

CSC 479 FINAL YEAR PROJECT

EDUCATIONAL DATA MINING

(SCE12-0140)

Supervisor: Assoc. Prof. Hui Siu Cheung

Examiner: Dr. Shen Zhiqi

Submitted in Partial Fulfilment of Requirements for

Bachelor (Honours) in Engineering (Computer Science)

from Nanyang Technological University

by

TOH JUN HAO

(U1021549H)

School of Computer Engineering

2012/2013

# Abstract

Mathematics is one of the core subjects which all Secondary School students have to take in order to further on with their tertiary education or qualify for the workforce. With multiple subjects to focus before the examinations, educators and parents are finding the quickest and fastest method to best equip the students to face the examination challenge.

Because of the demand for intelligent way of examination revision, the Education industry has already taken steps into investing on software systems that make revision more efficient. However, most systems do not offer features such as Tag, Topic Distribution Analysis, Topic Trend Analysis, and Question Clustering based on Tag Similarities.

These features are found to be very useful in allowing students to identify the major topics through topic trend and distribution analysis, identifying the knowledge needed to attempt questions through useful knowledge tags, and finding similar questions through Tag searches or Question Clustering tools.

The project aims to develop a web-based application that uses the “GCE Ordinary Level Additional Mathematics” subject as the platform for performing topic distribution and trend analysis, and also Tags and Clustering features. Dataset consists of questions and answers in text, images, and mathematical formula state. Visualization tools are also incorporated for better user readability.

# Acknowledgement

The author would like to express his sincere gratitude to his Project Supervisor, Associate Professor Hui Siu Cheung, for his wise advices and thoughtful criticisms while guiding him till the completion of project.

The author would also like to thank the friendly staffs from Database Technology Laboratory and Software Engineering Laboratory for their technical support and assistance.

Last but not least, the author would also like to thank Mr. Chia Zhou Hiang for his assistance during the initial phase of taking over the project, and Mr. Chen Jingci for his help in deploying the project to AppFog, not forgetting everyone else that provided inputs, ideas, feedbacks and moral support to the author throughout the project.

# Table of Content

[Abstract II](#_Toc352601797)

[Acknowledgement III](#_Toc352601798)

[Table of Content IV](#_Toc352601799)

[List of Figures VIII](#_Toc352601800)

[List of Tables XII](#_Toc352601801)

[1. Introduction 1](#_Toc352601802)

[1.1. Motivation 1](#_Toc352601803)

[1.2. Project Objective 2](#_Toc352601804)

[1.3. Prior Work 3](#_Toc352601805)

[1.4. Report Organization 4](#_Toc352601806)

[2. Related Work 5](#_Toc352601807)

[2.1. Web Framework (Django) 5](#_Toc352601808)

[2.2. Mathematics Formula Typesetter (LaTeX) 7](#_Toc352601809)

[2.3. Visualization Tools (Highcharts) 7](#_Toc352601810)

[2.4. Data Mining Process 9](#_Toc352601811)

[2.5. Clustering Algorithm 10](#_Toc352601812)

[2.5.1. K-Means 11](#_Toc352601813)

[2.6. User Interface (Bootstrap) 12](#_Toc352601814)

[2.7. Deployment (AppFog) 12](#_Toc352601815)

[3. System Overview 14](#_Toc352601816)

[3.1. Software Development Lifecycle (SDLC) 14](#_Toc352601817)

[3.2. Requirement Analysis 15](#_Toc352601818)

[3.3. System Design 16](#_Toc352601819)

[3.3.1. Use Case Diagram 16](#_Toc352601820)

[3.4. Database Design 17](#_Toc352601821)

[4. Topic Distribution and Trend Analysis 20](#_Toc352601822)

[4.1. Feature Objective 20](#_Toc352601823)

[4.2. Approach for Topic Distribution 20](#_Toc352601824)

[4.3. Approach for Topic Trend Analysis 23](#_Toc352601825)

[4.4. Discussions 26](#_Toc352601826)

[5. Tagging 27](#_Toc352601827)

[5.1. Feature Objective 27](#_Toc352601828)

[5.2. Types of Tags 27](#_Toc352601829)

[5.3. Approach for Implementation of Keyword Tag Data 28](#_Toc352601830)

[5.4. Approach for Implementation of Concept and Formula Tag Data 29](#_Toc352601831)

[5.5. Approach for Implementation of Tag Cloud Visualization 30](#_Toc352601832)

[5.6. Discussions 32](#_Toc352601833)

[5.6.1. Paper-Based Tag Cloud Findings 32](#_Toc352601834)

[5.6.2. Topic-Based Tag Cloud Findings 34](#_Toc352601835)

[6. Clustering for Question Similarity 37](#_Toc352601836)

[6.1. Feature Objective 37](#_Toc352601837)

[6.2. Approach for Data Pre-processing 37](#_Toc352601838)

[6.2.1. Vector-Space Model 37](#_Toc352601839)

[6.2.2. Euclidean Distance 38](#_Toc352601840)

[6.2.3. From Data Pre-Processing to Application of Clustering Algorithm 39](#_Toc352601841)

[6.3. Performance Analysis 40](#_Toc352601842)

[6.3.1. Analysis of Best K-Value (Number of Clusters) 40](#_Toc352601843)

[6.3.2. Discussion 52](#_Toc352601844)

[7. Tag Search 54](#_Toc352601845)

[7.1. Feature Objective 54](#_Toc352601846)

[7.2. Implementation Approach 54](#_Toc352601847)

[7.2.1. Tag Search Bar 54](#_Toc352601848)

[7.2.2. Search Result Page 55](#_Toc352601849)

[7.3. Discussion 55](#_Toc352601850)

[8. User Evaluation and Feedbacks 56](#_Toc352601851)

[8.1. Users’ Evaluation on Usefulness of View Paper 56](#_Toc352601852)

[8.2. Users’ Evaluation on Usefulness of Paper Tag Cloud 57](#_Toc352601853)

[8.3. Users’ Evaluation on Usefulness of Topic Distribution 57](#_Toc352601854)

[8.4. Users’ Evaluation on Usefulness of Topic Trend 58](#_Toc352601855)

[8.5. Users’ Evaluation on Usefulness of Topic Tag Cloud 59](#_Toc352601856)

[8.6. Users’ Evaluation on Usefulness of Question Clustering 59](#_Toc352601857)

[8.7. Users’ Evaluation on Usefulness of Tag Search 60](#_Toc352601858)

[8.8. Users’ Evaluation on Usefulness of the Overall System 60](#_Toc352601859)

[8.9. Additional Remarks 61](#_Toc352601860)

[8.10. Discussion 61](#_Toc352601861)

[9. Experimental Scenario 62](#_Toc352601862)

[9.1. Experiment Objective 62](#_Toc352601863)

[9.2. Predictions 62](#_Toc352601864)

[9.3. Prediction Results 65](#_Toc352601865)

[9.4. Discussion 66](#_Toc352601866)

[10. Conclusion and Future Works 67](#_Toc352601867)

[10.1. Conclusion 67](#_Toc352601868)

[10.2. Future Work 68](#_Toc352601869)

[10.2.1. Improvise on Tags 68](#_Toc352601870)

[10.2.2. Clustering Algorithms 69](#_Toc352601871)

[10.2.3. Improve on User Experience 69](#_Toc352601872)

[10.2.4. Ratings and Recommendations 69](#_Toc352601873)

[A. Appendix 70](#_Toc352601874)

[1. Web Design and Navigation 70](#_Toc352601875)

[I. Main Landing Page 70](#_Toc352601876)

[II. View Paper Selection Page 71](#_Toc352601877)

[III. View Paper (Paper View) 72](#_Toc352601878)

[IV. View Paper (Topic View) 74](#_Toc352601879)

[V. Analyzer Main Landing Page 75](#_Toc352601880)

[VI. Tag Cloud (Paper / Topic) 76](#_Toc352601881)

[VII. Topic Distribution 78](#_Toc352601882)

[VIII. Topic Trend 80](#_Toc352601883)

[IX. Question Clustering 82](#_Toc352601884)

[X. View Concepts 84](#_Toc352601885)

[XI. Tag Search 85](#_Toc352601886)

[XII. Admin Question Selection 87](#_Toc352601887)

[XIII. Admin Question Listing 88](#_Toc352601888)

[XIV. Admin Create/Modify Question Form 89](#_Toc352601889)

[XV. Admin Tag Listing 90](#_Toc352601890)

[XVI. Admin Create/Modify Tag Form 91](#_Toc352601891)

[XVII. Django Admin 92](#_Toc352601892)

[2. Tag List 95](#_Toc352601893)

[3. Topic Tag Cloud Full Evaluation 113](#_Toc352601894)

[4. Survey Question 129](#_Toc352601895)

[B. References 130](#_Toc352601896)

# List of Figures

[Figure 2‑1: MVC Design Overview [2] 5](#_Toc351211609)

[Figure 2‑2: An example of Django Model with defined Data Object 5](#_Toc351211610)

[Figure 2‑3: An example of Django Template binding with HTML Dropdown-List 6](#_Toc351211611)

[Figure 2‑4: Django Admin Interface Example 6](#_Toc351211612)

[Figure 2‑5: An example of Highchart’s Line graph 8](#_Toc351211613)

[Figure 2‑6: An example of Highchart’s Pie Chart 8](#_Toc351211614)

[Figure 2‑7: Illustration of Data Mining Process 9](#_Toc351211615)

[Figure 2‑8: Clustering Concept Illustration 10](#_Toc351211616)

[Figure 2‑9: K-Means Pseudo-Code [9] 12](#_Toc351211617)

[Figure 3‑1: Rapid Application Design Model 14](#_Toc351211618)

[Figure 3‑2: Use Case Diagram 16](#_Toc351211619)

[Figure 3‑3: ER Diagram 18](#_Toc351211620)

[Figure 3‑4: Database Diagram 19](#_Toc351211621)

[Figure 4‑1: Topic Distribution Analysis Steps 20](file:///C:\Users\HUIMIN\Desktop\FYP%20Report%20Draft.docx#_Toc351211622)

[Figure 4‑2: Topic Distribution Result (2010 November Example) 22](#_Toc351211623)

[Figure 4‑3: Topic Trend Analysis Steps (Method 1) 23](file:///C:\Users\HUIMIN\Desktop\FYP%20Report%20Draft.docx#_Toc351211624)

[Figure 4‑4: Topic Trend Analysis Steps (Method 2) 24](file:///C:\Users\HUIMIN\Desktop\FYP%20Report%20Draft.docx#_Toc351211625)

[Figure 4‑5: Topic Trend Analysis Method 1 Result (2008 Nov – 2010 Nov Example) 25](#_Toc351211626)

[Figure 4‑6: Topic Trend Analysis Method 2 Result (2008 Nov – 2010 Nov Example) 25](#_Toc351211627)

[Figure 5‑1: Stemming Keyword 29](file:///C:\Users\HUIMIN\Desktop\FYP%20Report%20Draft.docx#_Toc351211628)

[Figure 5‑2: Keyword Regeneration button 29](#_Toc351211629)

[Figure 5‑3: One-Tag Cloud Visual Display 30](#_Toc351211630)

[Figure 5‑4: An example of Multiple-Tags Cloud (Not Full List) 31](#_Toc351211631)

[Figure 5‑5: Multiple-Tags Cloud Implementation Steps 32](file:///C:\Users\HUIMIN\Desktop\FYP%20Report%20Draft.docx#_Toc351211632)

[Figure 5‑6: A One-Tag Cloud Example for 2010 November Paper 32](#_Toc351211633)

[Figure 5‑7: A Multiple-Tags (two-tags) Cloud Example for 2010 November Paper (Not Full List) 33](#_Toc351211634)

[Figure 5‑8: One-Tag Cloud of Quadratic Functions 34](#_Toc351211635)

[Figure 5‑9: Two-Tag Cloud for Quadratic Equations 35](#_Toc351211636)

[Figure 5‑10: Three-Tag Cloud for Quadratic Equations 35](#_Toc351211637)

[Figure 5‑11: Four-Tag Cloud for Quadratic Equations 36](#_Toc351211638)

[Figure 5‑12: One-Tag Cloud for Indices and Surds 37](#_Toc351211639)

[Figure 5‑13: Two-Tag Cloud for Indices and Surds 37](#_Toc351211640)

[Figure 5‑14: Three-Tag Cloud for Indices and Surds 37](#_Toc351211641)

[Figure 5‑15: One-Tag Cloud for Polynomials 38](#_Toc351211642)

[Figure 5‑16: Two-Tag Cloud for Polynomials 38](#_Toc351211643)

[Figure 5‑17: Three-Tag Cloud for Polynomials 39](#_Toc351211644)

[Figure 5‑18: Four-Tag Cloud for Polynomials 39](#_Toc351211645)

[Figure 5‑19: Five-Tag Cloud for Polynomials 39](#_Toc351211646)

[Figure 5‑20: One-Tag Cloud for Partial Fractions 40](#_Toc351211647)

[Figure 5‑21: Two-Tag Cloud for Partial Fractions 40](#_Toc351211648)

[Figure 5‑22: One-Tag Cloud for Simultaneous Equations 40](#_Toc351211649)

[Figure 5‑23: One-Tag Cloud for Binomial Theorem 41](#_Toc351211650)

[Figure 5‑24: Two-Tag Cloud for Binomial Theorem 41](#_Toc351211651)

[Figure 5‑25: Three-Tag Cloud for Binomial Theorem 41](#_Toc351211652)

[Figure 5‑26: One-Tag Cloud for Exponential and Logarithmic Functions 42](#_Toc351211653)

[Figure 5‑27: Two-Tag Cloud for Exponential and Logarithmic Functions 42](#_Toc351211654)

[Figure 5‑28: Three-Tag Cloud for Exponential and Logarithmic Functions 43](#_Toc351211655)

[Figure 5‑29: Four-Tag Cloud for Exponential and Logarithmic Functions 43](#_Toc351211656)

[Figure 5‑30: Five-Tag Cloud for Exponential and Logarithmic Functions 43](#_Toc351211657)

[Figure 5‑31: One-Tag Cloud for Trigonometry 44](#_Toc351211658)

[Figure 5‑32: Two-Tag Cloud for Trigonometry 45](#_Toc351211659)

[Figure 5‑33: Four-Tag Cloud for Trigonometry 45](#_Toc351211660)

[Figure 5‑34: One-Tag Cloud for Coordinate Geometry 46](#_Toc351211661)

[Figure 5‑35: Two-Tag Cloud for Coordinate Geometry 47](#_Toc351211662)

[Figure 5‑36: Five-Tag Cloud for Coordinate Geometry 48](#_Toc351211663)

[Figure 5‑37: One-Tag Cloud for Differentiation 48](#_Toc351211664)

[Figure 5‑38: Two-Tag Cloud for Differentiation 49](#_Toc351211665)

[Figure 5‑39: Five-Tag Cloud for Differentiation 50](#_Toc351211666)

[Figure 5‑40: One-Tag Cloud for Integration 50](#_Toc351211667)

[Figure 5‑41: Two-Tag Cloud for Integration 51](#_Toc351211668)

[Figure 5‑42: Four-Tag Cloud for Integration 52](#_Toc351211669)

[Figure 5‑43: Five-Tag Cloud for Integration 52](#_Toc351211670)

[Figure 5‑44: One-Tag Cloud for Matrices 53](#_Toc351211671)

[Figure 5‑45: Two-Tag Cloud on Matrices 53](#_Toc351211672)

[Figure 5‑46: Three-Tag Cloud for Matrices 54](#_Toc351211673)

[Figure 5‑47: Four-Tag Cloud for Matrices 54](#_Toc351211674)

[Figure 6‑1: From Document to Vector-Space 56](#_Toc351211675)

[Figure 6‑2: Euclidean Distance Illustration for Two Vectors 56](#_Toc351211676)

[Figure 6‑3: Illustrated Summary of the Procedures from Data to Result 57](#_Toc351211677)

[Figure 7‑1: Tag Search Bar 72](#_Toc351211678)

[Figure 8‑1: Survey Result for View Paper 74](#_Toc351211679)

[Figure 8‑2: Survey Result for Paper Tag Cloud 75](#_Toc351211680)

[Figure 8‑3: Survey Result for Topic Distribution 75](#_Toc351211681)

[Figure 8‑4: Survey Result for Topic Trend 76](#_Toc351211682)

[Figure 8‑5: Survey Result for Topic Tag Cloud 77](#_Toc351211683)

[Figure 8‑6: Survey Result for Question Clustering 77](#_Toc351211684)

[Figure 8‑7: Survey Result for Tag Search 78](#_Toc351211685)

[Figure 8‑8: Survey Result for Overall System Usefulness 78](#_Toc351211686)

[Figure 8‑9: Breakdown of Feature Ratings 79](#_Toc351211687)

[Figure A‑1: Main Landing Page 83](#_Toc351211688)

[Figure A‑2: View Paper Selection Page 84](#_Toc351211689)

[Figure A‑3: View Paper (Paper View) 85](#_Toc351211690)

[Figure A‑4: View Solution Before-and-After 86](#_Toc351211691)

[Figure A‑5: View Paper (Topic View) 87](#_Toc351211692)

[Figure A‑6: Analyzer Main Landing Page 88](#_Toc351211693)

[Figure A‑7: Tag Cloud Selection 89](#_Toc351211694)

[Figure A‑8: Tag Cloud Generated Result Page 90](#_Toc351211695)

[Figure A‑9: Topic Distribution Selection Page 91](#_Toc351211696)

[Figure A‑10: Topic Distribution Result Page 92](#_Toc351211697)

[Figure A‑11: Topic Trend Analysis Selection Page 93](#_Toc351211698)

[Figure A‑12: Topic Trend Analysis Result Page 94](#_Toc351211699)

[Figure A‑13: Question Clustering Selection Page 95](#_Toc351211700)

[Figure A‑14: Question Clustering Result Page 96](#_Toc351211701)

[Figure A‑15: View Concepts Page 97](#_Toc351211702)

[Figure A‑16: Tag Search bar 98](#_Toc351211703)

[Figure A‑17: Search Result (Search Tab) 98](#_Toc351211704)

[Figure A‑18: Search Result (Image Tab) 99](#_Toc351211705)

[Figure A‑19: Admin Question Selection Page 100](#_Toc351211706)

[Figure A‑20: Admin Question Listing Page 101](#_Toc351211707)

[Figure A‑21: Admin Question Form Page 102](#_Toc351211708)

[Figure A‑22: Admin Tag Listing page 103](#_Toc351211709)

[Figure A‑23: Admin Tag Form 104](#_Toc351211710)

[Figure A‑24 Django Admin Table Listing 105](#_Toc351211711)

[Figure A‑25: Django Admin Table Data Listing 106](#_Toc351211712)

[Figure A‑26: Django Admin Form 107](#_Toc351211713)

# List of Tables

[Table 2-1: Mathematical Equation to LaTeX 7](#_Toc351211714)

[Table 2-2: Django-Supported Cloud Services 13](#_Toc351211715)

[Table 3-1: Requirements of Implemented Systems 15](#_Toc351211716)

[Table 3-2: Relational Table Descriptions 17](#_Toc351211717)

[Table 4-1: Topic Distribution of 2010 November 21](#_Toc351211718)

[Table 4-2: Topic Trend Analysis Data Example (2008 Nov – 2010 Nov) 24](#_Toc351211719)

[Table 6-1: Clustering Performance Result for Unclassified 58](#_Toc351211720)

[Table 6-2: Clustering Performance Result for Quadratic Equations 59](#_Toc351211721)

[Table 6-3: Clustering Performance Result for Indices and Surds 59](#_Toc351211722)

[Table 6-4: Clustering Performance Result for Polynomials 60](#_Toc351211723)

[Table 6-5: Clustering Performance Result for Partial Fractions 60](#_Toc351211724)

[Table 6-6: Clustering Performance Result for Simultaneous Equation 61](#_Toc351211725)

[Table 6-7: Clustering Performance Result for Binomial Expansions 61](#_Toc351211726)

[Table 6-8: Clustering Performance Result for Exponential and Logarithmic Functions 62](#_Toc351211727)

[Table 6-9: Clustering Performance Result for Modulus Functions 62](#_Toc351211728)

[Table 6-10: Clustering Performance Result for Trigonometry 63](#_Toc351211729)

[Table 6-11: Clustering Performance Result for Coordinate Geometry 64](#_Toc351211730)

[Table 6-12: Clustering Performance Result for Plane Geometry 64](#_Toc351211731)

[Table 6-13: Clustering Performance Result for Differentiation 65](#_Toc351211732)

[Table 6-14: Clustering Performance Result for Integration 66](#_Toc351211733)

[Table 6-15: Clustering Performance Result for Kinematics 66](#_Toc351211734)

[Table 6-16: Set Language and Notations 67](#_Toc351211735)

[Table 6-17: Clustering Performance Result for Functions 67](#_Toc351211736)

[Table 6-18: Clustering Performance Result for Matrices 68](#_Toc351211737)

[Table 6-19: Clustering Performance Result for Vectors 68](#_Toc351211738)

[Table 6-20: Clustering Performance Result for Permutations and Combinations 69](#_Toc351211739)

[Table 6-21: Clustering Performance Result for Arithmetic Progression and Geometric Progression 69](#_Toc351211740)

[Table 6-22: Clustering Performance Result for Circular Measures 70](#_Toc351211741)

[Table A-1: Tag List for Generic 108](#_Toc351211742)

[Table A-2: Tag List for Set Language and Notations 108](#_Toc351211743)

[Table A-3: Tag List for Functions 108](#_Toc351211744)

[Table A-4: Tag List for Quadratic Functions 109](#_Toc351211745)

[Table A-5: Tag List for Indices and Surds 110](#_Toc351211746)

[Table A-6: Tag List for Polynomials 111](#_Toc351211747)

[Table A-7: Tag List for Simultaneous Equations 112](#_Toc351211748)

[Table A-8: Tag List for Logarithmic and Exponential Functions 112](#_Toc351211749)

[Table A-9: Tag List for Partial Fractions 113](#_Toc351211750)

[Table A-10: Tag List for Binomial Expansions 114](#_Toc351211751)

[Table A-11: Tag List for Modulus Functions 114](#_Toc351211752)

[Table A-12: Tag List for Trigonometry 114](#_Toc351211753)

[Table A-13: Tag List for Coordinate Geometry 116](#_Toc351211754)

[Table A-14: Tag List for Plane Geometry 117](#_Toc351211755)

[Table A-15: Tag List for Differentiation 118](#_Toc351211756)

[Table A-16: Tag List for Integration 120](#_Toc351211757)

[Table A-17: Tag List for Kinematics 122](#_Toc351211758)

[Table A-18: Tag List for Matrices 123](#_Toc351211759)

[Table A-19: Tag List for Vectors 123](#_Toc351211760)

[Table A-20: Tag List for Permutations and Combinations 124](#_Toc351211761)

[Table A-21: Tag List for Arithmetic Progression and Geometric Progression 124](#_Toc351211762)

[Table A-22: Tag List for Circular Measure 124](#_Toc351211763)

# Introduction

## Motivation

In Mathematics course, students or educators usually tackled the subject through textbooks, past year papers, and assessment books. The process can be tedious as significantly much amount of time might be spent in identifying the concepts needed for the question, and countlessly referring to the textbook for guidance. Students and educators also face problem in the format of the question collection which most of them will be most willing to invest in “Paper-based” assessment books and “Topic-based” assessment books for the ease of practising questions in full paper or practising questions based on topic classifications respectively.

Therefore, a web-based system called the MathAnalyzer was developed to enhance the efficiency of the teaching/learning for educators and students. The system has a vast amount of past year questions dated back as old as 1995, and is able to display questions by paper or topic.

This is not all, as the web-based system also comes with the addition of smart analysis tools that have the capabilities of analysing the topic distribution in a specific paper, analysing the topic trend within a range of papers, or tagging system which helps to search for keywords, concepts, and formula for each questions, and question clustering tool for students to select similar questions from clusters.

As such, educators and students can better make use of their time by revising efficiently with this system with the aid of its capabilities.

## Project Objective

The primary objective in this project is to develop an online web-based Mathematics assessment system that can help students in preparing for their examination in a short time.

To achieve the objective, statistical data from past year questions based on their topic focus and mark distribution are extracted and illustrated with visualization tools as historical topic distribution and topic trends for reference to students and educators on possible key topics to focus.

Tags, a metadata technique, would also be used as an approach to classify questions based on relevant keywords, mathematical concepts, and mathematical formulae. Users can then be able to identify patterns of tags via the Tag Cloud, or search for relevant questions using the Tag Search.

Data mining techniques such as clustering will also be explored to suggest similarity of questions to the users.

The scope of the project encompasses the following work:

1. Implement a simple topic trend and topic distribution tool by transforming the raw data of questions’ topic focus and mark weightage into statistical data that can be easily displayed into a visualization tool.
2. Build up a Tag library based on keywords, mathematical concepts, and formulae related to each question and its solution. Tag library should be displayed in a form of Tag Cloud to show the frequency counts for individual or multiple-combined tag(s). A search function should be implemented to search for relevant questions related to the tag(s) being searched.
3. Develop and explore clustering techniques such as K-Means on identifying question similarities.

## Prior Work

This project continues from the previous FYP candidate, Chia Zhou Hiang. The system consisted of the Django Framework implementation of the MathAnalyzer which allows user to browse questions and check solutions, and modifying the questions with the aid of a modification form. A set of mathematics records are stored in a SQL database.

The detailed information of the system was given as of below.

System Architecture

* Django Framework is the core framework for this web-based application project.
* Due to Django’s pure python environment, a Django-based AJAX library called Dajaxice is used as a core for asynchronous communication between Django server-side code and JavaScript code.

Database

* Database was based in MySQLdb which holds the record of Additional Mathematics question and solutions that classified into topics, sub-topics, and papers.
* Records were well-documented and have questions spanned from 1995 to 2010.

Web Application

* Web application was able to view questions by topic or paper, view solutions, and attempt question via a solution checking mechanism.
* The application also consists of a Question Modification section which allows user to modify question without the hassle of modifying it through the database or Django admin page.
* Solution checking functionality was removed due to the redundancy in this project.
* The User Interface (UI) is based on a very simple template from the Internet. Not much work on the UI done previously.

## Report Organization

The entire report is organized into 9 chapters and a brief overview of each chapter is provided below.

Chapter 1 starts with the motivation of the project and also stated the features of the current system. It then states the objective and scope of the project.

Chapter 2 discusses the related work on relevant areas of the implementation. It covers the Django web framework, LaTeX document markup language for Mathematical formulae, HighCharts Javascript visualization tool for graphical charts, Data Mining Process, K-Means Clustering Algorithm, Bootstrap UI framework, and also AppFog deployment.

Chapter 3 talks about the overview of the system. It discusses about the database design and also a brief walkthrough on the system.

Chapter 4 examines the approach of statistical techniques to show Topic Distribution and Topic Trend of the past years’ paper.

Chapter 5 discusses the approach of defining tags and presentation of tags into useful knowledge via Tag Cloud.

Chapter 6 examines the approach of using K-Means clustering algorithm to cluster questions into clusters of similar tags, and evaluates its performance.

Chapter 7 discusses the implementation of Tag Search and its performance in gathering useful knowledge for user using tags as search keyword.

Chapter 8 shows the user evaluation result from a group of users.

Chapter 9 touch on an experimental scenario to predict 2011 paper with existing tools and compare the result with the actual 2011 paper.

And last but not least, chapter 10 concludes the entire project and provides some recommendations on the future work of this project.

# Related Work

## Web Framework (Django)

Django [1] is a Python Web framework that encourages rapid application development, and is also the platform in which the project is built with. Django follows the Model-View-Controller (MVC) (See Figure 2-1) software architecture which encourages code reusability, and is also a tidier architecture for building large project.

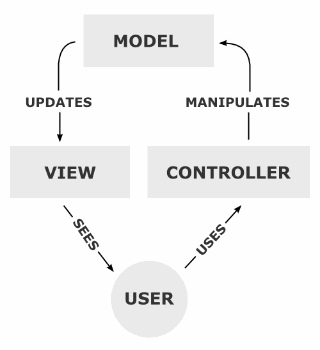


Figure ‑: MVC Design Overview [2]

Using Django context, Model (See Figure 2-2) is a section in which rich and dynamic database API is provided for users to define their data models entirely in Python, and Django then maps it automatically to the database. This further saves the time spent in scripting SQL queries.

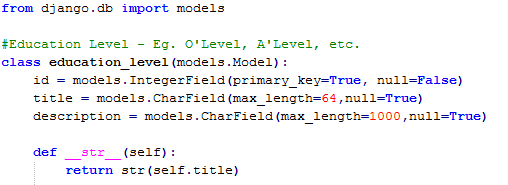


Figure ‑: An example of Django Model with defined Data Object

The View section allows users to define useful methods which requests related data from the Model, and eventually passes the data to the Controller via the powerful Django Template. Django Template is a template language which offers a separation between front-end HTML, and allow users to further express the data presentation to bind to the controllers.

Lastly, the Controller runs on HTML to allow user interactions as shown in Figure 2-3.

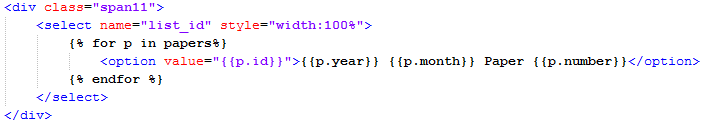


Figure ‑: An example of Django Template binding with HTML Dropdown-List

Django also automatically creates an admin interface for ease of managing data (See Figure 2-4).

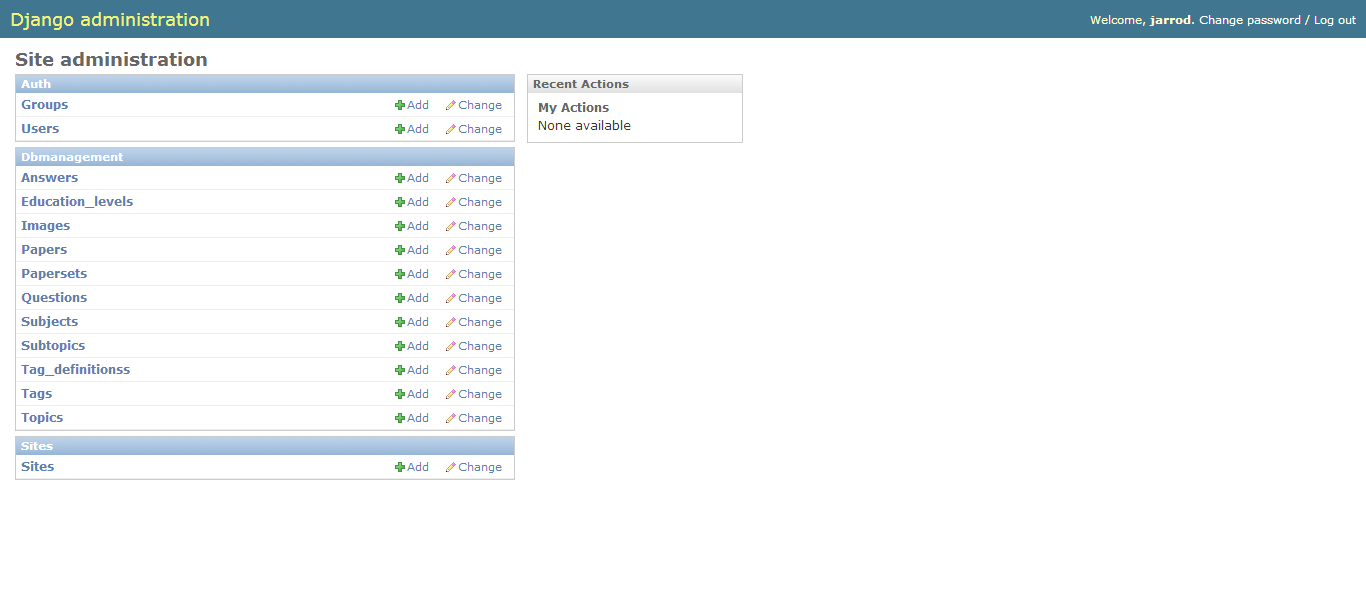


Figure ‑: Django Admin Interface Example

As a result, Django was chosen over other famous web framework like JSP, PHP or ASP.NET.

## Mathematics Formula Typesetter (LaTeX)

LaTeX [3] is a document markup language and also a document preparation system. It is an extension of TeX which also provides high-level of control on how text can be displayed. For this project, the document markup language is the focus as it resolves the issue of showing mathematical formulae in a web page.

An example of a mathematical equation done in LaTeX form is given in Table 2-1.

Table ‑: Mathematical Equation to LaTeX

|  |  |
| --- | --- |
| Before |  |
| After | \sqrt{\sin 2A + \cos^{2}B} |

The LaTeX markup language was implemented in presenting mathematical formulae in questions, answers, and also in some concept and formula tags.

## Visualization Tools (Highcharts)

As Django framework has no built-in API for charts, a third-party package is needed for displaying of Topic Distribution and Topic Trend’s results.

Initially, matplotlib [4], a python API package was chosen as the graphical source for charts, however the API package caused much system dependencies issue when the project was deployed to Cloud server.

Therefore, Highcharts [5] was replaced as the visualization tool for charts since it is a pure Javascript package, which resolves the issue of system dependencies.

Highcharts provides interactive features such as mouseover effect to trigger tooltip to show details of the data point, which matplotlib was lacking. This project implemented Highchart’s line graph and pie chart. Figure 2-5 shows an example of Highchart Line Graph, and Figure 2-6 shows an example of Highchart Pie Chart.

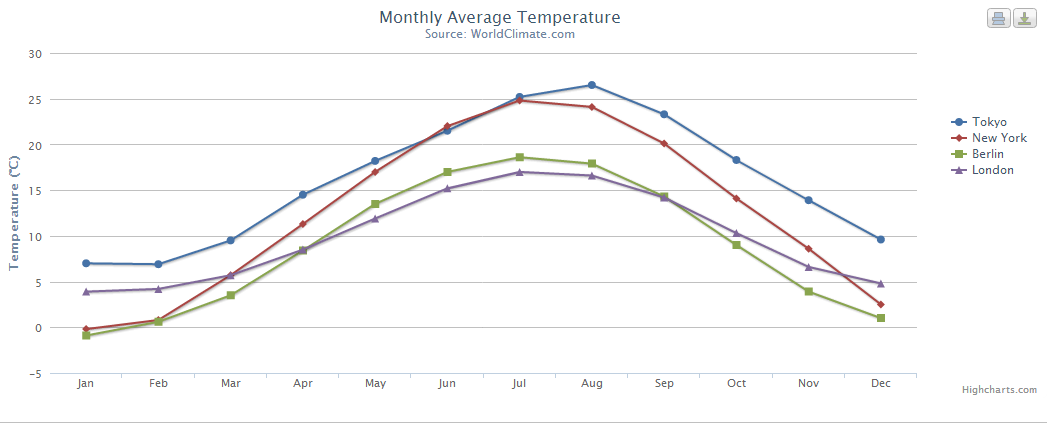


Figure ‑: An example of Highchart’s Line graph

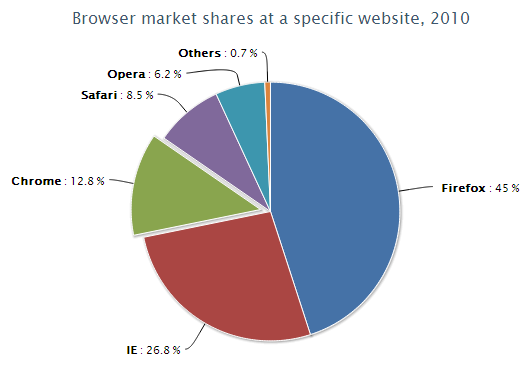


Figure ‑: An example of Highchart’s Pie Chart

## Data Mining Process

The Data Mining Process [6] consists of 5 stages as shown in Figure 2-7.

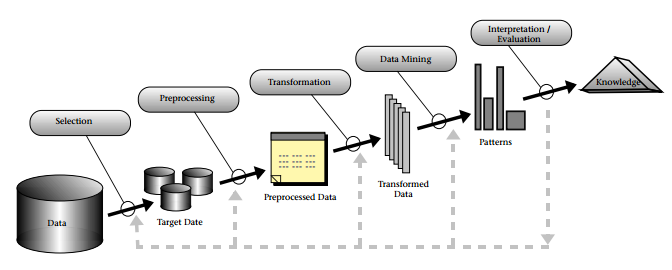


Figure ‑: Illustration of Data Mining Process

Stage 1: Data Selection

Data source can be from physical source (for example, Paper), or multiple sources (for example, data collated from various countries). Therefore, there is a need for data selection by organizing data into a single data set for further pre-processing.

Stage 2: Data Pre-processing

At this stage, data is pre-processed by removing noise and inconsistency. This stage is also known as data cleaning stage.

Stage 3: Data Transformation

After cleaning the data, relevant data attributes are then being selected by data reduction or projection.

Stage 4: Data Mining

Data mining techniques are then applied at this stage to extract useful knowledge.

Data mining techniques can be based on Descriptive Data Mining or Predictive Data Mining. Descriptive method uses human-interpretable pattern to describe the data, whereas Predictive method uses data variables to predict unknown or future values of other data. Descriptive can be Clustering, Association Rule Mining, or Pattern Mining, and Predictive can be Classification, Outlier Detection, or Regression.

For this project, only clustering technique will be experimented.

Stage 5: Knowledge Interpretation

Knowledge is then interpreted in this stage. In some cases, visualization tool might be needed for better interpretation. Performance of the data mining technique is also evaluation, and user might try different techniques and repeat the process for different set of knowledge discovery.

## Clustering Algorithm

Clustering [7] is the approach of finding groups of objects that are similar to each other and are different from objects of other groups. Figure 2-8 illustrates the clustering concept.

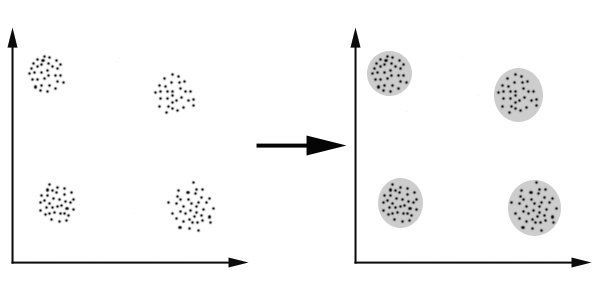


Figure ‑: Clustering Concept Illustration

There are many types of clustering algorithms to date, and some of the famous types are Partitional Clustering and Hierarchical Clustering. Partitional Clustering divides the data objects into the number of clusters defined, and each data object eventually belongs to a cluster. However for Hierarchical Clustering, each object started in a cluster (if there are objects, then there will be clusters). Then objects/clusters will merge with their closest neighbour until one cluster or -clusters are left. For this project, only Partitional Clustering will be used, particularly the K-Means.

Clustering is also described as an unsupervised machine learning technique whereby technique can be applied in an unlabelled dataset.

### K-Means

K-Means [8] is one of the simplest unsupervised learning algorithms that have been famous for a starter algorithm to learn when Clustering comes into picture. The in K-Means would represent the number of clusters to be used as defined by the user at the start of the algorithm.

The algorithm first defined centroids (one for each clusters) in a random location which each new run of the K-Means will generate different results due to the difference of location. The centroid is then recalculated each time and repositioned to the mean point of all the data objects located in the cluster. Algorithm will repeat the repositioning of centroid until no further changes for the location of the centroid.

The repositioning is done using the variance measure called the Sum of Squared Error (SSE), which is given by the following formula:

where is the data point in cluster and is the centroid for cluster (which is the mean of the cluster).

The summary of K-Means algorithm in pseudo-code is shown in Figure 2-9:

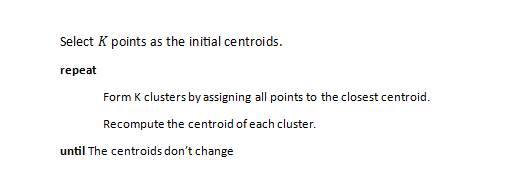


Figure ‑: K-Means Pseudo-Code [9]

K-Means is chosen as the clustering algorithm for this project as it is one of the simplest to implement, and its clustering is generally faster than other algorithms. More about the implementation will be discussed in Chapter 6.

## User Interface (Bootstrap)

Bootstrap [10] is one of the best open-sourced front-end framework for quick-and-easy user interface (UI) design. The Bootstrap package comes with Cascading Style sheet (CSS) template fully built to most needs a web designer will need, and also useful bundle of AJAX (Asynchronous JavaScript and XML) web components. Bootstrap was made not only for desktop, but also a mobile-friendly template that displays well in mobile devices, and inclusive of cross-browser support.

The UI of the project easily revamped into Bootstrap UI within few days of work. More details on the UI of the website can be referred to Appendix A: Web Design and Navigation.

## Deployment (AppFog)

After implementing the system, a suitable deployment server has to be sourced. Cloud services are chosen over the conventional PC hosting (e.g. using the Software Engineering Lab PC that the project was assigned) is the effortless access as Cloud service will be hosted on public domain rather than private domain restriction for Software Engineering Lab’s PC’s case.

Some of the options found for Cloud services are given in Table 2-2:

Table ‑: Django-Supported Cloud Services

|  |  |  |
| --- | --- | --- |
| Service Name | Pros | Cons |
| Appfog | * Up to 8 service instances * Offers MySql GUI support (phpmyadmin) * Wide range of third-party plugins (e.g. XeroundDB, MailGun) * Friendly GUI | * Does not have SSH * Does not have console |
| Pythonanywhere | * Offers built-in console, good for root-access * Pure-Python cloud (no system dependencies issue for Python Packages) | * MySql does not have GUI support like phpmyadmin * Running Insert statements on MySql console takes ages |
| Heroku | * Widely used * Wide range of third-party plugins (e.g. XeroundDB, MailGun) * Friendly GUI | * Database limit to 10K rows (One of the system’s table already has 6K rows) * Does not have SSH * Does not have console |
| Amazon Web Service | * Allows SSH * Widely used | * Immediate billing to credit card for excess usage |

After much considerations, the decision was to use AppFog since SSH and console are not a big issue to do without. Heroku can be a better option if not for its database limit to 10K rows.

# System Overview

## Software Development Lifecycle (SDLC)

Software Development Lifecycle is commonly used as a development model to aid system developers to plan their development cycle.

There are many methodologies that a developer team can follow to best suit their situation. Due to the limited time to develop the system, Rapid Application Development (RAD) [11] (See Figure 3-1) was chosen.

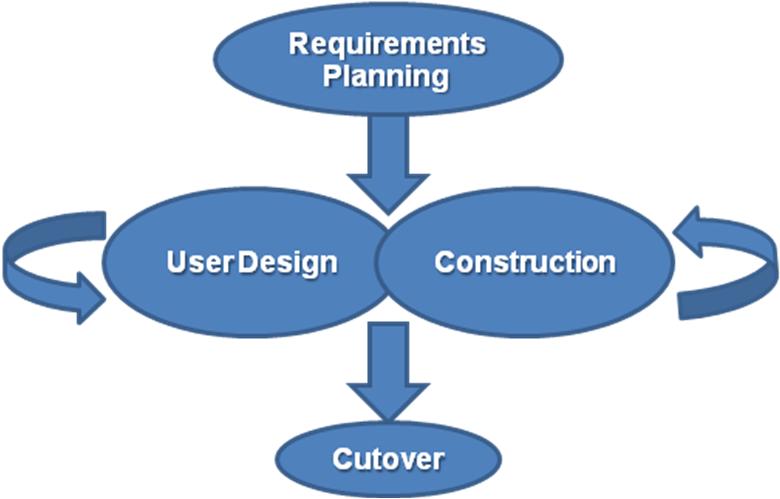


Figure ‑: Rapid Application Design Model

Rapid Application Design uses minimal planning to be in favour of rapid prototyping. With such, there are many iterations on the design and implementation phase throughout the project. The system was built functions by functions and reviewed each time a function was completed. This allows more flexibility on the requirements if there is no concrete requirement decided upfront.

However for report purpose, the subsequent requirements planning and system designs are delivered with the concept of fully built system to allow better understanding of the system.

## Requirement Analysis

Requirement Analysis is the first part of SDLC. During the Requirement Analysis stage, a few discussions were conducted and problems of the previous versions of Mathematical Data Analytical Systems were looked into, before deciding on the features that is going to be part of the system.

Apart from the selected features, other features such as Topic Classification, Topic Trend Predictions were also considered as requirements options.

After much discussion, the decided requirements are given in Table 3-1:

Table ‑: Requirements of Implemented Systems

|  |  |
| --- | --- |
| No. | Requirement Description |
| 1 | User must be able to view questions by Paper and by Topic. |
| 2 | User must be able to view solution from each question. |
| 3 | User must be able to modify questions. |
| 4 | Questions must be tagged with keywords, concepts and formulas to enable visualization of Tag Cloud by Paper or by Topic. |
| 5 | Tags must be able to be searched by a search function. Search function must display by questions and by images. |
| 6 | User must be able to modify existing tags. |
| 7 | User must be able to view Topic Distribution of a selected paper. |
| 8 | User must be able to view Topic Trend of a selected period of time. |
| 9 | User must be able to run a clustering algorithm to view question similarities. |

## System Design

### Use Case Diagram

A use case diagram was used to in the initial system design phase as it is one of the simplest to draw, and best to depict user interactions with the system. Figure 3-2 shows the use case diagram for the implemented system.

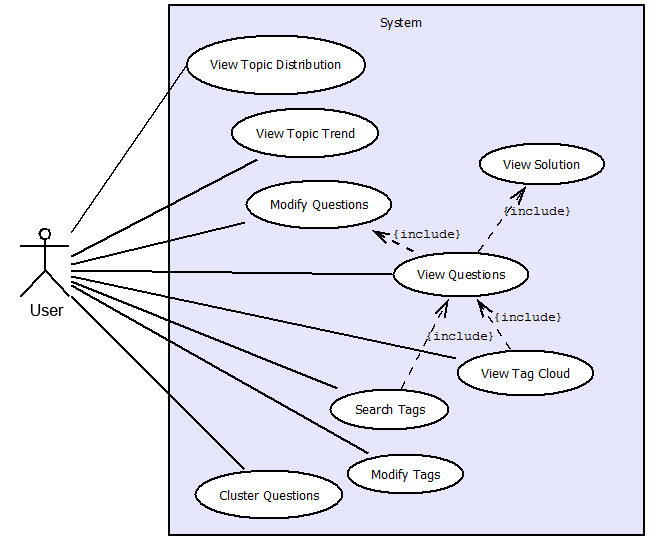


Figure ‑: Use Case Diagram

## Database Design

This section describes the database design adopted for the system. As mentioned earlier in Section 2.1, Django takes care of the database by mapping the object relations defined in Django’s Model file using the models API. The implementation of the model objects will not be discussed, but instead the overview of the design of the database will be the focus.

For this system, there are a total of 11 relational tables. Descriptions of these relational tables are given in Table 3-2. The ER Diagram is given in Figure 3-3. The Database Diagram is given in Figure 3-4.

Table ‑: Relational Table Descriptions

|  |  |
| --- | --- |
| Table Name | Description |
| Education\_Level | Contains the Education level’s data. For example, education level for Additional Mathematics is “GCE ‘O’ Level”. |
| Subject | Contains the subject’s data. |
| Paperset | Contains the paper set’s data. For each subject, there might be multiple examination papers. This table contains the attribute data for the set of examination papers. |
| Paper | Contains the attribute data for each individual examination paper. |
| Topic | Contains the topic data from a certain subject. |
| Subtopic | Contains the subtopic data from a certain topic. |
| Question | Contains the attribute data for each question |
| Answer | Contains the answer from a question. |
| Image | Contains the image path for images used in some questions and answers. |
| Tag\_Definitions | Contains the attribute data of a tag. |
| Tag | Contains the relationship mapping of Question and Tag\_Definitions |

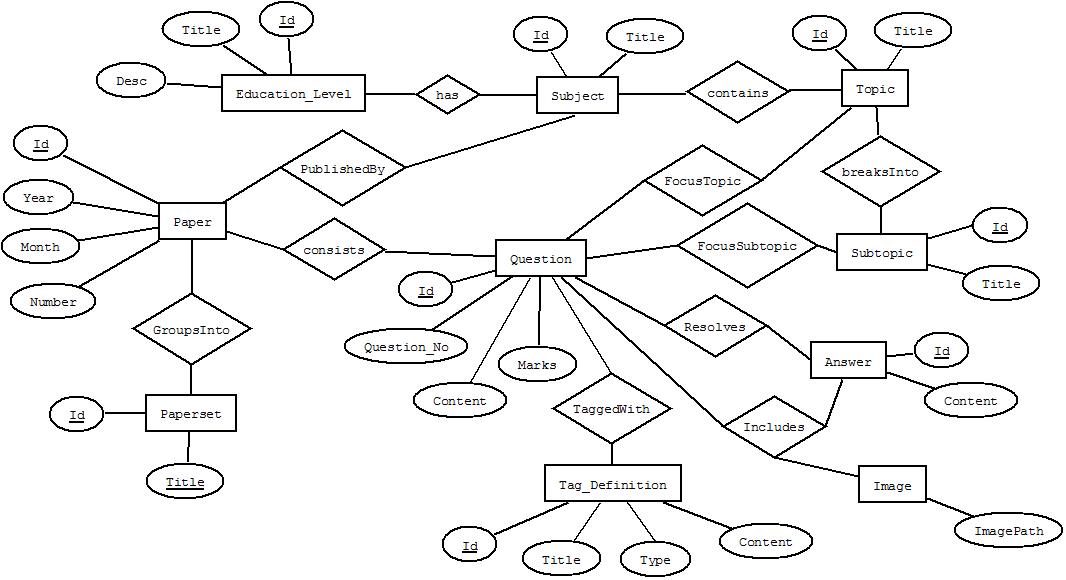


Figure ‑: ER Diagram

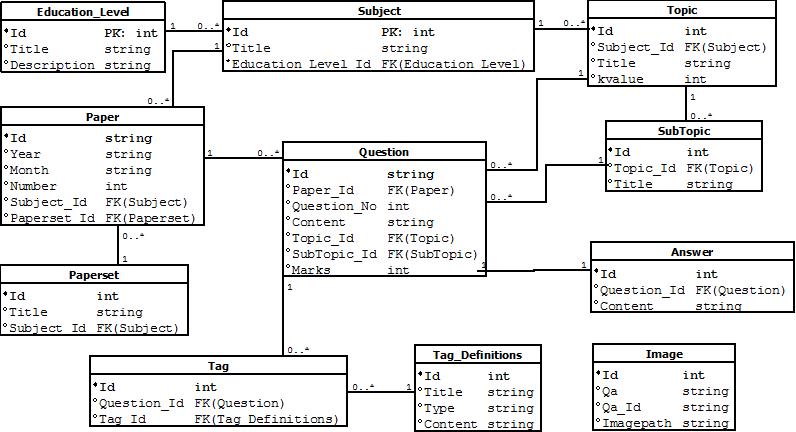


Figure ‑: Database Diagram

# Topic Distribution and Trend Analysis

## Feature Objective

Students and Educators dealing with Mathematics subject often ask themselves on the topics tested from the past years’ examination and wondering about the trend of the topics of the previous years. They resorted to do a manual analysis by looking up the past years’ examination papers and identify the topic for each question, and later plot the data to see the topic distribution and trend analysis.

Therefore, the objective of building a Topic Distribution and Trend Analysis function is to help students and educators to view topic distribution and trend of past years’ paper effectively without much hassle of going through the painful process of manual analysis.

## Approach for Topic Distribution

Topic distribution aims to analyse the topic breakdown for a paper. Therefore, the parameter needed will be the paper set. Additionally, topic distribution can be done in two methods:

1. Topic Distribution by Marks (Method 1)
2. Topic Distribution by Question count (Method 2)

Upon a paper set was selected, question objects related to the paper set are being retrieved as the data object, and each question object contains the topic and also the marks. The steps given in Figure 4-1 described the pre-processing of data before presenting to pie chart.

1. A data array was created, each array element with attributes stored with a unique *topic id*, *topic name*, and *data value* (for topic weightage later). Total array size equates to the total number of topics in the subject.
2. Questions are read one by one and its marks/count is updated to a respective topic object’s *data value*. A variable named “total” also stores the total marks/count from the questions.
3. Upon reading all the questions, each topic is then weighted by percentage of the total value in the variable “total”.

Figure 4‑1: Topic Distribution Analysis Steps

A summary of the percentage weightage is shown as equation below:

where represents the total questions in the topic, represents the total questions in the database, and represents either the marks or the question count data.

Using 2010 November Paper set as example, Table 4-1 shows the topic distributions.

Table ‑: Topic Distribution of 2010 November

|  |  |  |  |
| --- | --- | --- | --- |
|  | Topic Name | Marks | Weight |
| 1 | Quadratic Equations | 13 | 7.22% |
| 2 | Indices and Surds | 8 | 4.44% |
| 3 | Polynomials | 4 | 2.22% |
| 4 | Partial Fractions | 11 | 6.11% |
| 5 | Simultaneous Equations | 0 | 0.00% |
| 6 | Binomial Expansions | 8 | 4.44% |
| 7 | Exponential and Logarithmic Functions | 8 | 4.44% |
| 8 | Modulus Functions | 6 | 3.33% |
| 9 | Trigonometry | 29 | 16.11% |
| 10 | Coordinate Geometry | 28 | 15.56% |
| 11 | Plane Geometry | 10 | 5.56% |
| 12 | Differentiation | 44 | 24.44% |
| 13 | Integration | 0 | 0.00% |
| 14 | Kinematics | 11 | 6.11% |
| 15 | Set Language | 0 | 0.00% |
| 16 | Functions | 0 | 0.00% |
| 17 | Matrices | 0 | 0.00% |
| 18 | Vectors | 0 | 0.00% |
| 19 | Permutations and Combinations | 0 | 0.00% |
| 20 | Arithmetic Progression and Geometric Progression | 0 | 0.00% |
| 21 | Circular Measure | 0 | 0.00% |
|  | Total | 180 | 100% |

An example of the result of topic distribution of the 2010 November paper set is shown in Figure 4-2.

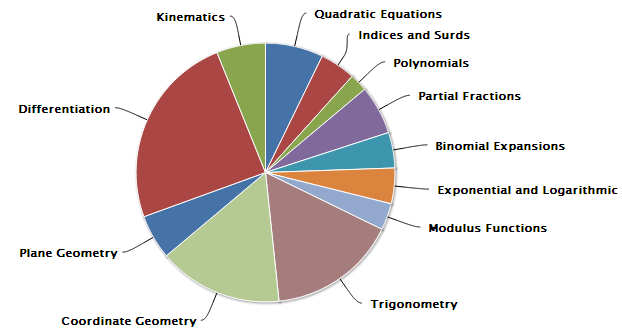


Figure ‑: Topic Distribution Result (2010 November Example)

## Approach for Topic Trend Analysis

For topic trend analysis, a range of paper set must be chosen as the range of analysis. User can choose topic(s) to analyse by checking the topics’ checkboxes. The topic trend can be displayed in two methods:

1. Topic Trend by Percentage (Percentage refers to marks weightage of the topic with regards to total paper’s marks) (Method 1)
2. Topic Trend by Question count (Method 2)

For method 1 (See Figure 4-3), the steps used were similar as the equation used for topic distribution. After data selection was made by retrieving the list of questions from the analysed paper set, the following steps were similar except for the parts in bold.

Figure 4‑3: Topic Trend Analysis Steps (Method 1)

1. A data array was created, each array element with attributes stored with a unique *topic id*, *topic name*, and *data value* (for topic weightage later). Total array size equates to the total number of **related** topics in the subject. **This data array represents a paper set.**
2. A **paper set’s** questions are read one by one and its marks/count is updated to a respective topic object’s *data value*. A variable named “total” also stores the total marks/count from the questions.
3. Upon reading all the questions, each topic is then weighted by percentage of the total value in the variable “total”.
4. **Repeat steps for the next paper set until there is no paper set left.**

A similar percentage weightage formula was also applied:

where represents the total questions in the topic, represents the total questions in the database, and represents either the marks or the question count data.

For method 2 (See Figure 4-4), a variation was that there will be no need for a percentage weightage. Therefore the following steps describe its procedure.

1. A data array was created, each array element with attributes stored with a unique *topic id*, *topic name*, and *data value* (for question count later). Total array size equates to the total number of related topics in the subject. This data array represents a paper set.
2. Each topic will query the *topic id* as the query parameter, inclusive of mandatory parameters such as the *paper set ids* involved for the range. The total questions returned will be the question count, and the *data value* is updated.
3. Repeat steps for the next paper set until there is no paper set left.

Figure 4‑4: Topic Trend Analysis Steps (Method 2)

As such, an example of the topic trend analysis pre-processed data was illustrated in Table 4-2.

Table ‑: Topic Trend Analysis Data Example (2008 Nov – 2010 Nov)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Selected Topics | 2008 | | 2009 | | 2010 | |
| Met. 1 | Met. 2 | Met. 1 | Met. 2 | Met. 1 | Met. 2 |
| Quadratic Equation | 13 (7.22%) | 2 | 6 (3.31%) | 1 | 13 (7.22%) | 2 |
| Trigonometry | 39 (21.66%) | 5 | 30 (16.57%) | 4 | 24 (16.11%) | 3 |
| Differentiation | 41 (22.78%) | 5 | 22 (17.68%) | 3 | 44 (24.44%) | 6 |
| Integration | 6 (3.33%) | 1 | 0 (0%) | 0 | 0 (0%) | 0 |
| Polynomials | 9(5%) | 1 | 4(2.21%) | 1 | 4(2.22%) | 1 |
| Total\* | 24 Qns, 180 marks | | 23 Qns, 181 marks | | 23 Qns, 180 marks | |

\*Note: Total does not equate to the 5 topics’ total. Total is still based on ALL topics total.

Method 1 (By Percentage) Graph Result:

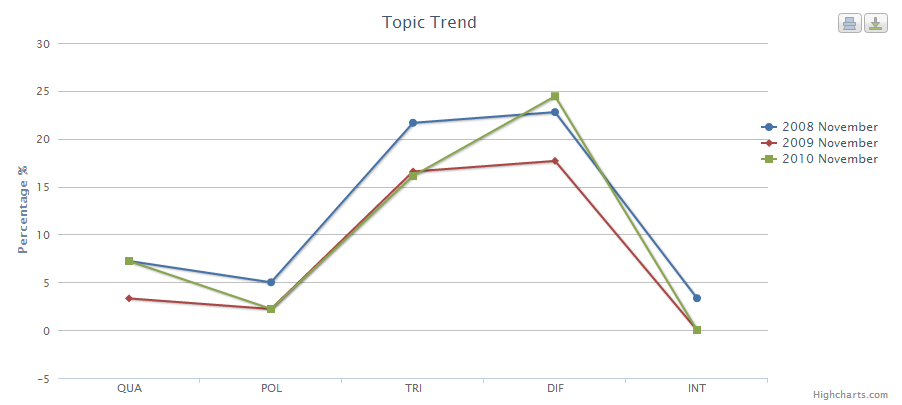


Figure ‑: Topic Trend Analysis Method 1 Result (2008 Nov – 2010 Nov Example)

Method 2 (By Question Count) Graph Result:

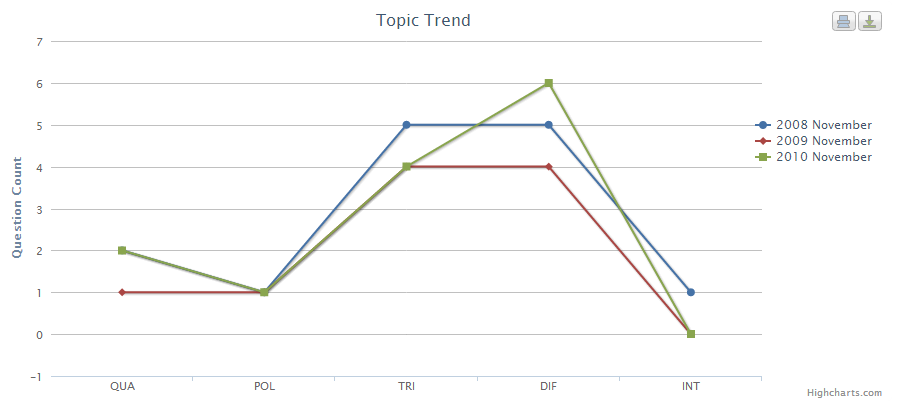


Figure ‑: Topic Trend Analysis Method 2 Result (2008 Nov – 2010 Nov Example)

## Discussions

The implementation of Topic Distribution and Topic Trend Analysis are indeed helpful in some ways such like students and educators no longer need to do a manual breakdown of the distribution or trend.

However, the knowledge extracted is not very novel, as it serves as just an information of the past distribution or trend. As spotted on the recent year distributions based on the 2008 November to 2010 November Topic Trend in Figure 4-6 (on previous page), the distributions vary each year, although the percentage shift is not huge, with about +/- 10% distribution difference for a topic. These percentage shifts will not affect much on a student as it is just a matter of e.g. 1 Differentiation question more, with 1 Trigonometry question deducted.

Nevertheless, having the information of Topic Distribution and Topic Trend is still a basic feature that a Mathematics Analysis system should have.

# Tagging

## Feature Objective

In normal online Mathematics Assessments systems, questions are usually displayed with just a single label of topic classification. The current MathAnalyzer system also comes with this feature when it was built together the basic functions of viewing and attempting questions. However, such topic label does not describe much of the question, and user still need to read through and digest, before understanding what the question is testing about.

Therefore, tags are a new form of labelling the questions to help user understands the question better. The feature’s aim is to implement a tag system with a set of useful tags in which user can understand the question just by reading its list of tags. On a higher level, user can understand what is within a paper or within a topic just by looking at the tag cloud and understanding the terms that appears more frequently.

## Types of Tags

There are three types of tags defined based on Additional Mathematics. The three types are: keyword, concept, and formula.

For keyword tags, they are the basic keywords that were identified as words that are important in describing a particular topic. For example, in “Quadratic Functions” topic, keywords might be “gradient”, “distance of line”, or “parallel”.

For concept and formula tags, they are based on the concepts and formula required for each topic.

## Approach for Implementation of Keyword Tag Data

Tag Data refers to the set of data in which are used as tag terms for Additional Mathematics question dataset.

**Generating of Keyword List**

Keyword Tags are based on the study of questions of each topic, and identifying the keywords which are frequent in appearing on the related questions. Refer to Appendix A-2 for Tag List.

**Stemming of Keywords**

Stemming [12] is a technique widely used in Information Retrieval whereby it is a process of reducing words into their base form, so that words such as “accelerate”, “accelerating”, “acceleration”, and “accelerates” can be grouped into a single base form, “accelerate” in this case. Stemming algorithms are developed for uses in Search Engines and other related Information Retrieval systems. However, for this project, stemming algorithms are not required.

Stemming is simply implemented using Regular Expression (Regex). Using the “accelerate” keyword example, an explanation of the implementation of Regex is described in Figure 5-1.

Base form: accelerate

Other forms: accelerates, accelerating, accelerated, acceleration

[1] accelerat.\*

- A period (.) is a wildcard that matches a single character. However using an asterisk (\*) shows that the single character can be multiple to no limit. Usage of asterisk can be dangerous as uninvited words can also be matched to the base form. An example would be “term.\*”, and an uninvited word would be “termite”.

[2] accelerate.?

- A question mark (?) is a symbol to represent that the period can be zero or one only. Therefore, the range considered will only be “accelerate”, “accelerated” or “accelerates”.

Figure 5‑1: Stemming Keyword

[3] accelerat.{0,3}

- Using {,} with specifying the minimum number of counts, and specifying maximum number of counts, it is one of the safest way to specify the number of characters that is allowed to vary. This form of regex solves the issue of stemming the word “accelerate” and its forms.

Full implementation of keyword stemming can be found in Appendix A-2 Tag List.

**Tagging of Keyword Tags to Questions**

Question tagging of keywords are done automatically by going to “Settings > Modify Tags” section and click on “Regenerate Keywords” button and the system will retrieve the list of tag keywords and tag each question if the keyword is found. The button should be triggered each time the user added a new keyword to the system to ensure the tag is applied to the question by regenerating the tags.

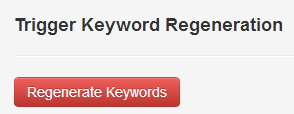


Figure ‑: Keyword Regeneration button

## Approach for Implementation of Concept and Formula Tag Data

**Generating the Concept and Formula List**

Concept and Formula Tag data are gathered by a manual mining of the New Syllabus Additional Mathematics 8th Edition [13] book and sieved out the important concepts and formulae that defines each topic.

**Tagging the Concept and Formula Tags to Questions**

Tagging of questions is not done manually as tagging concepts requires human inference of the question to determine the concepts involved. Tagging of formulae can be done automatically by regular expression, but the variation of formula is too great to conclude a set of fool-proof regular expressions. Due to the limitation of time for this project, a manual entry is chosen.

## Approach for Implementation of Tag Cloud Visualization

Tag Cloud is a common feature for tag visualization. It allows user to view the collection of tags and their frequencies in a glance. This section will describe the types of Tag Clouds implemented.

**One-Tag Cloud**

One-tag cloud refers to tag cloud displaying tags as a single term. Each tag exists as an individual term and the number of items being tagged (frequency) is being displayed. Figure 5-3 shows an example of One-tag cloud.

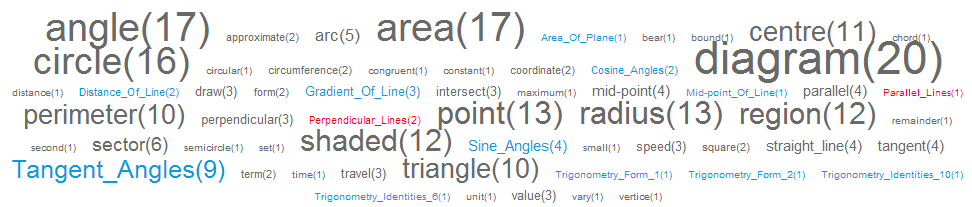


Figure ‑: One-Tag Cloud Visual Display

The font size is controlled by the frequency of the tag term, and the colour differs by the type of the tag. Grey represents Keyword Tags, Red represents Concept Tags, and Blue represents Formula Tags.

The typography of Tag Cloud was inspired from a source from Wikipedia on Creation of a Tag Cloud [14]. Below is the formula to determine the font size of each tag terms in a tag cloud.

where refers to display font size, refers to maximum font size, refers to minimum font size, refers to current term’s frequency, refers to maximum recorded term frequency, and refers to minimum recorded term frequency.

**Multiple-Tags Cloud**

Multiple-Tags Cloud refers to Tag Cloud with more than one tag term combined to show the number of items being tagged with the combination of the tag terms. For this project, a maximum combination allowed for the Tags was limited to 5.

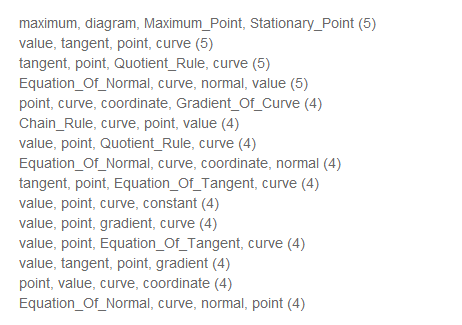


Figure ‑: An example of Multiple-Tags Cloud (Not Full List)

From Figure 5-4, it is obvious that the implementation of the multiple-tags cloud was different from the one-tag cloud. This is because that the one-tag cloud implementation will be very messy if the terms are long for the case of multiple-tags. Therefore, a list view display was preferred.

The implementation steps were described in Figure 5-5.

Figure 5‑5: Multiple-Tags Cloud Implementation Steps

1. Python’s Counter [18] class was assigned to each related tag. Counter object will then store the set of questions that was tagged to the tag.
2. Combinations [19] method was then called to compute the combinations for the tags with an integer parameter setting the size of combinations (number of tag combinations for the tag cloud).
3. The result was then sorted and the top 50 was displayed if the result was too long.

## Discussions

Apart from the usefulness of Tagging system on describing the question, the main objective is the knowledge discovery from the Tag Clouds after the implementation was done. On this section, the knowledge discovered will be highlighted in the discussion.

### Paper-Based Tag Cloud Findings

Using one-tag cloud, the information given is subtle and the more frequent ones are mostly keywords, which are understandable that the keywords may overlap on questions of different topics. The concepts and formulae are mostly frequency of 1 or 2, which is understandable that examiners are trying to test as much concepts and formulae rather than repeating testing on the same concepts and formulae.

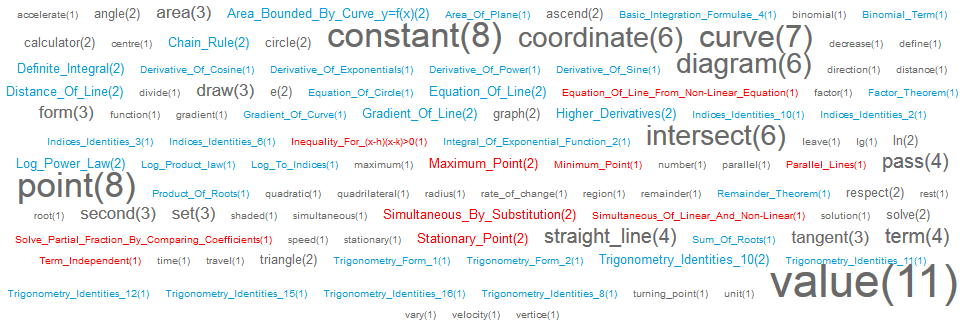


Figure ‑: A One-Tag Cloud Example for 2010 November Paper

Therefore, since the frequency of terms in one-tag cloud is already low, it is hard to find notable information from the multiple-tags cloud. From Figure 5-7, a two-tag cloud was generated from 2010 November paper, and it is clear that the frequency of terms are low, with 4 counts and 3 counts dominated by keywords describing possible chapters on “Coordinate Geometry”, “Differentiation” and “Integration”. The other notable information is that 2 formulae (Area\_Bounded\_By\_Curve\_y=f(x), Definite\_Integral) are having a frequency count of 2.



Figure ‑: A Multiple-Tags (two-tags) Cloud Example for 2010 November Paper (Not Full List)

The rest of the papers are not elaborated as the knowledge discovered is similar. Therefore, it is concluded that Paper-Based Tag Cloud is not a useful tool for data analysis, as there is no need for user to understand the concepts or formulae tested in that particular paper since the only patterns that the user will discover is that keywords such as “curve”, “point”, “coordinate”, “intersect” are commonly found in every paper, and familiar concepts and formulae from topics such as Differentiation, Integration, Coordinate Geometry, Simultaneous Equations, or Quadratic Functions are commonly appearing with frequency count of 1 to 2.

### Topic-Based Tag Cloud Findings

Topic-Based Tag Clouds are interesting in showing all the keywords, concepts and formulae in a nutshell to the user. In this section, the evaluation of findings was done with “Integration” topic as example. The rest of the topics can be referred in Appendix A-3.

**Integration**



Figure ‑: One-Tag Cloud for Integration

From the one-tag cloud in Figure 5-8, the top most common terms are area bounded by curve and definite integral. These two are usually an application question on finding area under curve, which uses definite integral. Therefore, the one-tag cloud shows the popularity on this type of question. Next, the other integration formulae are also very commonly used. This could be representing questions which ask for solution on finding integral of an equation. Differentiation formulae are also a commonly found, each with frequency between 3 to 9 counts. Differentiation questions on stationary points were also found to be involved with Integration. Trigonometry formulae are also used in minority questions.

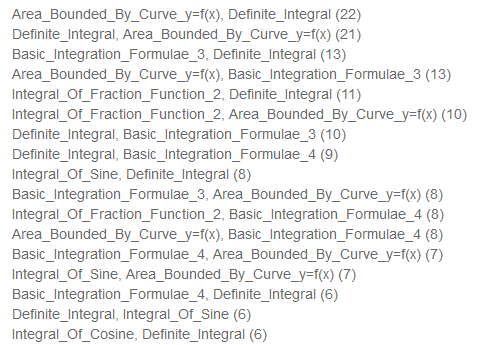


Figure ‑: Two-Tag Cloud for Integration

From the two-tag cloud in Figure 5-9, the Integration formulae still stand as the major formulae used in questions. The figure above showed that there are no sign of foreign formulae or concept in the top few most common terms in the list. The list was an incomplete generation, but the list was enough to show that the Integration formulae and concepts still stand as major setup of the question.

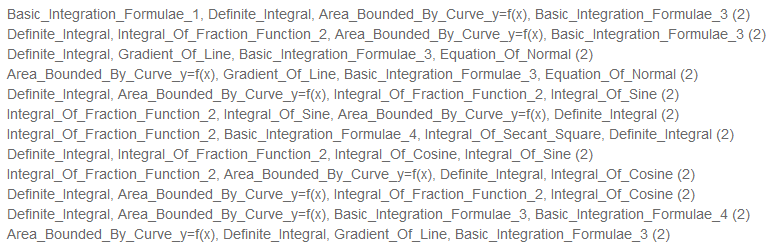


Figure ‑: Four-Tag Cloud for Integration

From the four-tag cloud in Figure 5-10, the Integration formulae still maintain its popularity in questions that are more complicated as assumed by the number of formulae used. Almost no foreign topic formulae are used on the top few most common in this list.

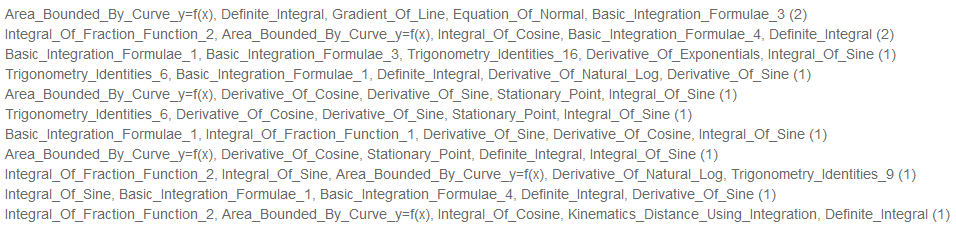


Figure ‑: Five-Tag Cloud for Integration

However on the five-tag cloud in Figure 5-11, more foreign topic formulae started to appear. This could be because that there are minor number of questions which have heavy marks distribution (12 – 16 marks), used a part-question to test foreign topic concept such as Differentiation or Trigonometry. Therefore, it is conclusive that Integration questions are usually pure on its concepts and formulae, with Area under curve and Definite Integral standing as major question format, but other Integration formulae are not neglected. On certain circumstance of question, a small portion of foreign concepts such as Differentiation and Trigonometry will be tested.

**Summary of Discussion**

After a detailed analysis of the Tag clouds and uses, it is concluded that the Paper Tag Cloud is not very useful in extracting useful knowledge of the tags. The tags only posed as a sign in which the topic exists in the paper, but do not show much on the benefits of the tags.

The Topic Tag Cloud however showed potentials on its benefits to aid study on Additional Mathematics. Students can instantly know the rough importance on each topic by looking at the frequency of formulae or concepts appearance on the past years’ records. Students can also gain insights on possible foreign topic’s formulae or concepts which might be needed in as well to solve the particular topic. This feature definitely can be improved to be more informative to the students. More on the future work suggestion on Topic Tag Cloud will be elaborated on Chapter 10.2.

# Clustering for Question Similarity

## Feature Objective

Clustering is a well-known technique in Data Mining field to find the similarities of data and grouped them into clusters. Due to the fact that data can be grouped into clusters based on similarities, it struck as an idea that questions in a Topic can be passed as data into the clustering algorithm to group questions into clusters. The best possible result achieved would show the different sets of question patterns formed into clusters. Therefore, students and educators will be boosted by identifying the different question patterns on each topic.

## Approach for Data Pre-processing

The data that were used for clustering are the tag terms that describe each question. Initially, only keyword tags were used for clustering. However since the concept and the formula tags better describes the question, it is also utilized in the clustering process as well.

As mentioned earlier in Chapter 2.5, the Clustering algorithm used will be K-Means. K-Means uses centroids that are distributed randomly to different locations and the data points will be compared to find the nearest cluster to be assigned to. Centroids are then repositioned to the centre of the data points until there are no movement to the centroids.

For the representing position of data points, the Vector-Space Model is used.

### Vector-Space Model

Vector-Space Model [15] is a technique used mostly in Information Retrieval mostly for comparing the similarity of documents. Therefore, each question is considered as a “document”. Vector-Space Model uses a Term-Frequency Table whereby each row on the table represents a document vector. The columns of Term-Frequency Table will be filled with all the Tags that exist in the system, and each document vector will represent a question with its column values representing whether the Tag term exists in the question. An example of the population of Term-Frequency Table is illustrated in Figure 6-1.

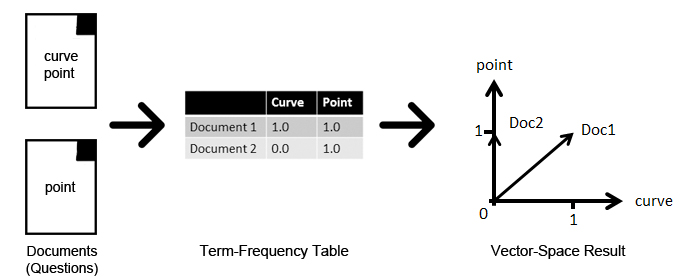


Figure ‑: From Document to Vector-Space

The illustration above was however a simplified example with just 2 terms. The actual versions will be dealing with over 200+ terms, which means the Vector-Space will be plotted into over 200+ dimensions. Therefore, an algorithm for comparing the document distance in the dimensions must be used. In this case, the Euclidean Distance was used.

### Euclidean Distance

Euclidean Distance [16] is the straight-line distance between two vectors. Supposedly, vector and are two vectors with dimensions, the Euclidean Distance can be derived by

where is the number of dimensions, and and are the kth attribute of data objects and respectively. Figure 6-2 shows an illustration of Euclidean Distance, , in a 2-D vector space.

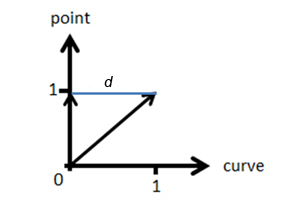


Figure ‑: Euclidean Distance Illustration for Two Vectors

### From Data Pre-Processing to Application of Clustering Algorithm

This section will describe the process of pre-processing of data to document vectors and the application process of K-Means clustering algorithm with the document vectors to get the clusters result.

User needs to define the topic in which to conduct the clustering on. After defining the topic, the relevant questions from the topic are extracted. The term-frequency table was formed by checking for the existence of each Tag term from the question’s tag list with the use of the complete list of existing tag list from the database. Since it is only checking for existence, 1 will represent Yes, and 0 will then represent No.

Once the term-frequency table was filled completely with all the relevant questions, a k-value used as the number of centroids to be generated and the K-Means algorithm was run. This k-value can either be defined by user, or from the database which kept the k-value that will get the optimal result. More on the selection of K-value will be evaluated later on Chapter 6.3.1.

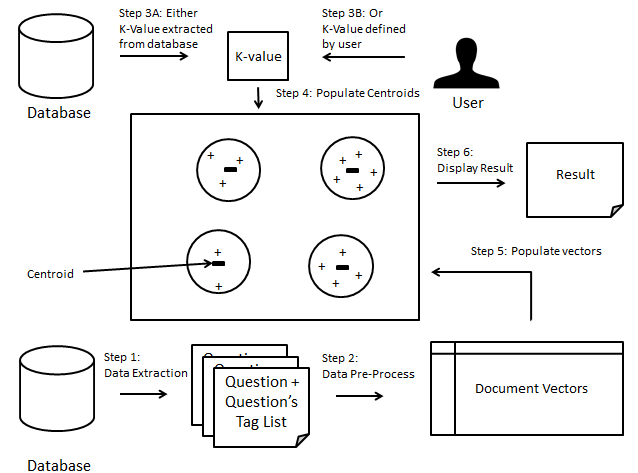


Figure ‑: Illustrated Summary of the Procedures from Data to Result

## Performance Analysis

### Analysis of Best K-Value (Number of Clusters)

Due to the fact that every topic may have different number of clusters in which generates the most number of good clusters, an experiment was conducted to find the performance of each topic against different number of clusters. The result of the experiment will get the best k-value for the topic, and also the best average accuracy of clustering the questions.

The experiment was conducted over 5 runs as each run uses a random set of centroids due to the K-Means clustering algorithm. Therefore, 5 runs are used and the average accuracy calculated to determine the best result.

The Accuracy was calculated by:

**Topic Name: Unclassified**

Number of Questions: 7

Unique Tags: 31

Best K-Value: 3

Concept and Formula Tags applied: No

Table ‑: Clustering Performance Result for Unclassified

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 50.00% | 50.00% | 50.00% | 50.00% | 50.00% |
| 3 | 100.00% | 100.00% | 33.33% | 33.33% | 66.66% | 66.66% |
| 4 | 50.00% | 0.00% | 50.00% | 50.00% | 50.00% | 40.00% |
| 5 | 40.00% | 40.00% | 40.00% | 40.00% | 40.00% | 40.00% |

**Topic Name: Quadratic Equations**

Number of Questions: 43

Unique Tags: 39

Best K-Value: 9

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Quadratic Equations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 50.00% | 0.00% | 50.00% | 50.00% | 30.00% |
| 3 | 33.33% | 66.67% | 66.67% | 0.00% | 33.33% | 40.00% |
| 4 | 50.00% | 50.00% | 75.00% | 50.00% | 75.00% | 60.00% |
| 5 | 60.00% | 80.00% | 60.00% | 60.00% | 60.00% | 64.00% |
| 6 | 66.67% | 66.67% | 50.00% | 66.67% | 83.33% | 66.67% |
| 7 | 57.14% | 85.71% | 57.14% | 57.14% | 71.43% | 65.71% |
| 8 | 62.50% | 50.00% | 50.00% | 50.00% | 50.00% | 52.50% |
| 9 | 66.67% | 77.78% | 55.56% | 66.67% | 77.78% | 68.89% |
| 10 | 50.00% | 60.00% | 60.00% | 40.00% | 50.00% | 52.00% |
| 11 | 54.54% | 45.46% | 63.63% | 27.28% | 45.46% | 47.27% |
| 12 | 58.33% | 58.33% | 58.33% | 58.33% | 33.33% | 53.33% |

**Topic Name: Indices and Surds**

Number of Questions: 19

Unique Tags: 32

Best K-Value: 8

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Indices and Surds

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 0.00% | 50.00% | 0.00% | 10.00% |
| 3 | 0.00% | 33.33% | 33.33% | 33.33% | 33.33% | 26.66% |
| 4 | 0.00% | 25.00% | 25.00% | 50.00% | 50.00% | 30.00% |
| 5 | 20.00% | 20.00% | 40.00% | 0.00% | 20.00% | 20.00% |
| 6 | 33.33% | 16.67% | 33.33% | 16.67% | 16.67% | 23.33% |
| 7 | 42.85% | 42.85% | 28.57% | 42.85% | 28.57% | 37.14% |
| 8 | 37.50% | 50.00% | 37.50% | 50.00% | 25.00% | 40.00% |
| 9 | 44.44% | 22.22% | 44.44% | 33.33% | 11.11% | 31.11% |
| 10 | 40.00% | 10.00% | 30.00% | 30.00% | 20.00% | 26.00% |

**Topic Name: Polynomials**

Number of Questions: 26

Unique Tags: 37

Best K-Value: 7

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Polynomials

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 50.00% | 50.00% | 0.00% | 50.00% | 40.00% |
| 3 | 66.67% | 33.33% | 66.67% | 33.33% | 33.33% | 46.67% |
| 4 | 50.00% | 75.00% | 75.00% | 75.00% | 0.00% | 55.00% |
| 5 | 80.00% | 60.00% | 80.00% | 60.00% | 40.00% | 64.00% |
| 6 | 66.67% | 50.00% | 50.00% | 83.33% | 66.67% | 63.33% |
| 7 | 85.71% | 57.14% | 71.43% | 57.14% | 71.43% | 68.57% |
| 8 | 37.50% | 50.00% | 62.50% | 75.00% | 75.00% | 60.00% |
| 9 | 44.44% | 55.56% | 66.67% | 66.67% | 66.67% | 60.00% |
| 10 | 60.00% | 50.00% | 50.00% | 70.00% | 50.00% | 56.00% |

**Topic Name: Partial Fractions**

Number of Questions: 3

Unique Tags: 11

Best K-Value: 2

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Partial Fractions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 0.00% | 50.00% | 50.00% | 0.00% | 30.00% |
| 3 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |

**Topic Name: Simultaneous Equation**

Number of Questions: 11

Unique Tags: 11

Best K-Value: 5

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Simultaneous Equation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 50.00% | 0.00% | 50.00% | 0.00% | 30.00% |
| 3 | 33.33% | 33.33% | 33.33% | 0.00% | 33.33% | 26.66% |
| 4 | 50.00% | 25.00% | 50.00% | 25.00% | 50.00% | 40.00% |
| 5 | 60.00% | 40.00% | 40.00% | 40.00% | 40.00% | 44.00% |
| 6 | 33.33% | 50.00% | 33.33% | 16.67% | 16.67% | 30.00% |

**Binomial Expansions**

Number of Questions: 24

Unique Tags: 11

Best K-Value: 6

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Binomial Expansions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 50.00% | 0.00% | 50.00% | 20.00% |
| 3 | 66.67% | 0.00% | 66.67% | 33.33% | 33.33% | 40.00% |
| 4 | 50.00% | 75.00% | 100.00% | 75.00% | 50.00% | 70.00% |
| 5 | 80.00% | 60.00% | 80.00% | 60.00% | 80.00% | 72.00% |
| 6 | 83.33% | 100.00% | 50.00% | 50.00% | 83.33% | 73.33% |
| 7 | 85.71% | 71.43% | 57.14% | 71.43% | 71.43% | 71.43% |
| 8 | 50.00% | 62.50% | 87.50% | 62.50% | 50.00% | 62.50% |
| 9 | 55.56% | 44.44% | 44.44% | 44.44% | 44.44% | 46.66% |

**Topic Name: Exponential and Logarithmic Functions**

Number of Questions: 34

Unique Tags: 50

Best K-Value: 8

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Exponential and Logarithmic Functions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 0.00% | 50.00% | 0.00% | 10.00% |
| 3 | 0.00% | 33.33% | 33.33% | 66.67% | 66.67% | 40.00% |
| 4 | 25.00% | 50.00% | 50.00% | 50.00% | 50.00% | 45.00% |
| 5 | 60.00% | 40.00% | 60.00% | 40.00% | 60.00% | 52.00% |
| 6 | 50.00% | 33.33% | 50.00% | 33.33% | 50.00% | 43.33% |
| 7 | 42.86% | 42.86% | 57.14% | 28.57% | 42.86% | 42.86% |
| 8 | 62.50% | 50.00% | 87.50% | 62.50% | 50.00% | 62.50% |
| 9 | 44.44% | 55.56% | 44.44% | 55.56% | 44.44% | 48.89% |
| 10 | 40.00% | 70.00% | 60.00% | 30.00% | 30.00% | 46.00% |

**Topic Name: Modulus Functions**

Number of Questions: 13

Unique Tags: 24

Best K-Value: 6

Concept and Formula Tags applied: No

Table ‑: Clustering Performance Result for Modulus Functions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 50.00% | 0.00% | 0.00% | 0.00% | 20.00% |
| 3 | 33.33% | 0.00% | 0.00% | 33.33% | 33.33% | 20.00% |
| 4 | 50.00% | 50.00% | 25.00% | 25.00% | 75.00% | 45.00% |
| 5 | 60.00% | 80.00% | 40.00% | 40.00% | 20.00% | 48.00% |
| 6 | 33.33% | 66.67% | 83.33% | 50.00% | 50.00% | 56.67% |
| 7 | 57.14% | 42.86% | 42.86% | 42.86% | 28.57% | 42.86% |
| 8 | 25.00% | 37.50% | 25.00% | 25.00% | 25.00% | 27.50% |

**Topic Name: Trigonometry**

Number of Questions: 73

Unique Tags: 76

Best K-Value: 11

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Trigonometry

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 50.00% | 50.00% | 0.00% | 20.00% |
| 3 | 33.33% | 33.33% | 0.00% | 0.00% | 0.00% | 13.33% |
| 4 | 50.00% | 25.00% | 50.00% | 25.00% | 0.00% | 30.00% |
| 5 | 40.00% | 20.00% | 60.00% | 20.00% | 20.00% | 32.00% |
| 6 | 0.00% | 33.33% | 16.67% | 33.33% | 16.67% | 20.00% |
| 7 | 28.57% | 14.29% | 14.29% | 28.57% | 42.86% | 25.72% |
| 8 | 37.50% | 37.50% | 37.50% | 37.50% | 25.00% | 35.00% |
| 9 | 22.22% | 33.33% | 33.33% | 55.56% | 44.44% | 37.78% |
| 10 | 40.00% | 30.00% | 50.00% | 30.00% | 40.00% | 38.00% |
| 11 | 45.45% | 54.54% | 45.45% | 45.45% | 36.36% | 45.45% |
| 12 | 33.33% | 41.67% | 33.33% | 33.33% | 33.33% | 35.00% |
| 13 | 46.15% | 23.07% | 61.54% | 38.46% | 30.77% | 40.00% |
| 14 | 35.71% | 28.57% | 35.71% | 50.00% | 35.71% | 37.14% |

**Topic Name: Coordinate Geometry**

Number of Questions: 65

Unique Tags: 81

Best K-Value: 11

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Coordinate Geometry

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 3 | 33.33% | 0.00% | 0.00% | 33.33% | 0.00% | 13.33% |
| 4 | 0.00% | 50.00% | 25.00% | 25.00% | 0.00% | 20.00% |
| 5 | 0.00% | 20.00% | 20.00% | 20.00% | 20.00% | 16.00% |
| 6 | 50.00% | 50.00% | 33.33% | 33.33% | 16.67% | 36.67% |
| 7 | 57.14% | 42.86% | 28.57% | 42.86% | 28.57% | 40.00% |
| 8 | 25.00% | 37.50% | 50.00% | 25.00% | 50.00% | 37.50% |
| 9 | 44.44% | 22.22% | 44.44% | 33.33% | 55.56% | 40.00% |
| 10 | 50.00% | 40.00% | 70.00% | 40.00% | 60.00% | 52.00% |
| 11 | 54.54% | 45.45% | 63.64% | 45.45% | 54.54% | 52.72% |
| 12 | 50.00% | 58.33% | 58.33% | 50.00% | 41.67% | 51.67% |
| 13 | 69.23% | 46.15% | 38.46% | 46.15% | 38.46% | 47.69% |
| 14 | 50.00% | 21.24% | 50.00% | 64.29% | 57.14% | 48.53% |
| 15 | 46.67% | 53.33% | 40.00% | 53.33% | 33.33% | 45.33% |

**Topic Name: Plane Geometry**

Number of Questions: 25

Unique Tags: 59

Best K-Value: 5

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Plane Geometry

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 3 | 0.00% | 0.00% | 33.33% | 33.33% | 0.00% | 13.33% |
| 4 | 25.00% | 50.00% | 50.00% | 0.00% | 25.00% | 30.00% |
| 5 | 60.00% | 40.00% | 40.00% | 20.00% | 40.00% | 40.00% |
| 6 | 50.00% | 16.67% | 16.67% | 33.33% | 33.33% | 30.00% |
| 7 | 14.29% | 14.29% | 42.86% | 28.57% | 28.57% | 25.72% |

**Topic Name: Differentiation**

Number of Questions: 109

Unique Tags: 100

Best K-Value: 15

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Differentiation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 3 | 33.33% | 0.00% | 0.00% | 0.00% | 33.33% | 13.33% |
| 4 | 0.00% | 0.00% | 0.00% | 50.00% | 0.00% | 10.00% |
| 5 | 40.00% | 20.00% | 0.00% | 20.00% | 0.00% | 16.00% |
| 6 | 33.33% | 0.00% | 16.67% | 16.67% | 0.00% | 13.33% |
| 7 | 0.00% | 14.29% | 14.29% | 14.29% | 0.00% | 8.57% |
| 8 | 12.50% | 37.50% | 12.50% | 12.50% | 0.00% | 15.00% |
| 9 | 11.11% | 22.22% | 22.22% | 0.00% | 11.11% | 13.33% |
| 10 | 0.00% | 40.00% | 20.00% | 30.00% | 40.00% | 26.00% |
| 11 | 18.18% | 36.36% | 36.36% | 27.27% | 27.27% | 29.09% |
| 12 | 33.33% | 0.00% | 25.00% | 16.67% | 25.00% | 20.00% |
| 13 | 38.46% | 30.77% | 46.15% | 23.07% | 38.46% | 35.38% |
| 14 | 21.42% | 50.00% | 42.86% | 28.57% | 35.71% | 35.71% |
| 15 | 46.67% | 40.00% | 33.33% | 33.33% | 46.67% | 40.00% |
| 16 | 31.25% | 25.00% | 25.00% | 43.75% | 31.25% | 31.25% |
| 17 | 41.18% | 35.29% | 35.29% | 41.18% | 35.29% | 37.65% |
| 18 | 33.33% | 44.44% | 27.78% | 22.22% | 38.89% | 33.33% |

**Topic Name: Integration**

Number of Questions: 57

Unique Tags: 80

Best K-Value: 6

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Integration

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 50.00% | 0.00% | 50.00% | 0.00% | 20.00% |
| 3 | 66.67% | 33.33% | 33.33% | 33.33% | 33.33% | 40.00% |
| 4 | 25.00% | 75.00% | 50.00% | 50.00% | 25.00% | 45.00% |
| 5 | 20.00% | 60.00% | 60.00% | 20.00% | 60.00% | 44.00% |
| 6 | 66.67% | 66.67% | 50.00% | 66.67% | 66.67% | 63.34% |
| 7 | 28.57% | 42.86% | 57.14% | 71.43% | 28.57% | 45.71% |
| 8 | 50.00% | 62.50% | 50.00% | 37.50% | 50.00% | 50.00% |

**Topic Name: Kinematics**

Number of Questions: 37

Unique Tags: 46

Best K-Value: 8

Concept and Formula Tags applied: No

Table ‑: Clustering Performance Result for Kinematics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 3 | 0.00% | 33.33% | 33.33% | 0.00% | 33.33% | 20.00% |
| 4 | 25.00% | 50.00% | 25.00% | 50.00% | 25.00% | 35.00% |
| 5 | 40.00% | 20.00% | 60.00% | 40.00% | 40.00% | 40.00% |
| 6 | 66.67% | 33.33% | 50.00% | 50.00% | 33.33% | 46.67% |
| 7 | 42.86% | 57.14% | 42.86% | 28.57% | 42.86% | 42.86% |
| 8 | 50.00% | 62.50% | 50.00% | 62.50% | 37.50% | 52.50% |
| 9 | 44.44% | 55.56% | 44.44% | 33.33% | 44.44% | 44.44% |
| 10 | 50.00% | 60.00% | 50.00% | 50.00% | 40.00% | 50.00% |
| 11 | 36.36% | 45.45% | 54.55% | 36.36% | 36.36% | 41.82% |

**Topic Name: Set Language and Notations**

Number of Questions: 8

Unique Tags: 15

Best K-Value: 3

Concept and Formula Tags applied: No

Table ‑: Set Language and Notations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 50.00% | 50.00% | 50.00% | 0.00% | 40.00% |
| 3 | 66.67% | 33.33% | 33.33% | 33.33% | 33.33% | 40.00% |
| 4 | 25.00% | 25.00% | 50.00% | 25.00% | 25.00% | 30.00% |
| 5 | 0.00% | 20.00% | 20.00% | 40.00% | 20.00% | 20.00% |

**Topic Name: Functions**

Number of Questions: 32

Unique Tags: 31

Best K-Value: 6

Concept and Formula Tags applied: No

Table ‑: Clustering Performance Result for Functions

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 3 | 66.67% | 0.00% | 33.33% | 0.00% | 0.00% | 20.00% |
| 4 | 25.00% | 25.00% | 25.00% | 25.00% | 25.00% | 25.00% |
| 5 | 40.00% | 40.00% | 20.00% | 0.00% | 60.00% | 32.00% |
| 6 | 33.33% | 50.00% | 16.67% | 50.00% | 50.00% | 40.00% |
| 7 | 42.86% | 28.57% | 42.86% | 14.29% | 28.57% | 31.43% |
| 8 | 12.50% | 25.00% | 37.50% | 12.50% | 25.00% | 22.50% |

**Topic Name: Matrices**

Number of Questions: 14

Unique Tags: 23

Best K-Value: 3

Concept and Formula Tags applied: Yes

Table ‑: Clustering Performance Result for Matrices

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 50.00% | 100.00% | 50.00% | 100.00% | 70.00% |
| 3 | 100.00% | 33.33% | 66.67% | 66.67% | 100.00% | 73.33% |
| 4 | 50.00% | 50.00% | 50.00% | 50.00% | 100.00% | 60.00% |
| 5 | 60.00% | 80.00% | 40.00% | 60.00% | 100.00% | 68.00% |
| 6 | 100.00% | 66.67% | 50.00% | 66.67% | 33.33% | 63.33% |

**Topic Name: Vectors**

Number of Questions: 37

Unique Tags: 45

Best K-Value: 4

Concept and Formula Tags applied: No

Table ‑: Clustering Performance Result for Vectors

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| 3 | 33.33% | 33.33% | 66.67% | 33.33% | 0.00% | 33.33% |
| 4 | 0.00% | 25.00% | 50.00% | 50.00% | 50.00% | 35.00% |
| 5 | 40.00% | 40.00% | 20.00% | 40.00% | 20.00% | 32.00% |
| 6 | 33.33% | 33.33% | 16.67% | 50.00% | 33.33% | 33.33% |
| 7 | 28.57% | 28.57% | 42.86% | 28.57% | 28.57% | 31.43% |

**Topic Name: Permutations and Combinations**

Number of Questions: 8

Unique Tags: 6

Best K-Value: 4

Concept and Formula Tags applied: No

Table ‑: Clustering Performance Result for Permutations and Combinations

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 0.00% | 50.00% | 50.00% | 0.00% | 30.00% |
| 3 | 33.33% | 0.00% | 66.67% | 33.33% | 33.33% | 33.33% |
| 4 | 50.00% | 50.00% | 25.00% | 50.00% | 25.00% | 40.00% |
| 5 | 20.00% | 40.00% | 20.00% | 20.00% | 0.00% | 20.00% |

**Topic Name: Arithmetic Progression and Geometric Progression**

Number of Questions: 11

Unique Tags: 17

Best K-Value: 3

Concept and Formula Tags applied: No

Table ‑: Clustering Performance Result for Arithmetic Progression and Geometric Progression

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 0.00% | 50.00% | 0.00% | 0.00% | 20.00% |
| 3 | 66.67% | 66.67% | 0.00% | 33.33% | 66.67% | 46.67% |
| 4 | 50.00% | 25.00% | 50.00% | 25.00% | 25.00% | 35.00% |
| 5 | 20.00% | 20.00% | 60.00% | 0.00% | 0.00% | 20.00% |

**Topic Name: Circular Measures**

Number of Questions: 8

Unique Tags: 24

Best K-Value: 4

Concept and Formula Tags applied: No

Table ‑: Clustering Performance Result for Circular Measures

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| K-Value | 1 | 2 | 3 | 4 | 5 | Avg. Accuracy |
| 2 | 50.00% | 50.00% | 0.00% | 0.00% | 50.00% | 30.00% |
| 3 | 33.33% | 66.67% | 0.00% | 33.33% | 66.67% | 40.00% |
| 4 | 75.00% | 50.00% | 25.00% | 75.00% | 25.00% | 50.00% |
| 5 | 20.00% | 60.00% | 60.00% | 40.00% | 40.00% | 44.00% |
| 6 | 33.33% | 33.33% | 33.33% | 33.33% | 33.33% | 33.33% |

### Discussion

Due to time limitation on the project, a second algorithm was not implemented to do a comparison with K-Means. Therefore, there is no comparison on whether K-Means clustering algorithm is the best algorithm for this feature. Nevertheless, the K-Means algorithm performed quite up to standard on some topics.

However, it was a disappointment on certain topics such as Differentiation, Plane Geometry and Indices and Surds, concept and formula tags were applied but the clustering accuracy still failed to meet an average of more than 50%. There could be a possibility in which the average accuracy was affected by keyword tags overshadowing the concept and formula tags. This issue was found too late on the implementation as the generating of concepts and formulae took about 2 months to execute (inclusive of tagging). If keywords were to be excluded from Clustering terms, all questions must be tagged with concepts and/or formulae to be able to feature in an accurate cluster. Else the untagged questions will fall into a cluster of untagged questions. This situation will affect quite a lot of topics which are currently not tagged with any concepts and formulae as the topics are obsoleted from the current syllabus (Tagging was based on current syllabus only).

The functionality of Multiple-Tag Cloud as mentioned earlier in Chapter 5, has a quite similar function to Question Clustering. Both functions use the Tags to find question similarities. For Multiple-Tag Cloud, the questions that satisfy the tag combinations were displayed according to their frequency count. The results are much better and ordered as compared to the Question Clustering function. The Multiple-Tag Cloud covers all possibilities of combinations, but the Question Clustering captures the best possible clusters as combinations. The latter function might restrict other possible matches which might gain a better frequency count, and moreover it also generated a few bad clusters which are a turn-off in the result display. Multiple-Tag Cloud on the other hand only generates the correct combinations.

In conclusion, the current Question Clustering function is not a very useful function. The Multiple-Tag Cloud function can be used as a replacement in finding Question Similarities based on Tag combinations. However, with improvement to the current concept and formula tags, the Question Clustering function may boost its accuracy and become a better use than before.

# Tag Search

## Feature Objective

With the Tag clouds being the only visualization for Tags, it makes it rigid for users to find what they need. Certain terms may exist only on one topic, and users might have to look through every single topic’s Tag Cloud to find the term which they are looking for. Multiple-Tag Cloud combinations also prevent user from keying the term combinations which they want, and manually searching the list in the Multiple-Tag Cloud may prove to be a bit difficult if there are too many terms.

Therefore, Tag Search is a feature in which user can type which term(s) they need, and the result will be shown with the questions that match their search, and also user can choose to filter the result based on topics. The topic filter can also be useful as information on the topics that the search query can be found.

Tag Search can also return images that match the search query. For some user that may want to work on a specific question that has diagram (e.g. graph diagram or diagram on shapes), user can key in specific tag terms to try look for image result that matches what they need.

## Implementation Approach

### Tag Search Bar

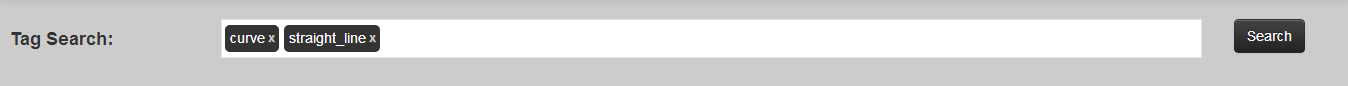


Figure ‑: Tag Search Bar

The Tag Search Bar was implemented using a jQuery plugin package called Tagit [17]. The search bar comprises of autocomplete function in which the autocomplete list was generated from the complete list of available tags. The user can click “Search” upon filling up their query.

### Search Result Page

The Search Result page consists of Search and Image tab. Screenshot of the page can be found in Appendix A-1(XI). The Search tab shows the result of all the questions that match the search term(s). The side-bar also allow user to filter the result according to topics. This also allow user to know how many different topics are retrieved according to the search term(s). This is an information in which Tag Clouds were unable to provide.

The Image tab displays the images related to the questions in which matched the search term(s). The result also comes with side-bar which allows user to filter the result according to topics.

## Discussion

The Tag Search implemented was a good addition to the overall Tag system. Tags can be searched flexibly instead of looking it up from the Tag Cloud. Furthermore, Tag Search allows result to be pulled from different topics, which is not seen in Tag Cloud. Thus, user can conduct a search and find out what topic(s) utilized the keyword, concept or formula.

The Images tab of the search result is currently generated by finding the existing images from questions that matched the search term(s). A possible way of creating a new set of tag list for each image can be considered as an alternative to generating search result from.

In conclusion, the Tag Search is still a useful tool for user to view result for questions and images related to the tag(s) that they are searching for.

# User Evaluation and Feedbacks

User evaluation was done through a demo of the system and followed by a survey. A sample of the survey is found at Appendix A-4.

A total of 20 people are being targeted and the result of the survey will be evaluated in this chapter.

## Users’ Evaluation on Usefulness of View Paper

Figure ‑: Survey Result for View Paper

The View Paper feature was voted “Useful” from the majority, with 3 voters voted slightly above average (Very Useful) and below average (Somewhat Useful), and 1 voter voted “Really Useful”. In general, most are satisfied with the functionalities given on View Paper.

## Users’ Evaluation on Usefulness of Paper Tag Cloud

Figure ‑: Survey Result for Paper Tag Cloud

The result voted by the users on the Paper Tag Cloud feature having most of the voters voted “Somewhat Useful”, followed by “Useful”. This result showed that users are also finding that the Paper Tag Cloud is not very useful in revealing knowledge to user from the tags involved with the paper.

## Users’ Evaluation on Usefulness of Topic Distribution

Figure ‑: Survey Result for Topic Distribution

The result for Topic Distribution shows most voters voted the feature as “Useful” and with a handful voting it as “Very Useful”. Therefore, in general the voters are satisfied with the feature.

## Users’ Evaluation on Usefulness of Topic Trend

Figure ‑: Survey Result for Topic Trend

The result for Topic Trend is somewhat the same as Topic Distribution, except that there is a user who voted “Somewhat Useful”. Most of the users voted “Useful” and some voted “Very Useful”. Therefore, users are generally happy with the Topic Trend feature.

## Users’ Evaluation on Usefulness of Topic Tag Cloud

Figure ‑: Survey Result for Topic Tag Cloud

The result for Topic Tag Cloud was one of the better ones with majority voted “Useful”, and some voted “Very Useful” and even “Really Useful”. Therefore, users rated it to be better than the rest of the features.

## Users’ Evaluation on Usefulness of Question Clustering

Figure ‑: Survey Result for Question Clustering

The result for Question Clustering shows that most voters voted “Useful”, with some voters begged to differ with “Somewhat Useful”, and some voted better with “Very Useful”. Overall, users are generally satisfied with the feature.

## Users’ Evaluation on Usefulness of Tag Search

Figure ‑: Survey Result for Tag Search

Tag Search has majority voted “Useful” and some voted “Very Useful” and “Really Useful”. Therefore, users are satisfied with the feature.

## Users’ Evaluation on Usefulness of the Overall System

Figure ‑: Survey Result for Overall System Usefulness

In terms of overall system usefulness, majority voted 7 out of 10, with some voted 6 and 5, and a handful also voted as high as 8. Therefore, the overall system achieved an average rating of 6.9 out of 10.

## Additional Remarks

Some users also contributed constructive feedbacks from the system. Below is the list of feedbacks from the users.

1. Can improve on the concepts and formulae for some topics.
2. Not very user-friendly for Secondary school students to adopt.
3. Can make it more interactive with a community.
4. Generate random questions for user to attempt when he/she chooses a topic.

## Discussion

Figure ‑: Breakdown of Feature Ratings

The survey result was further evaluated by calculating the average ratings for each feature by using the scale level of 1 to 5, with 1 being “Not Useful” and 5 being “Really Useful”. The result was shown in Figure 8-9, which shows that Topic Tag Cloud achieved the best rating, followed by Tag Search, and then Topic Distribution. Paper Tag Cloud did not fare well in the rating, and Question Clustering did below par as well.

Therefore, this feedback serves as a platform for future works to improve the system further. More on the recommended future works will be discussed in Chapter 10.2

# Experimental Scenario

## Experiment Objective

The objective of the experiment is to predict the major topics and concepts that will appear in 2011 Paper. Based on the existing tools developed, a prediction will be made and then compare with the actual 2011 Paper to see the accuracy.

The success of the experiment will depend on the accuracy level of the result.

## Predictions

**Major Topics**

To predict the major topics, the topic trend can be used to analyse the topic trend of the recent years. Therefore, a three-year period from 2008 to 2010 was chosen as the trend period to analyse.

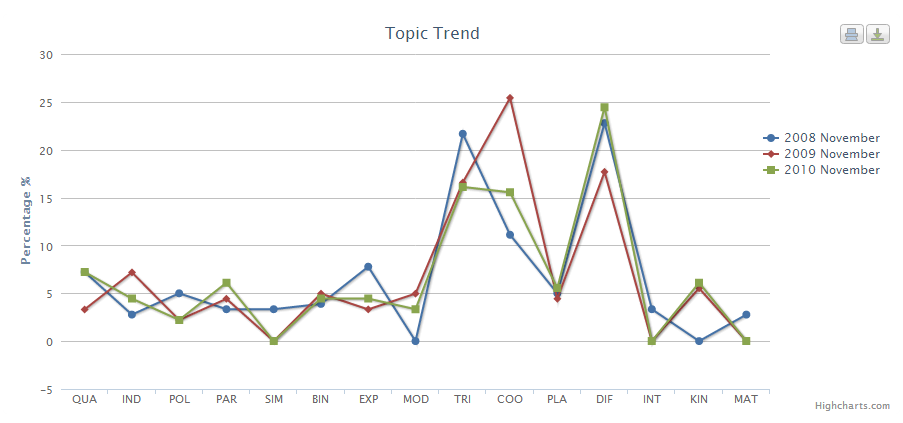


Figure ‑: Topic Trend from Year 2008 to 2010

Based on Figure 9-1, the questions on Differentiation, Coordinate Geometry, and Trigonometry are the major topics, with the rest of the topics carrying less than 10% of the overall weightage.

To verify, Topic distribution for 2010 was also displayed in Figure 9-2.

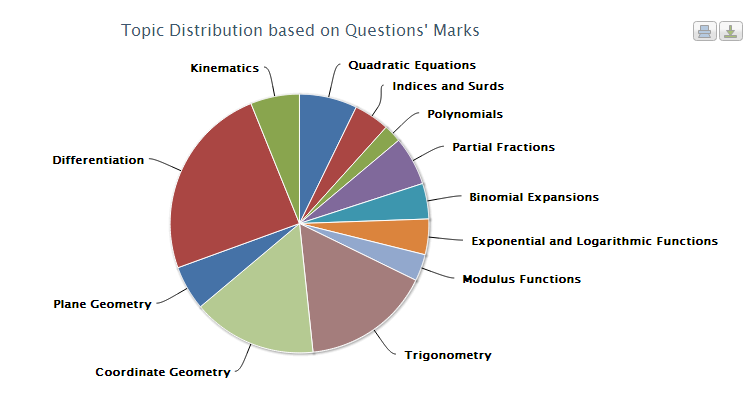


Figure ‑: Topic Distribution for Year 2010

The Figure 9-2 confirms that Differentiation, Coordinate Geometry and Trigonometry carries the highest weightage in the paper. Thus, it is a stable call to predict them as the upcoming major topics in the paper.

**Key Concepts**

Having known of the major topics, the key concepts should be clearly sitting around the major topics. Therefore, Topic Tag Cloud was used to generate each of the three topics to see the frequent occurrence of concepts and formulae.

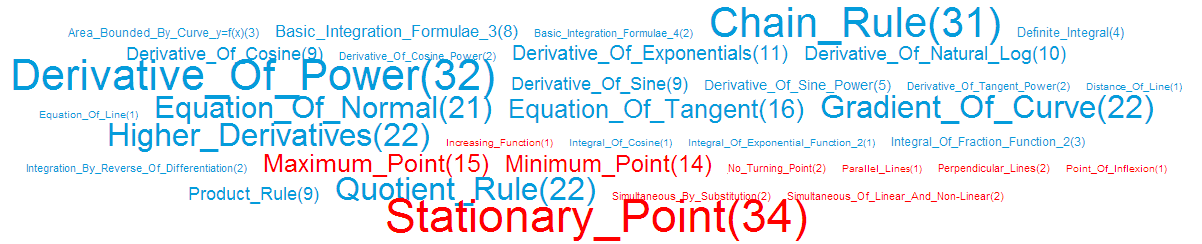


Figure ‑: Differentiation Concepts and Formulae Tag Cloud

From Figure 9-3, the diagram displays the frequent occurrence of concepts and formulae for Differentiation topic. For concepts, there is a significant occurrence of “Stationary\_Point”, “Maximum\_Point”, and “Minimum\_Point”. From knowledge, user can guess that it is likely that these concepts may overlap as they are from the sub-topic of “Maxima, Minima and Stationary Points”. Therefore, this can be a major concept. The frequent occurrence of certain formulae such as “Equation\_Of\_Normal”, “Equation\_Of\_Tangent”, “Gradient\_Of\_Curve” are general Differentiation techniques that are applied on graphs. These formulae frequent occurrence can be a likelihood that differentiation techniques are commonly used on question related to graphs. “Chain\_Rule” and “Quotient\_Rule” also occurs quite frequently, as they are some of the basic techniques for differentiating equations.

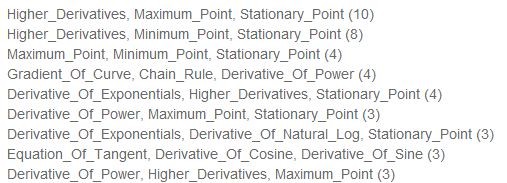


Figure ‑: Differentiation Concepts and Formulae in Multiple Tag Cloud

The multiple tag cloud in Figure 9-4 further concretes the prediction by showing that Maximum\_Point, Minimum\_Point and Stationary\_Point occurs together quite often in the top of the multiple tag cloud. This shows that complex questions involving Differentiation uses concepts from these area.

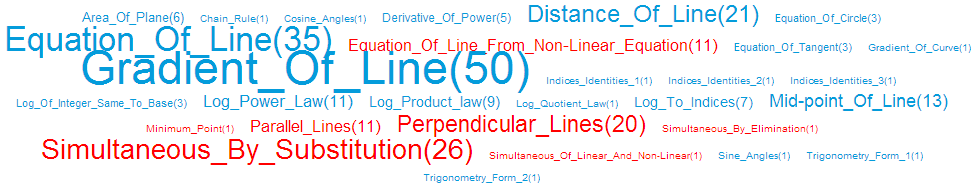


Figure ‑: Coordinate Geometry Concepts and Formulae Tag Cloud

Next, Figure 9-5 displays the concepts and formulae occurrence in Coordinate Geometry. It shows that “Gradient\_Of\_Line”, “Equation\_Of\_Line” and “Distance\_Of\_Line” occurs frequently for formulae, as they are the key formulae for Coordinate Geometry. Also, concepts on “Perpendicular\_Lines” and “Simultaneous\_By\_Substitution” shows that there are many questions asking for gradient of perpendicular lines for the earlier, and frequent usage of simultaneous equations on Coordinate Geometry equations also occurred.

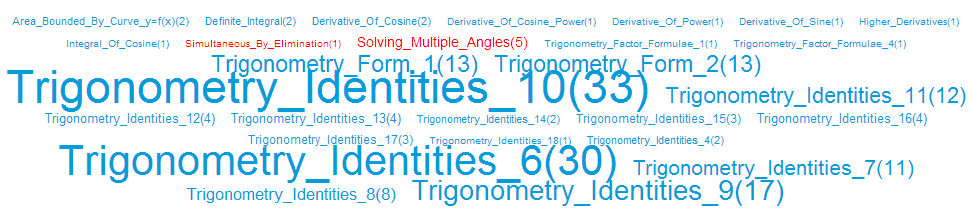


Figure ‑: Trigonometry Concepts and Formulae Tag Cloud

From Figure 9-6, it is shown that Trigonometry is largely populated by its Identities formulae. Therefore, it is definite that most of the Trigonometry questions will be based on solving the identities.

## Prediction Results

Before revealing the result, Table 9-1 is the summary of predictions made earlier on using the analysis tools from MathAnalyzer.

Table ‑: Summary of Predictions

|  |  |
| --- | --- |
| No. | Prediction |
| 1 | Differentiation is a major topic that covers close to 20% of overall weightage |
| 2 | Coordinate Geometry is a major topic that covers close to 20% of overall weightage. |
| 3 | Trigonometry is a major topic that covers close to 20% of overall weightage |
| 4 | Differentiation’s Stationary\_Point, Minimum\_Point, and Maximum\_Point are important concepts. |
| 5 | Differentiation’s Equation\_Of\_Normal, Equation\_Of\_Tangent, Gradient\_Of\_Curve, Chain\_Rule and Quotient\_Rule are important formulae. |
| 6 | Coordinate Geometry’s Gradient\_Of\_Line, Equation\_Of\_Line and Distance\_Of\_Line are important formulae. |
| 7 | Trigonometry’s Identities formulae are important formulae. |

Results were shown on Table 9-2.

|  |  |  |
| --- | --- | --- |
| No. |  | Verdict |
| 1 | Differentiation covered 33 marks out of 180 in Year 2011. It obtained a weightage of 18.33% | Hit! |
| 2 | Coordinate Geometry covered 29 marks out of 180 in Year 2011. It obtained a weightage of 16.11% | Close |
| 3 | Trigonometry covered 27 marks out of 180 in Year 2011. It obtained a weightage of 15% | Close |
| 4 | Differentiation’s Stationary\_Point, Minimum\_Point, and Maximum\_Point occurred in Paper 2 Question 6, weighing 9 marks. | Close |
| 5 | Differentiation’s Equation\_Of\_Normal, Equation\_Of\_Tangent, Gradient\_Of\_Curve, Chain\_Rule and Quotient\_Rule occurred in Paper 1 Question 9 weighing 6 marks. | Miss |
| 6 | Coordinate Geometry’s Gradient\_Of\_Line, Equation\_Of\_Line and Distance\_Of\_Line occurred in Paper 1 Question 2 & 11 and Paper 2 Question 11 weighing a total of 23 marks. | Hit! |
| 7 | Trigonometry’s Identities formulae occurred in Paper 1 Question 4, and Paper 2 Question 10 weighing a total of 16 marks. | Hit! |

## Discussion

Based on the result highlighted in Table 9-2, there is only 1 mis-prediction and 3 other close calls. These close calls however can be considered a success since the occurrence are still there and weighing a decent amount of marks. Other than that, most of the predictions are correct.

Therefore, the system is quite accurate in predicting future papers.

# Conclusion and Future Works

## Conclusion

After going through a series of implementations and discussions, the system’s feature usefulness are concluded below.

Topic Distribution is a tool for user to look at past years’ topic distribution. It prepares students to know what kind of weightage for each topic is expected. This allows students to plan their revisions properly and not over-focusing on certain topics that are not heavily weighted. The voters for the survey voted this feature as “Useful”.

Topic Trend is a tool for user to find out the trend of topics from a range of years. It does not show a predictive trend which is certain to predict future examination topic weightage as the topic weightage are always rising and falling. However, the rise and fall of each topic weightage is not drastic, which is still rather useful to show that there is a stable trend from the past few years that the topic is for example between 10%-15% weightage distribution. The voters for the survey voted this feature as “Useful”

Paper Tag Cloud is a tool for user to view the tags that are featured in the questions of a particular paper. The result of visualization does not tell much as the frequency for each tag was low (between 1 to 3). This is due to the fact that the paper will try to feature as much concept/formula as possible, therefore it is rare that concepts/formulae are repeated many times in a paper. Thus, Tag Cloud visualization is not a good tool on Paper mode. The voters voted “Somewhat Useful”, which is the lowest among other features.

Topic Tag Cloud is a tool for user to view the tags that are featured in the questions of a particular topic. The result of visualization is in fact very useful as user get to understand which concept/formula is the most important. The multiple-tag cloud option also allows user to spot some concepts/formulae that frequently go along with each other. With the additional feature to allow user to click on the link to study the questions with the combination of multiple tags, it serves as a good tool for user to spot question patterns from the particular topic. Thus, it is the highlight feature of the system. The voters voted close to “Very Useful” on average. This is the highest rated feature among the rest.

Question Clustering is a tool for user to look at a particular topic to see the questions being clustered according to Tag similarity. Due to the issue that the tags are not very mature (more work can be done to improve the tags), the clustering result was not very accurate on some topics. However, there are still a handful of good clusters within each topic. This shows that Question Clustering can be a useful tool if the tags can be refined. This tool uncovers the unknown knowledge of question similarity matching with its concepts and formulae. The voters voted this feature as “Useful”.

The final feature is the Tag Search, which user can key in tag(s) provided from the autocomplete field. The search result generates the questions and the images that are related to the query. A further addition of the “Topic Filters” also allows user to filter the search according to topics. The search feature opens a new way to visualize the tag(s) in an overall view as the Tag Cloud was broken into Paper and Topic view. The Topic Filter also provides user additional information that the tag(s) combination can be found in which topic. Thus, the voters voted this feature as “Useful”.

In conclusion, the system as a whole was voted 6.9 out of 10, which is a good result to show that the current work has a potential to allow students and educators understand the Additional Mathematics better than the conventional way. This serves as a platform for a future of smart learning in Mathematics.

## Future Work

Due to the time limitation on the project, there are some value-added features that are not implemented. Therefore, this section recommends some value-added features that can be implemented to improve the overall performance of the system.

### Improvise on Tags

The current tags that were implemented in the system was done solely on the author’s knowledge while referencing from the MOE-authorized Additional Mathematics textbook. Therefore, the concept and formula tags should go through multiple revisions, and recommended by the hands of educators that are directly involved with the subject.

The user evaluation was done by some educators from the subject who also added that the concept and formula tags can be improved.

Last but not least, the Keyword tags can be omitted as it serves little aid in knowledge discovery, and might even serve as a hindrance to the clustering result.

### Clustering Algorithms

The current K-Means clustering algorithm has proven to be relatively good in its performance but the accuracy may not be up to standard to be counted as a reliable tool. The problem with the accuracy may be due to the overpopulated keyword tags which steered the result, or might be due to K-Means algorithm. Therefore, system may try to introduce another clustering algorithm such as Agglomerative Hierarchical Clustering (AHC) algorithm to see if the accuracy will be improved.

### Improve on User Experience

From the user evaluation feedback, some users pointed out that the user experience for the system may not be simple enough for use by Secondary 3 or 4 students. This is partially due to the fact that the Settings and Controls for each features were relatively based on technical terms as the system is still purely built for prototype testing. However if the system is ready for commercial, system should be designed idiot-proof to user of young age and also aging educators whom might not be very IT-savvy.

### Ratings and Recommendations

Questions can be rated by community according to difficulty level, and allow filtering of question difficulty when viewing by Topics. This allows user to try questions that are comfortable for their level.

Each question display can also implement a recommendation section to recommend related questions that are similar to the current question. This can be easily done with the existence of tags. For example, top 3 questions with most similar tag comparison to the current question will be featured as recommendation. This will aid the user to link up to a similar question for more practices.

# Appendix

## Web Design and Navigation

Online version available at: <http://exampapers.ap01.aws.af.cm/>

### Main Landing Page

Upon arrival from the web link, the main landing page is displayed. User can choose to navigate using the navigation bar on the top, or read the brief detail of the 3 main sections of the web system located below the Tag Search bar before deciding on his/her navigation decision. User can also do a tag search using the Tag Search bar directly at the middle of the page.

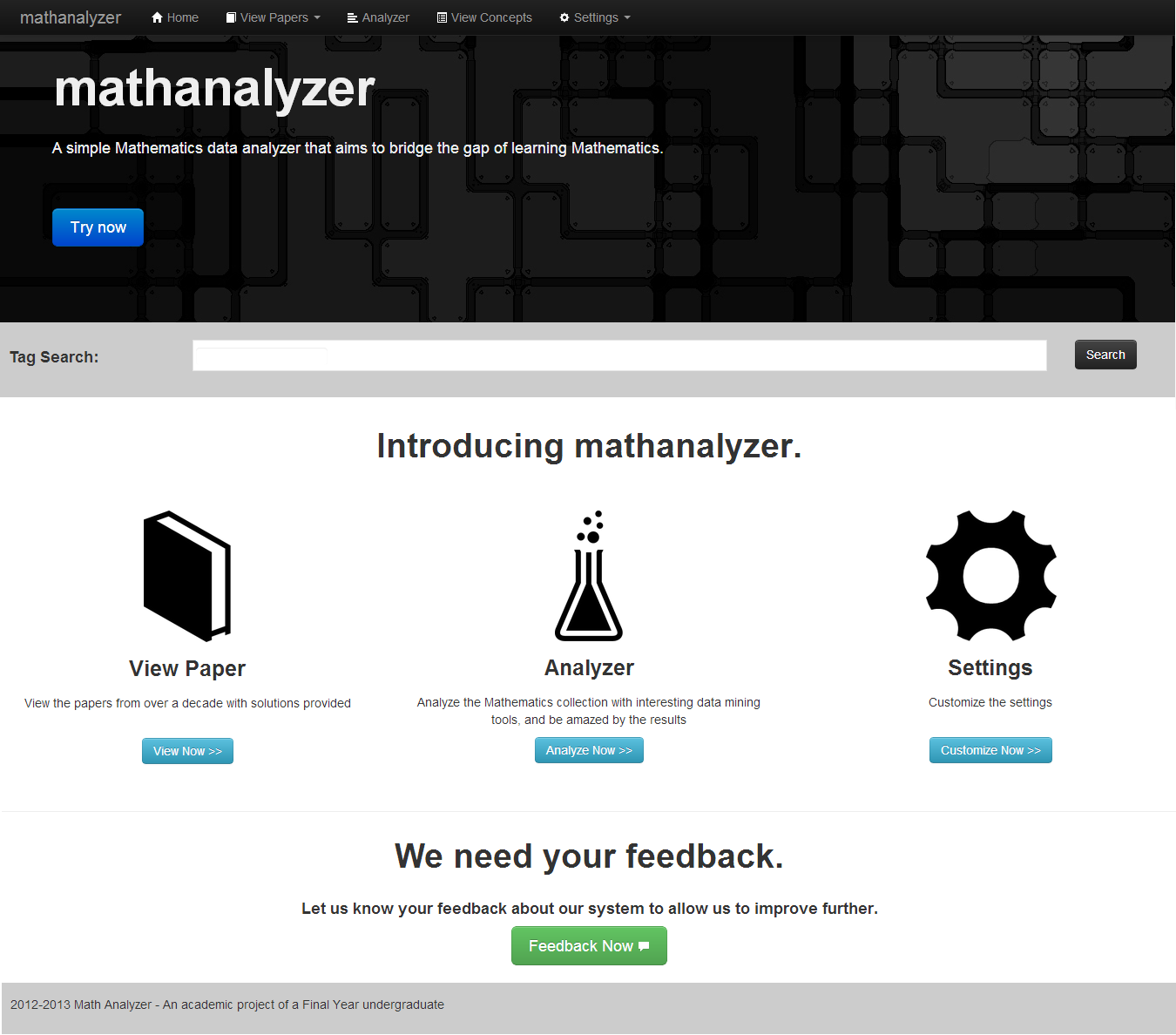


Figure ‑: Main Landing Page

### View Paper Selection Page

If user clicked on “View Paper” link, user will be directed to this selection page whereby user can choose to view questions by paper or by topic.

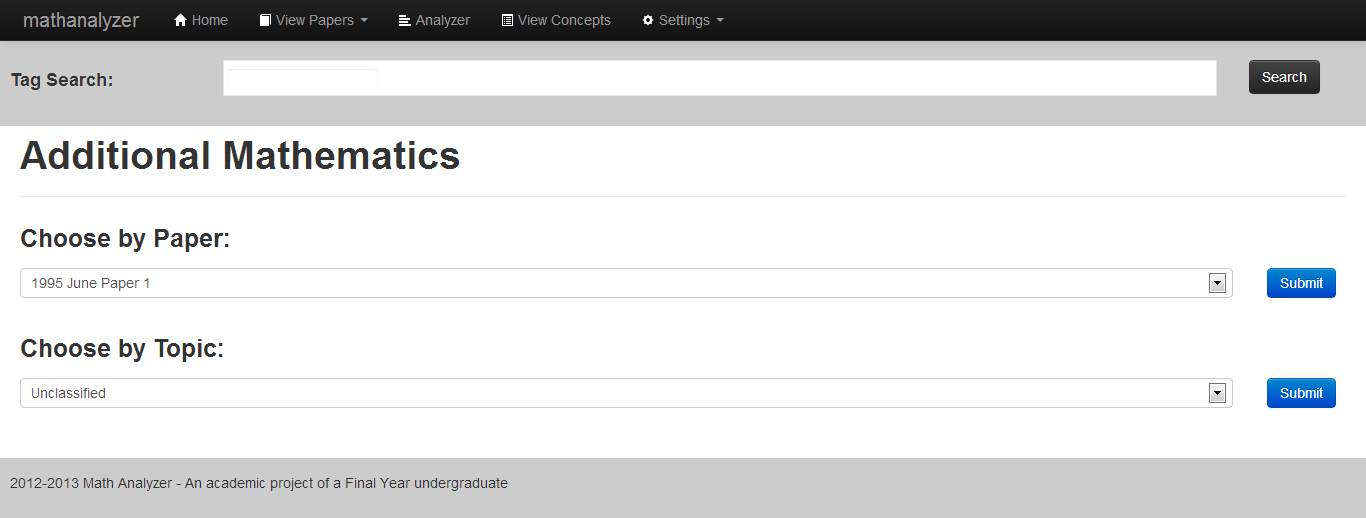


Figure ‑: View Paper Selection Page

### View Paper (Paper View)

When user chose a paper and clicked the “Submit” button, they will be navigated to the Paper view, with all the questions being listed in the format below.

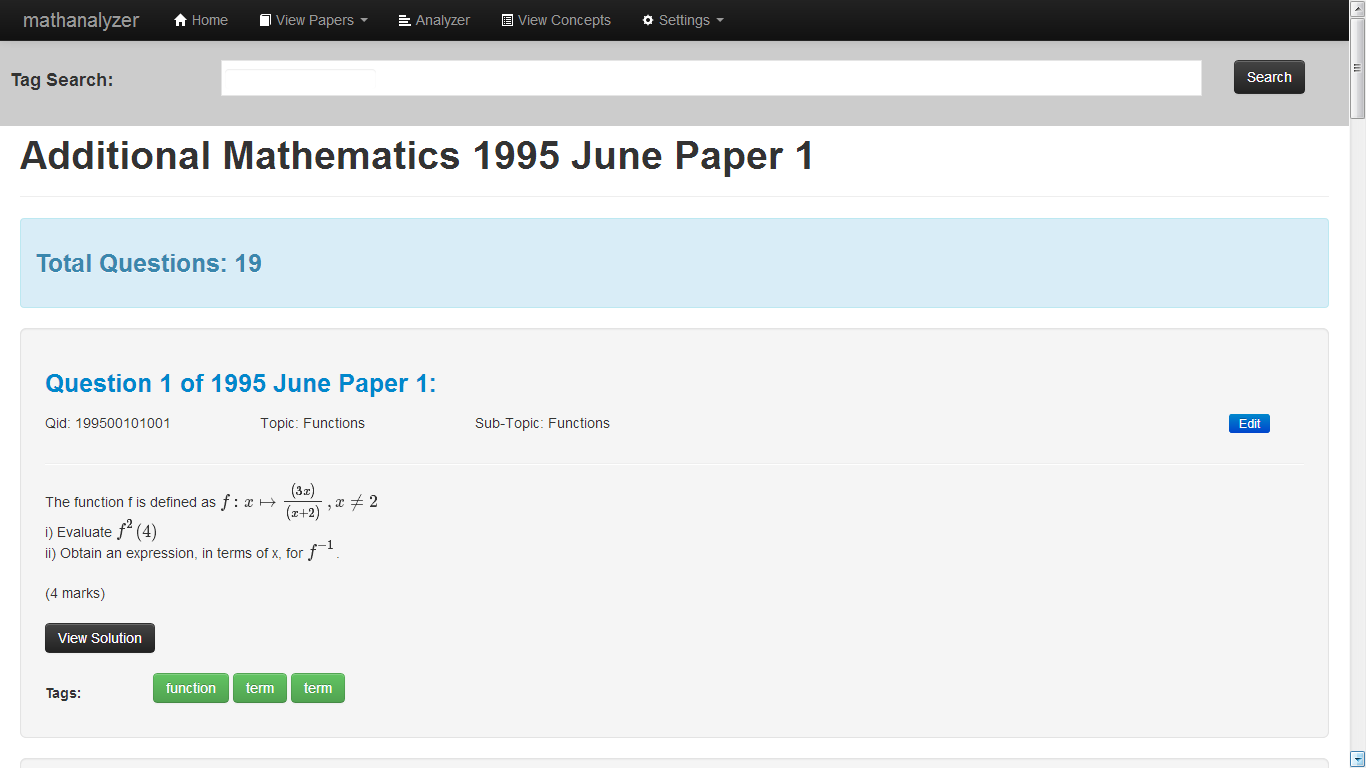


Figure ‑: View Paper (Paper View)

When “View Solution” button is clicked, the solution section will be collapsed below the button using AJAX.



Figure ‑: View Solution Before-and-After

User can also click “Edit” button for each question. Please refer to XIV: Admin Create/Modify Question Form for the navigation after “Edit” is being clicked.

### View Paper (Topic View)

For the Topic view, it is similar to Paper View, except that the questions are pulled based on the topic selected.

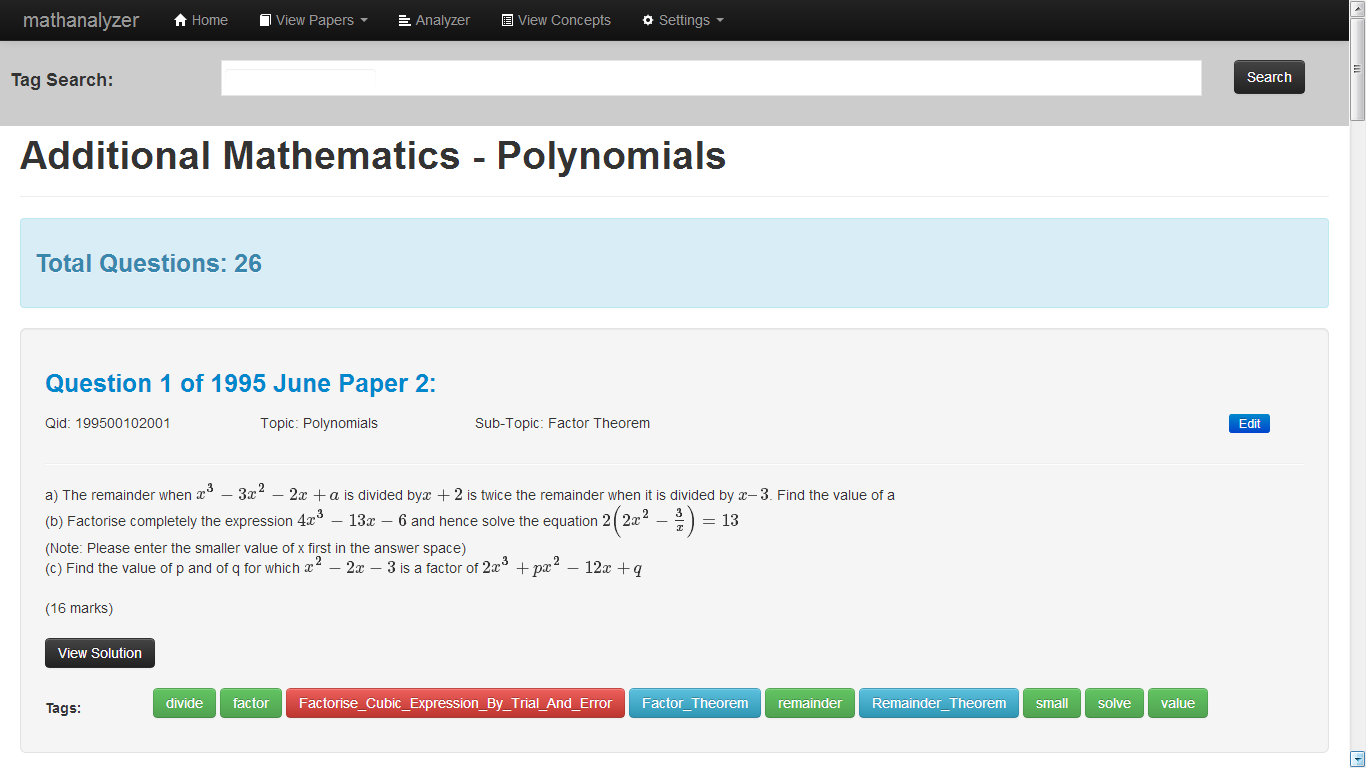


Figure ‑: View Paper (Topic View)

### Analyzer Main Landing Page

When user clicked on the “Analyzer” link, the page is navigated to the Analyzer main landing page. The landing page basically explains some of the description of the functions available, and user is able to select one of the functions on the side-bar on the left.

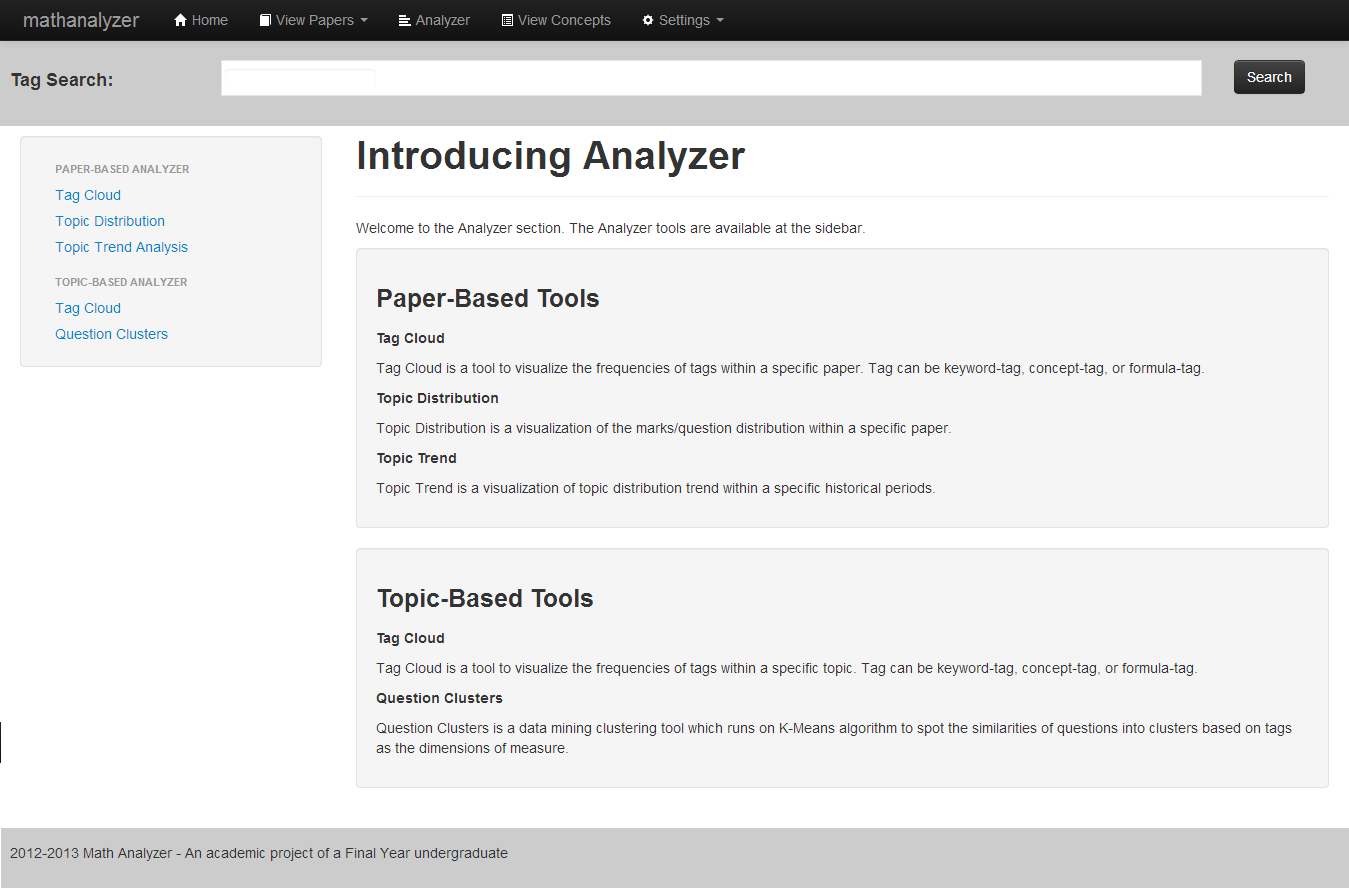


Figure ‑: Analyzer Main Landing Page

### Tag Cloud (Paper / Topic)

When user clicked on the “Tag Cloud” (applies to both Paper-based and Topic-based) link on the side-bar, the system navigated to the Tag Cloud page. User can choose the Paper/Topic to view, Combination (One for 1 tag, or Two for two-tags combination, etc.), and also filter the tag type. For this example, we used the Paper-Based Tag Cloud as illustration. (Topic-Based Tag Cloud looks similar as well)

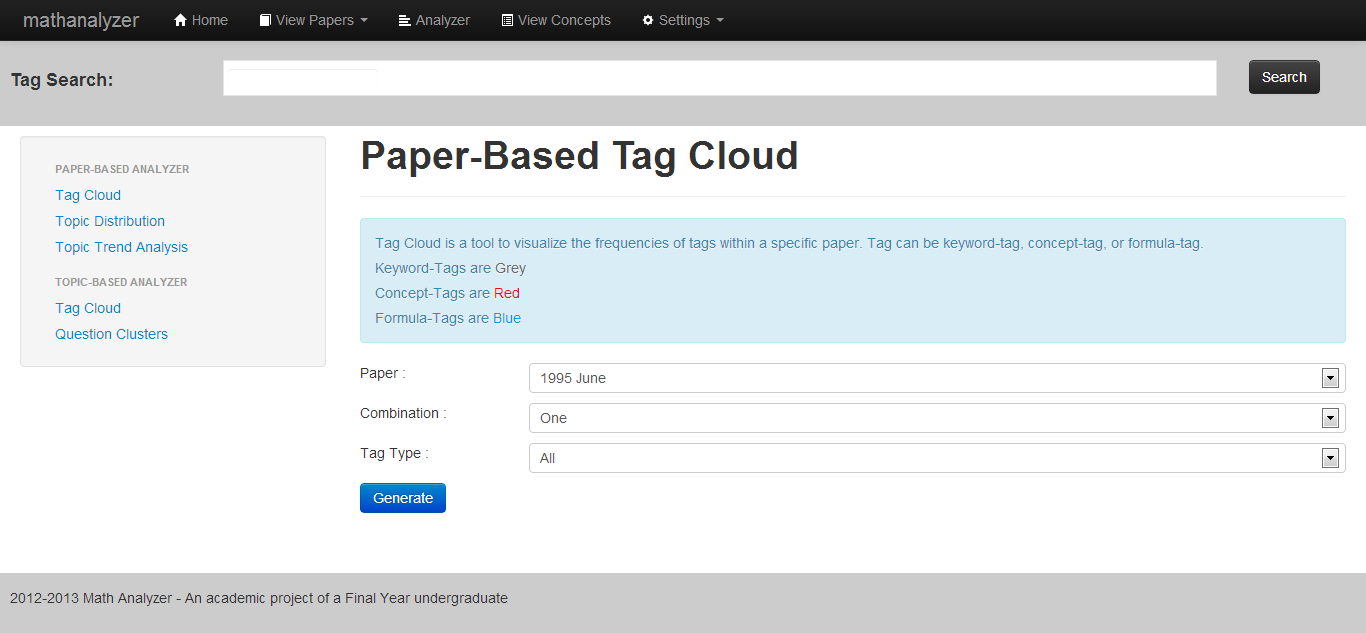


Figure ‑: Tag Cloud Selection

When user clicks “Submit” button, the tag cloud will be generated according to user’s specification. An accordion control was also generated if there are concept or formula keywords. The accordion will explain what does each tag title means.



Figure ‑: Tag Cloud Generated Result Page

### Topic Distribution

When user selects “Topic Distribution” on the side-bar, the page will be directed to Topic Distribution page. User can select the Paper, and also distribution by “Marks” or “Count” (Question Count).

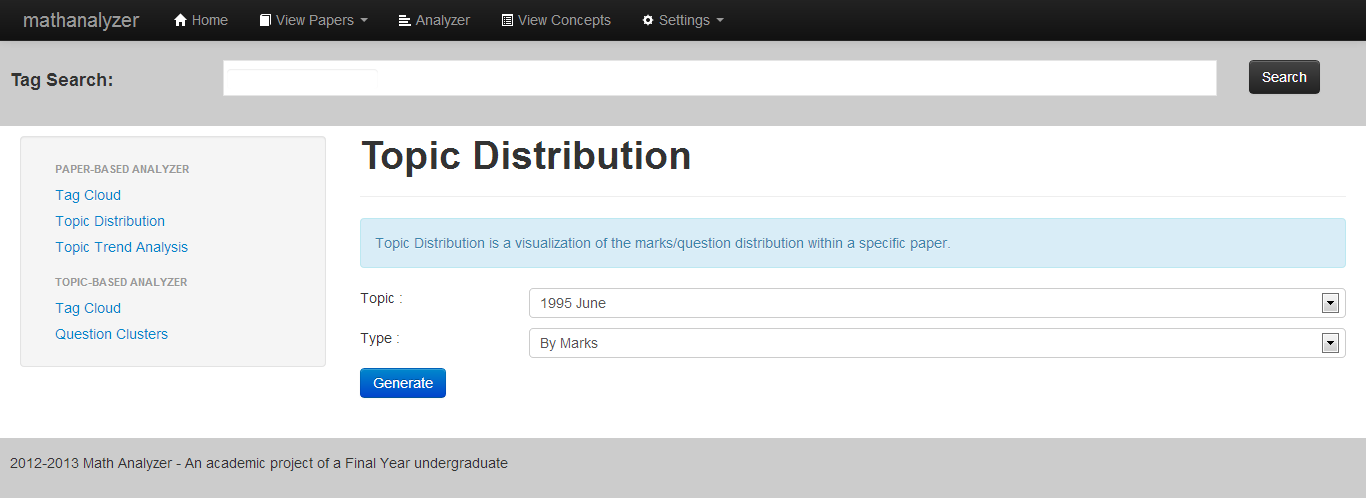


Figure ‑: Topic Distribution Selection Page

Upon clicking the “Generate” button, user can view the result of the distribution on a pie chart, accompanied by the list of topic links to view the exact question.

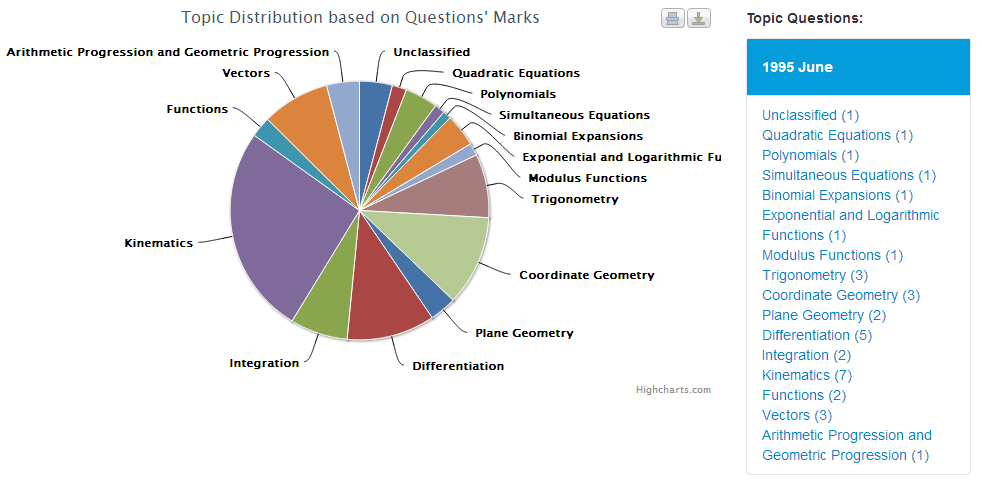


Figure ‑: Topic Distribution Result Page

### Topic Trend

When user clicks on the “Topic Trend Analysis” link on the side-bar, page is redirected to the Topic Trend Analysis page. User can select the period range, the topic to be included, and the analysis type (By percentage or By Question count). User can then click “Generate” button when done.

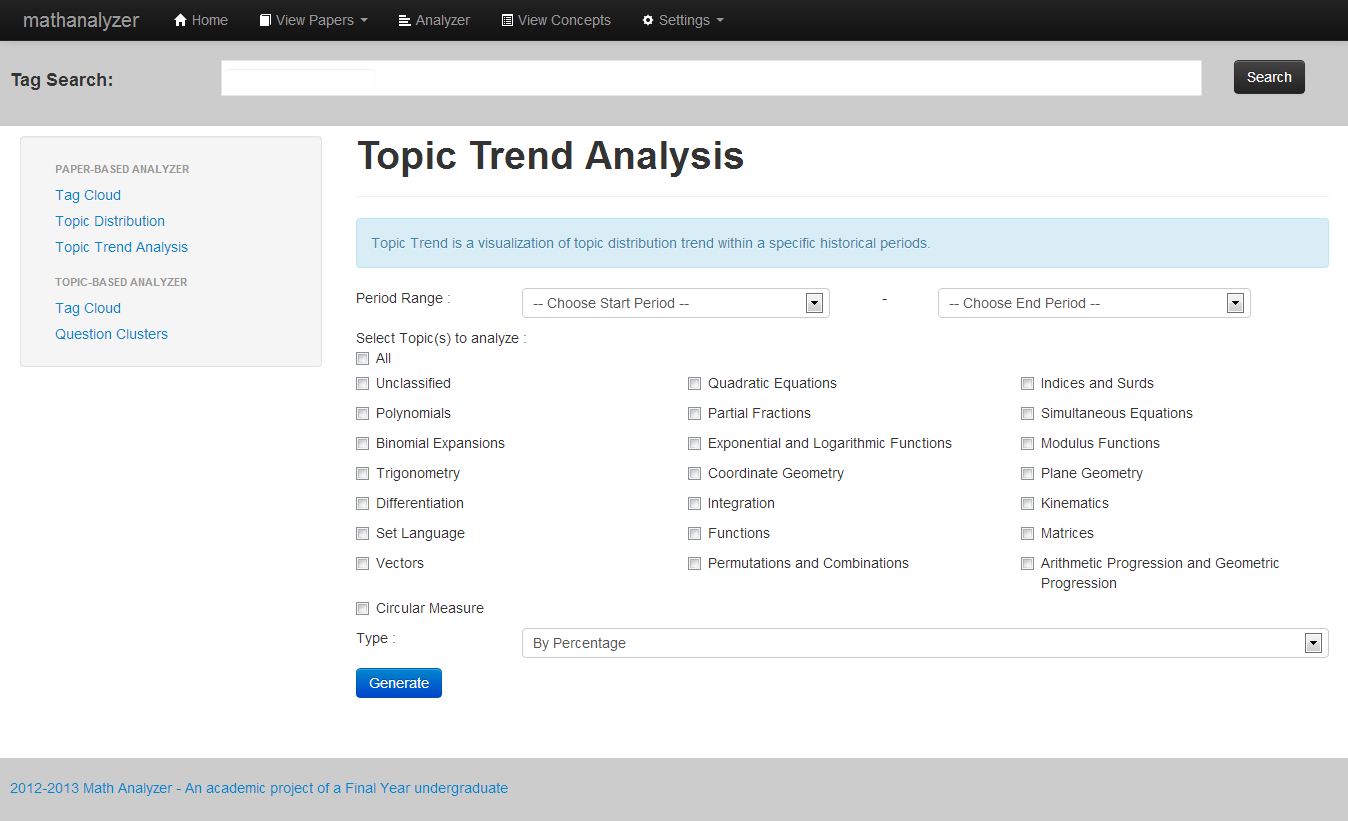


Figure ‑: Topic Trend Analysis Selection Page

Upon clicking the “Generate” button, the topic trend is generated in a line chart. The related papers are shown on the right bar with links directing to more details on each topic’s questions. Due to space constraint, the x-axis is labelled only the first 3 letter of the topic. User may need to refer to the legend below of the graph for more directions on the labelling meaning.

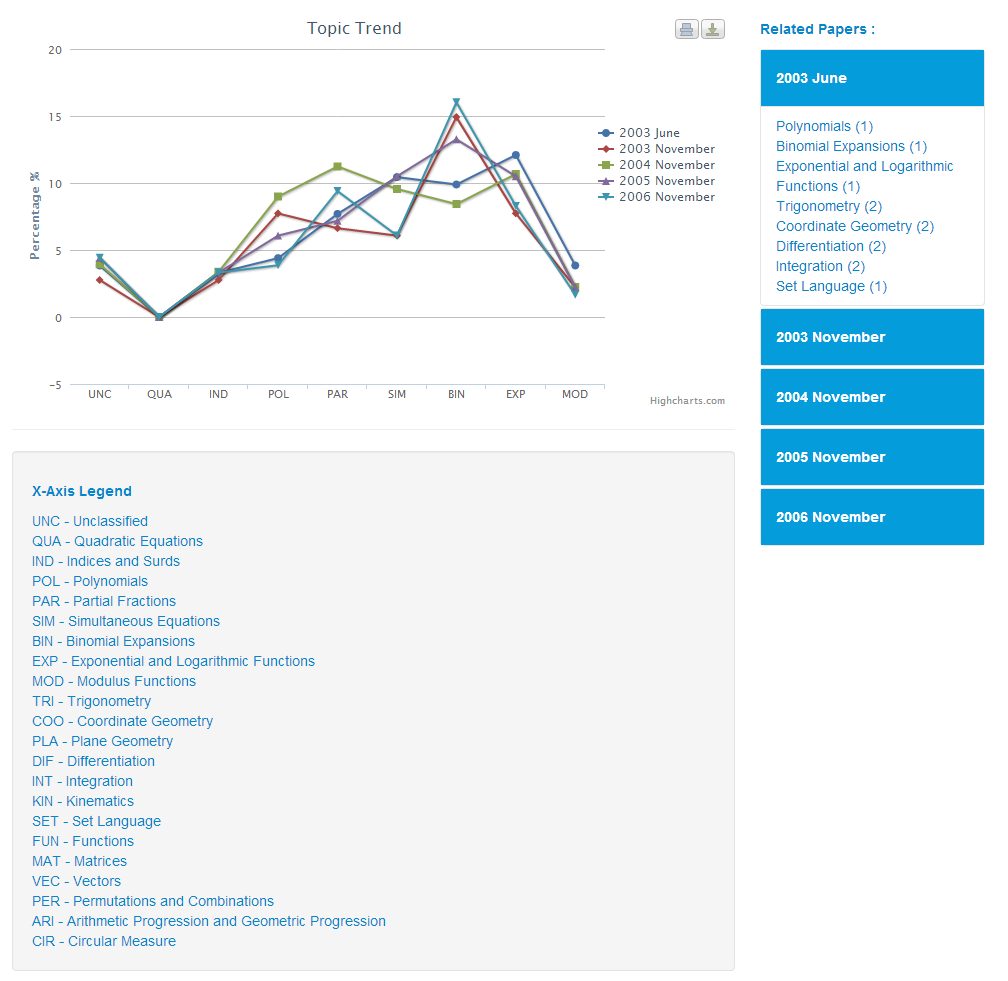


Figure ‑: Topic Trend Analysis Result Page

### Question Clustering

When user selects “Question Clusters” on the side-bar, the page redirects to the Question Clustering page. User can choose the topic to cluster, and can change the K value or leave it as default. User can then click “Submit” to generate the clustering result.

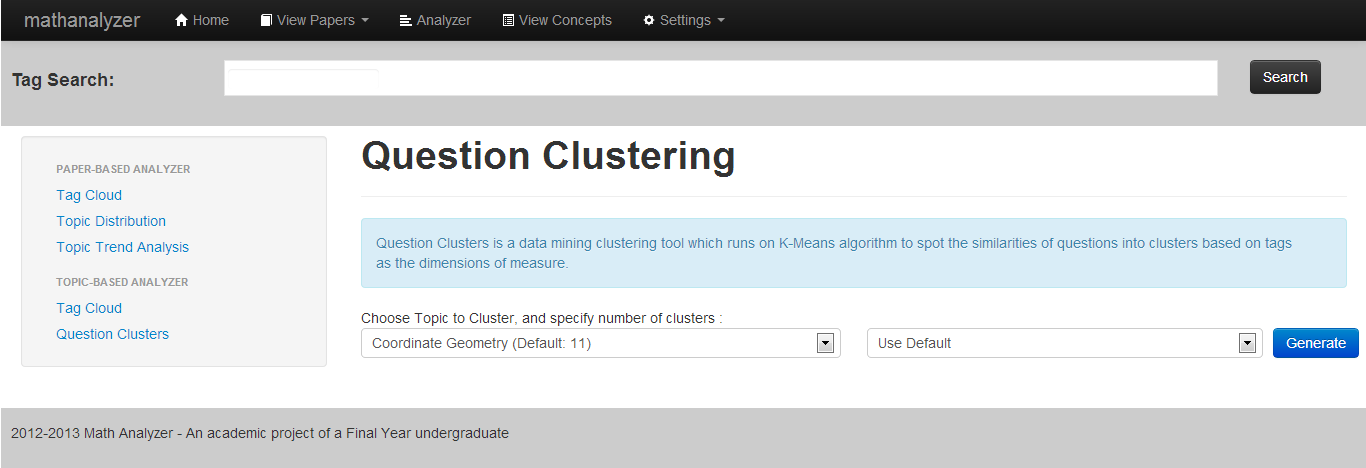


Figure ‑: Question Clustering Selection Page

Upon clicking “Submit”, the result will be generated. User can look at the list of result and see which cluster is good for gaining knowledge.

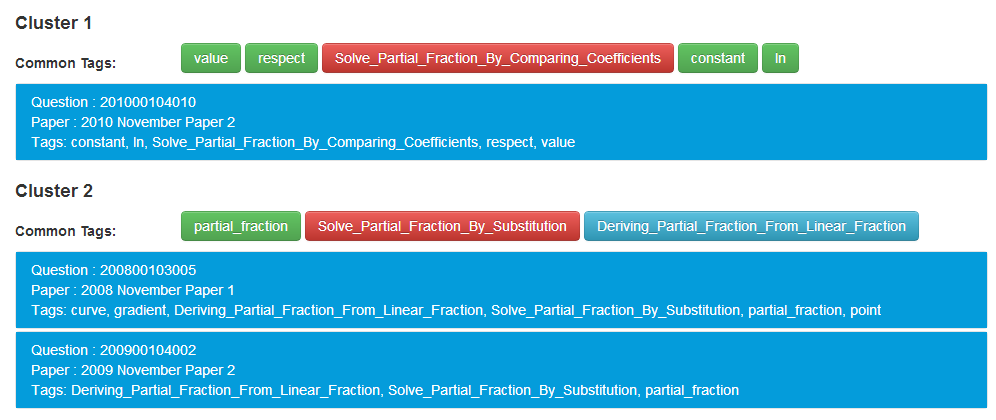


Figure ‑: Question Clustering Result Page

### View Concepts

View Concepts is basically allowing user to view concepts and formulae from a particular topic. This is just an informative function. Upon selecting a Topic, user can see the results below.



Figure ‑: View Concepts Page

### Tag Search

User can do a tag search anytime (except Admin section) with the Tag Search bar. User should stick to the autocomplete list to avoid bad search result.

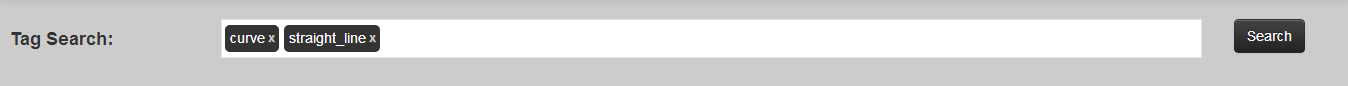


Figure ‑: Tag Search bar

Upon clicking “Search”, user will be brought to the Search Result page. Result is separated into two tabs, “Search” and “Image”. “Search” tab shows the questions as results, and “Image” tab shows the images as result.

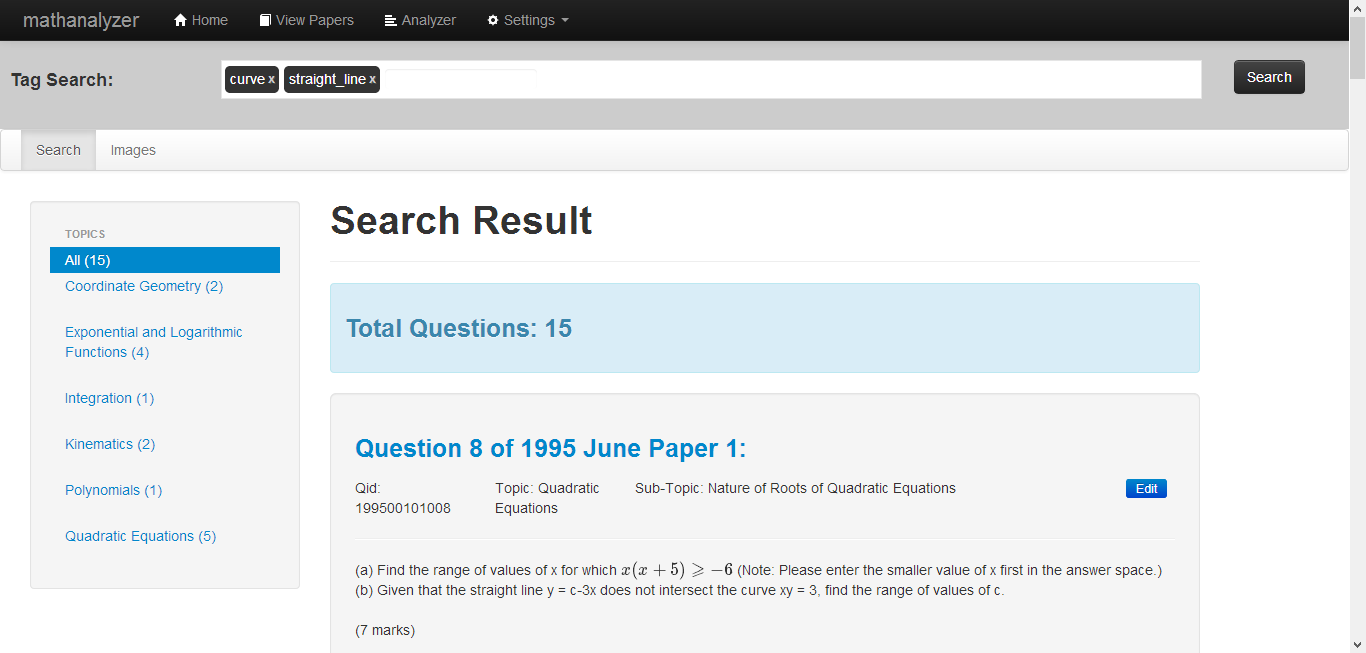


Figure ‑: Search Result (Search Tab)

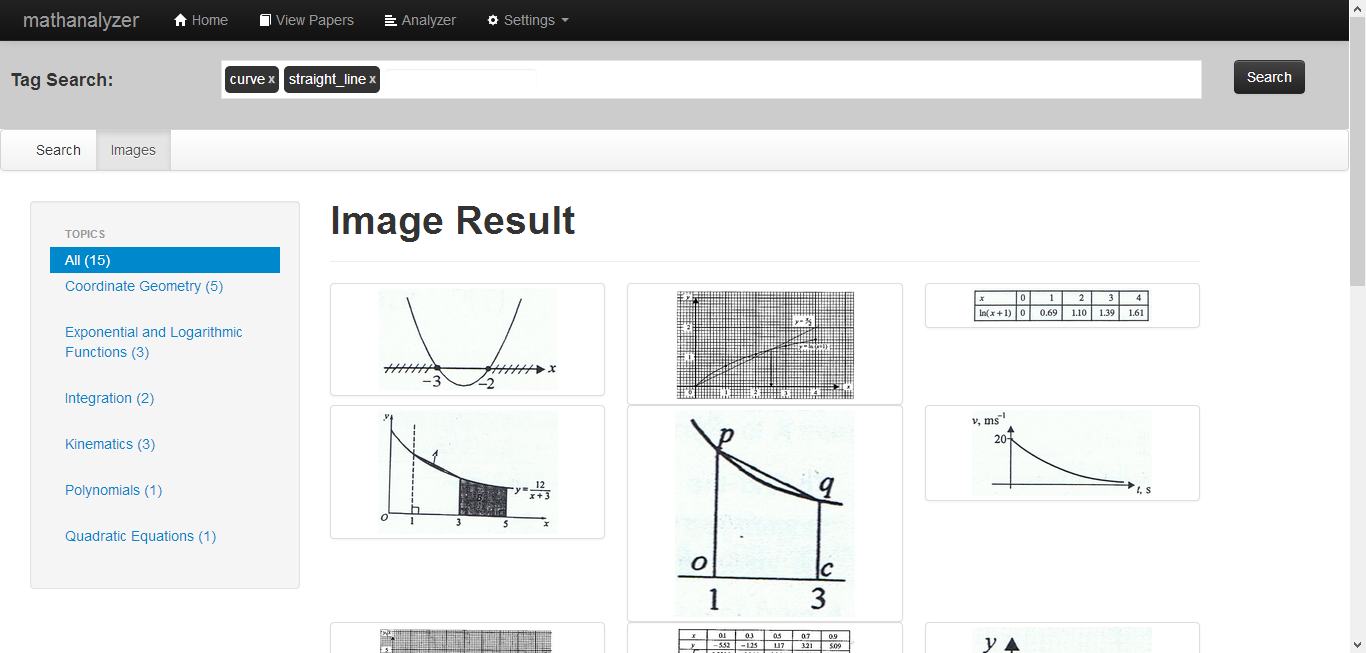


Figure ‑: Search Result (Image Tab)

### Admin Question Selection

When user clicked on “Modify Additional Math” link under the “Settings” section of the navigation bar, user will be redirected to the question selection page. User can either select a Paper or select a Topic to view the questions.



Figure ‑: Admin Question Selection Page

### Admin Question Listing

Upon selecting either a Paper or a Topic, the questions will be listed in the Question Listing page. User can click on “Modify” button on any of the question, or click “Insert New Question” for creating a new question.

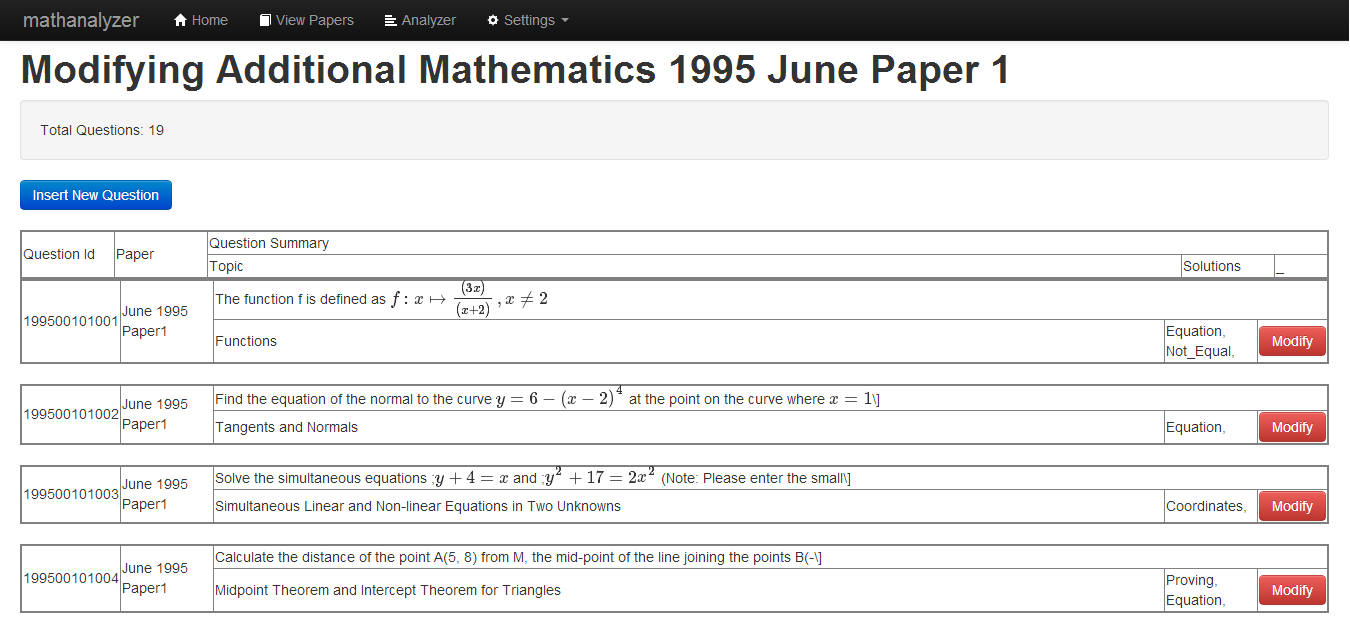


Figure ‑: Admin Question Listing Page

### Admin Create/Modify Question Form

When user clicks “Insert New Question” or “Modify” button, user will be redirected to the Question Form. User can then fill in any details they wish to add/change, and click “Submit” button.

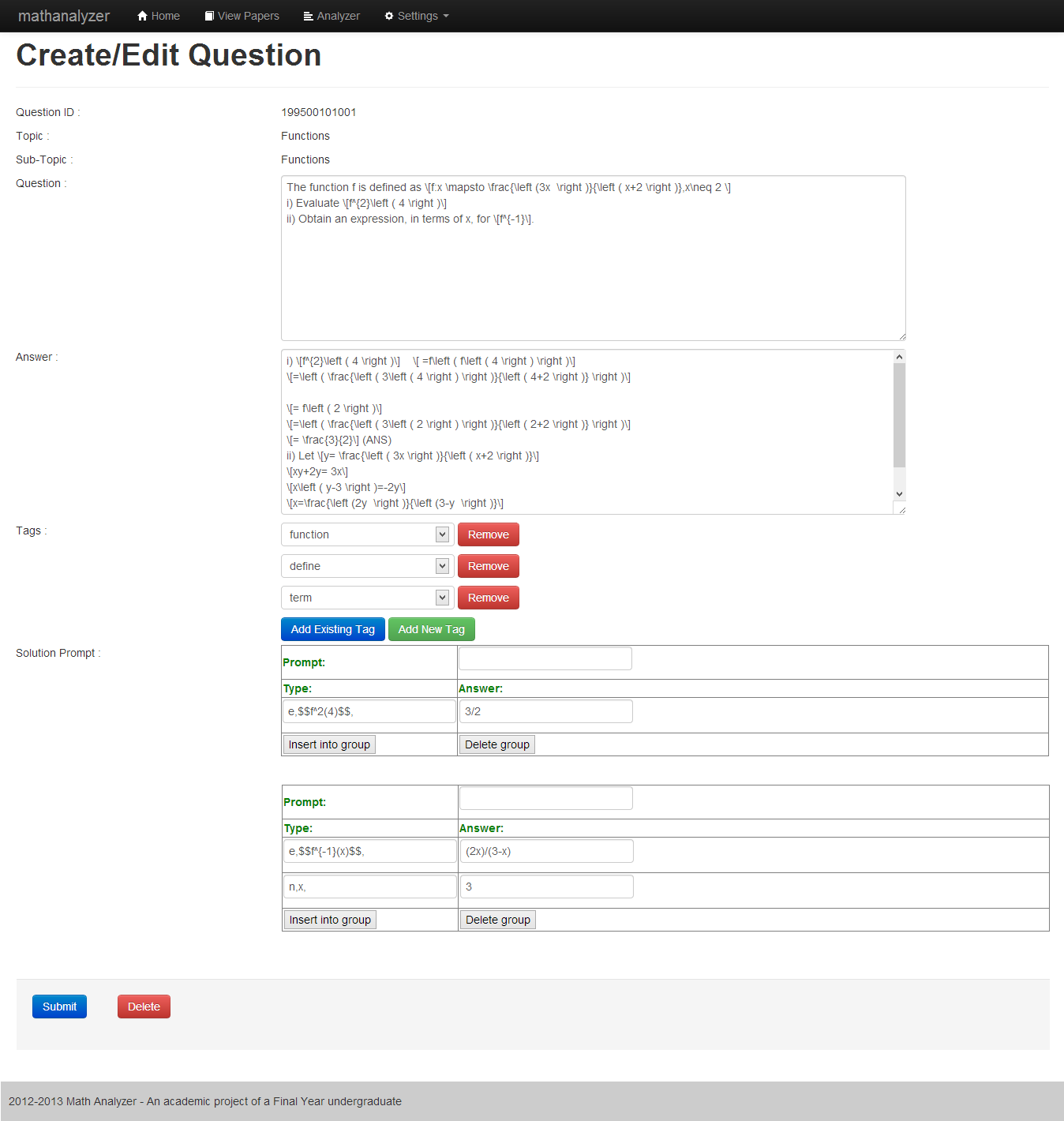


Figure ‑: Admin Question Form Page

### Admin Tag Listing

When user clicks on “Modify Tags” under the “Admin” section of the navigation bar, user will be navigated to the Tag Listing Page.

User can choose to filter the tag types under “Tag Type” selection list.

User can also regenerate keywords by clicking on “Regenerate Keywords” button. This is only works on keyword tags as any changes have to be committed by regenerating the question-tag relationships.

User can also create new tag by clicking “Add Tag” button or edit tag by clicking “Edit” button. User can also delete the tag by clicking the “Delete” button.

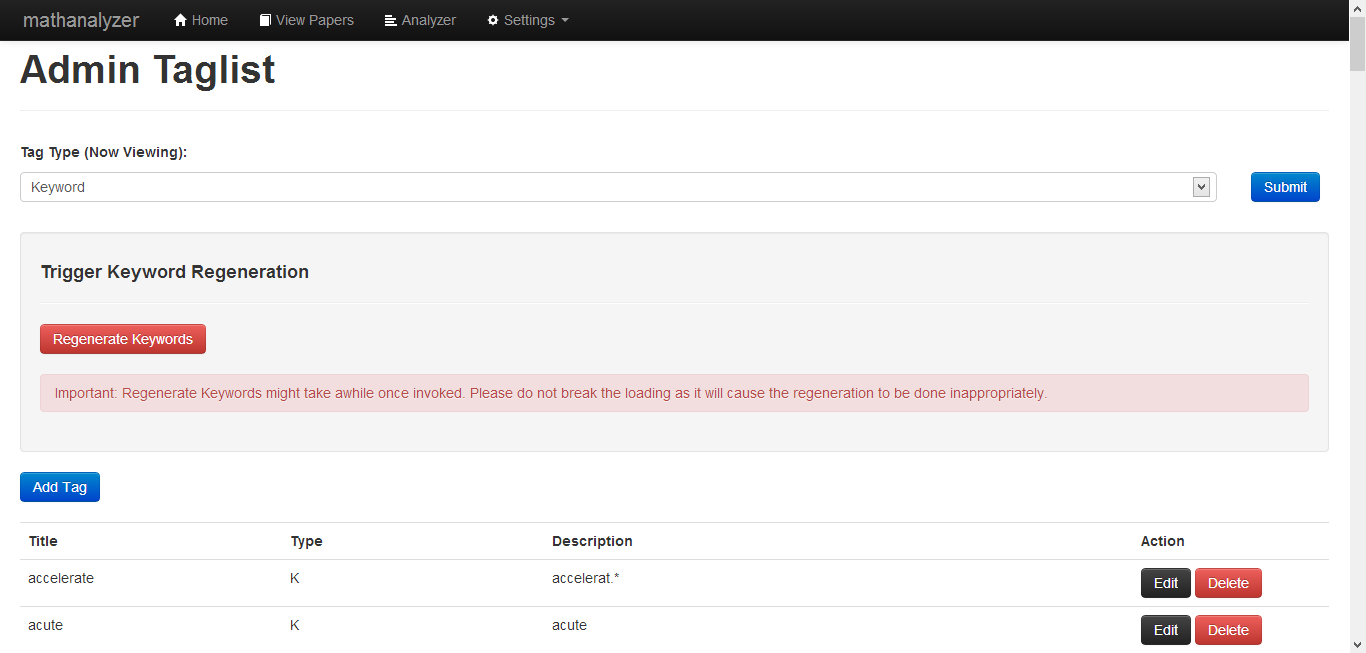


Figure ‑: Admin Tag Listing page

### Admin Create/Modify Tag Form

Upon clicking “Add Tag” or “Modify” button, page will display the Tag Form for user to add/edit data information. User can then click “Submit” to save.

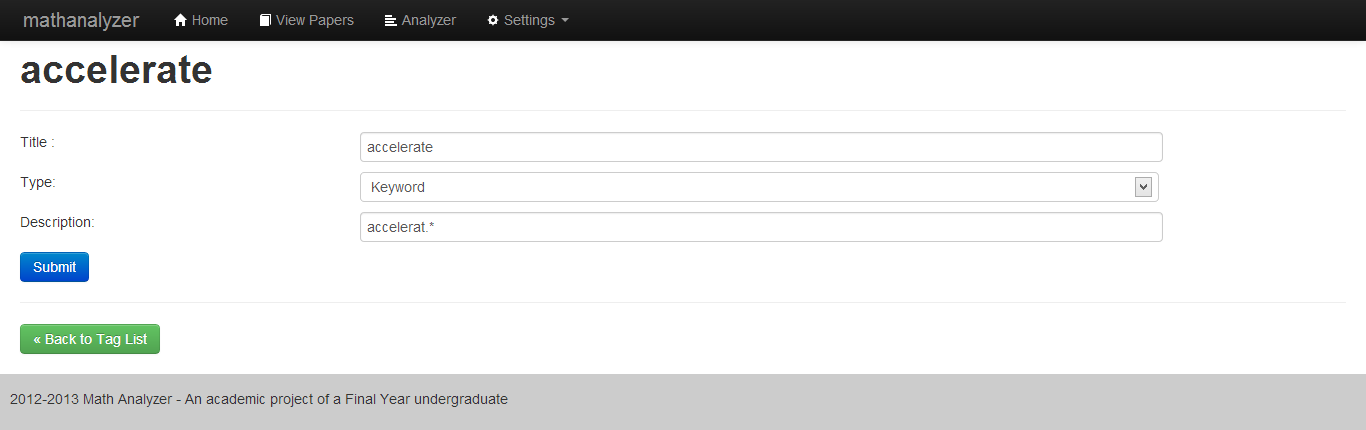


Figure ‑: Admin Tag Form

### Django Admin

This section only available for Admin user to change almost any data that can be found in the database. This Django Admin was mentioned earlier as the automatic generated Admin site once the database was synced. After logging in, user can see the list of table names.

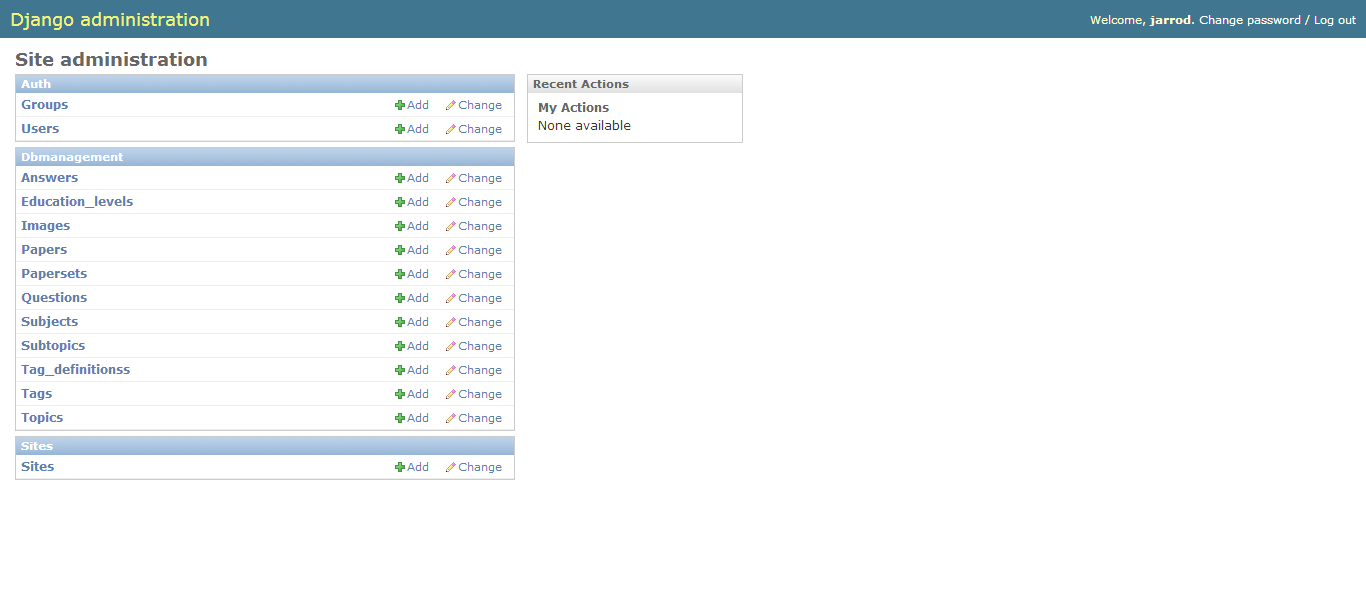


Figure ‑ Django Admin Table Listing

When a table is selected, user can then view the data within. User can choose to select item(s) and do an action, or add additional data by clicking the “Add …” button.



Figure ‑: Django Admin Table Data Listing

Lastly, if user intends to change or add a data item, a form will be loaded with all the relevant attributes. User can then make the changes and click “Save”

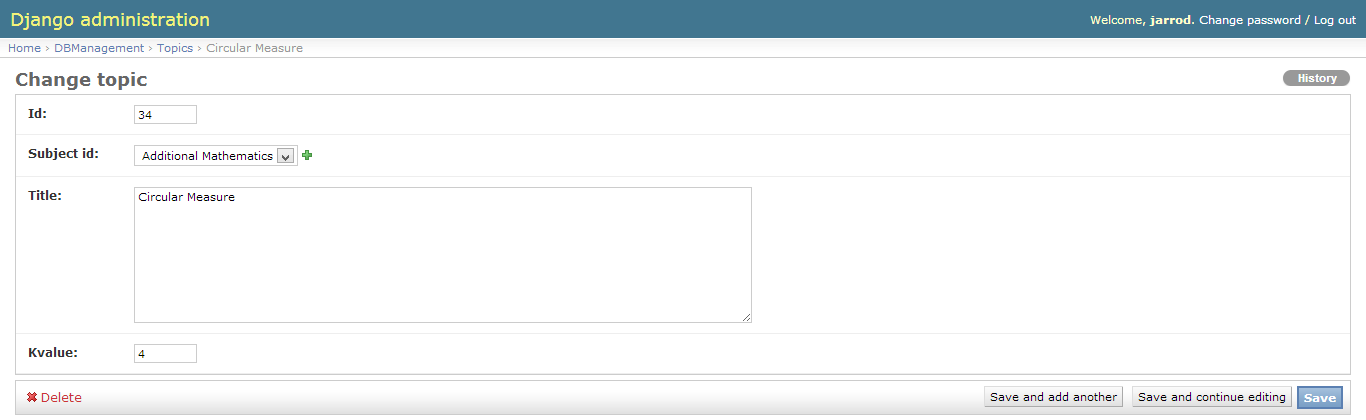


Figure ‑: Django Admin Form

## Tag List

Table ‑: Tag List for Generic

|  |  |  |
| --- | --- | --- |
| **Generic** | | |
| Title | Type | Content |
| area | Keyword | area.? |
| curve | Keyword | curve.? |
| diagram | Keyword | diagram.? |
| draw | Keyword | draw.? |
| graph | Keyword | graph.? |
| prove | Keyword | prov.{0,3} |
| sketch | Keyword | sketch.\* |
| value | Keyword | value.? |

Table ‑: Tag List for Set Language and Notations

|  |  |  |
| --- | --- | --- |
| **Set Language and Notations** | | |
| Title | Type | Content |
| element | Keyword | element.? |
| empty | Keyword | empty |
| set | Keyword | set.? |
| set\_notation | Keyword | set notation.? |
| subset | Keyword | subset.? |
| venn\_diagram | Keyword | venn diagram.? |

Table ‑: Tag List for Functions

|  |  |  |
| --- | --- | --- |
| **Functions** | | |
| Title | Type | Content |
| define | Keyword | defin.\* |
| domain | Keyword | domain.? |
| function | Keyword | function.? |
| inverse | Keyword | inverse |
| map | Keyword | map.? |

Table ‑: Tag List for Quadratic Functions

|  |  |  |
| --- | --- | --- |
| **Quadratic Functions** | | |
| Title | Type | Content |
| intersect | Keyword | intersect.\* |
| mid-point | Keyword | mid-?point.? |
| parallel | Keyword | parallel |
| perpendicular | Keyword | perpendicular |
| perpendicular\_bisector | Keyword | perpendicular bisector.? |
| quadratic | Keyword | quadratic |
| range\_of\_value | Keyword | range of value.? |
| real | Keyword | real |
| root | Keyword | root.? |
| straight\_line | Keyword | straight line.? |
| Absolute\_Values\_And\_Functions | Concept | Quadratic Equations - |
| Inequality\_For\_(x-h)(x-k)<0 | Concept | Quadratic Equations – For , when . |
| Inequality\_For\_(x-h)(x-k)>0 | Concept | Quadratic Equations - For , when , or . |
| Line\_Cuts\_Curve\_At\_Two\_Points | Concept | Quadratic Equations - When , Line cuts the curve at 2 distinct points |
| Line\_Does\_Not\_Intersects\_Curve | Concept | Quadratic Equations – When , Line does not intersect the curve. |
| Line\_Tangent\_To\_Curve | Concept | Quadratic Equations - When , Line is tangent to the curve. |
| Nature\_Of\_Roots\_For\_Line\_And\_Curve | Concept | Quadratic Equations - Nature of the roots depends on the value of the discriminant after solving the line and curve equations simultaneously. |
| No\_Real\_Roots | Concept | Quadratic Equations - When , there are no real roots. |
| Product\_Of\_Roots | Formula | Quadratic Equations - (Product of Roots) |
| Real\_And\_Distinct\_Roots | Concept | Quadratic Equations - When , there are 2 real and distinct roots. |
| Real\_And\_Equal\_Roots | Concept | Quadratic Equations - When , there are 2 real and equal roots. |
| Sum\_Of\_Roots | Formula | Quadratic Equations - (Sum of Roots) |

Table ‑: Tag List for Indices and Surds

|  |  |  |
| --- | --- | --- |
| **Indices and Surds** | | |
| Title | Type | Content |
| calculator | Keyword | calculator |
| form | Keyword | form.? |
| simplify | Keyword | simplify |
| surd | Keyword | surd.? |
| Indices\_Identities\_1 | Formula | Indices - , where a, m, n are positive integers |
| Indices\_Identities\_2 | Formula | Indices - , where a, m, n are positive integers |
| Indices\_Identities\_3 | Formula | Indices - , where a, m, n are positive integers |
| Indices\_Identities\_4 | Formula | Indices - , where a, b, m, n are positive integers |
| Indices\_Identities\_5 | Formula | Indices - , where a, b, m, n are positive integers |
| Indices\_Identities\_6 | Formula | Indices - , where a is positive integer |
| Indices\_Identities\_7 | Formula | Indices - , where a, n are positive integers |
| Indices\_Identities\_8 | Formula | Indices - , where a, n are positive integers |
| Indices\_Identities\_9 | Formula | Indices - , where a, m, n are positive integers |
| Indices\_Identities\_10 | Formula | Indices - If then where and |
| Surds\_Identities\_1 | Formula | Surds - , where |
| Surds\_Identities\_2 | Formula | Surds - , where , |
| Surds\_Identities\_3 | Formula | Surds - |

Table ‑: Tag List for Polynomials

|  |  |  |
| --- | --- | --- |
| **Polynomials** | | |
| Title | Type | Content |
| divide | Keyword | divid.\* |
| factor | Keyword | factor.\* |
| leave | Keyword | leav.\* |
| polynomial | Keyword | polynomial |
| power | Keyword | power |
| Factor\_Theorem | Formula | Polynomial - , where , is a factor of the polynomial (Factor Theorem) |
| Factorise\_Cubic\_Expression\_By\_Comparing\_Coefficient | Concept | Polynomial - If one of the factor is given, by comparing coefficient, the other factor must be second degree of the form |
| Factorise\_Cubic\_Expression\_By\_Long\_Division | Concept | Polynomial - Using Long division |
| Factorise\_Cubic\_Expression\_By\_Trial\_And\_Error | Concept | Polynomial - Using Trial and Error, factor of should give remainder of zero. |
| Remainder\_Theorem | Formula | Polynomial - When the polynomial is divided by where , the remainder R is given by R = . (Remainder Theorem) |

Table ‑: Tag List for Simultaneous Equations

|  |  |  |
| --- | --- | --- |
| **Simultaneous Equations** | | |
| Title | Type | Content |
| pair | Keyword | pair.? |
| simultaneous | Keyword | simultaneous |
| solution | Keyword | solution.? |
| solve | Keyword | solv.? |
| Simultaneous\_By\_Elimination | Concept | Simultaneous - Express one unknown of equal value in both equations, then eliminate the unknown to get the value of the other known (Elimination) |
| Simultaneous\_By\_Substitution | Concept | Simultaneous - Express one unknown in terms of the other, then substitute the result to solve for the unknown (Substitution) |
| Simultaneous\_Of\_Linear\_And\_Non-Linear | Concept | Simultaneous - Express one unknown in terms of the other with the linear equation, then substitute the result to the non-linear equation to solve it |

Table ‑: Tag List for Logarithmic and Exponential Functions

|  |  |  |
| --- | --- | --- |
| **Logarithmic and Exponential Functions** | | |
| Title | Type | Content |
| e | Keyword | e\^.\* |
| lg | Keyword | lg.\* |
| ln | Keyword | ln.\* |
| Log\_Change\_Of\_Base | Formula | Logarithms -, where and , and (Change of Base) |
| Log\_Of\_Integer\_Same\_To\_Base | Formula | Logarithms - |
| Log\_Of\_One | Formula | Logarithms - |
| Log\_Power\_Law | Formula | Logarithms - (Power Law) |
| Log\_Product\_law | Formula | Logarithms - (Product Law) |
| Log\_Quotient\_Law | Formula | Logarithms - (Quotient Law) |
| Log\_To\_Indices | Formula | Logarithms - means |

Table ‑: Tag List for Partial Fractions

|  |  |  |
| --- | --- | --- |
| **Partial Fractions** | | |
| Title | Type | Content |
| partial\_fraction | Keyword | partial fraction.? |
| Deriving\_Partial\_Fraction\_From\_Linear\_Fraction | Formula | Partial Fraction - Case 1: Linear Fraction e.g. , express in |
| Deriving\_Partial\_Fraction\_From\_Quadratic\_Factor | Formula | Partial Fraction - Case 3: Quadratic Factor which does not factorise e.g. , express in |
| Deriving\_Partial\_Fraction\_From\_Repeated\_Fraction | Formula | Partial Fraction - Case 2: Repeated Fraction e.g. , express in |
| Solve\_Partial\_Fraction\_By\_Comparing\_Coefficients | Concept | Partial Fraction - Solving unknown constants by comparing coefficients of like terms |
| Solve\_Partial\_Fraction\_By\_Substitution | Concept | Partial Fraction - Solving unknown constants by substituting values of x |

Table ‑: Tag List for Binomial Expansions

|  |  |  |
| --- | --- | --- |
| **Binomial Expansions** | | |
| Title | Type | Content |
| ascend | Keyword | ascend.\* |
| binomial | Keyword | binomial |
| coefficient | Keyword | coefficient.\* |
| descend | Keyword | descend.\* |
| remainder | Keyword | remainder.? |
| term | Keyword | term.? |
| Binomial\_Term | Formula | Binomial - th term is |
| Binomial\_Theorem | Formula | Binomial - |
| Term\_Independent | Concept | Binomial - Term Independent refers to the constant term in the binomial expansion. |

Table ‑: Tag List for Modulus Functions

|  |  |  |
| --- | --- | --- |
| **Modulus Functions** | | |
| Title | Type | Content |
| modulus | Keyword | | \*| |

Table ‑: Tag List for Trigonometry

|  |  |  |
| --- | --- | --- |
| **Trigonometry** | | |
| Title | Type | Content |
| acute | Keyword | acute |
| amplitude | Keyword | amplitude |
| angle | Keyword | angle.? |
| identity | Keyword | identit.\* |
| measure | Keyword | measur.\* |
| period | Keyword | period.? |
| Cosine\_Angles | Formula | Trigonometry - , |
| Positive\_Negative\_Angles | Concept | Trigonometry - Angles measured anticlockwise from the positive direction of x-axis are positive while angles measured clockwise from the positive direction of the x-axis are negative |
| Sine\_Angles | Formula | Trigonometry - , |
| Solving\_Multiple\_Angles | Concept | Trigonometry - Equations involving multiple angles are usually solved by converting the equation into another equation that contains only one trigonometrical ratio or by using the values from calculators. |
| Tangent\_Angles | Formula | Trigonometry - , |
| Trigonometry\_Factor\_Formulae\_1 | Formula | Trigonometry - |
| Trigonometry\_Factor\_Formulae\_2 | Formula | Trigonometry – |
| Trigonometry\_Factor\_Formulae\_3 | Formula | Trigonometry - |
| Trigonometry\_Factor\_Formulae\_4 | Formula | Trigonometry - |
| Trigonometry\_Form\_1 | Formula | Trigonometry – where , , and are positive and is acute |
| Trigonometry\_Form\_2 | Formula | Trigonometry - where , , and are positive and is acute |
| Trigonometry\_Identities\_1 | Formula | Trigonometry - |
| Trigonometry\_Identities\_10 | Formula | Trigonometry - |
| Trigonometry\_Identities\_11 | Formula | Trigonometry - |
| Trigonometry\_Identities\_12 | Formula | Trigonometry - |
| Trigonometry\_Identities\_13 | Formula | Trigonometry - |
| Trigonometry\_Identities\_14 | Formula | Trigonometry – |
| Trigonometry\_Identities\_15 | Formula | Trigonometry - |
| Trigonometry\_Identities\_16 | Formula | Trigonometry - |
| Trigonometry\_Identities\_17 | Formula | Trigonometry - |
| Trigonometry\_Identities\_18 | Formula | Trigonometry - |
| Trigonometry\_Identities\_2 | Formula | Trigonometry - |
| Trigonometry\_Identities\_3 | Formula | Trigonometry - |
| Trigonometry\_Identities\_4 | Formula | Trigonometry - |
| Trigonometry\_Identities\_5 | Formula | Trigonometry - |
| Trigonometry\_Identities\_6 | Formula | Trigonometry - |
| Trigonometry\_Identities\_7 | Formula | Trigonometry - |
| Trigonometry\_Identities\_8 | Formula | Trigonometry - |
| Trigonometry\_Identities\_9 | Formula | Trigonometry - |

Table ‑: Tag List for Coordinate Geometry

|  |  |  |
| --- | --- | --- |
| **Coordinate Geometry** | | |
| Title | Type | Content |
| cartesian | Keyword | cartesian |
| coordinate | Keyword | coordinate.? |
| point | Keyword | point.? |
| vertice | Keyword | vert.\* |
| Distance\_Of\_Line | Formula | Coordinate Geometry - Distance = |
| Equation\_Of\_Circle | Formula | Coordinate Geometry - Equation of Circle = , where is the centre and is the radius. |
| Equation\_Of\_Line | Formula | Coordinate Geometry - Equation of Line = , where = gradient of line. |
| Equation\_Of\_Line\_From\_Non-Linear\_Equation | Concept | Coordinate Geometry - Express equation in form of where X and Y are expressions in x and/or y. |
| Gradient\_Of\_Line | Formula | Coordinate Geometry - Gradient = |
| Mid-point\_Of\_Line | Formula | Coordinate Geometry - Mid-point = |
| Parallel\_Lines | Concept | Coordinate Geometry - Line and are parallel if . |
| Perpendicular\_Lines | Concept | Coordinate Geometry - Line and are perpendicular if . |

Table ‑: Tag List for Plane Geometry

|  |  |  |
| --- | --- | --- |
| **Plane Geometry** | | |
| Title | Type | Content |
| congruent | Keyword | congruen.\* |
| parallelogram | Keyword | parallelogram.? |
| quadrilateral | Keyword | quadrilateral |
| square | Keyword | square.? |
| trapezium | Keyword | trapezium.? |
| triangle | Keyword | triangle.? |
| Area\_Of\_Plane | Formula | Plane Geometry - Area of Plane = |

Table ‑: Tag List for Differentiation

|  |  |  |
| --- | --- | --- |
| **Differentiation** | | |
| Title | Type | Content |
| approximate | Keyword | approximat.\* |
| constant | Keyword | constant.? |
| decrease | Keyword | decreas.\* |
| differentiate | Keyword | differentiat.\* |
| distinct\_point | Keyword | distinct point.? |
| gradient | Keyword | gradient.? |
| increase | Keyword | increas.\* |
| instant | Keyword | instant.? |
| maximum | Keyword | maximum |
| minimum | Keyword | minimum |
| normal | Keyword | normal |
| percentage | Keyword | percentage |
| rate\_of\_change | Keyword | rate of change.? |
| respect | Keyword | respect |
| small | Keyword | small.\* |
| stationary | Keyword | stationary |
| tangent | Keyword | tangent |
| turning\_point | Keyword | turning point.? |
| vary | Keyword | vary.\* |
| volume | Keyword | volume |
| Chain\_Rule | Formula | Differentiation - (Chain Rule) |
| Decreasing\_Function | Concept | Differentiation - Decreasing functions has . |
| Derivative\_Of\_Cosine | Formula | Differentiation - |
| Derivative\_Of\_Cosine\_Power | Formula | Differentiation - |
| Derivative\_Of\_Exponentials | Formula | Differentiation - |
| Derivative\_Of\_Natural\_Log | Formula | Differentiation - |
| Derivative\_Of\_Power | Formula | Differentiation - |
| Derivative\_Of\_Sine | Formula | Differentiation - |
| Derivative\_Of\_Sine\_Power | Formula | Differentiation - |
| Derivative\_Of\_Tangent | Formula | Differentiation - |
| Derivative\_Of\_Tangent\_Power | Formula | Differentiation - |
| Equation\_Of\_Normal | Formula | Differentiation - (Equation of Normal) |
| Equation\_Of\_Tangent | Formula | Differentiation - (Equation of Tangent) |
| Gradient\_Of\_Curve | Formula | Differentiation - (Gradient of Curve) |
| Higher\_Derivatives | Formula | Differentiation - |
| Increasing\_Function | Concept | Differentiation - Increasing functions has . |
| Maximum\_Point | Concept | Differentiation - If , then it is a maximum point. |
| Minimum\_Point | Concept | Differentiation - If , then it is a minimum point. |
| Motion\_With\_Variable\_Velocity\_And\_Acceleration | Concept | Differentiation - Motion with Variable Velocity and Acceleration |
| No\_Turning\_Point | Concept | Differentiation - If has no solution, then there is no turning point. |
| Point\_Of\_Inflexion | Concept | Differentiation - If does not change as it passes through the stationary point, then it is a point of inflexion. |
| Product\_Rule | Formula | Differentiation - (Product Rule) |
| Quotient\_Rule | Formula | Differentiation - (Quotient Rule) |
| Stationary\_Point | Concept | Differentiation - When , it is a Stationary Point. |

Table ‑: Tag List for Integration

|  |  |  |
| --- | --- | --- |
| **Integration** | | |
| Title | Type | Content |
| bound | Keyword | bound.\* |
| enclose | Keyword | enclos.\* |
| region | Keyword | region.? |
| rotate | Keyword | rotat.\* |
| shaded | Keyword | shaded |
| Area\_Bounded\_By\_Curve\_x=f(y) | Formula | Integration - The area bounded by the curve , the y-axis and the lines and is given by |
| Area\_Bounded\_By\_Curve\_y=f(x) | Formula | Integration - The area bounded by the curve , the x-axis and the lines and is given by |
| Basic\_Integration\_Formulae\_1 | Formula | Integration - |
| Basic\_Integration\_Formulae\_2 | Formula | Integration - |
| Basic\_Integration\_Formulae\_3 | Formula | Integration - |
| Basic\_Integration\_Formulae\_4 | Formula | Integration - |
| Definite\_Integral | Formula | Integration - If , then |
| Integral\_Of\_Cosine | Formula | Integration - |
| Integral\_Of\_Exponential\_Function\_1 | Formula | Integration - |
| Integral\_Of\_Exponential\_Function\_2 | Formula | Integration - |
| Integral\_Of\_Exponential\_Function\_3 | Formula | Integration - |
| Integral\_Of\_Fraction\_Function\_1 | Formula | Integration - |
| Integral\_Of\_Fraction\_Function\_2 | Formula | Integration - |
| Integral\_Of\_Fraction\_Function\_3 | Formula | Integration - |
| Integral\_Of\_Secant\_Square | Formula | Integration - |
| Integral\_Of\_Sine | Formula | Integration - |
| Integration\_By\_Reverse\_Of\_Differentiation | Formula | Integration - If, then where is a constant. |
| Kinematics\_Distance\_Using\_Integration | Formula | Integration - If the velocity of a particle is given as a function of time , i.e. , then will give the expression for the distance covered. |
| Kinematics\_Velocity\_Using\_Integration | Formula | Integration - If the acceleration of a particle is given as a function of time , i.e. , then will give the expression for the velocity. |

Table ‑: Tag List for Kinematics

|  |  |  |
| --- | --- | --- |
| **Kinematics** | | |
| Title | Type | Content |
| acceleration | Keyword | accelerat.\* |
| decelerate | Keyword | decelerat.\* |
| displacement | Keyword | displacement.? |
| distance | Keyword | distance.? |
| instantaneous | Keyword | instantaneous |
| move | Keyword | mov.{0,3} |
| pass | Keyword | pass.{0,3} |
| reach | Keyword | reach.\* |
| rest | Keyword | rest.? |
| retard | Keyword | retard.\* |
| second | Keyword | second.? |
| speed | Keyword | speed.? |
| time | Keyword | time |
| travel | Keyword | travel.\* |
| uniform | Keyword | uniform |
| velocity | Keyword | velocity |
| way | Keyword | way |
| wind | Keyword | wind |

Table ‑: Tag List for Matrices

|  |  |  |
| --- | --- | --- |
| **Matrices** | | |
| Title | Type | Concept |
| matrice | Keyword | matri.{0,3} |
| matrix\_multiplication | Keyword | matrix multiplication.? |
| product | Keyword | product |
| Identity\_Matrix | Formula | Matrix - |
| Inverse\_Matrix | Formula | Matrix - If , provided det |
| Matrix\_Determinant | Formula | Matrix - If , det |
| Scalar\_Matrix\_Multiplication | Formula | Matrix - If |
| Simultaneous\_By\_Matrix | Concept | Matrix - A pair of simultaneous equations can be expressed in matrix form as . Then the solution is , if det . |
| Two\_Matrices\_Multiplication | Formula | Matrix - If , then |

Table ‑: Tag List for Vectors

|  |  |  |
| --- | --- | --- |
| **Vectors** | | |
| Title | Type | Content |
| bear | Keyword | bear.\* |
| collinear | Keyword | collinear |
| direction | Keyword | direction.? |
| due | Keyword | due |
| east | Keyword | east |
| magnitude | Keyword | magnitude.? |
| north | Keyword | north |
| origin | Keyword | origin.? |
| position | Keyword | position.? |
| scalar | Keyword | scalar |
| unit | Keyword | unit.? |
| vector | Keyword | vector.? |

Table ‑: Tag List for Permutations and Combinations

|  |  |  |
| --- | --- | --- |
| **Permutations and Combinations** | | |
| Title | Type | Content |
| arrange | Keyword | arrang.\* |
| number | Keyword | number.? |

Table ‑: Tag List for Arithmetic Progression and Geometric Progression

|  |  |  |
| --- | --- | --- |
| **Arithmetic Progression and Geometric Progression** | | |
| Title | Type | Content |
| arithmetic\_progression | Keyword | arithmetic progression.? |
| common\_ratio | Keyword | common ratio.? |
| geometric\_progression | Keyword | geometric progression.? |

Table ‑: Tag List for Circular Measure

|  |  |  |
| --- | --- | --- |
| **Circular Measure** | | |
| Title | Type | Content |
| arc | Keyword | arc |
| centre | Keyword | centre |
| chord | Keyword | chord.? |
| circle | Keyword | circle.? |
| circular | Keyword | circular |
| circumference | Keyword | circumference.? |
| perimeter | Keyword | perimeter.? |
| radian | Keyword | radian |
| radius | Keyword | radius |
| sector | Keyword | sector.? |
| semicircle | Keyword | semi-?circle.? |

## Topic Tag Cloud Full Evaluation

**Quadratic Functions**

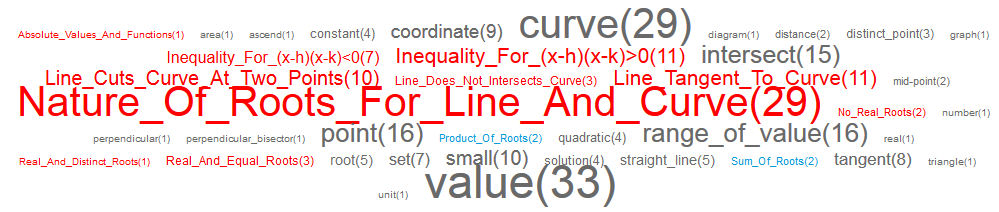


Figure ‑: One-Tag Cloud of Quadratic Functions

From Figure A-27, it is clear that questions on quadratic functions deal a lot on finding the nature of roots. Questions describing line cutting curve at two points or tangent to the curve are common as well, and finding quadratic inequalities are not scarce too. However, one-tag cloud only shows the term frequency of each tag, it is still unknown whether these top most common tags might have relationships with each other.

Note: Multiple-tags cloud may not show full detail of the generated list as term frequency of 1 might not be considered useful information (unless for case of four-tag or five-tag clouds).

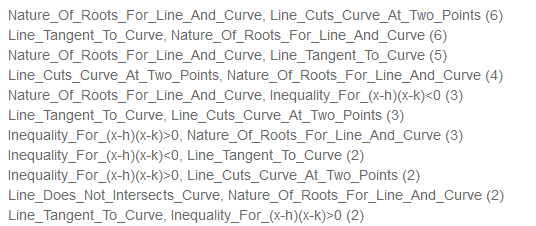


Figure ‑: Two-Tag Cloud for Quadratic Equations

Figure A-28 showed a two-tag cloud generated with a filtered criteria of “Concept and Formula only”. Exclusion of keyword tags from the tag cloud makes the list more concise as most of the keywords are not helping much in understanding the gist of the topic. From the two-tag cloud, it is confirmed that there are indeed relationships between the most common tags. Top few on the lists are describing “nature of roots” with combination of “line cuts the curve at 2 points” or “line is tangent to the curve”. There are also some less frequent items describing “nature of roots” with quadratic inequalities.

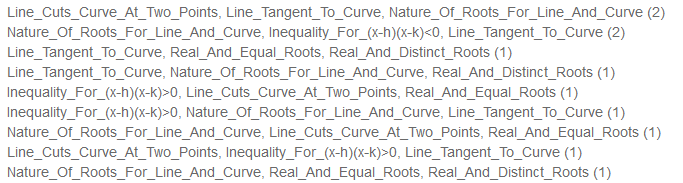


Figure ‑: Three-Tag Cloud for Quadratic Equations

From Figure A-29, the top most common item reduced from the previous 6 (from two-tag) to 2. Though tags appearing on the three-tag cloud are still the common tags such as “nature of roots”, “line tangent to curve”, “line cuts curve at two points”, “inequalities”, the three-tag cloud probably tells us that the questions are further tweaked with inclusion of other concepts to make it a little more unique than before. However, the list still shows that the popular questions revolve around root properties, with line cutting on 2 points or line tangent to curve, and quadratic inequalities.

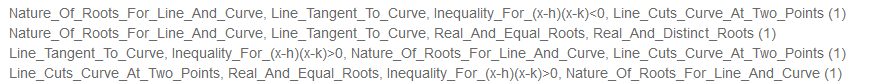


Figure ‑: Four-Tag Cloud for Quadratic Equations

The final piece of tag cloud on quadratic equations was shown in Figure A-30. There is no trace of any terms in five-tag cloud. This tag cloud further confirms the previous assumption that the popular questions still revolves around root properties, line cutting on 2 points, line tangent to curve, and quadratic inequalities.

**Indices and Surds**

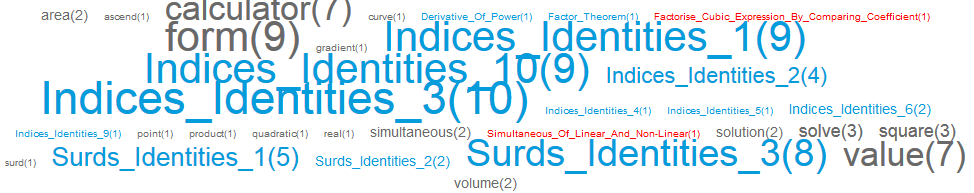


Figure ‑: One-Tag Cloud for Indices and Surds

From the one-tag cloud in Figure A-31, it is clear that indices and surds is a chapter purely on its identity formulae. However, there are traces of foreign topics’ concepts and formulae like Polynomial’s Factor Theorem and Simultaneous Equation’s concept of solving linear and non-linear equations simultaneously, are applied to the questions as well, but as a rare case.

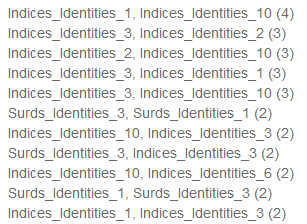


Figure ‑: Two-Tag Cloud for Indices and Surds

Two-tag cloud in Figure A-32 showed mostly formulae about identity.

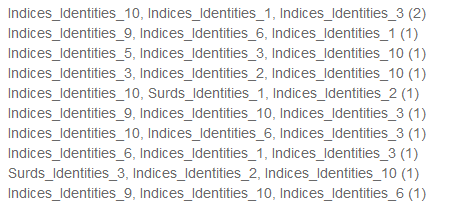


Figure ‑: Three-Tag Cloud for Indices and Surds

The three-tag cloud in Figure A-33 showed clearly the domination of questions on identities. Similar cases also shown on four-tag and five-tag cloud. Therefore, it is conclusive that the chapter is mostly about identities.

**Polynomials**

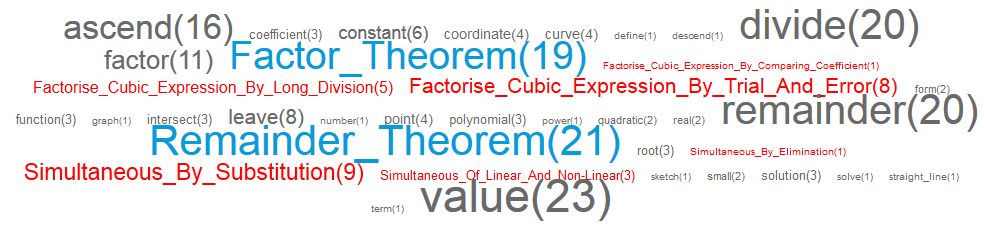


Figure ‑: One-Tag Cloud for Polynomials

The one-tag cloud on polynomial in Figure A-34 shows that the major concepts tested are Factor and Remainder Theorem, as both of their formulae ranked the highest, together with keyword “divide”, “remainder” and “factor” which are commonly used to describe Factor Theorem and Remainder Theorem.

There are also a smaller portion of questions testing on factorising cubic expressions, and concepts on Simultaneous Equations are commonly used in Polynomial topic too.

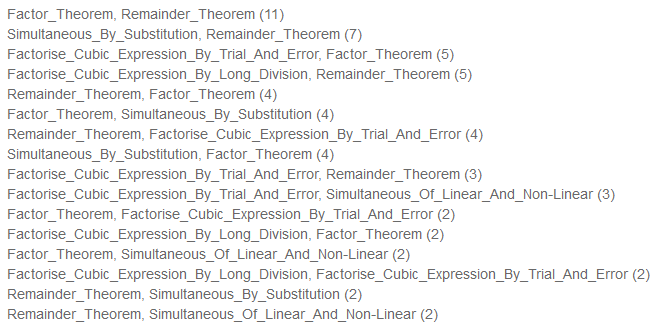


Figure ‑: Two-Tag Cloud for Polynomials

The two-tag cloud in Figure A-35 further shows that the questions are mostly focusing on factor theorem and remainder theorem. Simultaneous equations are commonly used to solve questions together with the 2 theorems and factorising cubic expressions are commonly together with the theorems too.

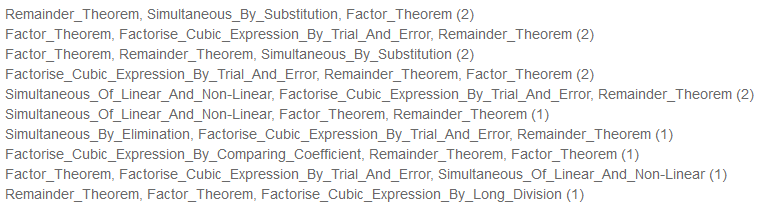


Figure ‑: Three-Tag Cloud for Polynomials

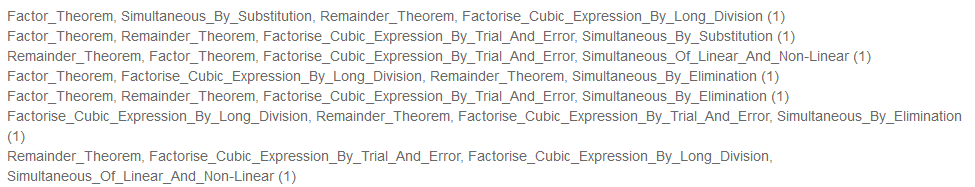


Figure ‑: Four-Tag Cloud for Polynomials



Figure ‑: Five-Tag Cloud for Polynomials

The three-tag (Figure A-36), four-tag (Figure A-37), five-tag (Figure A-38) clouds further proved that factor theorem, remainder theorem are the main focus for Polynomial questions, and concepts and formulae on factorising cubic expression and simultaneous equations are also crucial for some of the questions too.

**Partial Fractions**



Figure ‑: One-Tag Cloud for Partial Fractions



Figure ‑: Two-Tag Cloud for Partial Fractions

For Partial Fractions, due to the fact that it is a young topic debuted on 2008 November paper, not much questions were tested on it. It is a fact that every year since 2008, there are 1 question on partial fraction, and each time, at least 1 of the concept was tested. From figure A-39 on one-tag cloud, it is shown that all the concepts and formulae on Partial Fractions are already tested.

Figure A-40 showed two-tag cloud with a record of 2 concepts/formulae on Partial Fractions. This proved that the question may not just test on a single concept/formula only. Multiple concepts/formulae on partial fraction may be required. And currently, this question stands independent from other topic as no foreign topic’s concept or formula is found in the tag cloud.

**Simultaneous Equations**

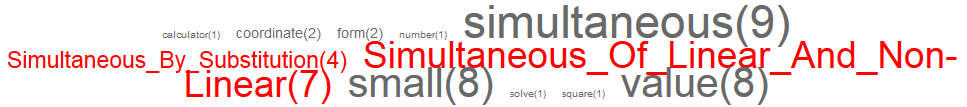


Figure ‑: One-Tag Cloud for Simultaneous Equations

Simultaneous Equation is a straight-forward topic with few concepts. There is no formula (blue tags) involved as seen from Figure A-41. There are no multiple-tag clouds involved in simultaneous equations and therefore we can conclude that the concepts are independently applied (no conjunctive usage).

**Binomial Theorem**



Figure ‑: One-Tag Cloud for Binomial Theorem

From the one-tag cloud in Figure A-42, the terms showed that Binomial Theorem are largely testing on the formulae for finding its term and for listing the whole binomial series. Concepts such as term independent featured on some questions as well. Simultaneous equation technique of simultaneous by substitution is also tested on binomial questions.

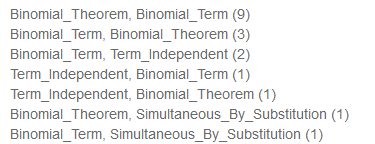


Figure ‑: Two-Tag Cloud for Binomial Theorem



Figure ‑: Three-Tag Cloud for Binomial Theorem

From Figure A-43, the two-tag cloud showed that most of the questions for Binomial Theorem are testing on the binomial theorem and finding binomial term. There are some questions testing on term independent and elements of simultaneous equations as well.

From Figure A-44, it is deduced that the term independent and simultaneous equation are not tested separately, but in conjunction to binomial theorem and binomial term.

**Exponential and Logarithmic Functions**

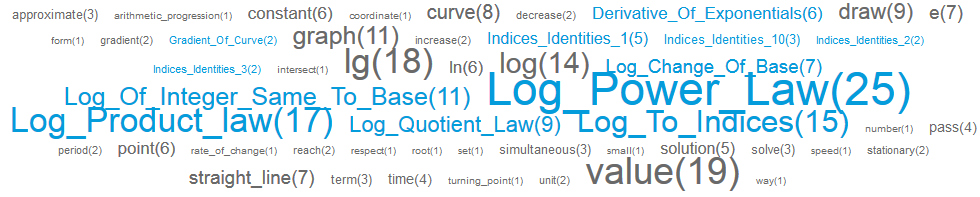


Figure ‑: One-Tag Cloud for Exponential and Logarithmic Functions

The one-tag cloud in Figure A-45 showed that Exponential and Logarithmic Functions topic are focusing mostly on logarithmic identities. Exponential functions also covers derivative of exponentials which is a formula from Differentiation. Indices identities are also an important aspect in solving questions of Exponential and Logarithmic Functions.

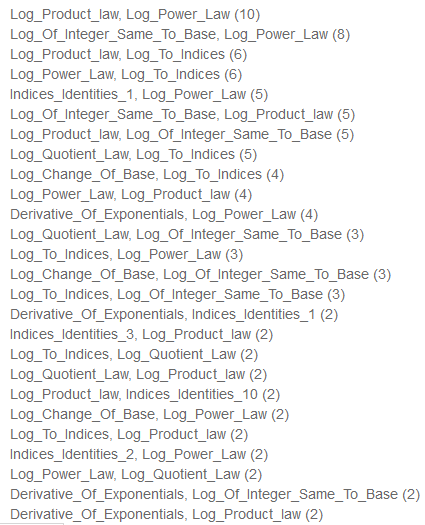


Figure ‑: Two-Tag Cloud for Exponential and Logarithmic Functions

From the two-tag cloud in Figure A-46, it is known that the top few on the lists are focusing on pure logarithmic formulae. However, indices identities and derivative of exponential function also covers a smaller proportion on the lower half of the list.

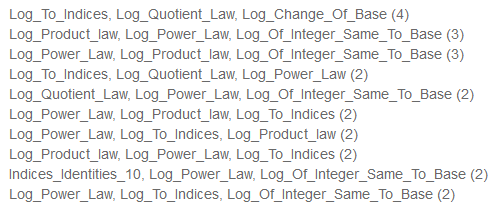


Figure ‑: Three-Tag Cloud for Exponential and Logarithmic Functions

From the three-tag cloud in Figure A-47, the Logarithmic formulae still stands the higher percentage, and stints of Indices identities at the bottom half of the list.

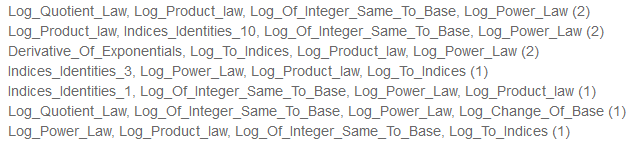


Figure ‑: Four-Tag Cloud for Exponential and Logarithmic Functions

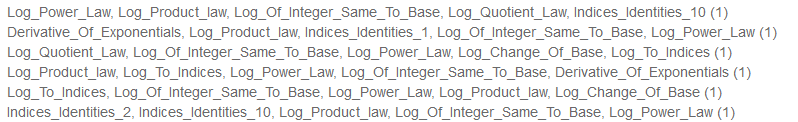


Figure ‑: Five-Tag Cloud for Exponential and Logarithmic Functions

From Figure A-48 and Figure A-49 above, it is still the same that the Logarithmic formulae covers the majority, while such formulae has some indices identities and derivatives of exponential functions involved to form a trickier question.

**Trigonometry**

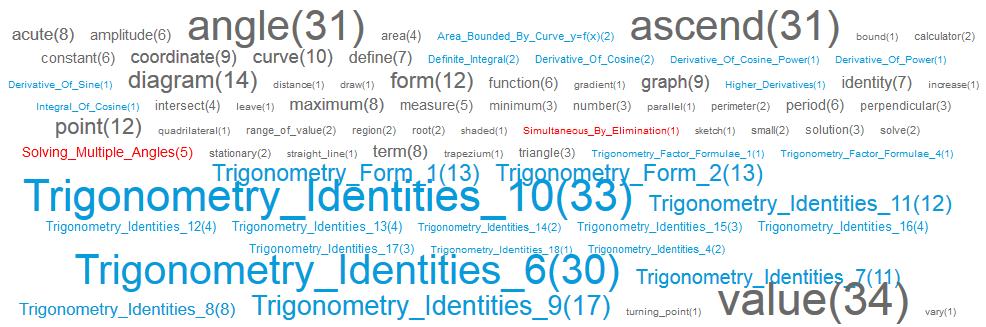


Figure ‑: One-Tag Cloud for Trigonometry

The one-tag cloud in Figure A-50 showed that trigonometry consists mostly on trigonometry identities, and also trigonometry form formulae. Some questions also require concepts on solving multiple angles using trigonometry. Foreign topics such as Integration’s area bounded by curve, definite integral and integral of cosine, and Differentiation’s derivative of sine/cosine and derivative of cosine power and derivative of power are formulae that are also required.

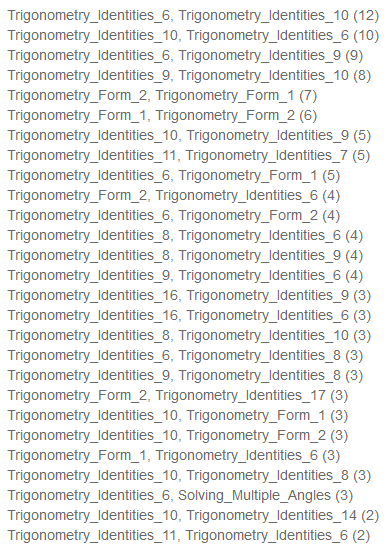


Figure ‑: Two-Tag Cloud for Trigonometry

For two-tag cloud in Figure A-51, the incomplete list already showed that it is occupied mostly by trigonometry formulae. This shows the importance of the formulae as they are used in most trigonometry questions. This also further shows that the Differentiation and Integration formulae exist on the minority part of the total questions. The pattern is also similar to three-tag cloud. Therefore, three-tag cloud will not be elaborated.

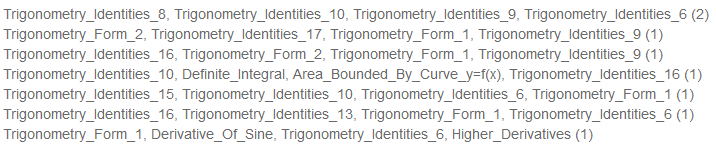


Figure ‑: Four-Tag Cloud for Trigonometry

When multiple-tag cloud is chosen as four-tag or five-tag, a slight improvement on involvement of foreign topics can be seen. The Figure A-52 is a four-tag cloud, and traces of Integration’s formulae were found on the 4th record of the list. Differentiation formulae can also be found on the last record of the list. This shows that the involvement of foreign topics on Trigonometry is not large, but a small portion of questions could see involvement of such formulae together with trigonometry formulae.

**Coordinate Geometry**

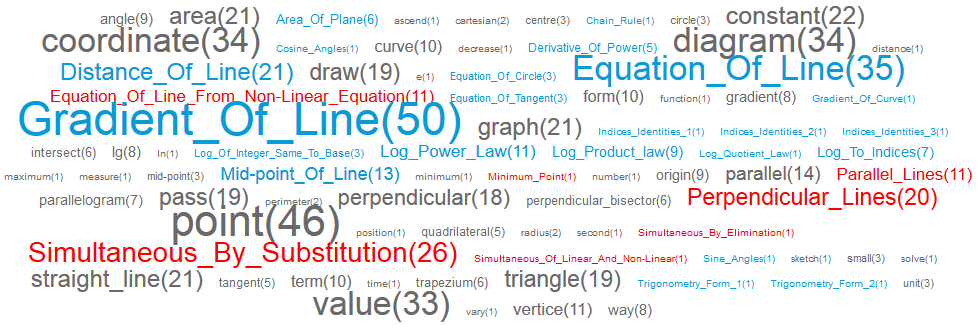


Figure ‑: One-Tag Cloud for Coordinate Geometry

The one-tag cloud of Coordinate Geometry in Figure A-53 showed most frequent terms as the gradient of line and equation of line formulae from the topic itself. However, a generous proportion of formulae and concepts also derived from foreign topics as well. Logarithmic formulae are seen as mid-size fonts, representing its importance in the questions as well. Differentiation, Trigonometry, Indices, and Simultaneous Equations are among the few which showed stints of existence as well. This tag cloud can clearly showed that Coordinate Geometry is definitely not a standalone topic which can be revised purely with the topic’s concepts and formulae alone. It requires a lot of concepts and formulae from many different topics as well to solve the questions.



Figure ‑: Two-Tag Cloud for Coordinate Geometry

From the two-tag cloud in Figure A-54, it is further proven that some of the important formulae for Coordinate Geometry are equation of line, distance of line and gradient of line. Simultaneous equation also featured in large portion of the questions, and Logarithmic formulae also featured in the lower list which also covers a decent number of questions (7 questions and below).

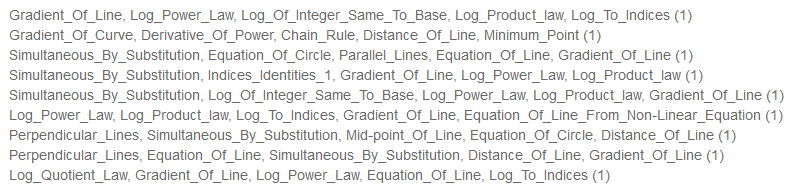


Figure ‑: Five-Tag Cloud for Coordinate Geometry

Coordinate Geometry’s multiple-tag clouds in Figure A-55 further show dominance of Coordinate Geometry’s concepts and formulae until four-tag and five-tag whereby more foreign concepts and formulae took root as the top most frequent terms. This could be due to the fact that Coordinate Geometry normally feature in questions with high marks distribution (12-16 marks) and these questions are more inclined in adding foreign concepts and formulae in conjunction to the main topic. Coordinate Geometry is a topic in which are easily combined with topics like Trigonometry, Plane Geometry, Differentiation, Integration, Indices and Simultaneous Equations.

**Differentiation**

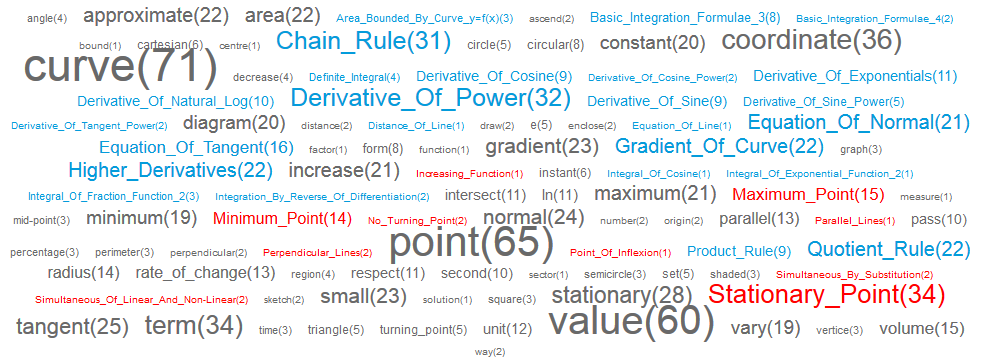


Figure ‑: One-Tag Cloud for Differentiation

The one-tag cloud on Differentiation in Figure A-56 showed that differentiation formulae stood as the major components of the questions, however formulae from Integration, Coordinate Geometry, and Simultaneous Equations are also covering some of the minority parts of the questions. Basic Integration Formulae stood at a frequency as high as 8, which shows that Integration is quite important in solving Differentiation problems.

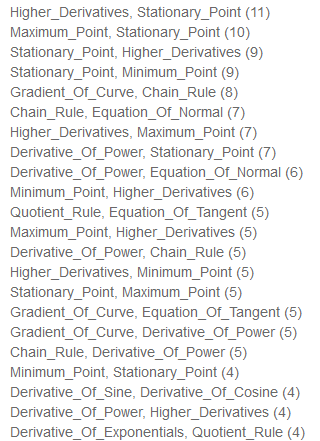


Figure ‑: Two-Tag Cloud for Differentiation

Using multiple-tag clouds, terms are more prominent in showing the patterns of questions from Differentiation topics. The two-tag cloud in Figure A-57 showed that questions on stationary points (maximum or minimum points) are the most common type of questions, followed by questions on curve focusing on gradient and finding equation of normal or equation of tangent. Formulae on Chain rule, Quotient rule are often applied together with some derivative formulae as well. Some of the less popular but important formulae involved derivative of trigonometry functions (sine, cosine, tangents) and derivative of exponential functions.

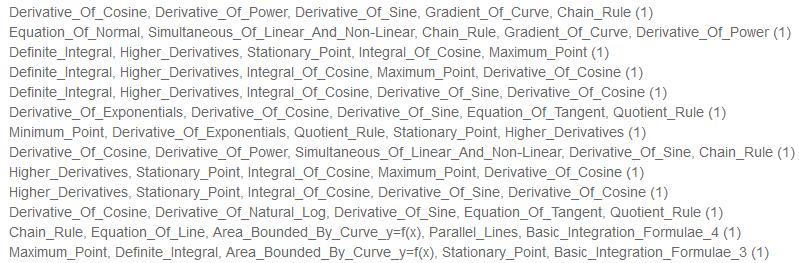


Figure ‑: Five-Tag Cloud for Differentiation

As multiple-tag cloud increases its combinations, the less prominent formulae such as derivative of exponential functions and derivative of trigonometry rises. These formulae also accompanied by formulae from foreign topics such as Integration and Simultaneous Equation. This can be deduced that Differentiation questions have different patterns, and some pattern of questions involved tougher derivative formulae on functions such as trigonometry and exponential functions. Integration and other foreign topics are also added to add spice to the difficulty of question could be another assumption deduced. Nevertheless, these questions occupies the minority part as previous observation on two-tag cloud showed that questions on stationary points are still a more popular focus for Differentiation.

**Matrices**

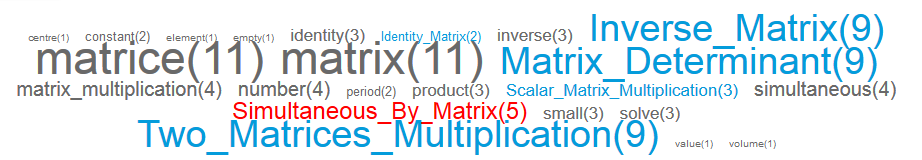


Figure ‑: One-Tag Cloud for Matrices

From the one-tag cloud in Figure A-59, it is clear that matrices are purely dealing with concepts and formulae from its topic. Most formulae and concepts are quite proportional in frequency count, which means that the questions might either be asking the formulae all together, or formulae and concepts are rotated well each year. Therefore, multiple-tags will be assessed to find out the unknown.

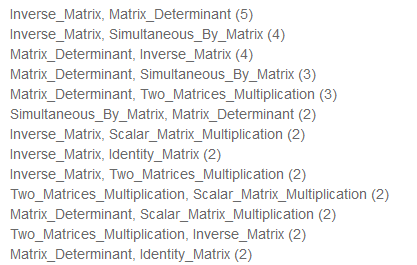


Figure ‑: Two-Tag Cloud on Matrices

From the two-tag cloud in Figure A-60, it is proven that there are intersections between formulae in questions. Inverse Matrix is one of the more popular formulae to be used, as well as simultaneous by matrix. Simultaneous by Matrix is known to be used on case study of a scenario, and solving the table of data using Simultaneous by Matrix.

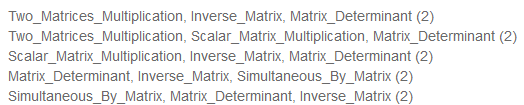


Figure ‑: Three-Tag Cloud for Matrices

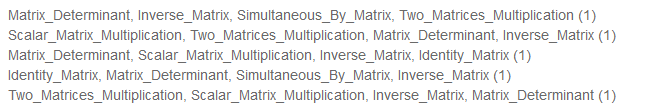


Figure ‑: Four-Tag Cloud for Matrices

The three-tag (Figure A-61) and four-tag (Figure A-62) further showed the different combinations of formulae that are involved on questions. This proven that the formulae and concepts can be used in conjunction with other formulae or concepts.

## Survey Question

1. For survey authenticity, please indicate your email address. (We will not use this email address for any purpose)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Under View Paper, how would you rate the usefulness?

Not Useful Somewhat Useful Useful Very Useful Really Useful

c

c

c

c

c

1. Under Paper Tag Cloud, how would you rate the usefulness?

Not Useful Somewhat Useful Useful Very Useful Really Useful

c

c

c

c

c

1. Under Topic Distribution, how would you rate the usefulness?

Not Useful Somewhat Useful Useful Very Useful Really Useful

c

c

c

c

c

1. Under Topic Trend, how would you rate the usefulness?

Not Useful Somewhat Useful Useful Very Useful Really Useful

c

c

c

c

c

1. Under Topic Tag Cloud, how would you rate the usefulness?

Not Useful Somewhat Useful Useful Very Useful Really Useful

c

c

c

c

c

1. Under Question Clustering, how would you rate the usefulness?

Not Useful Somewhat Useful Useful Very Useful Really Useful

c

c

c

c

c

1. Under Tag Search, how would you rate the usefulness?

Not Useful Somewhat Useful Useful Very Useful Really Useful

c

c

c

c

c

1. Overall, how useful do you rate our system? (Indicate in scale of 1 for Lousy to 10 for Excellent): 1 2 3 4 5 6 7 8 9 10
2. (Optional) Lastly, let us know if you have any suggestion that can improve our system.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# References

|  |  |
| --- | --- |
| [1] | D. S. Foundation, “django | The Web framework for perfectionists with deadlines.,” 2005-2013. [Online]. Available: https://www.djangoproject.com/. [Accessed 1 March 2013]. |
| [2] | W.-w.-a. a. R. Frey, “The model, view, and controller (MVC) pattern relative to the user.,” May 2010. [Online]. Available: http://en.wikipedia.org/wiki/File:MVC-Process.png. [Accessed 1 March 2013]. |
| [3] | “LaTeX - A document preparation system,” LaTeX Project Public License (LPPL), 10 January 2010. [Online]. Available: http://latex-project.org/. [Accessed 1 March 2013]. |
| [4] | J. D. Hunter, “Matplotlib: A 2D graphics environment,” *Computing In Science & Engineering,* vol. 9, no. 3, pp. 90-95, 2007. |
| [5] | H. S. AS, “Highcharts - Interactive Javascript charts for your web projects,” Highsoft Solutions AS, [Online]. Available: http://www.highcharts.com/. [Accessed 1 March 2013]. |
| [6] | G. P.-S. a. P. S. Usama Fayyad, “From Data Mining to Knowledge Discovery in Databases,” *AI Magazine,* pp. 37-54, 1997. |
| [7] | A. P. M. Matteucci, “A Tutorial on Clustering Algorithms - Clustering Introduction,” [Online]. Available: http://home.dei.polimi.it/matteucc/Clustering/tutorial\_html/index.html. [Accessed 1 March 2013]. |
| [8] | A. P. M. Matteucci, “A Tutorial on Clustering Algorithm - K-Means Clustering,” [Online]. Available: http://home.dei.polimi.it/matteucc/Clustering/tutorial\_html/kmeans.html. [Accessed 1 March 2013]. |
| [9] | S. Hoi, “Chapter 6 - Basic Clustering,” *Advanced Data Mining Techniques Lecture,* p. 20, 2013. |
| [10] | Bootstrap, “Bootstrap - Sleek, intuitive, and powerful front-end framework for faster and easier web development.,” Bootstrap, [Online]. Available: http://twitter.github.com/bootstrap/. [Accessed 1 March 2013]. |
| [11] | “Wikipedia - Software Development Process (Rapid Application Development),” Wikimedia Foundation, Inc, 26 February 2013. [Online]. Available: http://en.wikipedia.org/wiki/Software\_development\_process#Rapid\_application\_development. [Accessed 1 March 2013]. |
| [12] | I. Wikimedia Foundation, “Stemming - Wikipedia,” Wikimedia Foundation, Inc, 25 February 2013. [Online]. Available: http://en.wikipedia.org/wiki/Stemming. [Accessed 1 March 2013]. |
| [13] | L. C. Y. Teh Keng Seng, New Syllabus Additonal Mathematics 8th Edition, Singapore: Shinglee Publishers Pte Ltd, 2009. |
| [14] | I. Wikimedia Foundation, “Tag Cloud - Creation of a Tag Cloud,” Wikimedia Foundation, Inc., 21 February 2013. [Online]. Available: http://en.wikipedia.org/wiki/Tag\_cloud#Creation\_of\_a\_tag\_cloud. [Accessed 2 March 2013]. |
| [15] | A. Langville, “The Linear Algebra Behind Search Engines,” 2005. [Online]. Available: http://mathdl.maa.org/mathDL/4/?pa=content&sa=viewDocument&nodeId=636&pf=. [Accessed 3 March 2013]. |
| [16] | S. Hoi, “Euclidean Distance,” *Advanced Data Management Techniques Chapter 2,* pp. 62-63, 2013. |
| [17] | M. Hailwood, “jQuery Tagit – A jQuery tagging plugin,” 2010. [Online]. Available: http://jquery.webspirited.com/2011/02/jquery-tagit-a-jquery-tagging-plugin/. [Accessed 3 March 2013]. |
| [18] | P. S. Foundation, “Python Documentation - Collections - High performance container datatypes,” Python Software Foundation, 2013. [Online]. Available: http://docs.python.org/2/library/collections.html#collections.Counter. [Accessed 2 March 2013]. |
| [19] | P. S. Foundation, “Python Documentation - itertools,” Python Software Foundation, 2013. [Online]. Available: http://docs.python.org/2/library/itertools.html#itertools.combinations. [Accessed 2 March 2013]. |