Penn State University
SWENG 545 Data Mining
Midterm 1

Student Name: Shashi Singh
Angel ID: sks235

HONOR STATEMENT:

I have completed my work according to the principle of Academic Integrity. I have neither given nor received any unauthorized aid on this assignment/examination.

Date: 04/11/2015 Initial: Shashi Singh

The following are the rules relating to this take home-exam. Any questions about interpretation of problems should be addressed to me.

1. Once you have downloaded the exam from ANGEL, you may not discuss it in any way with anyone until the exam period is over.
2. You may use a word processor or computer in preparing your answers. You may also write in pen/pencil and scan for submission.
3. You NEED NOT worry about margins, fonts, etc. on the exam as long as it is easily readable. Suggested font size is between 11 and 14 points.
4. Any violation of the rules regarding consultation with others including family members will be considered honor code violations.
5. If any problems arise during this exam, email me. I cannot make exceptions or give extensions.

Setting:

The College of Arts and Sciences is a fast-growing entity at the University of Diploma Printing in Romanigstan. Within this College, the Department of Arts consists of twelve full-time faculty, six adjunct instructors, and approximately fifteen graduate teaching assistants. Many faculty are nationally and internationally renowned and are well-established professional artists and designers with numerous exhibitions, commissions, and awards to their credit. The Department currently has 240 undergraduate majors within its BA and BFA programs spanning a number of areas including art of walking, retail floristry, puppetry, adventure recreation, fishing sciences, ant nesting habits, coin slot technology, and turfgrass management. It offers approximately 50 different courses to art and non-art majors alike. In addition, art education students enroll in a significant number of art courses for their Bachelor of Science in Art Education degree offered cooperatively through the Department of Learning, Teaching, and Curriculum.

You, as a new faculty in the Department of Arts, were assigned to the huge task of reviewing and redesigning the BFA curriculum. Your help is needed for the following reasons:

- Stagnant enrollment in the last five years
- Changes in student demographics and interests.
- Shifts in disciplinary approaches

The main tasks of curriculum redesign are

- Reducing the number of credit hours from 39 to 33
- Defining concentration areas for elective courses
- Reducing the number of offered electives

To evaluate the opportunity of grouping elective classes in concentration areas, the chair of the committee provided you with the following pieces of information (attached in ANGEL)

- List of current course offered
- List of all student enrollments for the last three years

Exam problem

1. What elective courses should be retained in the course offerings?
2. Recommend sets of three or four courses that can be grouped in concentration areas.

Notes:

- Your answers should be based on analyzing the provided data.
- Explain your answers and support them by results and/or screen captures.
- Try to follow the Guidelines for Data Mining Experiments from your term project.

Scoring

Scoring will be based on:

- Appropriateness and correctness of experiment (70%)
- Discussion (20%)
- Conclusions (10%)

Particularly good discussions may result in 2% extra credit.

Question 1: What elective courses should be retained in the course offerings?

Answer:

Approach:

Before analyzing the solution, in the CRISP – Cross Industry Standard Process for Data Mining — two of steps after business understanding is data understanding and data pre-processing.

Data Understanding:

Looking at the data based on the data provided (Midterm1data) and courses (Midterm1Courses) offered, the following observations have been given below:

- Duplicate Data: There are some records which are duplicate vis-à-vis for each field.
- Redundant or Do Not Need type data: There are some records which are for the students enrolled in the core courses. These are not required for the problems which need to be answered. In addition, there are records related to the foundation courses also needs to be removed as these are redundant data in solving the given problem.
- Header Data: Header records was recognized so that actual and header records could be differentiated.
- Case of Data: Data given were in the mixed case particularly for the Couse Name. These need to be the same case to avoid error in data mining process.
- Courses not on the Courses Offered doc: There are some records related to the courses which are not listed in the courses offered as BFA majored. These records are redundant is solving the given task.

Pre-Processing or Cleaning of Data:

With regards to the data preprocessing, data cleaning, data transformation, and data reduction are done. The reason for the preprocessing because the data was dirty in terms of duplicates, redundant, and inconsistent, and not in correct format to perform the analysis.

For the given data following data cleaning techniques have been considered or reviewed. The used technique is influenced by the data provided (Midterm1data) and courses (Midterm1Courses) offered.

- Case of the all values stored in the "coursename" column has been changed to the upper case.
- Remove Duplicates: The duplicate rows are removed. There are 1249 duplicate records.

- Then the new excel sheet which contain data after removing duplicates and upper case course name is imported into SQL Server.

- Then import all the courses data into SQL Server table. This was done in expedite data pre-processing such as deleting redundant record or finding the valid record for the analysis. The Course table would help in joining the record from the Midterm1data table.

- Imported student course data into another table which removes the record having courses not related to the analysis of the problem. For example, the record related to the foundation, core and not related to ones given for BFA major are not included into this table.

- After cleaning the using the join, there are 489 records into a new cleaned table. They are BFA elective courses' records.

- Also tokenize Semester Name like 'Fall 2002' into Semester Name and Year. In addition, assigned numeric value to semester name to create time series chart to analyses the trend line.

- Renamed the course name in the given data as the name of not consistent with the name given in the master course list. For example, 19TH-CENTURY BRITISH LITERATURE is as 19TH-CENT BRITISH LIT in the Midterm1data file. Similarly, AUGUSTAN CULTRAL REVOLUTION, EARLY MESOPOTAM HISTORY/SOCIETY, CONTEMPORARY SOCIO THEORY names are incorrect as well in the data file.
- There is duplicate record in the course list (Midterm1Courses) for the BFA major as well: AMERICAN HEALTH POLICY and FRANCE & THE EUROP.UNION.
- Actually there are 31 elective courses for the BFA major.

What elective courses should be retained in the course offerings?

To answer this question, time series chart of each course are analyzed for the pattern and outlier are also looked into to find out which courses should be retained to; in other words, which courses should not be retained.

Based on the pivot table chart, the one colored in red clearly shows as – outlier. Students have registered just once after that no student registered into these courses. Similarly, the one colored in yellow may need to analyze further before making any decision to discontinue.

On the other hand the courses colored in green should be continued.

Course Name	NumberOfRegistration
19TH-CENTURY BRITISH LITERATURE	**5**
Fall 2000	2
Fall 2001	2
Fall 2002	5
Fall 2004	3
Spring 2001	9
Spring 2002	6
Spring 2005	5

Summer 2002	7
20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY	**6**
Fall 2000	3
Fall 2003	4
Spring 2002	10
Spring 2005	9
Summer 2002	5
AESTHETICS	**6**
Spring 2002	3
Summer 2001	9
AFRICAN-AMERICAN LIT: AFRICAN-AMER LIT:CHANGE	**19**
Spring 2001	19
AMERICAN HEALTH POLICY	**6**
Fall 2001	3
Fall 2002	2
Fall 2003	4
Spring 2001	8
Spring 2003	5
Spring 2005	9
Summer 2002	11
AMERICAN SOUTH 1861-PRES	**5**
Fall 2001	3
Fall 2005	3
Spring 2003	10
Spring 2005	1
Summer 2001	7
ANALYZING THE POL WORLD	**6**
Summer 2002	6
ART - ANCIENT TO 1945	**8**
Fall 2002	8
ART AND RELIGION	**3**
Spring 2003	3
AUGUSTAN CULTRAL REVOLUTION	**4**
Spring 2002	2
Spring 2005	6
BECOMING HUMAN	**8**
Fall 2000	2
Fall 2003	5
Spring 2005	16
Summer 2003	9
BRITISH POETRY 1660-1914	**6**
Summer 2000	8
Summer 2003	4
BUSINESS GERMAN: A MICRO PERSPECTIVE	**8**
Fall 2004	8
CELL. BIOL. & BIOCHEM.	**9**
Spring 2005	9
COMM AND THE PRESIDENCY	**4**
Fall 2005	4
COMMUNICATIONS INTERNSHP	**11**
Fall 2003	1

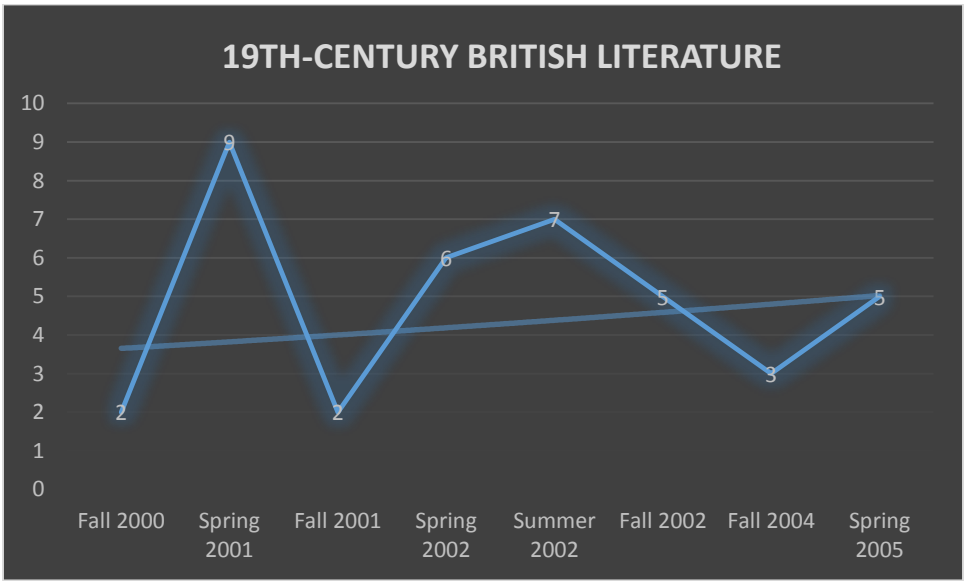
Spring 2004	20
Summer 2001	10
Summer 2004	13
COMPARATIVE POLITICS	**5**
Fall 2003	1
Fall 2005	7
Spring 2001	2
Spring 2005	8
Summer 2003	5
CONTEMP ART - 1945 TO PRESENT	**4**
Fall 2001	4
CONTEMPORARY POL.THOUGHT	**9**
Fall 2002	5
Fall 2003	3
Spring 2002	10
Spring 2005	18
Summer 2003	8
CONTEMPORARY SOCIO THEORY	**1**
Spring 2001	1
DEVIL'S PACT LIT/FILM	**3**
Fall 2001	2
Spring 2001	1
Spring 2003	8
Spring 2005	1
EARLY MESOPOTAM HISTORY/SOCIETY	**1**
Fall 2000	1
ELEMENTARY ARABIC II	**8**
Fall 2002	1
Fall 2004	3
Spring 2001	6
Spring 2004	18
Spring 2005	13
Summer 2002	12
Summer 2003	2
Summer 2004	9
ELEMENTARY GERMAN 1	**1**
Spring 2004	1
ENVIRONMENTAL STUDIES RESEARCH SEMINAR JUNIOR LEVEL	**5**
Summer 2000	5
Summer 2001	5
EUROPE IN A WIDER WORLD	**3**
Spring 2001	1
Spring 2002	2
Summer 2001	4
Summer 2004	4
EVIDENCED BASED CRIME AND JUSTICE POLICY	**1**
Fall 2001	1
FRANCE & THE EUROP.UNION	**8**
Spring 2002	5
Summer 2002	11
FRENCH THOUGHT SINCE 1945	**7**

Summer 2002	7
FRESHWATER ECOLOGY	**10**
Fall 2003	1
Spring 2001	11
Summer 2000	12
Summer 2002	17

From the given pivot chart, further analysis of the following elective courses in BFA program is given below:

1. **19TH-CENTURY BRITISH LITERATURE**

The time series graph of this course shows an increasing linear trend, which means that this course should be retained.

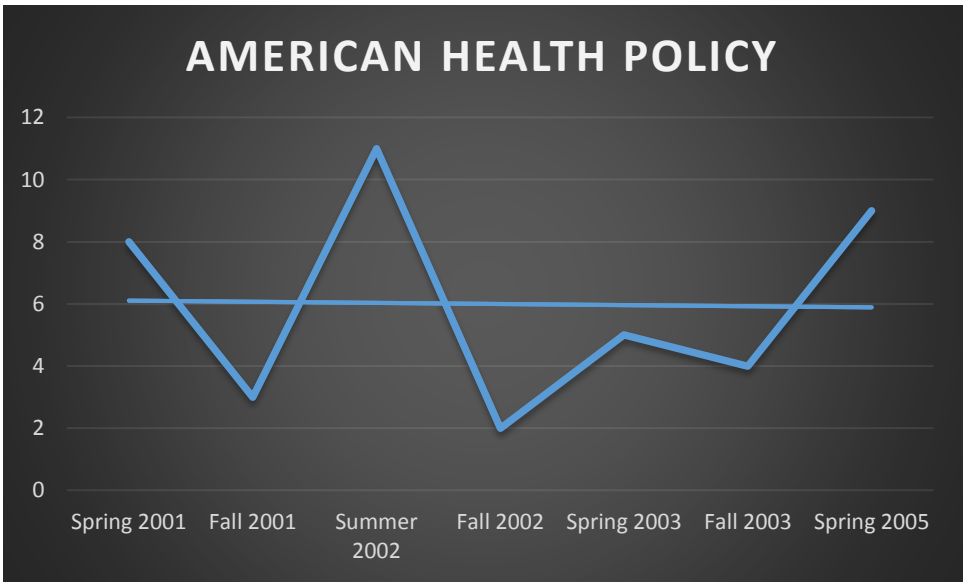


2. **20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY**

Based on the time series chart of this course, the linear rising trend suggests that this course should be retained.

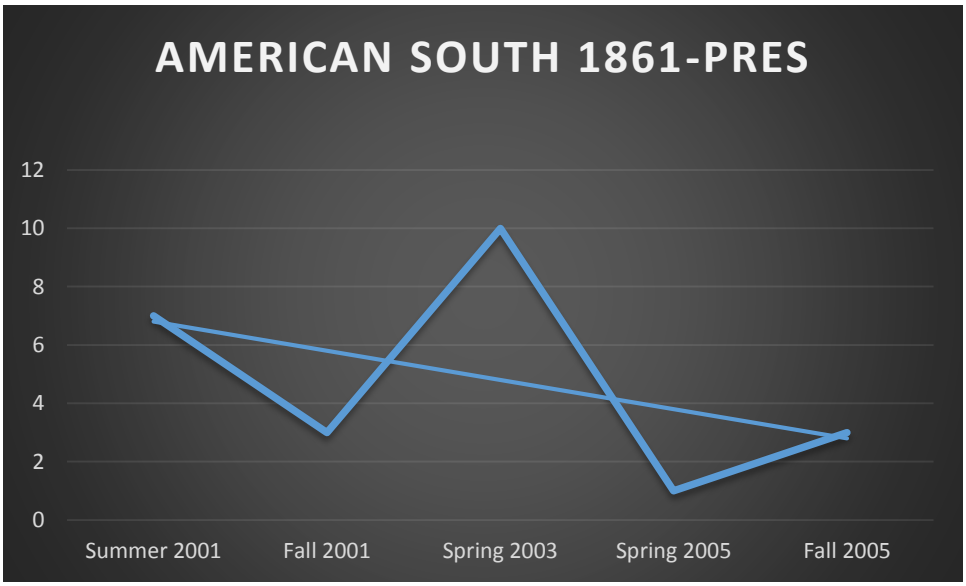
3. **AMERICAN HEALTH POLICY**

Based on the time series chart of this course, the steady constant trend suggests that this course should be retained.



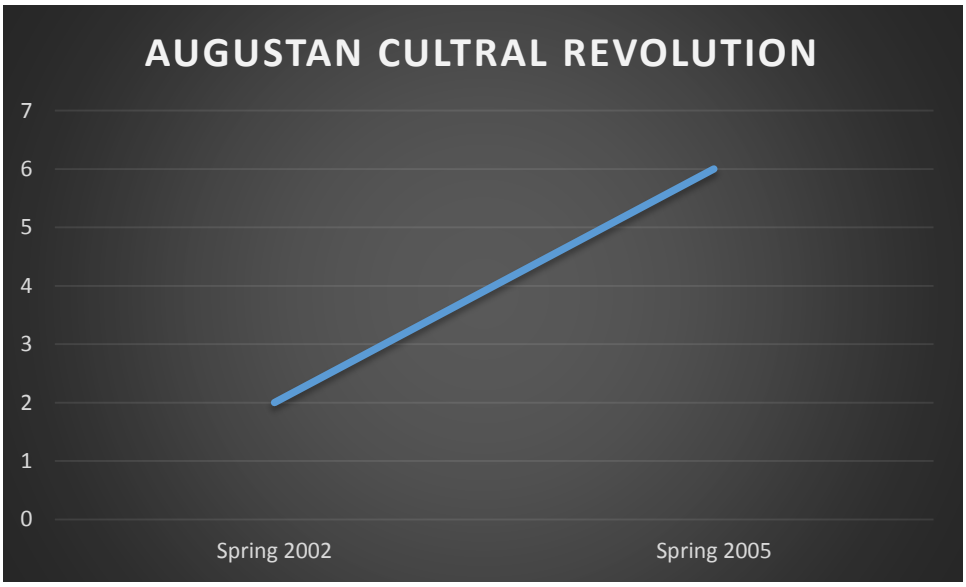
4. AMERICAN SOUTH 1861-PRES

Based on the time series chart of this course, the downward linear trend suggests that this course should be analyzed further before making decision of retaining it.



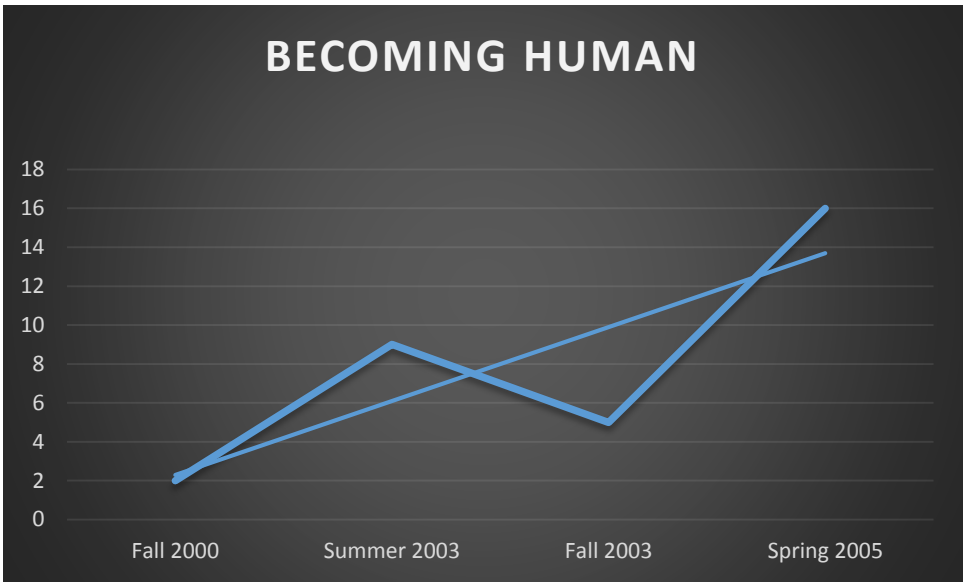
5. AUGUSTAN CULTRAL REVOLUTION

Based on the trend this course should be retained. It seems that this course has been added recently.



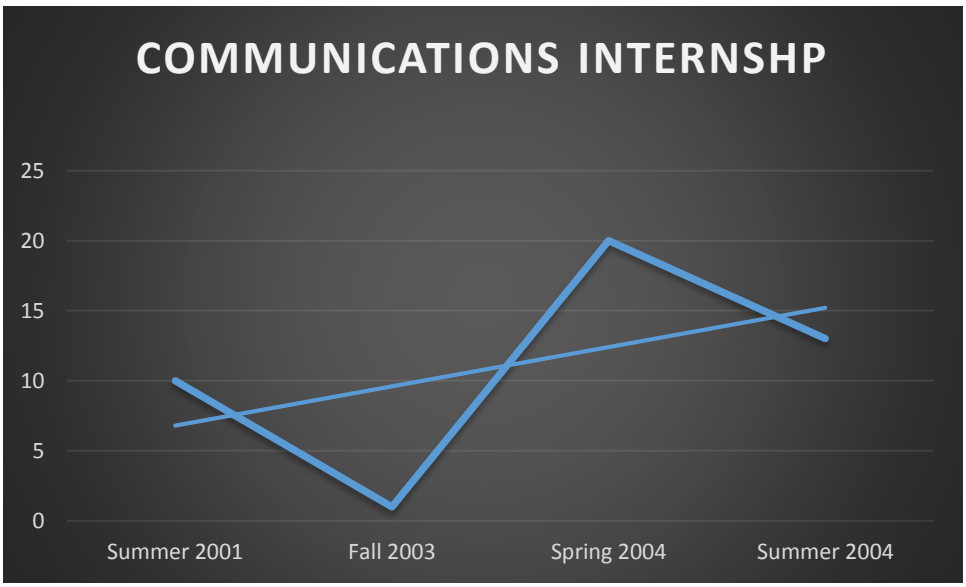
6. **BECOMING HUMAN**

Based on the trend this course should be retained.



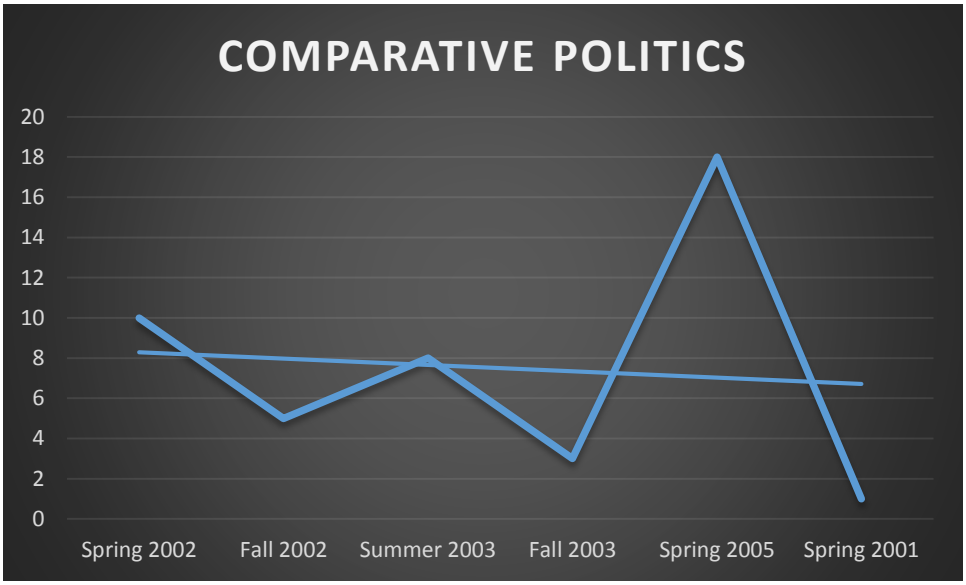
7. **COMMUNICATIONS INTERNSHP**

Based on the trend this course should be retained.



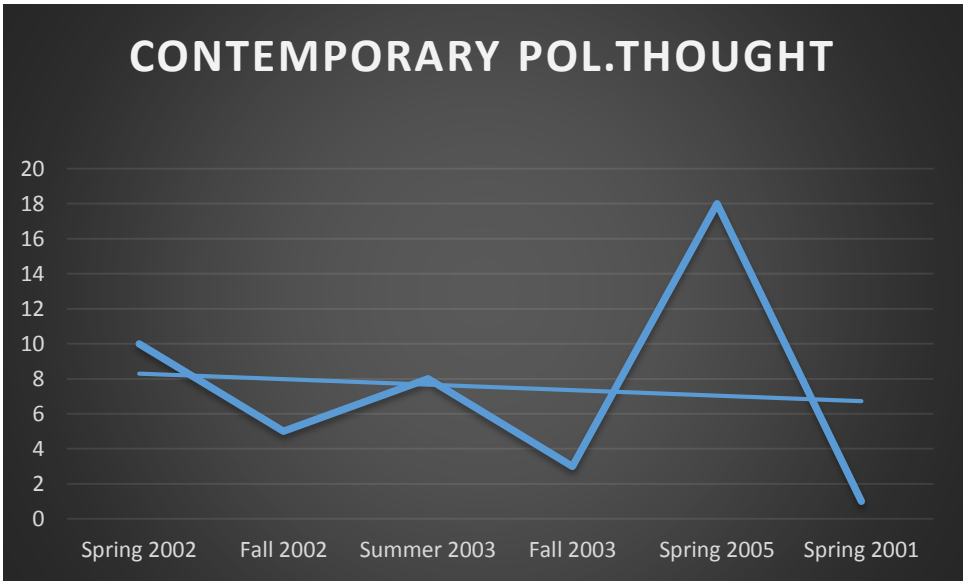
8. COMPARATIVE POLITICS

Based on the time series chart of this course, the very slow downward trend suggests that this course should be looked into further detail before taking decision of discontinued it.



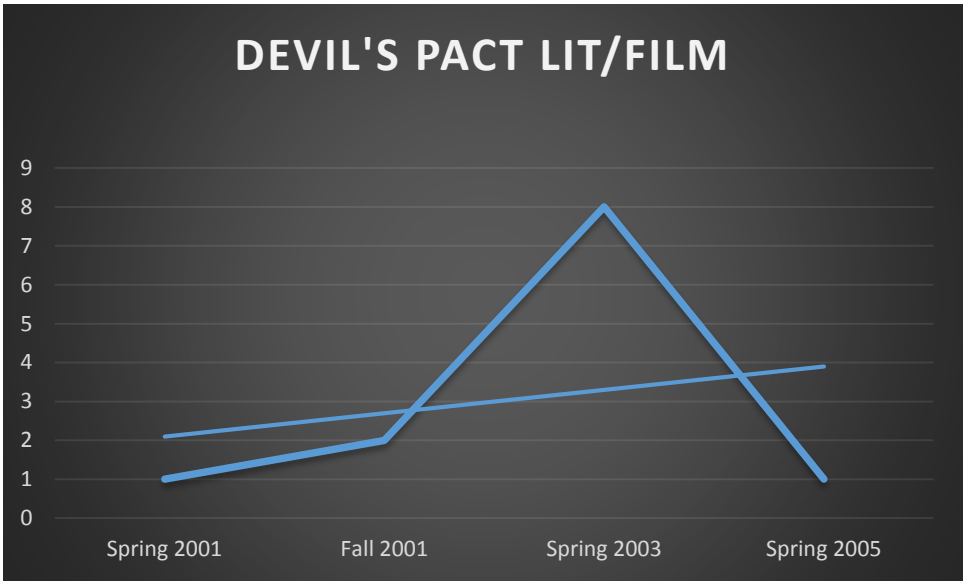
9. CONTEMPORARY POL.THOUGHT

Based on the time series chart of this course, the very slow downward trend suggests that this course should be looked into further detail before taking decision of discontinued it.



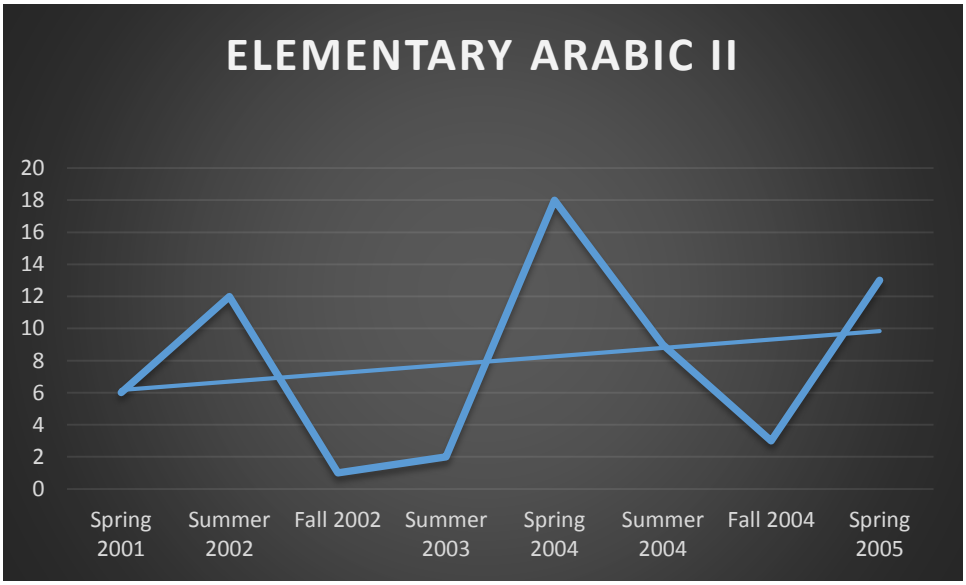
10. DEVIL'S PACT LIT/FILM

Based on the trend this course should be retained.



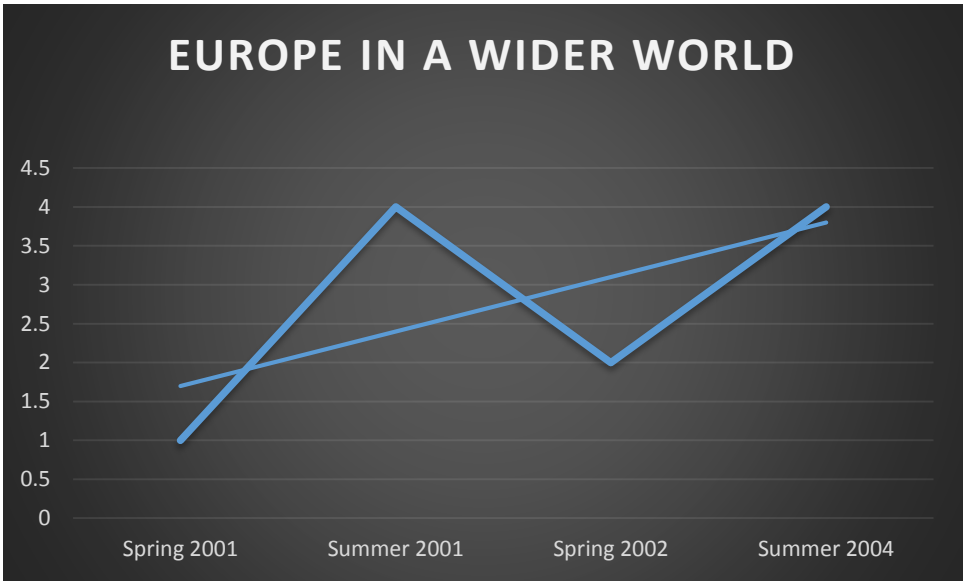
11. ELEMENTARY ARABIC II

Based on the trend this course should be retained.



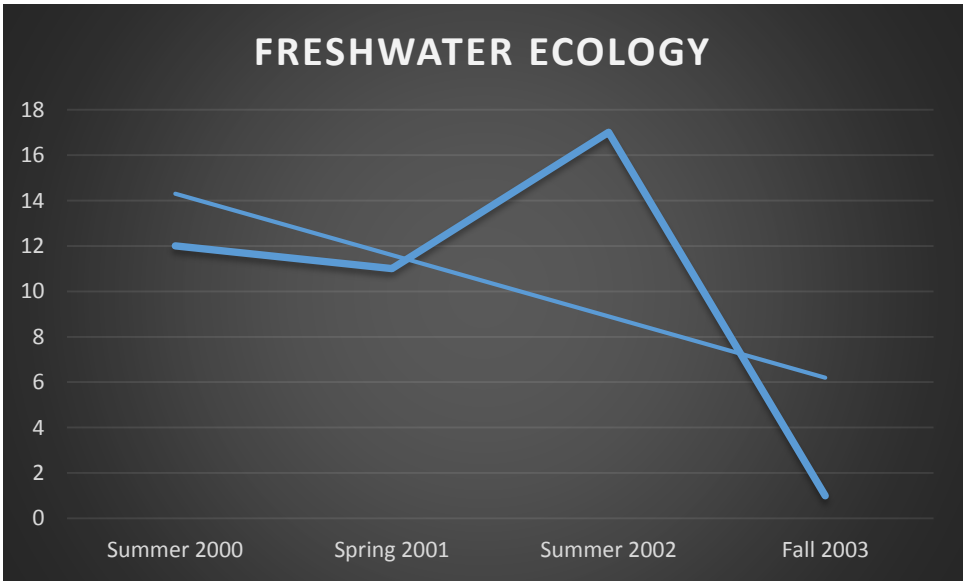
12. EUROPE IN A WIDER WORLD

Based on the trend this course should be retained.



13. FRESHWATER ECOLOGY

Based on the time series chart of this course, the very slow downward trend suggests that this course should be looked into further detail before taking decision of discontinued it.



From the analysis of the pivot chart and time series chart of the various BFA program course. It cab be concluded that following courses should be retained:

- 19TH-CENTURY BRITISH LITERATURE
- 20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY
- AMERICAN HEALTH POLICY
- AUGUSTAN CULTRAL REVOLUTION
- BECOMING HUMAN
- COMMUNICATIONS INTERNSHP
- DEVIL'S PACT LIT/FILM
- ELEMENTARY ARABIC II
- EUROPE IN A WIDER WORLD

Question 2: Recommend sets of three or four courses that can be grouped in concentration areas.

Answer:

This question is also answered based on CRISP. The data understanding and pre-processing has been utilized from the question number 1.

The Association Rule or Market Basket Analysis has been used in recommendation of item set or group of courses for the elective courses.

There are two sets of three courses that can be grouped together. The finding is based on using Association Analysis on the Elective courses with minimum support of 5 and minimum probability of 30%.

Recommended Set #1:

- AMERICAN SOUTH 1861-PRES,
- 20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY,
- 19TH-CENTURY BRITISH LITERATURE

Association Rules:

1. {AMERICAN SOUTH 1861-PRES, 20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY}-> {19TH-CENTURY BRITISH LITERATURE}

   Probability: 0.833
   Importance: 0.771311825853338

2. {AMERICAN SOUTH 1861-PRES, 19TH-CENTURY BRITISH LITERATURE }->{ 20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY}

   Probability: 0.833
   Importance: 0.771311825853338

3. {20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY, 19TH-CENTURY BRITISH LITERATURE} -> {AMERICAN SOUTH 1861-PRES}

   Probability: 0.556
   Importance: 1.00627150944333

Recommended Set #2:

- AMERICAN HEALTH POLICY
- FRESHWATER ECOLOGY
- CONTEMPORARY POL.THOUGHT

Association Rules:

1. {FRESHWATER ECOLOGY, CONTEMPORARY POL.THOUGHT} -> {AMERICAN

HEALTH POLICY}

Probability: 0.625
Importance: 0.55863116308587

2. {CONTEMPORARY POL.THOUGHT, AMERICAN HEALTH POLICY}-> {FRESHWATER ECOLOGY}

   Probability: 0. 556
   Importance: 0.608331500771297

3. {FRESHWATER ECOLOGY , AMERICAN HEALTH POLICY }-> {CONTEMPORARY POL.THOUGHT}

   Probability: 0. 313
   Importance: 0. 313333778361275

There no item set of 4 to be recommended for even minimum support level equals 2.

Excel Microsoft Data Mining Screen-Shots Showing Associations of the recommended sets: