



POSTECH
교육혁신센터

Big Data 첫 걸음 (Intro to Big Data)

담당 교수 유환조 / 조교 이동하

강좌개요

본 강좌는 인공지능과 더불어 소프트웨어 중심의 미래사회에서 지능의 고도화에 핵심적인 역할을 하는 빅 데이터에 대한 일반적인 이해와 초보적인 방법론을 학습하는 강좌이다.

참여교수 소개

- 유환조 / 포스텍 컴퓨터공학과 교수
- 2004년 미국 일리노이 대학 컴퓨터공학 박사
- 2004년 ~ 2007년 미국 아이오와 대학 교수
- 2008년 ~ 현 포스텍 교수
- 빅 데이터 분야 가장 권위있는 국제 학회인 ACM KDD에 한국인 최초 & 최다 논문 발표
- Information Sciences 저널 부 편집장
- Neurocomputing 저널 부 편집장
- ACM KDD, IEEE ICDM, ACM CIKM, ACM SIGMOD, AAAI, IJCAI 등 인공지능 및 빅 데이터 관련 국제학회에서 프로그램 커미티 멤버로 활동 중

운영조교(TA) 소개

- 이동하 / 컴퓨터공학과 석박사 통합과정 대학원생
- 연구주제: 싱글 머신에서의 디스크 기반 기계 학습 알고리즘 최적화 연구

I. 강좌 일반 - 1

강좌명	빅 데이터 첫걸음	교수자	유환조 교수 (hwanjoyu@postech.ac.kr)
	Intro to Big Data		개발조교-이동하, 운영조교-이동하
학 습 목 표	본 강좌는 <ul style="list-style-type: none"> • 빅 데이터 분석의 기초가 되는 개념들을 습득하고 실습한다. • 빅 데이터의 저장, 관리, 검색, 분석을 위한 SQL을 공부한다. • 빅 데이터의 분산 저장 및 분석을 위한 맵-리듀스를 공부한다. • 빅 데이터 분석의 기초가 되는 통계 및 최적화 기법을 익힌다. • 기초적인 데이터 마이닝, 기계학습 알고리즘 및 통계적 모델링 방법을 공부한다. 		
수 강 대 상	<ul style="list-style-type: none"> ▪ 빅 데이터를 본격적으로 공부하고자 하는 컴퓨터 전공 학생 ▪ 빅 데이터에 대한 일반적인 이해가 필요한 타 전공 대학생 ▪ 빅 데이터에 대한 기초적인 이해가 필요한 재직 일반인 	학 습 시 간	주당 6시간
평 가 준	중간고사 25%, 기말고사 25%, 과제(2회) 20%, 주차별 퀴즈 30%	이 수 건	60%이상
교 재	<ul style="list-style-type: none"> ▪ 별도 제공 		

I. 강좌 일반 - 2

● 선수학습 요건(prerequisite)

- 예: 컴퓨터 공학 개론을 이수한 자
- SQL의 기초적인 이해수준 이상인 자...

● 학습자

- 학습주제가 널리 회자되어 다양한 계층의 학습자가 참여할 것으로 예상
- 전산 분야에 대한 상당한 이해를 갖고 Big Data 학습을 시작하는 수강생이 일부 있을 것으로 예상 (실습을 통한 수준 높은 학습 요구)
- 전산분야 기초지식은 부족하지만 Big Data에 대한 개념적인 이해를 필요로 하는 일반인 또는 비전공 대학생들이 다수 있을 것으로 예상(개념적인 이해에 대한 학습 유도)

● 학습 주제 관련

- 사회적으로 관심이 큰 분야이어서 많은 수강생들이 우선적으로 참여는 할 것으로 예상
- 실습 등이 포함된 전공기초 성격이 강하여 조기 이탈자가 많을 것으로 우려됨.

II. 내용구성 - 1

주·차시	학습주제	담당교수	주제별 수업방법/자료	평가	학습자료
1	1 빅데이터 정의	유환조	<ul style="list-style-type: none"> • 빅데이터 응용, 정의, 역사 • 빅데이터 관련 분야 	• 차시별 퀴즈	• 강의노트(PDF)
	2 빅데이터의 Technical Challenge	유환조	<ul style="list-style-type: none"> • 빅데이터 시스템 구조 • 빅데이터의 Technical Challenge 	• 차시별 퀴즈	• 강의노트(PDF)
2	1 데이터 모델	유환조	<ul style="list-style-type: none"> • 데이터 저장 및 표현 • 데이터 모델의 요소 	• 차시별 퀴즈	• 강의노트(PDF)
	2 관계형 데이터 모델	유환조	<ul style="list-style-type: none"> • 관계형 데이터 모델 • 데이터 속성 • Relational Algebra 	• 차시별 퀴즈	• 강의노트(PDF)
3	1 Relational Algebra	유환조	<ul style="list-style-type: none"> • Set vs. Bag • Basic Relational Algebra Operations • SQL 	• 차시별 퀴즈	• 강의노트(PDF)
	2 SQLite 실습	유환조	<ul style="list-style-type: none"> • SQLite를 이용한 Basic Relational Algebra and SQL 실습 	• 차시별 퀴즈	• 강의노트(PDF)
4	1 Extended Relational Algebra	유환조	<ul style="list-style-type: none"> • Extended Relational Algebra Operations • SQLite를 이용한 Extended Relational Algebra and SQL 실습 	• 차시별 퀴즈	• 강의노트(PDF)
	2 Relational Query Processing	유환조	<ul style="list-style-type: none"> • Query Plan Generation • View and Index 	• 차시별 퀴즈	• 강의노트(PDF)

II. 내용구성 - 2

주·차시	학습주제	담당교수	주제별 수업방법/자료	평가	학습자료
5	1 SQL을 이용한 데이터 분석	유환조	• SQL을 이용한 데이터 분석 연습	• 차시별 퀴즈	• 강의노트(PDF)
	2 빅데이터 처리	유환조	• 빅데이터의 크기 • 빅데이터 읽기 • 빅데이터 쓰기	• 차시별 퀴즈	• 강의노트(PDF)
6	1 Map-Reduce I	유환조	• Map function • Reduce function	• 차시별 퀴즈	• 강의노트(PDF)
	2 Map-Reduce II	유환조	• Map Reduce를 이용한 문서 처리 • Map Reduce를 이용한 Join • Map Reduce를 이용한 Matrix Multiplication	• 차시별 퀴즈	• 강의노트(PDF)
7	1 빅데이터 파일 시스템	유환조	• 빅데이터를 위한 Cluster Computing • 빅데이터 파일 시스템 • Combiner	• 차시별 퀴즈	• 강의노트(PDF)
	2 Parallel Data Processing vs. Distributed Data Processing	유환조	• Parallel Data Processing과 Distributed Data Processing의 비교	• 차시별 퀴즈	• 강의노트(PDF)
				• 중간고사	
8	1 Map Reduce vs. Databases	유환조	• Hadoop vs. RDBMS • Selection, Aggregate, Join Task 비교 • 장단점 비교	• 차시별 퀴즈	• 강의노트(PDF)
	2 NoSQL	유환조	• NoSQL 시스템 비교 • Two-Phase Commit • Eventual Consistency	• 차시별 퀴즈	• 강의노트(PDF)

II. 내용구성 - 3

주·차시	학습주제	담당교수	주제별 수업방법/자료	평가	학습자료	
9	1	NoSQL	유환조	<ul style="list-style-type: none"> • Document Store and CouchDB • Big Table and Hbase 	• 차시별 퀴즈	• 강의노트(PDF)
	2	Statistics for Big Data	유환조	<ul style="list-style-type: none"> • Hypothesis Test • P-value, Effect Size • Publication Bias 	• 차시별 퀴즈	• 강의노트(PDF)
10	1	Statistics for Big Data	유환조	<ul style="list-style-type: none"> • Benford's law • Multiple Hypothesis Testing 	• 차시별 퀴즈	• 강의노트(PDF)
	2	Machine Learning Basics	유환조	<ul style="list-style-type: none"> • Supervised vs. Unsupervised Learning • Overfitting • Validation 	• 차시별 퀴즈	• 강의노트(PDF)
11	1	Decision Trees I	유환조	<ul style="list-style-type: none"> • Decision Tree Induction • Entropy and Information Gain 	• 차시별 퀴즈	• 강의노트(PDF)
	2	Decision Trees II	유환조	<ul style="list-style-type: none"> • Tree Pruning • Decision Boundary 	• 차시별 퀴즈	• 강의노트(PDF)
12	1	Weka 실습: Decision Trees	유환조	<ul style="list-style-type: none"> • Weka 실습 : 데이터 입력 • Weka 실습 : 모델 학습 	• 차시별 퀴즈	• 강의노트(PDF)
	2	Bayesian Classification	유환조	<ul style="list-style-type: none"> • Bayesian Theorem • Naïve Bayes Classification 	• 차시별 퀴즈	• 강의노트(PDF)

II. 내용구성 - 4

주·차시	학습주제	담당교수	주제별 수업방법/자료	평가	학습자료
13	1 Ensemble Methods	유환조	<ul style="list-style-type: none"> • Bagging • Boosting • Random Forest 	• 차시별 퀴즈	• 강의노트(PDF)
	2 Artificial Neural Network	유환조	<ul style="list-style-type: none"> • Linear Classification Function • Perceptron • Multi-layer Perceptron 	• 차시별 퀴즈	• 강의노트(PDF)
14	1 Support Vector Machine I	유환조	<ul style="list-style-type: none"> • Linear Support Vector Machine • Primal Problem • Dual Problem 	• 차시별 퀴즈	• 강의노트(PDF)
	2 Support Vector Machine II	유환조	<ul style="list-style-type: none"> • Nonlinear Support Vector Machine • Kernel Trick • Multi-Class Classification 	• 차시별 퀴즈	• 강의노트(PDF)
				• 기말고사	