University of North Carolina at Charlotte
DSBA/ITIS 6010 Text Mining and Information Retrieval

Credits: 3 Credit Hours

Days, Time/Location:
2:00pm - 4:45pm on Fridays at Center City 901, totally 14 lessons with the first lesson on Jan 14 (Friday), 2022.

Due to the surge of COVID-19, the first two lessons will be online via Zoom.

Zoom Meeting ID: 918 4125 6169
Passcode: 6010

Course Description:
The availability of text data has created unprecedented opportunities to leverage computational and statistical approaches to turn data into actionable knowledge. This course covers general techniques for analyzing large amounts of text data as well as basic techniques for information retrieval.

The current technology of natural language processing has not yet reached a point to enable a computer to precisely understand natural language text, but text mining (TM) techniques with a wide range of statistical and heuristic approaches have been developed over the past few decades. They are usually very robust and can be applied to analyze and manage text data in any natural language, and about any topic. This course intends to provide a systematic introduction to many of these approaches, such as word association mining, topic modeling, and text classification. On the other hand, information retrieval (IR) is a relatively mature and well-established field. We will introduce the contemporary retrieval models as well as their evaluations.

We will offer Python code examples which contain implementations of many techniques discussed in this course. Homework exercises are designed based on Python to help students acquire practical skills of experimenting with the learned techniques and applying them to solve real-world application problems.

The required background knowledge to take this course is minimal since the it is intended to be mostly self-contained. However, students are expected to have basic knowledge about computer science, particularly some programming language, and be comfortable with some basic concepts in probability and statistics such as conditional probability and parameter estimation.

Faculty Information:  Xi (Sunshine) Niu, Ph.D., Associate Professor
Office: Woodward 310G
Email: xniu2@uncc.edu
Office Hours: by appointment
Teaching Assistant: Mr. Chieh Wu  
Email: cwu21@uncc.edu

Textbooks

Title: Text Data Management and Analytics: A Practical Introduction to Information Retrieval and Text Mining  
Author(s): ChengXiang Zhai and Sean Massung  
Publisher: ACM and Morgan & Claypool Publishers  
Year: 2016

Title: Introduction to Information Retrieval  
Author(s): Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze  
Publisher: Cambridge  
Year: 2008

Evaluation Methods:

Course grading will be based on these activities.

<table>
<thead>
<tr>
<th>Activities</th>
<th>Point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bootcamp course (online asynchronous) on Python offered by School of Data Science (SDS) (complete by March 15, 2022)</td>
<td>10 points</td>
</tr>
<tr>
<td>In-Class Quizzes</td>
<td>2 points x 13 = 26 points</td>
</tr>
<tr>
<td>After-Class Homework</td>
<td>5 points x 13 = 65 points</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>101 points</strong></td>
</tr>
</tbody>
</table>

Grade Scale:
A = 90 points – 101 points  
B = 80 points – 89 points  
C = 70 points – 79 points  
U = Below 70 points
### Weekly Lesson Schedule:

<table>
<thead>
<tr>
<th>Date</th>
<th>Contents</th>
</tr>
</thead>
</table>
| Jan 14 | Syllabus  
Lesson 1: Basic Concepts  
Getting started with Python |
| Jan 21 | Lesson 2: Word Association Mining I: Paradigmatic Relation               |
| Jan 28 | Lesson 3: Word Association Mining II: Syntagmatic Relation               |
| Feb 4  | Lesson 4: Topic Modeling I                                               |
| Feb 11 | Lesson 5: Topic Modeling II                                              |
| Feb 18 | Lesson 6: Text Classification                                            |
| Feb 25 | Lesson 7: Retrieval Model I: Boolean Retrieval                           |
| Mar 4  | Lesson 8: Retrieval Model II: Vector Space Model 1                        |
| Mar 11 | Spring Break, no classes                                                |
| Mar 18 | Lesson 9: Retrieval Model II: Vector Space Model 2                        |
| Mar 25 | Lesson 10: Evaluations in Information Retrieval                           |
| Apr 1  | Lesson 11: Retrieval Model III: Probabilistic Information Retrieval      |
| Apr 8  | Lesson 12: Retrieval Model III: Language Models for Information Retrieval|
| Apr 15 | Health Break, no classes                                                |
| Apr 22 | Lesson 13: Retrieval Model IV: Web Search 1                              |
| Apr 29 | Lesson 14: Retrieval Model IV: Web Search 2                              |

### Course Policies:

#### Course Credit Workload:
This 3-credit course requires 9-12 hours effort (including the class time) for this course each week for approximately 14 weeks. Efforts may include but is not limited to: required reading, homework assignments, and studying for quizzes.

#### Late Submissions:
For assignments, unexcused late submission (according to the Canvas timestamp and the “late” flag) will receive a grade of 0. You should plan sufficiently for completing and submitting assignments. Should an emergency arise that greatly disrupts one’s ability to complete an assignment, please send an email to Dr. Niu before the due date with a plan for submission after the due date. You need to receive Dr. Niu’s permission for late submission.

#### Copyright and Permissions:
My lectures and course materials, including presentations, quizzes, homework problems and answers, and similar materials, are protected by copyright. I am the exclusive owner of copyright in those materials I create. I encourage you to take notes and make copies of course materials for your own educational use. However, you may not, nor may you knowingly allow others to reproduce or distribute lecture notes and course materials publicly without my express written consent. This includes providing materials to commercial course material suppliers such as CourseHero, Chegg, and other similar
services. Students who publicly distribute or display or help others publicly distribute or display copies or modified copies of an instructor's course materials may be in violation of University Policy 406, The Code of Student Responsibility.

**COVID-19 Policy:**
It is the current policy of UNC Charlotte that as a condition of on-campus enrollment, all students are required to engage in safe behaviors to avoid the spread of COVID-19 in the 49er community. Such behaviors specifically include the requirement that all students properly wear CDC-compliant face coverings in all indoor spaces on campus, including classrooms and labs, regardless of vaccination status. Failure to comply with this policy in the classroom or lab may result in dismissal from the current class session. If the student refuses to leave the classroom or lab after being dismissed, the student may be referred to the Office of Student Conduct and Academic Integrity for charges under the Code of Student Responsibility.

Students are expected to attend every class and remain in class for the duration of the session when it is safe to do so in accordance with university guidance regarding COVID-19. Failure to attend class or arriving late may impact your ability to achieve course objectives which could affect your course grade. An absence, excused or unexcused, does not relieve a student of any course requirement. Regular class attendance is a student's obligation, as is a responsibility for all the work of class meetings, including tests and written tasks. Any unexcused absence or excessive tardiness may result in a loss of participation points.

Students are encouraged to work directly with their instructors regarding their absence(s). For absences related to COVID-19, please adhere to the following:

- **Complete your Niner Health Check** each morning.
- **Do not come to class if you are sick.** Please protect your health and the health of others by staying home. Contact your healthcare provider if you believe you are ill.
- **If you are sick:** If you test positive or are evaluated by a healthcare provider for symptoms of COVID-19, indicate so on your Niner Health Check to alert the University. Submit a copy of your Niner Health Check notification email to your instructors. Upon learning that you have tested positive or have been diagnosed for symptoms of COVID-19, either from your reporting or from Student health Center testing or diagnosis, representatives from Emergency Management and/or the Student Health Center will follow up with you, and your instructors will be notified of the need for accommodations, as necessary.
- **If you have been exposed** to COVID-19 positive individuals and/or have been notified to self-quarantine due to exposure, indicate so on your Niner Health Check to alert the University. Representatives from Emergency Management and/or the Student Health Center will follow up with you as necessary. Submit a copy of your Niner Health Check notification email to your instructors. If you need any additional support verifying your absence after you have communicated with your professors, contact Student Assistance and Support Services.
To return to class after being absent due to a period of **self-quarantine**, students should submit a copy of their Niner Health Check clearance email to their instructor(s). To return to class after being absent due to a COVID-19 **diagnosis**, students should submit an online request form to Student Assistance and Support Services (SASS). Supporting documentation can be attached directly to the request form and should be from a student’s health care provider or the Student Health Center, clearly indicating the dates of absences and the date the student is able to return to class. Instructors will be notified of such absences.

If you are absent from class as a result of a COVID-19 diagnosis or quarantine, as instructor I will help you continue to make progress in the course. The final decision for approval of all absences and missed work is determined by the instructor.