Büyük Veri Analitiği

Bölüm 10

Dr. Büşra EMİR
İzmir Katip Çelebi Üniversitesi Tıp Fakültesi Biyoistatistik AD., busra.emir@ikcu.edu.tr

Bu Ünitede Neler Öğreneceksiniz?

Bu üniteyi tamamladıktan sonra;

- Veri işleme metodolojisi,
- Makine öğrenmesi ve derin öğrenme destekli analitik algoritmalar,
- Gerçek zamanlı veri filtreleme,
- Veri görselleştirmede kullanılan teknolojiler hakkında bilgi sahibi olacaksınız.
- Google Colaboratory bulut servisinde PySpark ile Spark MLlib kütüphanesi kullanılarak bir uygulama gerçekleştirebileceksiniz.

Hedefler

- Veri yapısına ve veri işleme yöntemine göre geliştirilen Apache Spark bileşenlerini tanımlamak
- Makine öğrenmesi ve derin öğrenme destekli analitik algoritmaları ve kullanım alanlarını belirlemek
- Programlama bilgisi gerektirmeyen görselleştirme araçlarını ve geliştiriciler için görselleştirme araçlarını belirleyebilmek.

veri görselleştirme aracı olan Tableu ve geliştiriciler için Seaborn, Matplotlib kütüphaneleri ücretsiz olarak en çok tercih edilen yazılımları oluşturmaktadır.

İleri okumalar için öneriler

- Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers Inc.
- Drabas, T., & Lee, D. (2017). *Learning PySpark*. Packt Publishing Ltd.
- Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. Springer open.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- Guller, M. (2015). *Big data analytics with Spark: A practitioner's guide to using Spark for large scale data analysis*. Apress.
- Marconi, K., & Lehmann, H. (Eds.). (2014). *Big data and health analytics*. Crc Press.
- Singh, P., & Singh. (2004). *Machine Learning with PySpark*. Apress.

KAYNAKLAR

1. Ohlhorst, F. J. (2012). *Big data analytics: turning big data into big money* (Vol. 65). John Wiley & Sons.
2. Cielen, D., & Meysman, A. (2016). *Introducing data science: big data,* machine learning, and more, using Python tools. Simon and Schuster.
3. Laney, D., & Kart, L. (2012). *Emerging role of the data scientist and the art of data science.* Gartner Group. White paper.
4. Assefi, M., Behravesh, E., Liu, G., & Tafti, A. P. (2017, December). Big data machine learning using apache spark MLlib. In 2017 ieee international conference on big data (big data) (pp. 3492-3498). *IEEE*.
5. Parashar,X.Li, and Chandra, S., *Advanced computational infrastructures for parallel and distributed applications*. John Wiley & Sons, 2010, vol. 66.
6. Talia, D. "Toward cloud-based big-data analytics," *IEEE* Computer Science, pp. 98–101, 2013.
7. Agrawal, D., Das, S and A. El Abbadi, "Big data and cloud computing: current state and future opportunities," in Proceedings of the 14th *International Conference on Extending Database Technology*. ACM, 2011, pp. 530–533.
8. Abbasi, S. Sarker, and Chiang, R. "Big data research in information systems: Toward an inclusive research agenda," Journal of the Association for Information Systems, vol. 17, no. 2, p. 3, 2016.
9. Archenaa, J. and Anita, E. M. "Interactive big data management in health- care using spark," *in Proceedings of the 3rd International Symposium on Big Data and Cloud Computing* Challenges (ISBCC–16). Springer, 2016, pp. 265–272.
10. Tafti, E. LaRose, E., Badger, J. C., Kleiman, R. and Peissig, P. "Machine learning-as-a-service and its application to medical informatics," *in International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2017, pp. 206–219.
11. Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C. and Iyengar, S. "Computational health informatics in the big data age: a survey," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 12, 2016.
12. Wiewiorka, M. S., Messina, A., Pacholewska, A., Maffioletti, S., Gawrysiak, P. and Okoniewski, M. J. "Sparkseq: fast, scalable, cloud- ready tool for the interactive genomic data analysis with nucleotide precision," *Bioinformatics*, p. btu343, 2014.
13. Masseroli, M., Pinoli, P. F., Venco, A., Kaitoua, Jalili, V., Palluzzi, F., Muller, H. and Ceri, S.

"Genometric query language: a novel approach to large-scale genomic data management," *Bioinformatics*, vol. 31, no. 12, pp. 1881–1888, 2015.
14. Ding, D., Wu, D., & Yu, F. (2016, August). An overview on cloud computing platform spark for Human Genome mining. In 2016 IEEE International Conference on Mechatronics and Automation (pp. 2605-2610). IEEE.
15. Ryan, J. (2016). Rapidminer for text analytic fundamentals. *Text Mining and Visualization: Case Studies Using Open-Source Tools*, *40*, 1.
16. Rong, C. (2011, July). Using Mahout for clustering Wikipedia's latest articles: A comparison between k-means and fuzzy c-means in the cloud. In 2011 *IEEE Third International Conference on Cloud Computing Technology and Science* (pp. 565-569). IEEE.
17. Tafti, A. P., Badger, J., LaRose, E., Shirzadi, E., Mahnke, A., Mayer, J., ... & Peissig, P. (2017). Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR medical informatics*, 5(4), e9170.
18. García-Pablos, A., Cuadros, M., & Rigau, G. (2015, June). V3: Unsupervised aspect based sentiment analysis for semeval2015 task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation* (SemEval 2015) (pp. 714-718).
19. Demirezen, M. U. (2020), Büyük Veri ve Büyük Veri İşleme Mimarileri, Şeref Sağıroğlu, Mustafa Umut Demirezen (Ed.), Yapay Zekâ ve Büyük Veri: Teknolojiler, Yaklaşımlar ve Uygulamalar, Seri 1, (s. 61-104), Ankara: Nobel Akademi.
20. Databricks, "What is Apache SparkTM?", 2021, https://databricks.com/spark/about/, (28.08.2021).
21. Penchikala, S. (2018). Big data processing with apache spark. Lulu. com.
22. Apache Spark Ekosistemi, http://itechseeker.com/tutorials/apache-spark/thanh-phan-cua-apache-spark/ (27.08.2021)
23. Penchikala S., "Big Data Processing with Apache Spark – Part 1: Introduction". https://www.infoq.com/articles/apache-spark-introduction, (27.08.2021).
24. Apache Spark – Tutorial, "Apache Spark – Introduction", 2021, https://www.tutorialspoint.com/apache_spark/apache_spark_introduction.htm, (01.02.2021).
25. Alpaydın, E. (2011). *Yapay öğrenme*. Boğaziçi Üniversitesi Yayınevi.
26. Somani, A. K., & Deka, G. C. (Eds.). (2017). Big data analytics: Tools and technology for effective planning. CRC Press.
27. Chollet, F. 2017. Deep *learning with Python*. Manning Publications Company.
28. Goodfellow, Ian, Yoshua Bengio, ve Aaron Courville. 2016. "*Deep Learning*".
29. Sewak, M., Karim, M. R., & Pujari, P. (2018). *Practical convolutional neural networks: implement advanced deep learning models using Python*. Packt Publishing Ltd.
30. Kinsley, H.& Kukiela, D. , Neural Network From Scratch in Python, 2021.
31. X. Meng, J. Bradley, B. Yuvaz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen et al., "Mllib: Machine learning in apache spark," JMLR, vol. 17, no. 34, pp. 1–7, 2016.
32. Anıl, U. & Akcayol, M. A. (2019). Akan Veri Karakterizasyonu, Üretimi Ve Analitiği Üzerine Kapsamlı Bir İnceleme. *Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 12(1), 379-410.
33. Aggarwal, C.C. 2007. "Data streams: models and algorithms", *Springer Science & Business Media*.
34. Leskovec, J., Rajaraman, A., Ullman, J.D. 2014. "Mining of massive datasets",*Cambridge University Press*.
35. Friendly, M. (2008). A brief history of data visualization. In Handbook of data visualization (pp. 15-56). Springer, Berlin, Heidelberg.
36. Anıl, U & Akcayol, M. A. (2018). Tavsiye sistemlerinde büyük verinin kullanımı üzerine kapsamlı bir inceleme. *Marmara Fen Bilimleri Dergisi*, 30(4), 339-357.
37. Wang, L., Wang, G., & Alexander, C. A. (2015). Big data and visualization: methods, challenges and technology progress. Digital Technologies, 1(1), 33-38.

38. Fox, P., & Hendler, J. (2011). Changing the equation on scientific data visualization. Science, 331(6018), 705-708.
39. Buzan, T., & Griffiths, C. (2013). *Mind Maps for Business 2nd ed. Using the ultimate thinking tool to revolutionise how you work*. Pearson UK.
40. Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). A. J. Hey (Ed.). Redmond, WA: Microsoft research.
41. Caldarola, E. G., & Rinaldi, A. M. (2017). Big Data Visualization Tools: A Survey-The New Paradigms, Methodologies and Tools for Large Data Sets Visualization in DATA (pp. 296-305).
42. Cebeci, H. I. (2017). Tableu Eğitim Dökümanları, Sakarya Universitesi.
43. Eken, S. (2020). Büyük Verinin İnteraktif Görselleştirilmesi: Tableau Üzerine Öğrenci Deneyimleri. *Avrupa Bilim ve Teknoloji Dergisi*, (18), 262-271. DOI: 10.31590/ejosat.659823