

## Course Syllabus Part I

### DSC 550 - Data Mining

3 Credit Hours

---

#### Course Description

Data can often contain patterns and anomalies that only emerge at large scale. In this course, students explore techniques to mine and analyze large datasets to discover useful knowledge. Text mining, unstructured data, social networks, and other types of unsupervised data mining methods for data science are included.

#### Course Prerequisites

Recommend DSC 510 or equivalent.

---

#### Course Objectives

Students who successfully complete this course should be able to:

1. Explain data mining techniques available for analyzing big data
2. Transform data in preparation for data mining algorithms
3. Construct natural language models for text analysis
4. Recommend statistical approaches for evaluating big data
5. Formulate implementation and automation strategies for data mining projects

---

#### Grading Scale

93 – 100% = A	87 – 89% = B+	77 – 79% = C+	67 – 69% = D+
90 – 92% = A-	83 – 86% = B	73 – 76% = C	63 – 66% = D
	80 – 82% = B-	70 – 72% = C-	60 – 62% = D-
			0 – 59% = F

---

#### Topic Outline

1. Text Preprocessing, Transformation, and Vectorization
  - a. Identifying paragraphs, sentences, and words
  - b. Segmenting text into tokens and n-grams
  - c. Feature extraction, transformation, and selection
2. Text Classification
  - a. Naive Bayes
  - b. Logistic Regression
  - c. Cross-validation
  - d. Model validation and evaluation

3. Handling Categorical Data, Text, Dates & Times
4. Topic Modeling and Document Similarity
  - a. Latent Dirichlet Allocation
  - b. Latent Semantic Analysis
5. Context Aware Text Analysis
  - a. Part of speech (POS) tagging
  - b. Keyphrase extraction
  - c. Named-entity resolution
  - d. Word vectors
6. Graph Analysis
  - a. Social network analysis
  - b. Graph algorithms
  - c. Graph visualization
7. Unsupervised Learning
  - a. Principal Components Analysis
  - b. Hierarchical Clustering
  - c. Frequent Pattern Mining
  - d. Collaborative Filtering
8. Model Evaluation and Selection
9. Real-World Implementation
  - a. Organizing a data mining project
  - b. Data infrastructure
  - c. Scaling
  - d. Ethics
10. Advanced Topics in Data Mining
  - a. Neural networks and deep learning
  - b. Bayesian networks
  - c. Image, video, and speech processing

**Course Syllabus Part II**  
**DSC 540 – Data Preparation**  
**3 Credit Hours**

---

**Course Resources**

**Course Text(s)**

*Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning (1st Edition).*

Benjamin Bengfort, Rebecca Bilbro, & Tony Ojeda.

O'Reilly Media

**ISBN-13:** 978-1491963043

**ISBN-10:** 1491963042

*Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning (1st Edition).*

Chris Albon

O'Reilly Media

**ISBN-13:** 978-1491989388

**ISBN-10:** 1491989386

### Required Resources

In this course, you will need to be able to:

- Access the Internet.
- Access Cyberactive.
- Access to Github
- Collaborate Online via Video and Voice.
- Collaborate while writing a single document.
- Submit a Word Document.
- Access to GitHub account.
- Python programming environment using PyCharm, Anaconda, and Jupyter Notebook.

---

### Course Schedule

Week	Topic	Assigned Reading
1	Text preprocessing, transformation, vectorization, and feature extraction	<i>Machine Learning with Python</i> Chapter 1-4 <i>Applied Text Analysis</i> Chapter 3
2	Text classification	<i>Machine Learning with Python</i> Chapter 16 and 18 <i>Applied Text Analysis with Python</i> Chapter 5
3	Handling Categorical Data, Text, Dates & Times	<i>Machine Learning with Python</i> Chapters 5-7
4	Topic modeling and document similarity	<i>Applied Text Analysis with Python</i> Chapter 6
5	Context aware text analysis	<i>Applied Text Analysis with Python</i> Chapter 7
6	Graph analysis and social networks	<i>Applied Text Analysis with Python</i> Chapter 8 and 9

7	Unsupervised learning part 1: Collaborative filtering and frequent pattern mining	<i>Machine Learning with Python</i> Chapter 9 <i>Applied Text Analysis with Python</i> Chapter 10
8	Unsupervised learning part 2: Hierarchical clustering and dimensionality reduction	<i>Machine Learning with Python</i> Chapter 10 and 19
9	Model Evaluation and Selection	<i>Machine Learning with Python</i> Chapter 11 and 12
10	Practical considerations including project organization, infrastructure, scaling, and ethics	<i>Machine Learning with Python</i> Chapter 21 <i>Applied Text Analysis with Python</i> Chapter 11
11	Advanced topics in data mining	<i>Machine Learning with Python</i> Chapter 8 and 20 <i>Applied Text Analysis with Python</i> Chapter 12
12	Final project	

### Course Activities

In this section of the syllabus, I will describe what we will be doing in each of the activities for each week. Specifically, I will be describing your deliverables – those items you need to submit at or before the deadline. You can find more detail on grading criteria for each category by viewing its detailed rubric.

### Written Assignments

Each week, you will be assigned a written assignment aligning to the weekly reading and topics, which is due to the discussion board. This post must be 500 words minimum and contain at least two credible sources. It should be written with an introduction, body and conclusion and cited references.

### Exercises

Each week, you will be assigned an exercise or series of exercises based on the weekly topic to complete and submit to the assignment link. These are not group assignments to complete and should be done on your own. However, if you have questions about a specific exercise, you are encouraged to use the discussion board to discuss with your classmates, without completing the assignment together.

### Discussion

Each week, you will be making 2 discussion posts in the specified forums. These two posts can either be responses to a fellow classmate or they can be something you found interesting in the reading/homework or something you didn't understand or agree with.

Each post must be a minimum of 250 words and contain at least one credible source. These responses should be "substantive" which means more than, "Neat!" or "Good job!" They should also not contain jargon or be a post that boils down to you reposting the same thing you're commenting on in a different way.

**Quizzes**

Each week, there will be a short 10-20 question quiz. The weekly quiz is not timed and is open book and open note.

**Term Project**

This course has one major programming assignment. In this project, the students create a fully functional Python program that applies knowledge and skills from previous lessons.

---

**Grade and Point Breakdown**

<u>Component:</u>	<u>Percentage</u>	<u>Point Value</u>	<u>Number of Times</u>	<u>Total</u>
Discussion	30%	60 Points	2 Times per Week for 12 Weeks	720
Written Assignment	20%	40 Points	1 Submission Per Week for 12 Weeks	480
Exercises	20%	40 Points	1 Submission Per Week for 12 Weeks	480
Quizzes	10%	20 Points	1 Quiz Per Week for 12 Weeks	240
Term Project	20%	480 Points	1 Submission for Term	480
			Total Points	2400

---

**Late Work**

Late work is not accepted unless arrangements are made with the instructor for very special, unavoidable circumstances. If you do not alert the professor before or shortly after something that will make you late, the chances of special arrangements are much lower. If in doubt, please email as soon as possible.

**Participation**

Students are expected to login often and contribute to the class on a regular basis, including posting to the discussion board, submitting assignments, and participating in group activities as required. If you have specific participation requirements related to your educational funding or student status, you are expected to monitor your own participation to ensure you are in compliance with those requirements.

**Expectations for Students**

- Students should expect to spend approximately 10-15 hours per week to complete the activities and assignments in this course.
- Students will log in as often as needed to complete their assignments and progress through the course.
- Students will treat their classmates and the instructor with respect and courtesy.
- Students are responsible for keeping current with the reading assignments and coming to class prepared to discuss the work assigned.
- Students are responsible for knowing what assignments are due and when.
- Students will submit only their own work and will not commit plagiarism or other acts of academic dishonesty.
- Students will contact the instructor as soon as personal problems arise that may affect the

student's ability to complete assignments on time.

**Expectations for Faculty**

- The instructor will treat all students with respect and courtesy.
- The instructor will make grading criteria clear and follow the criteria scrupulously in evaluating student work.
- The instructor will provide feedback about student work within 6 days of due dates (or 24 hours prior to the next due date)—feedback that helps the student learn and improve.
- The instructor will respond to all student messages within 48 hours.