Büyük Veri Analizine Giriş (BLM442)
Süleyman Eken, Dr., Kocaeli Üniversitesi, Bilgisayar Mühendisliği

Ders programı: Cuma saat 10:00-13:00 (I. Ogr), Cuma saat 17:00-20:00 (II. Ogr)

Tanıtım: İnternetin yaygınlaşması ve bilhassa sosyal medyanın gündelik hayatımızın her alanına temas etmesiyle beraber yapısal olmayan veride artış oldu: web kayıtları, videolar, konuşma kayıtları, resimler, e-postalar, Tweetler vb. Bunların yanısıra güvenilir ve ucuz bir şekilde bu büyük miktardaki veriyi depoloma, verimli bir şekilde analiz etme ve nihayetinde anlamlı bilgi çıkarabilme kabiliyetine sahibiz. Bu dersin temel amacı bu büyük verinin depolanması, manipüle edilmesi ve analiz edilmesinde kullanılan yöntem ve teknolojileri öğrencilere tanıtmaktır.

Ön koşullar: Orta seviye/üzeri Python veya Java bilgisinin olması önerilir. Keşif aktiviteleri Python, Scala, Java veya R'da yapılmalıdır. Farklı açık kaynak teknolojilerin yüklenmesi açısından Linux işletim sistemi tercih edilebilir. Hesaplama gücü yüksek bilgisayarlarla çalışmak daha hızlı sonuçlar almanız açısından iyi olacaktır.

Ders notları: https://suleymaneken.github.io/teaching/ sayfasından konu işlenmeden önce yayınlanacaktır. Derste not tutmakta zorlananların sunum handout'larını çıktı olarak derste bulundurmalarında fayda vardır. Ders notu haricinde ekstra kullanılacak diğer eğitim materyalleri sizlere dağıtılacaktır.

Notlandırma: Keşif aktiviteleri (KA), genellikle her hafta işlenen konulara paralel olarak verilecek öğrendiğiniz konular hakkında çeşitli keşifler/öğrenimler sağlayacağınız uygulamalardır. Verildikten sonra 1 hafta içinde teslim edilmelidir, aksi takdirde alınmayacaktır. Karşılıklı Geri Bildirimler (KGB) ise geliştirecek olduğunuz projelerin, rastgele seçilen bir arkadaşınız tarafından belirli kriterlere göre değerlendirilmesi ve geri bildirim yapılması anlamına gelir. Geri bildirim alan proje bir hafta sonra geri bildirim checklist'i ile beraber nihai notlandırma için teslim edilir.

Ders notu oranları şu şekildedir: %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.

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.

Haberleşme: Dersle ilgili tüm soru ve cevaplar piazza (https://piazza.com/) üzerinden olacaktır. Piazza kodu dersi alanlar için daha sonra paylaşılacaktır.

Kaynaklar:

- 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

Konular listesi

Hafta	Konular	Okuma parçaları	Dokümanlar/pdfs
1 (22 subat)	Ders oryantasyonu Büyük veriye giriş	Link1, Link2, Link3, Link4	
2	Elektronik Tablolar (Spreadsheets) Kullanarak Veri Analizi ve Görselleştirme	Link1, Link2, Link3, Link4 Görsellestirme Hataları	
2	**Keşif Aktivitesi (KA) 1:** Verilecek spreadsheet üzerinde birtakım isterler gerçeklenecek		
2	Tableau kullanarak gelişmiş görselleştirme	Tableau Referanslar	
	Karşılıklı Geri Bildirim (KGB)1: Seçilecek olan bir spreadsheet üzerinde veri analizi ve görselleştirme		
3	İlişkisel veritabanları ve SQL	Link1, Link2 Project Jupyter home page	
3	İleri SQL		
3	**KA 2:** SQL Uygulamaları		
4	Pyhton'a giriş, built-in veri yapıları, built-in fonksiyonlar	Link1, Link2, Link3, Link4	
4	**KA 3:** Python temeller ve veri yapıları üzerine		
5	Python veri analizi ve görselleştirme	Pandas intro Pyplot intro	
5	**KA 4:** pandas & plotlib		
6	Lineer Cebir, İstatistik, Olasık Temeller		
7	Makine Öğrenmesi - Regression	Link1, Link2, Link3, Link4	
7	Makine Öğrenmesi – Sınıflandırma ve Kümeleme	Link1, Link2, Link3, Link4	
8	Python Kullanarak Makine Öğrenmesi	Link1, Link2	
8	**KA 5:** scikit-learn		
	KGB 2: scikit-learn uygulaması		

13-21 Nisan	Ara sınav Haftası	
9	Apache Hadoop Tasarım Kalıpları	Link
9	**KA 6:** Tasarım kalıbı uygulaması	
10	**Misafir Katılımcı (Amazon):** Amazon Big Data Platforms and Services	
11	Apache Spark, Spark ML, Akan Veri Analizi	Web, Awesomes
11	**KA 7:** Akan veri üzerine uygulama	
12	NoSQL Veritabanları, Ağ Analizi, Graf Veritabanları, Neo4j	Awesomes
12	**KA 8:** Neo4j uygulama	Neo4j
13	Yapısal olmayan veri analizi, metin analizi	Link1, Link2, Link3, Link4
14	Evrişimsel Sinir Ağları ve Tensor Flow	Awesome-deep-learning Link1, Link2, Link3
10-18 Haziran **Final Sınavı**		

Akla gelebilecek sorular ve yanıtları:

1- Bu kadar ders aktivitesi çok değil mi?

Öğrenilen konularla ilgili daha çok keşif yapmak açısından konulmuştur. Harcayacağınız emek notlandırmada dikkate alınmıştır.

2- Ön koşullarda geçen "orta seviye/üzeri Python veya Java bilgisine" sahip değilsem sıkıntı olur mu?

En azından Python 3. haftdan itibaren yüzeysel olarak bahsedileceğinden çok sıkıntı olmaz; fakat aktiviteleri tam gerçekleyememe sorunu ile karşılaşabilirsiniz.

3- Akşam 20.00'den önce servisim var dersi almalı mıyım?

Gece dersini seçenlerin 20.00'ye kadar ders olacağını düşünerek seçmelerini istirham ederim.

4- Sınavda nasıl sorular gelebilir?

(i) Python kodu ekran çıktısı nedir?

```
def func(x):
        print(2*x)

func(5)
func(4)
```

(ii) RDD'in SparkContext kullanılarak nasıl oluşturalabileceğine dair birkaç örnek verin.
(iii) Lazy olarak değerlendirilen RDD'nin anlamı nedir?
(iv) Neo4j'deki düğümler, ilişkiler, özellikler ve etiketler gibi yapı taşlarının rolünü açıklayın.

5- Derste yoklama alınacak mı?

Devam koşulu

MADDE 18 – (1) Öğrenci, ilk kez kayıt yaptırdığı teorik derslerin en az %70'ine, diğer öğretim türlerinin de en az %80'ine devam etmek zorundadır. Laboratuvar ve uygulamalı derslerden başarısızlık durumunda devam şartı tekrar aranır.

(2) Öğrencilerin derse devamları, sorumlu öğretim elemanı tarafından yoklamalarla imza karşılığı tespit edilir. Devam durumu yarıyıl/yılın son haftasında ilgili öğretim elemanı tarafından öğrenci bilgi sisteminde öğrenciye ilan edilir.

http://odb.kocaeli.edu.tr/dosyalar/yonetmelikler/KOU_On_Lisans_ve_Lisans_Egitim_ve_Ogretim_Yonetmeligi-2016.pdf