USC Viterbi School of Engineering

INF 553: Foundations and Applications of Data Mining

Syllabus

Units: 4

Term — Day — Time: Fall 2018, F – 2:00-5:20 pm

Location: THH 212

Instructors: Rafael Ferreira da Silva, PhD
Anoop Kumar, PhD
Office: TBD
Regular Office Hours: Friday after class

Contact Info: rafsilva@isi.edu
anoopk@isi.edu

Course Producer: TBD
Office: TBD
Office Hours: TBD
Contact Info: TBD

I. Catalogue Course Description

Data mining and machine learning algorithms for analyzing very large data sets. Emphasis on Map Reduce. Case studies.

II. Expanded Course Description

Data mining is a foundational piece of the data analytics skill set. At a high level, it allows the analyst to discover patterns in data, and transform it into a usable product. The course will teach data mining algorithms for analyzing very large data sets. It will have an applied focus, in that it is meant for preparing students to utilize topics in data mining to solve real world problems.

III. Recommended Preparation

INF 550, INF 551 and INF 552. Knowledge of probability, linear algebra, basic programming, and some machine learning.

A basic understanding engineering principles is required, including basic programming skills; familiarity with the Python language is desirable. Most assignments are designed for the Unix

environment; basic Unix skills will make programming assignments much easier. Students will need sufficient mathematical background, including probability, statistics, and linear algebra. Some knowledge of machine learning is helpful, but not required.

IV. Course Notes

The course will be run as a lecture class with student participation strongly encouraged. There are weekly readings and students are encouraged to do the readings prior to the discussion in class. All of the course materials, including the readings, lecture slides, home works will be posted online

V. Technological Proficiency and Hardware/Software Required

Students are expected to know how to program in a language such as Python. Students are also expected to have their own laptop or desktop computer where they can install and run software to do the weekly homework assignments.

VI. Required Readings and Supplementary Materials

- Rajaraman, J. Leskovec and J. D. Ullman, *Mining of Massive Datasets*
  - Cambridge University Press, 2012.
  - Available free at: http://infolab.stanford.edu/~ullman/mmds.html

In addition to the textbook, students may be given additional reading materials such as research papers. Students are responsible for all assigned reading assignments.

VII. Grading Structure

Homework Assignments: There will be 4 homework assignments. The assignments must be done individually. Each assignment is graded on a scale of 0-100 and the specific rubric for each assignment is given in the assignment.

Project: There is a project to be delivered and presented at the end of the semester. A project proposal is required to be submitted within about the first month of class. It is expected that students deliver the project one week previous to the end of class and perform a presentation of the project to the class. The project will be conducted by small groups of students (size of the groups TBD).

Grade breakdown:

Homework	40%
Project Proposal	10%
Project:	50%
Total	100%

Grades will range from A through F. The following is the breakdown for grading:

[93 – 100] = A	[73 – 77) = C
[90 – 93) = A-	[70 –73) = C-
[87 – 90) = B+	[67 – 70) = D+
[83 – 87) = B	[63 – 67) = D
[80 – 83) = B-	[60 – 63) = D-
[77 – 80) =C+	Below 60 is an F

Note that [90, 93) means that your score is greater than or equal to 90 but less than 93. Note that every point in your coursework counts. We will strictly follow the above cut-off and NO roundup will be performed. Note that grades are NOT negotiable!

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. You can submit homework up to one week late, but you will lose 20% of the possible points for the assignment. After one week, the assignment cannot be submitted.

VIII. Course Schedule: A Weekly Breakdown

	Topic	Readings and Assignments	Deliverables/Due Dates
Week 1 – **8/24**	Introduction to Data Mining, MapReduce	Ch1: Data Mining and Ch2: Large-Scale File Systems and Map-Reduce	
Week 2 – **8/31**	MapReduce (cont.) Introduction to Data Mining tools (Spark, Tensorflow, etc.)	Ch2: Large-Scale File Systems and Map-Reduce	
Week 3 – **9/7**	Shingling, Minhashing, Locality Sensitive Hashing	Ch3: Finding Similar Items	Homework 1 assigned
Week 4 – **9/14**	Shingling, Minhashing, Locality Sensitive Hashing Recommendation Systems: Content-based and Collaborative Filtering	Ch3: Finding Similar Items Ch9: Recommendation systems	
Week 5 – **9/21**	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems	Homework 1 due, homework 2 assigned
Week 6 – **9/28**	Recommendation Systems: Content-based and Collaborative Filtering	Ch9: Recommendation systems	Project Proposal due
Week 7 – **10/5**	Frequent itemsets and Association rules	Ch6: Frequent itemsets	Homework 2 due, homework 3 assigned

Week 8 – **10/12**	Link Analysis: PageRank, Web spam and TrustRank, Random Walks with Restarts	Ch5: Link Analysis	
Week 9 – **10/19**	Clustering	Ch7: Clustering	Homework 3 due, homework 4 assigned
Week 10 – **10/26**	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	
Week 11 – **11/2**	Analysis of Massive Graphs (Social Networks)	Ch10: Analysis of Social Networks	
Week 12 – **11/9**	Web Advertising	Ch8: Advertising on the Web	Homework 4 due
Week 13 – **11/16**	Mining data streams	Ch4: Mining data streams	
Week 14 – **11/23**	**Thanksgiving Holiday (no class)**		
Week 15 – **11/30**	Buffer or advanced topics		Project due
Week 16 – **12/7**	**Project Presentation 2-5:20pm, same classroom**		Project presentation due

IX. Statement on Academic Conduct and Support Systems

Academic Conduct

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards* https://policy.usc.edu/student/scampus/part-b/. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the *Office of Equity and Diversity* http://equity.usc.edu or to the *Department of Public Safety* http://adminopsnet.usc.edu/department/department-public-safety. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. *The Relationship and Sexual Violence Prevention Services* http://engemannshc.usc.edu/rsvp/ provides 24/7 confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the *American Language Institute* http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students. *The Office of Disability Services and Programs*

http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, *USC Emergency Information* http://emergency.usc.edu will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.