University of Central Florida
Department of Electrical Engineering & Computer Science

CAP 6307 Advanced Text Mining
Fall 20xx

Syllabus

Credits: 3(3,0)
Class Meetings: TBD

Instructor:
Fei Liu
Email: **feiliu@cs.ucf.edu**
Office: HEC-217
Office Hours: TBD

Course Objective:

This course presents current methods for extracting knowledge from unstructured text collections. There are two parts of this course. One part is the class lectures. The instructor will cover classification, clustering, named entity recognition, information extraction, topic modeling, and summarization. The second part is a collaborative learning experience. Students are required to read state-of-the-art research publications and present them in class, in order to hone their critical technical reading skills, oral presentation skills, and written communication skills. A midterm, final project, and several programming assignments will be required. There is no final exam.

Course outcomes and topics:

1) **Outcome 1 [Text Classification]:** Students will understand and be able to apply current approaches for text classification, including Naïve Bayes, logistic regression, and other statistical classification methods.

2) **Outcome 2 [Text Clustering]**: Students will understand and be able to apply current approaches for text clustering, including word and phrase-based clustering and probabilistic methods.

3) **Outcome 3 [Text Summarization]**: Students will understand and be able to apply state-of-the-art approaches for document summarization, including topic-based, frequency-driven, Bayesian, and machine learning methods.

4) **Outcome 4 [Information Extraction]:** Students will understand and be able to apply current approaches for named entity recognition and relation extraction, including unsupervised methods.

5) **Outcome 5** [**Topic Modeling**]: Students will understand and be able to apply current approaches for topic modeling, including probabilistic latent semantic analysis, latent Dirichlet allocation, and others.

Prerequisites:
CAP 5610 (Machine Learning)

Required Textbook:
Introduction to Information Retrieval. C. Manning P. Raghavan and H. Schutze. Cambridge, 2008. ISBN 0521865719, 9780521865715.

Recommended Textbook:
Speech and Language Processing (Second Edition), Daniel Jurafsky and James H. Martin. Pearson, 2008. ISBN-13: 978-0131873216 or ISBN-10: 0131873210.

Grading Policy:
(30%) **Programming assignments**
(20%) **Mid-Term Exam**
(15%) **Class Presentation**
(30%) **Final Project**
(5%) **Class Participation** – class attendance and participation in discussions

Makeup Exam Policy:
Makeup exams will not be given except in the most exceptional circumstances, as judged in the sole discretion of the instructor. Please give the instructor as much advance notice as possible for a missed exam.

Course Policies:
- **Programming Assignments:** The course is not intended to teach a programming language. Students should be acquainted with at least one programming language (e.g., C++, Java, Matlab, Python, Octave) and use it to solve the programming assignments. You are required to submit a brief description of how you implement the algorithm in the language you choose, the parameter setting, your test protocol, as well as the result you obtain. Source codes shall be submitted on Webcourses.

- **Final Project:** Students will form groups of 3-4 people to collaborate on the final project. It shall be a research problem that is related to text mining. Instructor will suggest some topics for the project. A project proposal shall be submitted to instructor for review before it is approved. Project proposal must specify the project team members, the roles of each member, and description of the research problem.

- **Webcourses:** The instructor will use Webcourses to post announcements, assignments, grades, the schedule and syllabus, and other information related to this course. Please be sure to check Webcourses frequently for any important information about the course.

- **Late submission:** Students have 24 hours to submit late assignment for half credit. Submissions more than 24 hours late will not be accepted unless there are extenuating circumstances (e.g., medical emergency -- where doctor notes will be required).

- **Communication with Course Instructor/TA:** For questions related to textbook, lecture slides, assignments, midterm, final project, and many others, students are encouraged to post their questions on Piazza and get the fastest response. For other questions, students are encouraged to

come to office hours. If you may have a confidential or personal question, please email the instructor/TA with 'CAP 6307' in the subject line.

• **Academic Integrity:** Plagiarism or cheating of any kind on an examination or programming assignment will not be tolerated. It may result in a zero (0) score for that assignment and may, depending on the severity of the case, lead to an "F" for the entire course. It may also be subject to appropriate referral to the Office of Student Conduct for further action. Please refer to the Golden Rule (http://www.goldenrule.sdes.ucf.edu/) in UCF's Student Handbook for further information. I will assume for this course that you will adhere to the academic creed of this University and will maintain the highest standards of academic integrity. In other words, don't cheat by giving answers to others or taking them from anyone else, and do not assist or enable others to do so. I will also adhere to the highest standards of academic integrity, so please do not ask or expect me to change your grade by bending or breaking rules for you that will not also apply to everyone else in the class.

Important Dates:
- **Classes begin: TBD**
- **Withdrawal deadline: TBD.**
- **Classes end: TBD**
- **Final Exam date: TBD**

Syllabus Quiz
All faculty are required to document students' academic activity at the beginning of each course. In order to document that you began this course, please complete the Syllabus Quiz activity by the end of the first week of classes or as soon as possible after adding the course. Failure to do so may result in a delay in the disbursement of your financial aid. You are required to complete the quiz even if you are not receiving financial aid.

Students with Disabilities:
Students with disabilities who have special testing or other needs are required to contact the Student Disability Services (SDS) office at the beginning of the semester in order to make special arrangements before requesting accommodations from the instructor. SDS is located in the Student Resource Center, Room 132, and can be reached at (407) 823-2371, TTY/TDD only phone (407) 823-2116.

Copyright
This course may contain copyright protected materials such as audio or video clips, images, text materials, etc. These items are being used with regard to the Fair Use doctrine in order to enhance the learning environment. Please do not copy, duplicate, download or distribute these items. The use of these materials is strictly reserved for this online classroom environment and your use only. All copyrighted materials are credited to the copyright holder.

Third-Party Software and FERPA
During this course you might have the opportunity to use public online services and/or software applications sometimes called third-party software such as a blog or wiki. While some of these could be required assignments, you need not make any personally identifying information on a public site. Do not post or provide any private information about yourself or your classmates. Where appropriate you may use a pseudonym or nickname. Some written assignments posted publicly may require personal reflection/comments, but the assignments will not require you to

disclose any personally identity-sensitive information. If you have any concerns about this, please contact your instructor.

***** This Syllabus is subject to change at any time and in any manner.**