**강의개요**

Big data for Bioinformatics

매우 큰 데이터를 어떻게 빠르게 처리할 수 있을까? 한 대의 머신에 들어가지 않는 데이터에 대한 분석은 어떻게 수행할 수 있을까? 최근 생성되는 바이오정보 데이터의 크기는 점점 커지고 있지만, 저장 공간과 시간의 제약으로 이러한 빅데이터를 하나의 머신으로 처리하는데 많은 어려움이 따른다.

본 강의에서는 대용량 데이터를 효과적이고 빠르게 처리하기 위해 널리 쓰이고 있는 플랫폼인 Hadoop과 Spark을 설명한다. Hadoop과 Spark의 개요, Hadoop/Spark 프로그래밍 방법, 텍스트 및 그래프 데이터 처리에의 응용을 설명하며 이를 통해 빅데이터를 빠르고 손쉽게 처리할 수 있는 기법을 배우고, 이를 실제 바이오 데이터에 적용하여 효율적이고 효과적인 분석을 할 수 있는 핵심 역량을 갖추는 것을 목표로 한다.

강의는 다음의 내용을 포함한다:

* Hadoop/Spark 개요
* Hadoop/Spark Programming
* 텍스트 데이터 처리에의 응용
* 그래프 데이터 처리에의 응용

\*참고강의교재:

러닝 스파크 (제이펍, 2015)

\*교육생준비물:

노트북 (메모리 8GB 이상, 디스크 여유공간 30GB 이상)

\* 강의: 강 유 교수 (서울대학교 컴퓨터공학부), 박하명 박사 (서울대학교 컴퓨터공학부)

\*조교: 조민용 (chominyong@gmail.com)

**Korean Society of Bioinformatics**

**BIML 2019**

|  |  |  |
| --- | --- | --- |
|  | | |
|  | | |
| **시 간** | **발 표 내 용** | **연 자** |
| 09:00-09:10 | 등 록 | |
| 09:10-09:30 | 개 회 사 | |
| 9:30-10:45 | MapReduce 개요 및 프로그래밍 | 강 유 |
|  | 휴식 및 질의응답 | |
| 11:15-12:30 | Spark 개요 및 프로그래밍 | 강 유 |
| 12:30–13:30 | 점 심 | |
| 13:30-15:10 | MapReduce 실습: 텍스트 분석 기법 | 박 하명 |
|  | 휴식 및 질의응답 | |
| 15:30-17:10 | Spark 실습: 텍스트 분석 기법 | 박 하명 |
|  | 휴식 및 질의응답 | |
| 17:30-19:10 | Spark 실습: 그래프 분석 기법 | 박 하명 |

**Curriculum Vitae**

**Speaker Name: U Kang, Ph.D.**

▶**Personal Info**



Name U Kang

Title Associate professor

Affiliation Seoul National University

▶**Contact Information**

Address 301-502, 1 Gwanak Ro Gwanak-Gu, Seoul 08826

Email ukang@snu.ac.kr

Phone Number 02-880-7254

**Research interest :** Data Mining, Big Data, Machine Learning

**Educational Experience**

2003 B.S., Seoul National University

2009 M.S., Carnegie Mellon University

2012 Ph.D., Carnegie Mellon University

**Professional Experience**

2012 Post-doc research fellow, Carnegie Mellon University

2013-2015 Assistant Professor, Department of Computer Science, KAIST

2015- Associate Professor, Department of Computer Science and Engineering, Seoul National University.

**Selected Publications (5 maximum)**

1. U Kang, Charalampos E. Tsourakakis, and Christos Faloutsos, PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations, IEEE International Conference on Data Mining (ICDM) 2009, Florida, Miami, USA.

2. Ha-Myung Park, Sung-Hyon Myaeng, and U Kang, PTE: Enumerating Trillion Triangles On Distributed System, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2016, San Francisco, USA.

3. Jinhong Jung, Namyong Park, Lee Sael, and U Kang, BePI: Fast and Memory-Efficient Method for Billion-Scale Random Walk with Restart, ACM International Conference on Management of Data (SIGMOD) 2017, Chicago, IL, USA.

4. Jaemin Yoo, Saehan Jo, and U Kang, Supervised Belief Propagation: Scalable Supervised Inference on Attributed Networks, IEEE International Conference on Data Mining (ICDM) 2017, New Orleans, USA.

5. Jun-gi Jang, Dongjin Choi, Jinhong Jung, and U Kang, Zoom-SVD: Fast and Memory Efficient Method for Extracting Key Patterns in an Arbitrary Time Range, ACM International Conference on Information and Knowledge Management (CIKM) 2018, Lingotto, Turin, Italy.

**Speaker Name: Ha-Myung Park , Ph.D.**

▶**Personal Info**



Name Ha-Myung Park

Title Research Fellow

Affiliation Seoul National University

▶**Contact Information**

Address Room 314, 138 Building, Seoul National Univ.

Email hamyung.park@snu.ac.kr

Phone Number 010-6778-6636

**Research interest :** Large scale graph mining, scalable machine learning and data mining

**Educational Experience**

2011 B.S., SungKyunKwan University

2013 M.S., KAIST

2018 Ph.D., KAIST

**Professional Experience**

2018-2019 Post-doc research fellow, Seoul National University

**Selected Publications (5 maximum)**

1. Ha-Myung Park, Francesco Silvestri, Rasmus Pagh, Chin-wan Chung, Sung-Hyon Myaeng, and U Kang, Enumerating Trillion Subgraphs On Distributed Systems, ACM Transactions on Knowledge Discovery from Data (TKDD), 12(6), 2018

2. Ha-Myung Park, Namyong Park, Sung-Hyon Myaeng, U Kang, Partition Aware Connected Component Computation in Distributed Systems, IEEE International Conference on Data Mining (ICDM), 2016

3. Ha-Myung Park, Sung-Hyon Myaeng , U Kang, PTE: Enumerating Trillion Triangles On Distributed Systems, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2016

4. Ha-Myung Park, Francesco Silvestri, U Kang, Rasmus Pagh, MapReduce Triangle Enumeration With Guarentees, ACM International Conference on Information and Knowledge Management (CIKM), 2014

5. Ha-Myung Park, Chin-Wan Chung, An Efficient MapReduce Algorithm for Counting Triangles in a Very Large graph, ACM International Conference on Information and Knowledge Management (CIKM), 2013