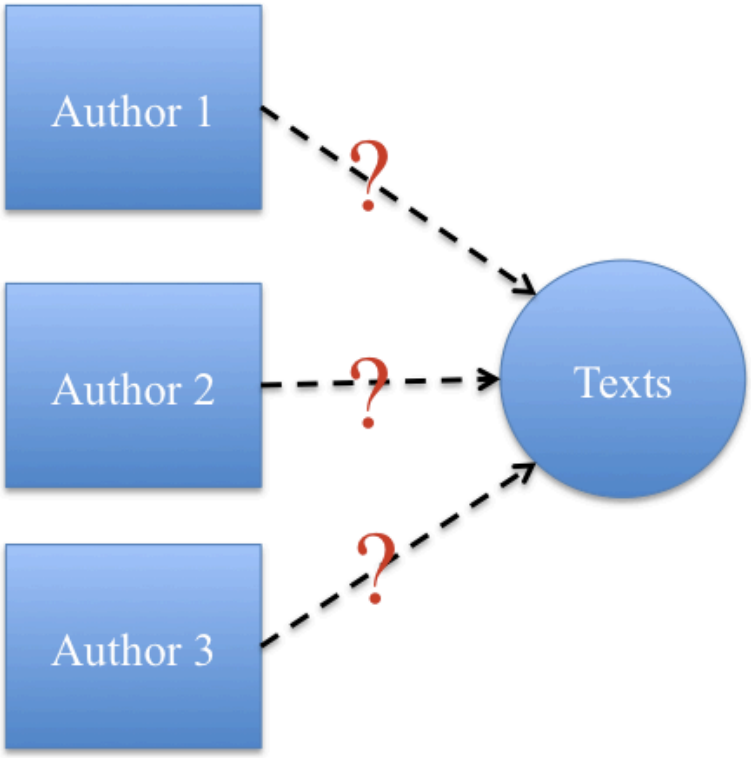
The process of analyzing data from different perspectives and summarizing it into useful information.

Applications:

- Plagiarism analysis
- Authorship identification
- Near-duplicate detection

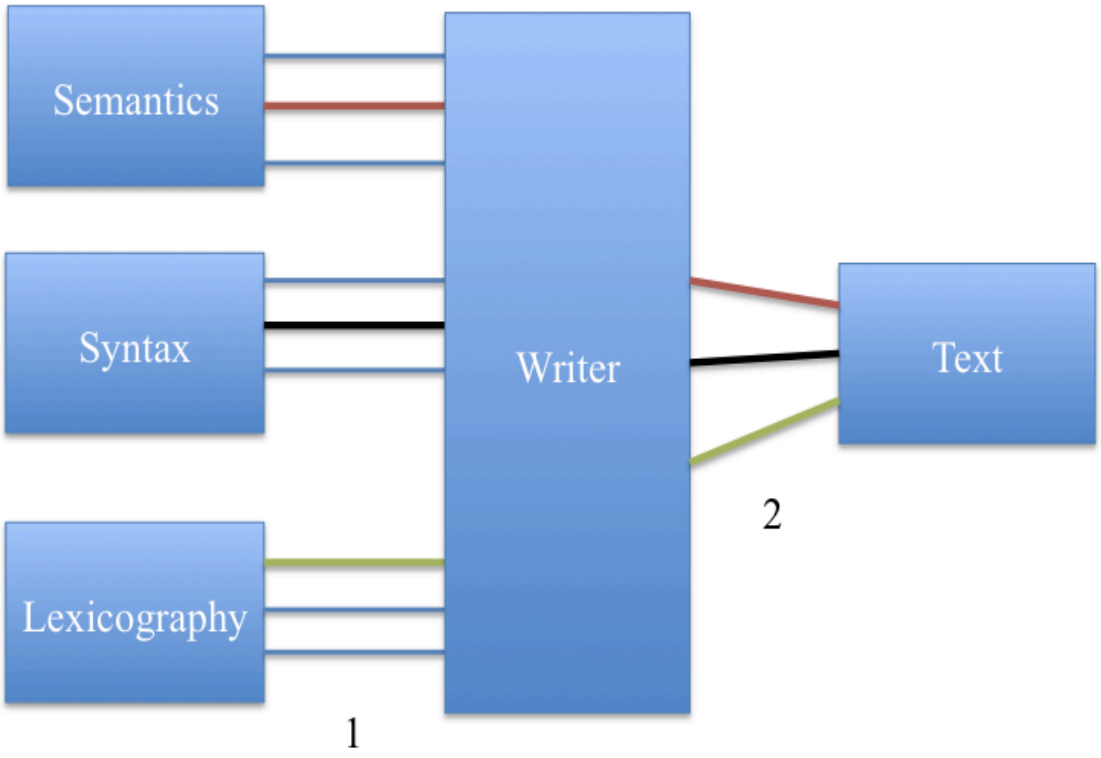
Two types of Authorship Identification problems

Attribution Problem



Verification Problem

The Basis for Authorship Attribution



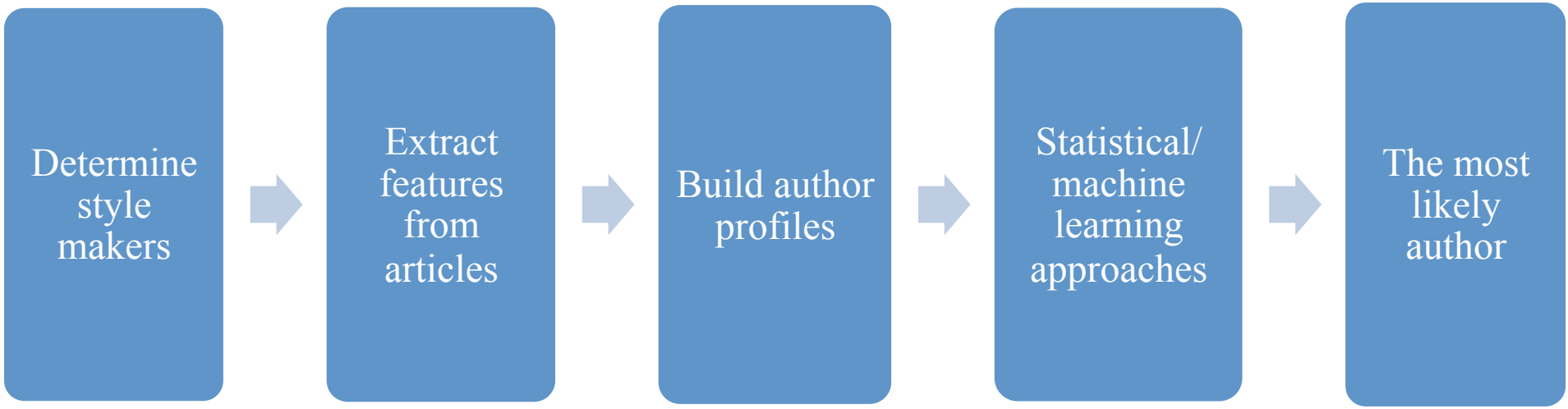
1. Writers are offered significant amount of choices on how to write a text.

2. Each specific text carries the fingerprint (Style marker) of its creator.

The Assumptions

- A specific single author.
- The author made the choices.
- The author is consistent in his/her preferred choices.
- These choices are present and could be detected in all texts written by that creator.

General Approach for Authorship Attribution

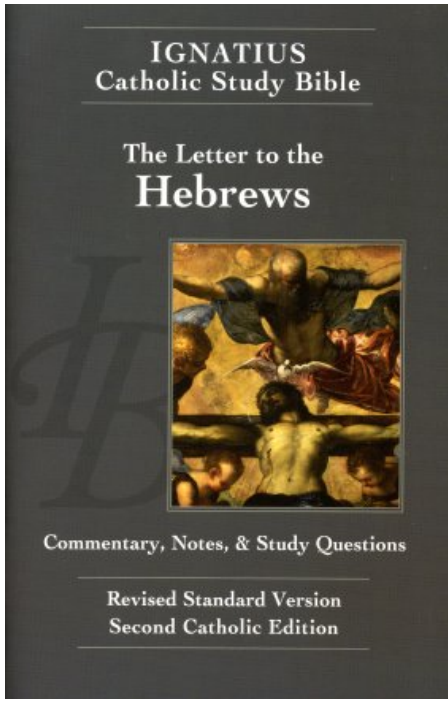


Background information for this project

The **Letter to the Hebrews** (or The **Epistle to the Hebrews**) is one of the books in the New Testament. Its real author is unknown.

Scholars have sought to identify the author of Hebrews since the time of Origen (185 - 256 AD.)

There are 6 candiates that could be the author of Hebrews.



Motivation for this project

Social:

- Result from this project might contribute to the study of New Testment.
- Can be used in Forensic investigations as well as cirminal investigations.

Ethical:

- Used in Plagiarism analysis to prevent someone stealing the credit of other people's work.

Project Objectives

- Implement two different algorithms in authorship attribution based on character level and word level.
- Compare performance of the two approaches.
- Apply our methods to identify the author of the letter to Hebrews.

Past Research

- 1887 - Mendenhall uses characteristic curve of composition
- 1983 - Smith uses stylometrics measurement
- …
- 2004 - Sabordo uses data compression technique
- 2005 - Talis uses Trigram Markov model with Multiple Discriminant Analysis(MDA)
- 2010 - Jie et al. use three algorithms: Function Word Frequency, Word Recurrence Interval and Trigram Markov Model

Authorship Attribution

Base

Advance

Features		Required tools and resources
Lexical	Token-based Vocabulary richness	word length, sentence length, letter, digits
Character	character types character n-grams	character, letter, digits
Syntactic	Part-of-speech(POS) Chunks Sentence and phrase structure	text chunker, sentence splitter, tokenizer
Semantic	Synonyms Semantic dependencies	text chunker, sentence splitter, tokenizer

Proposal Algorithms

- Two proposed algorithms:

- Common N-gram (CNG)

- Maximal Frequency Word Sequences

Common N-gram (CNG)

- Features:
  - Character level
  - Convert character n-grams into byte level

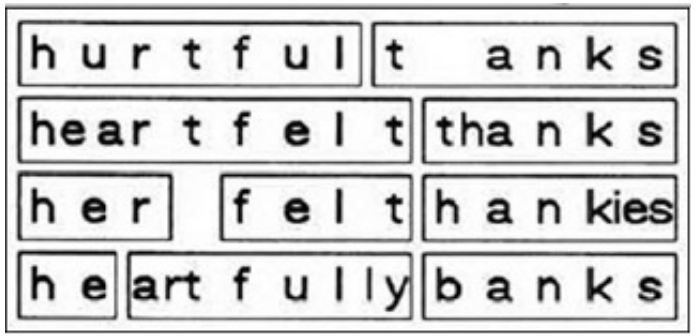
	"Adelaide"
Unigrams	A, d, e, l, a, I, d, e
Bi-grams	Ad, de, el, la, ai, id, de
Tri-grams	Ade, del, ela, lai, aid, ide

character n-grams

- Why CNG?
  - Language independent
  - Relatively higher accuracy (Keselj et al. 2004)

Maximal Frequency Word Sequences (MMFWS)

- Features:
  - Word level
  - Similar to the word N-gram based approach
  - A special kind of word sequence



- Why MMFWS?
  - Without using any sophisticated linguistic analysis of texts
  - Not very sensitive to the size of documents

Support Vector Machine (SVM)

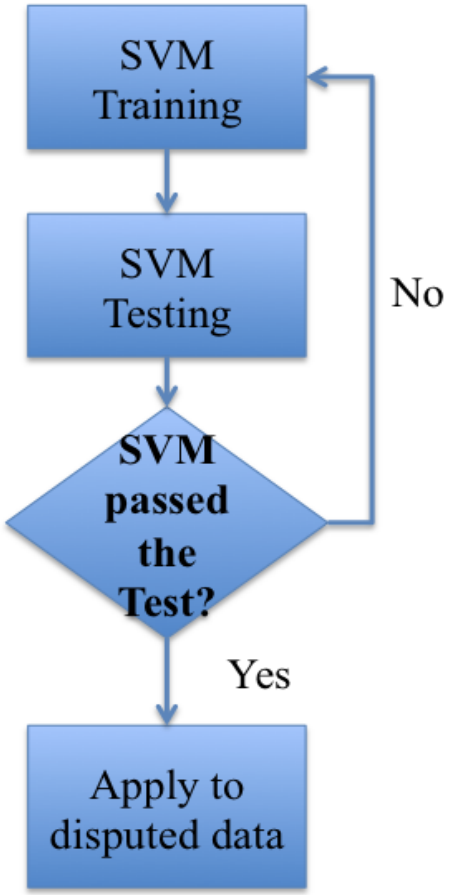
- What is Support Vector Machine:

  - A type of machine learning that classify data

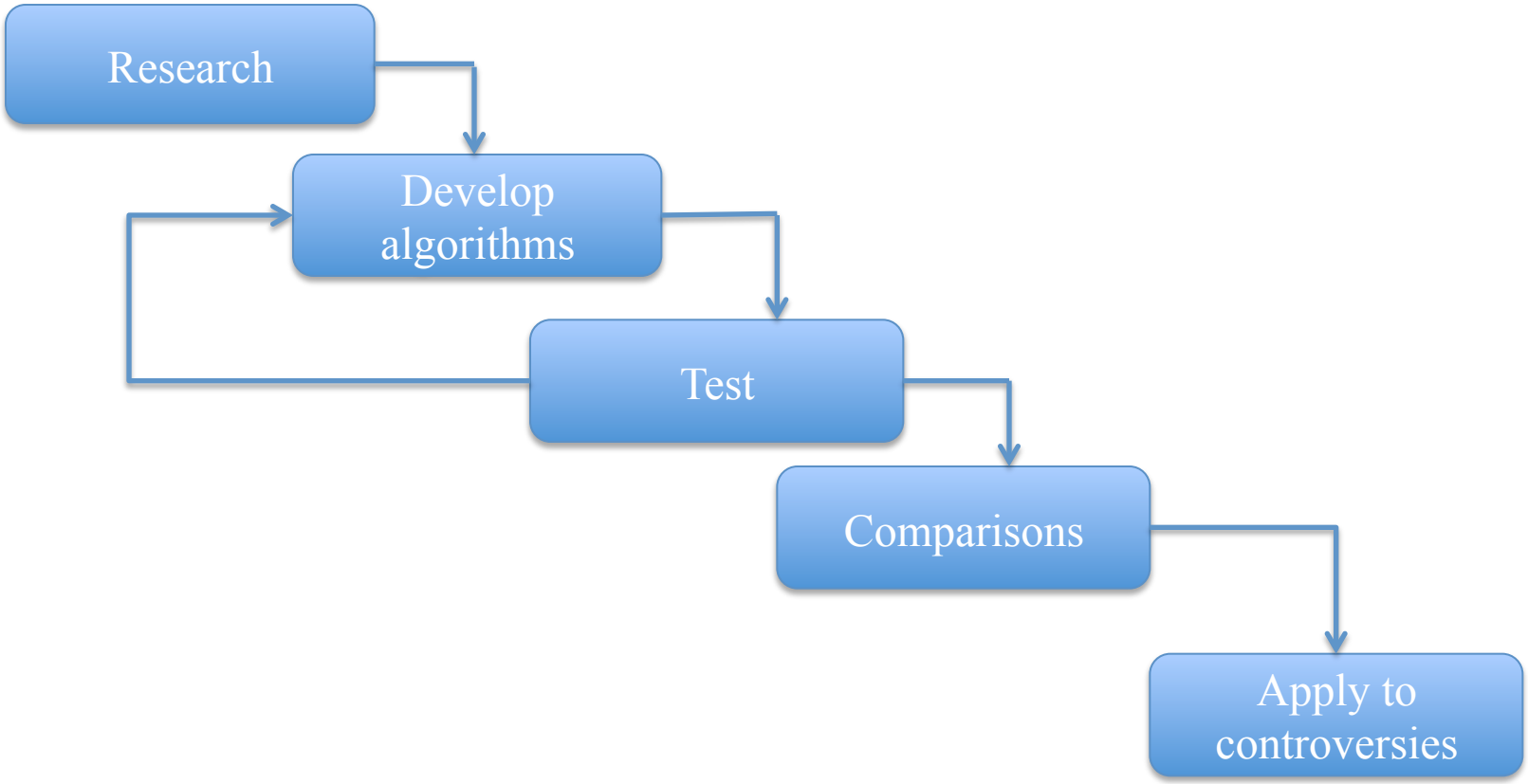
- The aim of SVM:

  - To find the decision boundary (hyperplane) that maximises the margin to the data points

- Why Support Vector Machine:
  - Success rate of prediction of the model
  - Easy to implement



Flow chart

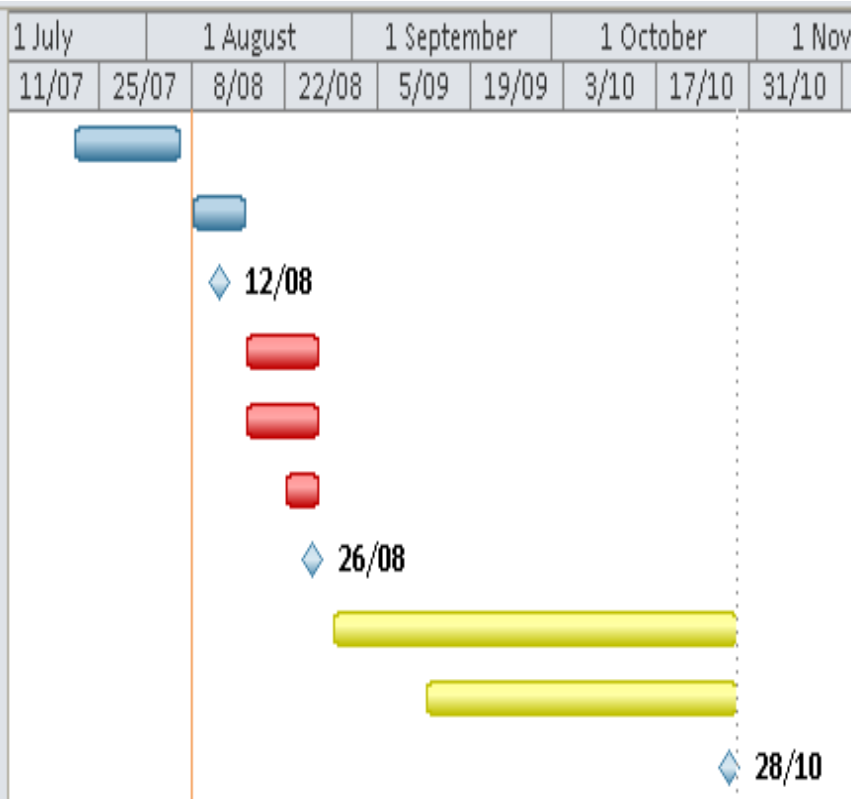


Project Plan

Important Date	Due Date
Proposal Seminar	12th Aug 2011
Stage 1 Report Due	26th Aug 2011
Stage 2 Report Due	28th Oct 2011
Exhibition Information Due	9th Mar 2012
Seminar Day	20th Apr 2012
Exhibition Briefing	4th May 2012
Final Report Due	25th May 2012
Poster Due	29th May 2012
Project Exhibition	1st Jun 2012

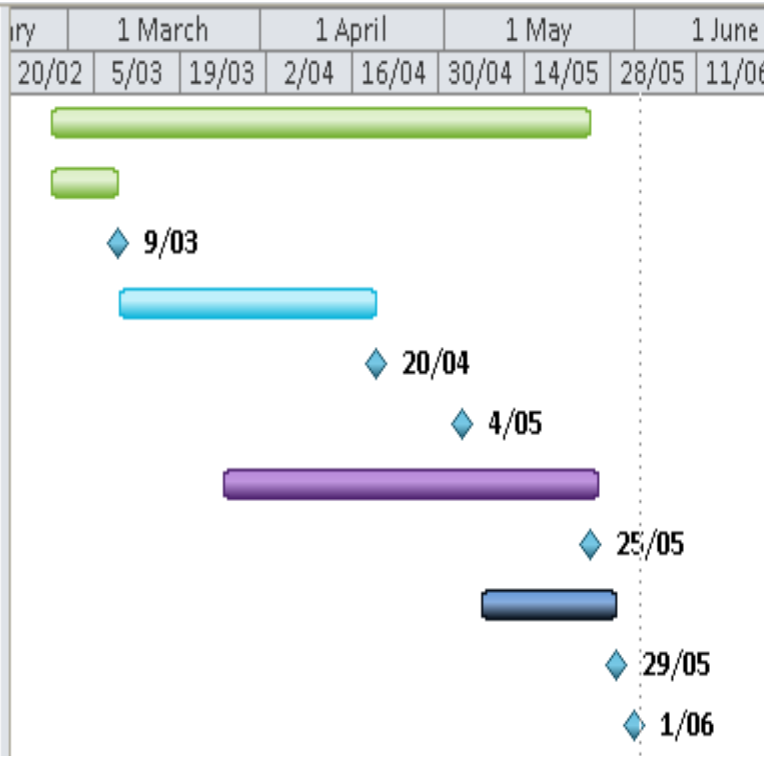
Gantt Chart

Task Name	Duration	Start	Finish
Research	2.4 wks	Thu 21/07/11	Fri 5/08/11
Proposal Seminar	6 days	Mon 8/08/11	Mon 15/08/11
Proposal Seminar	0 days	Fri 12/08/11	Fri 12/08/11
research 2	1.8 wks	Tue 16/08/11	Fri 26/08/11
New methods design	1.8 wks	Tue 16/08/11	Fri 26/08/11
Stage 1 Progress report	1 wk	Mon 22/08/11	Fri 26/08/11
Stage 1 Progress report d	0 days	Fri 26/08/11	Fri 26/08/11
Develop project	9 wks	Mon 29/08/11	Fri 28/10/11
Develop Stage 2 report	35 days	Mon 12/09/11	Fri 28/10/11
Stage 2 report due	0 days	Fri 28/10/11	Fri 28/10/11



Gantt Chart

Task Name	Duration	Start	Finish
Develop project	64 days	Mon 27/02/12	Thu 24/05/12
Prepare Exhibition	9 days	Mon 27/02/12	Thu 8/03/12
Exhibition information d	0 days	Fri 9/03/12	Fri 9/03/12
Prepare Seminar	30 days	Fri 9/03/12	Thu 19/04/12
Final Seminar	0 days	Fri 20/04/12	Fri 20/04/12
Exhibition briefing	0 days	Fri 4/05/12	Fri 4/05/12
final report	9 wks	Mon 26/03/12	Fri 25/05/12
final report due	0 days	Fri 25/05/12	Fri 25/05/12
Prepare exhibition	16 days	Mon 7/05/12	Mon 28/05/12
poster due	0 days	Tue 29/05/12	Tue 29/05/12
exhibition	0 days	Fri 1/06/12	Fri 1/06/12



Monitoring Mechanism

- Meeting
  - Meeting with supervisors once a week
  - Meeting with group members twice a week
- Monitoring
  - Create the "checklist" in group
  - Example :

Task	student	Expected due date	Actual due date

Group Member Roles

	Kai He	Yan Xie	Zhaokun Wang
Programmer	✔	✔	✔
Secretary		✔	
Document Manager	✔		
Group Leader			✔

Risk Assessments

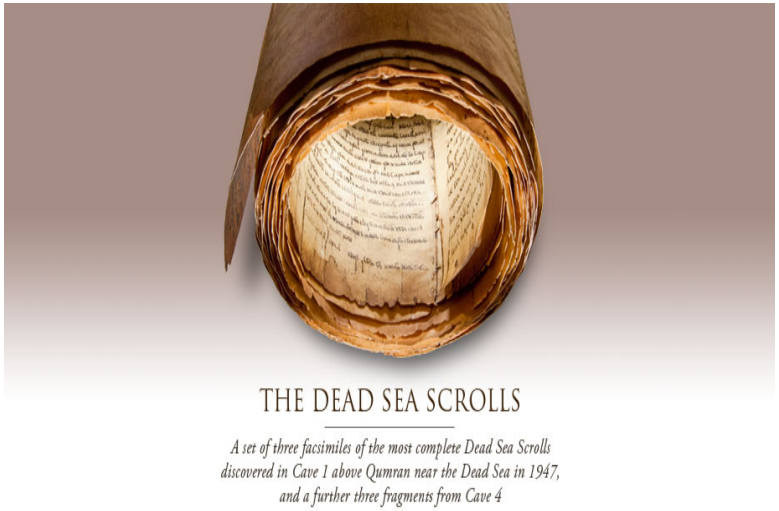
Risk	Priority	Probability Rating	Impact Rating	Preventive Measures
Wrong Software Function	90 (High Risk)	10	9	Proper design before writing the code
Behind Schedule	45	5	9	Regularity monitor the project schedule
Inefficient Resources	36	6	6	Find more information
Unfamiliar with Programming Language	30	5	6	Understand the basic programming language
Absence of Team Members	24	3	8	Keep in touch with each members
Wrong Direction of Project	20	2	10	Discuss with supervisor
Data Lost	20	2	10	Regularity save the files

Project Budget

- Budget: $250 per student
- Allocated Total Budget: $750
  - Printing and Binding: $100
  - E-resource Purchase: $150
  - Software Purchase: $150
  - Prepare Exhibition: $200
  - Others: $100
- Expected Total Expenses: $700

Deliverable

- Academic document
- Software package
- Suggestions for further research



Reference

Peng, F., Schuurmans, D., Keselj, V. & Wang, S. (2004). Augmenting Naïve Bayes Classifiers with Statistical Languages Models. Information Retrieval, vol. 7, 317-345. Kluwer Academic Publishers. 2004.

Stamatatos, E., Fakotakis, N. & Kokkinakis,G. Computer-Based Authorship Attribution Without Lexical Measures. Computers and the Humanities 35: 193-214, 2001. Kluwer Academic Publishers. 2001

Stamatatos, E. (2006b). Ensemble-based author identification using character n-grams. In Proceedings of the 3rd International Workshop on Text-based Information Retrieval (pp. 41-46).

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Proceedings of the 10th European Conference on Machine Learning (pp. 137-142).

Thank you for your attention

Any Questions ?