Syllabus: CS585-002/685-004 Fall 2016 - "Digital Assets at Scale"				
Week	**Date**	**Assignments Due**		**Topic**
1	Thurs 8/25	N/A	N/A	(First day of class) Course overview and introduction
2	Tues 8/30 Thurs 9/1	**HW1 "Selfie"**	**Lab Notes**	The **genesis** of web crawling
3	Tues 9/6 Thurs 9/8	**HW2 "Scrape"**	**Lab Notes**	Web crawler **architectures**
4	Tues 9/13 Thurs 9/15	N/A	**Lab Notes**	Web **crawling**, scraping, mining
5	Tues 9/20 Thurs 9/22	**HW3 "Crawl"**	**Lab Notes**	**Open source** frameworks
6	Tues 9/27 Thurs 9/29	N/A	**Lab Notes**	Ranking: **PageRank**
7	Tues 10/4 Thurs 10/6	**HW 4 "Rank"**	**Lab Notes**	Web search and information retrieval: queries, indices, and **vector spaces**
8	Tues 10/11 Thurs 10/13	N/A	**Lab Notes**	Web search and information retrieval: Similarity, classification, and **clustering**
9	Tues 10/18 Thurs 10/20	**Quiz 1**	**Lab Notes**	Web **spam**
10	Tues 10/25 Thurs 10/27	N/A	**Lab Notes**	**Tools** for Achieving **Scale**: MapReduce, BigTable, Sawzall
11	Tues 11/1 Thurs 11/3	**HW5 "Search"**	**Lab Notes**	**Architectures** for Achieving **Scale**: data centers, global storage
12	Tues 11/8 Thurs 11/10	N/A	**Lab Notes**	(No Class Tuesday: Presidential Election day) Scale: **Knowledge**
13	Tues 11/15 Thurs 11/17	**HW6 M1 "Scale"**	**Lab Notes**	Scale: **Geo**
14	Tues 11/22 Thurs 11/24	N/A	**N/A**	No Class: THANKSGIVING
15	Tues 11/29 Thurs 12/1	**Quiz 2**	**Lab Notes**	Scale: **Images**
16	Tues 12/6 Thurs 12/8	N/A	**Lab Notes**	(Last week of class) Frontiers and Challenges
Final	Thurs 12/15	**HW6 M2 "Scale"**	N/A	**Final Exam Meeting Time: 1pm Thursday December 15**

Reading List

Week 1: Launch

Week 2: Genesis

- Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine." Computer networks and ISDN systems 30.1 (1998): 107-117.
  http://www.sciencedirect.com/science/article/pii/S016975529800110X
- Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) 46.5 (1999): 604-632.
  http://dl.acm.org/citation.cfm?id=324140
- Li, Yanhong. "Toward a qualitative search engine." IEEE Internet Computing, July-Aug. 1998, Vol.2(4) pp. 24-29.
  http://ieeexplore.ieee.org.ezproxy.uky.edu/stamp/stamp.jsp?tp=&arnumber=707687

Week 3: Web Crawler Architectures

- Soumen Chakrabarti, "Mining the Web: Discovering Knowledge from Hypertext Data", Morgan-Kauffman, 2002. ISBN 1-55860-754-4. Chapters 1, 2.
  http://www.cse.iitb.ac.in/~soumen/mining-the-web/
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008. Chapters 19,20.
  http://www-nlp.stanford.edu/IR-book/

Week 4: Web Crawling, Scraping, Mining

- Jeffrey Kenneth Hirschey, Symbiotic Relationships: Pragmatic Acceptance of Data Scraping, 29 Berkeley Tech. L.J. (2014).
  http://scholarship.law.berkeley.edu/btlj/vol29/iss4/16
- The Atlantic Monthly: How Netflix Reverse Engineered Hollywood
  http://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

Week 5: Open Source Frameworks

- Wget, the GNU tool (command line).
  https://en.wikipedia.org/wiki/Wget
- Norconex HTTP Collector: Open-Source Enterprise Web Crawler (Java).
  http://www.norconex.com/collectors/collector-http/
- Scrapy Developers, "Scrapy Documentation: Release 1.1.0", 7/13, 2016 (Python).
  https://media.readthedocs.org/pdf/scrapy/1.1/scrapy.pdf

Week 6: Ranking

- Lawrence Page, Sergay Brin, Rageev Motwani, and Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Infolab Technical Report 1999-66, 1999.
  http://ilpubs.stanford.edu:8090/422/
- Austin, David. "How Google Finds Your Needle in the Web's Haystack". Feature Column, American Mathematical Society.
  http://www.ams.org/samplings/feature-column/fcarc-pagerank

Week 7: Search and Information Retrieval: Queries, Indices, and Vector Spaces

- Soumen Chakrabarti, "Mining the Web: Discovering Knowledge from Hypertext Data", Morgan-Kauffman, 2002. ISBN 1-55860-754-4. Chapter 3.
  http://www.cse.iitb.ac.in/~soumen/mining-the-web/

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008. Chapter 6.
  http://www-nlp.stanford.edu/IR-book/

Week 8: Search and Information Retrieval: Similarity, Classification, and Clustering

- Soumen Chakrabarti, "Mining the Web: Discovering Knowledge from Hypertext Data", Morgan-Kauffman, 2002. ISBN 1-55860-754-4. Chapter 4.
  http://www.cse.iitb.ac.in/~soumen/mining-the-web/
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "Introduction to Information Retrieval", Cambridge University Press. 2008. Chapter 14.
  http://www-nlp.stanford.edu/IR-book/

Week 9: Web Spamming and Web Abuse

- Zoltan Gyöngyi and Hector Garcia-Molina, "Web Spam Taxonomy", 2005
  http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.5304
- Kirill Levenchenko et al., "Click Trajectories: End-to-End Analysis of the Spam Value Chain", in Proc. IEEE Symposium on Security and Privacy, pp. 431-446, 2011.
  https://cseweb.ucsd.edu/~savage/papers/Oakland11.pdf

Week 10: Tools at Scale

- Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.
  http://research.google.com/archive/mapreduce.html
- Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. "Bigtable: A Distributed Storage System for Structured Data", OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November, 2006.
  http://research.google.com/archive/bigtable.html
- Rob Pike, Sean Dorward, Robert Griesemer, Sean Quinlan. "Interpreting the Data: Parallel Analysis with Sawzall", Scientific Programming Journal, Special Issue on Grids and Worldwide Computing Programming Models and Infrastructure 13:4, pp. 227-298, 2005.
  http://research.google.com/archive/sawzall.html

Week 11: Scale – Data Centers, Global Storage

- Stephen Levy, "Google Throws Open Doors to Its Top-Secret Data Center", Wired Magazine, 2012.
  http://www.wired.com/2012/10/ff-inside-google-data-center/
- James C. Corbett, Jeffrey Dean, et al., "Spanner: Google's Globally-Distributed Database", 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12), 2012.
  http://static.googleusercontent.com/media/research.google.com/en/us/archive/spanner-osdi2012.pdf

Week 12: Scale – Knowledge

- Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books", Science, 2010.
  http://stevenpinker.com/files/pinker/files/michel_et_al_quantitative_analysis_of_culture_science_2011.pdf
- Kurt Bollacker et al., "A Platform for Scalable, Collaborative, Structured Information Integration", AAAI 2007.

http://aaaipress.org/Papers/Workshops/2007/WS-07-14/WS07-14-004.pdf
http://bits.blogs.nytimes.com/2010/07/16/google-buys-metaweb-to-improve-search-results/

Week 13: Scale – Geo

- Dragomir Anguelov and Carole Dulong and Daniel Filip and Christian Frueh and Stéphane Lafon and Richard Lyon and Abhijit Ogale and Luc Vincent and Josh Weaver, "Google Street View: Capturing the World at Street Level", Computer, vol. 43, 2010.
  http://static.googleusercontent.com/media/research.google.com/en/us/pubs/archive/36899.pdf
- Alexis C. Madrigal, "How Google Builds its Maps – and What It Means for the Future of Everything", The Atlantic, September 6, 2012.
  http://www.theatlantic.com/technology/archive/2012/09/how-google-builds-its-maps-and-what-it-means-for-the-future-of-everything/261913/
- Bryan Klingner, David Martin, James Roseborough, "Street View Motion-fromStructure-from-Motion", Proc. ICCV, 2013.
  http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41413.pdf

Week 14: Thanksgiving

Week 15: Scale – Images

- Beaver, Doug and Kumar, Sanjeev and Li, Harry C. and Sobel, Jason and Vajgel, Peter, "Finding a Needle in Haystack: Facebook's Photo Storage", Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation, OSDI'10, 2010.
  http://dl.acm.org/citation.cfm?id=1924947
- Le, Q.V., "Building high-level features using large scale unsupervised learning," *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* , vol., no., pp.8595,8598, 26-31 May 2013
  http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6639343

Week 16: Frontiers and Challenges

- Machine Learning
  http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
- Virtual Reality
  http://www.wired.com/2016/04/magic-leap-vr/