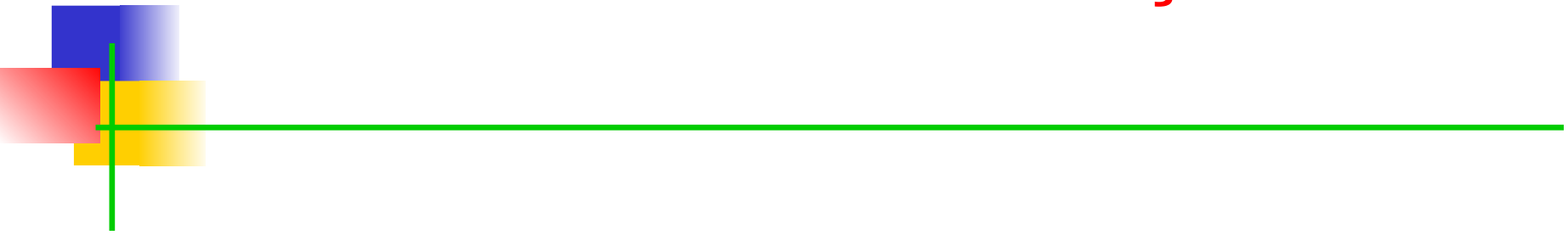
BLM 442: Büyük Veri Analizine Giriş



Dr. Süleyman Eken

Kocaeli Üniversitesi
Bilgisayar Mühendisliği

2018-2019 Bahar

Sunum Planı

- Ders ne ile ilgili?
- Dersi kimler almalı? & Kimler almamalı?
- Keşif çalışmları, karşılıklı geri bildirimler ve sınavlar
- Ders takibi ve bir takım bilgiler
- Ders konuları

Ders ne ile ilgilidir?

Büyük veri araç ve tekniklerinin temellerini öğrenmek isteyen ve öğrendiklerini kendi çalışma alanlarına (bitirme, iş vs) uygulayan lisans öğrencilerine yöneliktir. Bilim, teknoloji, işletme, tıp, politika ve tümünü içine alan bu alanlardaki en büyük keşif ve kararlarının çoğu, büyük veri setlerini analiz etme temeline dayanmaktadır. Bu ders büyük veri analizine geniş bir bakış açısı sağlar. Veritabanlarını içeren veri analizi teknikleri; makine öğrenmesi ve veri görselleştirme; elektronik tablolar, Tableau, ilişkisel veritabanları ve SQL ve Python içeren veri analiz araçları; lineer cebir, olasılık ve istatistik temelleri; Apache Hadoop tasarım kalıpları; Amazon büyük veri platform ve araçları; Apache Spark ile akan veri analizi; ağ analizi, çizge veritabanları üzerinde veri analizi; yapılandırılmamış verilere giriş ve derin öğrenmeye giriş. Ders kapsamındaki araçlar uygulamalı anlatılacak olup yapacağınız keşiflere yol gösterici olacaktır. Ön koşullar: Orta seviye/üzeri Python veya Java bilgisinin olması önerilir. Keşif aktiviteleri Python, Scala, Java veya R'da yapılmalıdır.

Ders ne ile ilgilidir?

Büyük veri araç ve tekniklerinin temellerini öğrenmek isteyen ve öğrendiklerini kendi çalışma alanlarına (bitirme, iş vs) uygulayan lisans öğrencilerine yöneliktir. Bilim, teknoloji, işletme, tıp, politika ve tümünü içine alan bu alanlardaki en büyük keşif ve kararlarının çoğu, büyük veri setlerini analiz etme temeline dayanmaktadır. Bu ders büyük veri analizine geniş bir bakış açısı sağlar. Veritabanlarını içeren veri analizi teknikleri; makine öğrenmesi ve veri görselleştirme; elektronik tablolar, Tableau, ilişkisel veritabanları ve SQL ve Python içeren veri analiz araçları; lineer cebir, olasılık ve istatistik temelleri; Apache Hadoop tasarım kalıpları; Amazon büyük veri platform ve araçları; Apache Spark ile akan veri analizi; ağ analizi, çizge veritabanları üzerinde veri analizi; yapılandırılmamış verilere giriş ve derin öğrenmeye giriş. Ders kapsamındaki araçlar uygulamalı anlatılacak olup yapacağınız keşiflere yol gösterici olacaktır. Ön koşullar: Orta seviye/üzeri Python veya Java bilgisinin olması önerilir. Keşif aktiviteleri Python, Scala, Java veya R'da yapılmalıdır.

Dersi kimler almalı?

Büyük veri araç ve tekniklerinin temellerini öğrenmek isteyen ve öğrendiklerini kendi çalışma alanlarına (bitirme, iş vs) uygulayan lisans öğrencilerine yöneliktir. Bilim, teknoloji, işletme, tıp, politika ve tümünü içine alan bu alanlardaki en büyük keşif ve kararlarının çoğu, büyük veri setlerini analiz etme temeline dayanmaktadır. Bu ders büyük veri analizine geniş bir bakış açısı sağlar. Veritabanlarını içeren veri analizi teknikleri; makine öğrenmesi ve veri görselleştirme; elektronik tablolar, Tableau, ilişkisel veritabanları ve SQL ve Python içeren veri analiz araçları; lineer cebir, olasılık ve istatistik temelleri; Apache Hadoop tasarım kalıpları; Amazon büyük veri platform ve araçları; Apache Spark ile akan veri analizi; ağ analizi, çizge veritabanları üzerinde veri analizi; yapılandırılmamış verilere giriş ve derin öğrenmeye giriş. Ders kapsamındaki araçlar uygulamalı anlatılacak olup yapacağınız keşiflere yol gösterici olacaktır. Ön koşullar: Orta seviye/üzeri Python veya Java bilgisinin olması önerilir. Keşif aktiviteleri Python, Scala, Java veya R'da yapılmalıdır.

Dersi kimler almamalı?

Yanlış yerde olduğunuzu düşünüyor ve tüm bunlara ne gerek var diyorsanız danışmanınıza dersi hemen bırakmayı talep ediniz.

Keşif aktiviteleri & Geri bildirimler

Konular	Hafta	Teslim
Seçilecek spreadsheet üzerinde birtakım problemler (analiz, gorsellestime iceren) gerçeklenecek (csv, excel vs)	1 Mart	8 Mart
SQL Uygulamaları	8 Mart	15 Mart
Python temeller ve veri yapıları üzerine	15 Mart	22 Mart
pandas & plotlib	22 Mart	29 Mart
scikit-learn	12 Nisan	26 Nisan
Apache Hadoop Tasarım kalıbı uygulaması	26 Nisan	3 Mayıs
Akan veri üzerine uygulama	10 Mayıs	17 Mayıs
Neo4j uygulama	17 Mayıs	24 Mayıs

Konular	Hafta	Teslim
• Tableau kullanarak seçilecek olan bir spreadsheet üzerinde veri analizi ve görselleştirme • İnteraktif CV hazırlama	1 Mart	22 Mart
scikit-learn uygulaması	12 Nisan	7 Haziran

Sınavlar

Sınav	Hafta	Şekli
Vize	13-21 Nisan	Klasik
Final	10-18 Haziran	Klasik

Notlandırma

- %28 (ara sınav), %16 (KAs), %16 (KGBs), %40 (final).
- Ara sınav ve final sınavları; öğrenilen programlama dili/teknolojiyi ilgilendiren sorular, kod/pseuco-code yazımı üzerine olacaktır.
- Her KA'nın genel ortalamaya katkısı: 2 puan
- Her KGB'nin genel ortalamaya katkısı: 8 puan
- Geç teslim politikası: 1 gün %10, 2 gün %30, geri kalan günler dikkate alınmaz.

Ders Kaynakları

- Tableau Your Data! : Fast and Easy Visual Analysis with Tableau Software, 2nd Edition, Dan Murray, January 2016
- Python for Data Analysis, 2nd Edition Data Wrangling with Pandas, NumPy, and IPython, William McKinney, 2017
- Learning scikit-learn: Machine Learning in Python– November 25, 2013, Raúl Garreta, Guillermo Moncecchi
- Building Machine Learning Systems with Python, Willi Richert, Luis Pedro Coelho, 2013
- MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems, 2nd Edition, Donald Miner, Adam Shook, February 25, 2017
- Learning Spark : Lightning-Fast Big Data Analysis, Holden Karau, Andy Kowinski, Mark Hamstra, Matei Zaharia, 01 Nov 2015
- Pro Spark Streaming: The Zen of Real-Time Analytics Using Apache Spark, 1st ed. Edition, Zubair Nabi, June 14, 2016
- Graph Databases, Second Edition, Ian Robinson, Jim Webber, and Emil Eifrem, June 2015

Haberleşme kanallarımız

- Genelde Görüntü İşleme laboratuvarındayım.
- Mail yoluyla bir gün öncesinden randevu (suleyman.eken@kocaeli.edu.tr)
- Web sitesi: https://suleymaneken.github.io/teaching/
- Piazza
  - Duyurular
  - Sorular ve Cevaplar (özel ve genel)
  - Tartışma
  - class link: piazza.com/kocaeli_university/spring2019/blm442/home
  - code: blm442

İntihal

- Netten alınacak kısmi kod parçaları önceden kod içinde/raporda belirtilmek ve soru sorulduğunda cevaplanması durumunda sıkıntı çıkarmayacaktır.
- (i) Internet kaynağını belirtmeyen/açıklayamayan/üzerinde geliştirme yapmayan veya (ii) birbirleriyle benzer/aynı çalışma teslim edenlerin aktiviteleri sıfır üzerinden değerlendirilecektir.

Konular

Hafta 1: Büyük veri analizine genel bakış

- Dağıtık sistemlere kısa temas
- Büyük veri nedir?
- Büyük veri ile neler yapılabilir?
- Büyük veri nasıl işlenir?
- Nereden başlamalı?
- Sertifika programları, iş ilanları vs

Hafta 2: Elektronik Tablolar (Spreadsheets) Kullanarak Veri Analizi ve Görselleştirme

- Elektronik tablo oluşturma ve veri ile doldurma
- Kolay görüntüleme için veri biçimlendirme
- Formüllerle veri üzerinde işlem (toplama, ort, filtreleme, fonksiyonlar vs)
- Veriyi tutma, paylaşma, modifiye etme vs
- Google E-Tablolar'daki grafik türleri

Hafta 2: Tableau kullanarak gelişmiş görselleştirme

- Tableau Desktop and Tableau Reader
- Popüler bağımsız veri görselleştirme ve analitik aracı
- Elektronik tablolara benzer görselleştirme türleri (daha sezgisel ve kullanımı daha kolay)
  - Etkileşimli görselleştirmeler oluşturma yeteneği
  - Etkileşimli görselleştirme ve yayınlayabilme
- Grafik türleri
- İnteraktif kontrol paneli (dashboard)

Hafta 3: İlişkisel veritabanları ve temel SQL

- Oracle, Microsoft SQL Server, IBM DB2, vd (ticari)
- MySQL, SQLite, PostgreSQL, vd (açık kaynak)
- Tablo & Elektronik tablolar
- Veri oluşturma ve yükleme
- Sorgular, joins, veri modifiyesi
- ipython-sql: Jupyter Notebook vasıtasıyla İlişkisel veritabanına erişim

Hafta 4: Pyhton'a giriş, built-in veri yapıları, built-in fonksiyonlar

- Modüller, aritmetik, fonksiyonlar
- Strings, lists, tuples, dictionaries, sets veri yapıları
- Koşul ifadeler ve döngüler
- Sıralama, list comprehensions, generators, iterators
- Nesne yönelimli programlama
- Functional tools, enumerate, zip vs.

Hafta 5: pandas

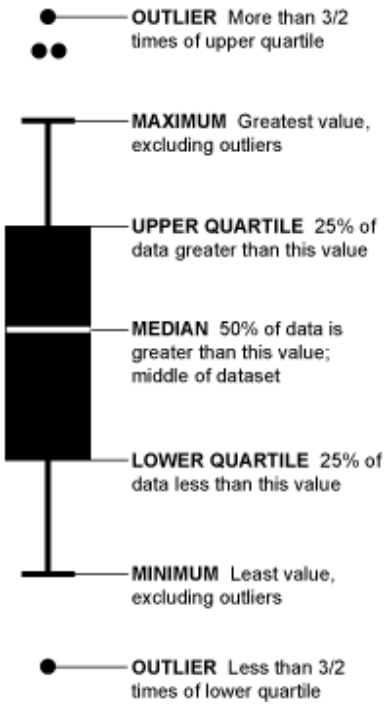
- Pandas veri analizi için bir Python paketidir.
- Veri setlerinin manipülasyonunu ve analizini basitleştiren built-in veri yapıları sağlar.
- Series, DataFrame
- Satır, sütun seçme ve indeksleme, groupby, merge
- Dosyadan okuma, dosyaya yazma
- Zaman serileri analizi

Hafta 5: Görselleştirme

- Matplotlib: 2D görselleştirme aracı
- Bar, line charts
- Scatter plots
- Seaborn: istatistiksel veri görselleştirme aracı
- Tematik haritalar
  - Folium, Basemap, Cartopy, Iris, …
- Diğer görselleştirme araçları
  - Bokeh (interaktif plots), plotly, …

Hafta 6: Lineer Cebir, İstatistik, Olasık Temeller

- Lineer Cebir
  - Vektörler, Matrisler, SciPy
- İstatitstik
  - mean, varyans -veri saçılımı-
  - standart sapma, min, max, median,
  - mode, korelasyon matrisi vs.
- Olasılık
  - Koşullu olasılık, bayes teoremi, normal dağılım, merkezi limit teoremi vs.



Büyük veri araçları ve teknikleri

- Temel veri manipülasyon ve analizi
  - Iyi tanımlanmış hesaplamalar yaptırma veya iyi tanımlanmış sorular sorma (queries)
- Veri madenciliği
  - Veri içindeki desenleri (patterns) arama
- Makine öğrenmesi
  - Veriyi kullanarak model oluşturma ve tahmin yapma
- Veri görselleştirme
  - Verinin grafiksel yorumu
- Veri toplama ve hazırlama

Makine öğrenmesi

- Supervised machine learning
  - Set of labeled examples to learn from: training data
  - Develop model from training data
  - Use model to make predictions about new data
- Unsupervised machine learning
  - Unlabeled data, look for patterns or structure (similar to data mining)
- Reinforcement learning
  - Improve model as new data arrives
- Semi-supervised learning
  - Labeled + unlabeled
- Active learning
  - Semi-supervised, ask user for labels

Hafta 7: Regresyon

- Supervised
- Training data, each example:
  - Set of predictor values -“independent variables”
  - Numeric output value -“dependent variable”
- Model is function from predictors to output
  - Use model to predict output value for new predictor values
- Example
  - Predictors: mother height, father height, current age
  - Output: height

Hafta 7: Diğer ML tipleri

- Classification
  - Like regression except output values are labels or categories
  - Example
    - Predictor values: age, gender, income, profession
    - Output value: buyer, non-buyer
- Clustering
  - Unsupervised
  - Group data into sets of items similar to each other
  - Example -group customers based on spending patterns

Hafta 8: Python Kullanarak Makine Öğrenmesi

- Scikit-learn açık kaynak kodlu bir Python kütüphanesidir.
- Bir dizi makine öğrenmesi, ön işleme, çapraz doğrulama ve görselleştirme algoritmaları sunar.
- Ön işleme
  - Standardization, Normalization, Binarization, Encoding Categorical Features, Imputing Missing Values, and etc.
- Model oluşturma
  - Supervised Learning Estimators (LR, SVM, NB, KNN, and etc.)
  - Unsupervised Learning Estimators (PCA, K-mean, and etc.s)
- Model oturtma (fitting) ve tahmin
- Performans değerlendirme
  - Sınıflandırma: doğruluk, karışıklık matrisi, kesinlik, f1-score vs.
  - Regresyon metrics: MSE, MAE, and etc.
  - Kümeleme: Homogeneity, V-measure, and etc.
- Model ayarlama (tuning)

Hafta 9: Apache Hadoop Tasarım Kalıpları

- Summarization Patterns
  - Numerical Summarization
  - Inverted Index Summarization
  - Counting with Counters
- Filtering Patterns
  - Bloom Filtering
  - Distinct
- Data Organization Patterns
  - Partitioning
- Join Patterns
  - Reduce Side Join
  - Replicated Join
  - Cartesian Product

Hafta 10: Amazon Big Data Platforms and Services

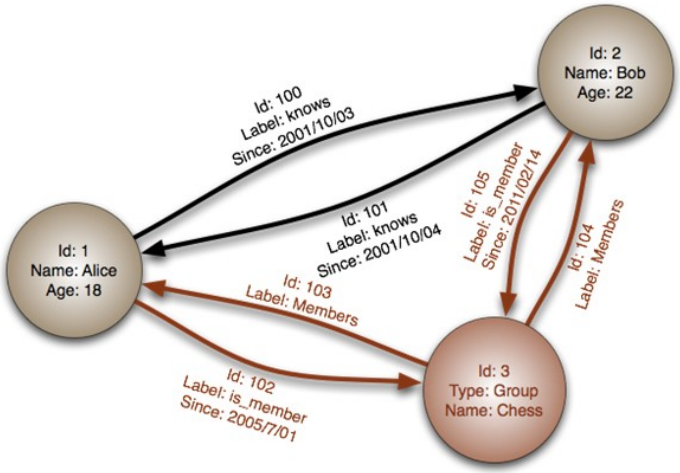
Misafir Katılımcı (Amazon)

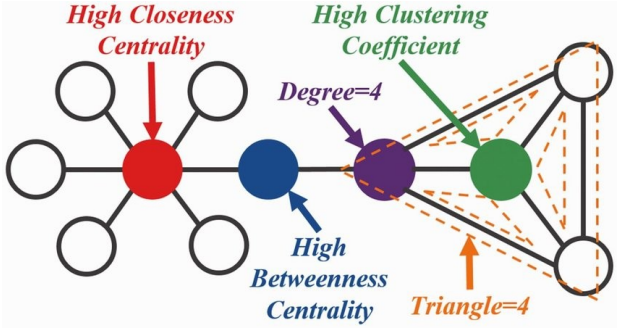
Hafta 11: Apache Spark, Spark ML, Akan Veri Analizi

- Apache Spark Nedir?
  - RDD, dönüşüm ve aksiyonlar, RDD sürekliliği
- Spark Uygulama Geliştirme, Shell'den çalıştırma
  - Demo 1: Scala, Java (wordcount)
- Spark Kütüphaneleri
  - Spark SQL, MLlib, Streaming, GraphX
  - Demo 2: Streaming uygulama (NetworkWordCount)
- Big Learning
  - Spark ML paketi, algoritmaları, pipeline yapısı,
  - Demo 3: Sınıflandırma (Naive Bayes)

Hafta 12: NoSQL Veritabanları, Ağ Analizi, Graf Veritabanları, Neo4j

- NoSQL Veritabanları çeşitleri
- Graf/çizge veritabanları nedir, nerelerde kullanılır?
- Bir çizge veritabanı: Neo4j
- Ağ analizi, çizge özellikleri
  - Betweennes, closeness centrality
  - Eigenvector centrality
  - Directed graphs and PageRank





Hafta 13: Yapısal olmayan veri analizi, metin analizi

- Yapısal olmayan veri: text, görüntü, video vs
  - Küçük –tweets
  - Orta –e-postalar, ürün yorumu dokümanları
  - Büyük–belgeler
  - Çok büyük –kitaplar, corpus vs.
- Sorgulamaya yönelik: “Metin analitiği”, “Metin madenciliği”
- Anlamaya yönelik: “Doğal dil işleme/anlama”
  - Arama motorları, Spam sınıflandırma, Doküman özetleme, Dil çevirileri, Sesten metne veya metinden sese dönüşüm vs.
- Görüntü analizi
  - Spesifikasyonlara göre görüntüleri sıralama (ranking), görüntü geri getirimi
  - Sınıflandırma, etiketleme vs.
- Video = görüntü dizisi + ses akışı

Hafta 14: Evrişimsel Sinir Ağları ve Tensor Flow

Kütüphane	Geliştirici	Programlandığı Dil	Özellikleri
TensorFlow	Google	Python	• Hızlı derleme yapabilmektedir. • TensorBoard ile görselleştirme yapabilmektedir • Veri ve model paralelliği sağlar. • GPU veya CPU'da paralel çalışabilmektedir.
Caffe	Berkeley Vision and Learning Center (BVLC)	Python	• İleri beslemeli ağlar ve görüntü işleme konularında hızlıdır. • Hassas ayarlama (finetuning) için önceden eğitilmiş modelleri vardır. • Hiçbir kod yazmadan model eğitilebilir. • Python arayüzü oldukça kullanışlıdır. • GPU desteği vardır.
Caffe2	Facebook	Python	• Python API ile C++ desteği sağlar. • Berkeley yazılım dağıtım lisansı vardır. • GPU desteği vardır.
Torch/PyTorch		Lua/Python	• Birçok modüler parçayı birleştirmek kolaydır. • Yeni katmanları yazıp GPU üzerinde çalıştırması kolaydır. • Çokça önceden eğitilmiş model vardır.
Keras	Francois Chollet-Google	Python	• Torch kütüphanesinden esinlenilmiş sezgisel bir API'dir. • Theano, TensorFlow, Deeplearning4j ve CNTK arkaplanda kullanmaktadır. • Hızlı büyüyen bir yapısı vardır. • GPU veya CPU'da paralel çalışabilmektedir.
MxNet	Pedro Domingos, Amazon AWS	R, Python ve Julia	• Çoklu GPU desteği vardır. • Eğitim hızı ve verimliliği yüksektir. • Yeni katmanlar eklemek kolaydır.
CNTK	Microsoft	C++	Geri planda Python API kullanımına izin verir.
KNet	Deniz Yüret, Koç Üniversitesi	Julia	• Kolay anlaşılır olması ve ifade gücü yüksektir. • GPU Desteği vardır.
Theano	Montreal Institute for Learning Algorithms (MILA) Lab.	Python	Yoshua Bengio tarafından Eylül 2017'de resmi olarak Theano kütüphanesnin geliştirilmeye devam edilmeyeceği duyuruldu.