PYTHON İLE METİN MADENCİLİĞİ VE DOĞAL DİL İŞLEME

Yazar
Yılmaz KAYA

ISBN
978-625-375-086-2

Kitap Adı
Python ile Metin Madenciliği ve Doğal Dil İşleme

Yazar
Yılmaz KAYA
ORCID iD: 0000-0001-5167-1101

Yayın Koordinatörü
Yasin DİLMEN

Sayfa ve Kapak Tasarımı
Akademisyen Dizgi Ünitesi

Yayıncı Sertifika No
47518

Baskı ve Cilt
Vadi Matbaacılık

Bisac Code
EDU039000

DOI
10.37609/akya.3298

Kütüphane Kimlik Kartı
Kaya, Yılmaz.
Python ile Metin Madenciliği ve Doğal Dil İşleme / Yılmaz Kaya.
Ankara : Akademisyen Yayınevi Kitabevi, 2024.
287 s. : şekil. ; 160x235 mm.
Kaynakça var.
ISBN 9786253750862

GENEL DAĞITIM
Akademisyen Kitabevi A.Ş.

Halk Sokak 5 / A Yenişehir / Ankara
Tel: 0312 431 16 33
siparis@akademisyen.com

www.akademisyen.com

ÖNSÖZ

Bu kitap, veri işleme ve analiz dünyasında son yıllarda giderek önem kazanan iki temel konuya odaklanmaktadır: Metin Madenciliği ve Doğal Dil İşleme (DDİ). Her gün dijital dünyada milyonlarca yapılandırılmamış veri üretilirken, bu verilerden anlamlı sonuçlar çıkarma ihtiyacı hem akademik araştırmalar hem de ticari uygulamalar için giderek daha kritik bir hale gelmiştir. **Python ile Metin Madenciliği ve Doğal Dil İşleme** adlı bu kitap, Python programlama dilinin güçlü araçlarıyla bu zorlu sürecin nasıl yönetilebileceğine dair kapsamlı bir rehber sunmaktadır.

Kitabın ilk bölümleri, Python'un temel programlama prensiplerine ayrılmıştır. Python'da değişkenler, veri türleri, döngüler, kontrol ifadeleri ve nesneye yönelik programlama (OOP) gibi konular, okuyuculara metin madenciliği ve DDİ uygulamaları için sağlam bir temel kazandırmaktadır.

Metin madenciliği ve DDİ'nin temelleri üzerine yoğunlaşan bölümlerde, sosyal medya gönderilerinden araştırma makalelerine kadar geniş bir veri yelpazesinde kullanılan yöntemler ele alınmaktadır. Bu bağlamda, Varlık İsmi Tanıma (NER), metin özetleme, tokenizasyon, metin sınıflandırma ve metin normalizasyonu gibi tekniklerin nasıl uygulanacağı Python kütüphaneleriyle detaylı bir şekilde anlatılmaktadır. NLTK, spaCy, Gensim ve Hugging Face Transformers gibi popüler kütüphanelerin gücünden yararlanılarak, pratik örnekler üzerinden bu süreçlerin nasıl hayata geçirileceği açıklanmaktadır.

Kitap ayrıca, öznitelik çıkarımı ve metin görselleştirme gibi daha ileri düzey konulara da geniş yer ayırmaktadır. Metin verilerinin sayısal temsillerini oluşturmak için kullanılan Kelime Çantası (Bag of Words), N-gram, TF-IDF gibi tekniklerin yanı sıra, daha derinlemesine analizler için 1D-Yerel İkili Örüntüler (1D-LBP) ve motif örüntüleri gibi yöntemler incelenmektedir. Görselleştirme teknikleriyle de metinlerin gizli kalmış anlamları, kelime bulutları ve ağ grafikleri gibi yöntemlerle görsel hale getirilmektedir.

Makine öğrenmesi ve derin öğrenme bölümleri ise sınıflandırma uygulamaları üzerine odaklanmaktadır. Random Forest, Naive Bayes ve K-Nearest Neighbors gibi klasik makine öğrenmesi algoritmalarının yanı sıra, LSTM, GRU ve

1D-CNN gibi derin öğrenme modelleriyle metin sınıflandırma problemlerinin nasıl çözüleceği uygulamalı örneklerle anlatılmaktadır.

Son olarak, günümüzde doğal dil işleme teknolojilerinde devrim yaratan **Büyük Dil Modelleri (LLMs)** üzerinde durulmuştur. GPT, BERT ve Transformer mimarisi gibi modellerin Python ile nasıl entegre edilerek metin üretimi, metin tamamlama, soru-cevap sistemleri ve dil çevirisi gibi görevlerde nasıl kullanılabileceği örneklerle açıklanmıştır.

Bu kitap, hem teorik bilgileri hem de pratik uygulamaları bir araya getirerek, Python ile metin madenciliği ve doğal dil işleme projeleri geliştirmek isteyen okuyucular için yol gösterici bir kaynak olacaktır. Faydalı olması dileğiyle...

Doç. Dr. Yılmaz KAYA

TEŞEKKÜR

Bu kitabın yazılmasında her türlü anlayışı gösteren sevgili eşime, çocuklarıma, beni bu günlere getiren annem ve babama gösterdikleri her türlü sabır, anlayış ve destekten dolayı sonsuz teşekkür ederim.

Ayrıca bu kitabın hazırlanmasında desteklerini eksik etmeyen Akademsiyen Yayınevi çalışanlarına çok teşekkür ederim.

Doç. Dr. Yılmaz KAYA
Batman Üniversitesi

İÇİNDEKİLER

BÖLÜM 1 **PYTHON PROGRAMLAMA TEMELLERİ 1**

1.1. Python Nedir? 1
1.2. Python Temelleri 2
1.3. Operatörler 2
1.4. Koşullu İfadeler (if-else) 3
1.5. Tekrarlayan İşlemler: Döngüler 5
1.7. Hata Yönetimi ve Hataların Denetimi 10
1.8. Fonksiyonlar ve Kullanımı 13
1.11. Hazır Fonksiyonlar 18
1.12. Dosyalama İşlemleri 30
1.13. Modüller ve Paketler 32
1.14. Nesneye Yönelik Programlama (Object-Oriented Programming - OOP) 33

BÖLÜM 2 **TEMEL KAVRAMLAR 39**

2.1. Doğal Dil İşleme (DDİ) 39
2.2. Metin Madenciliği 40
2.3. Makine Öğrenmesi (Machine Learning, ML) 41
2.4. Dilbilim 41
2.5. Sözdizimsel (sentaktik) analiz 43
2.6. Anlambilimsel (semantik) analiz 44
2.7. Kelimeler 46
2.8. Kök Bulma (Stemming) 47
2.9. Lemmatizasyon (Lemmatization) 49
2.10. DDİ Uygulama Alanları 53
2.11. Önemsiz Kelimeler (StopWords) 54
2.12. Öznitelikler (Features) ve Öznitelik Çıkarma Nedir? 56
2.13. Gözetimli Öğrenme (Supervised Learning) 57
2.14. Gözetimsiz Öğrenme 59
2.15. Pekiştirmeli öğrenme (Reinforcement Learning) 60
2.16. Kümeleme (Clustering) 62
2.17. Sınıflandırma 67
2.18. Uzman Sistemler 70
2.19. Part of Speech (POS) 73
2.20. Vec2Word 75
2.21. DDİ ve Metin Madenciliği için Python Kütüphaneleri 77

BÖLÜM 3 **METİN MADENCİLİĞİ VE DOĞAL DİL İŞLEME: TEMEL TEKNİKLER 83**
3.1. Varlık İsmi Tanıma (Named Entity Recognition - NER) 83
3.2. Metin Normalizasyonu (Text Normalization) 85
3.3. Tokenize İşlemi 89
3.4. Metin Sınıflandırma 92
3.5. Metin Özetleme 96

BÖLÜM 4 **ÖZNİTELİK ÇIKARIM YAKLAŞIMLARI 99**
4.1. Kelime Çantası (Bag of Words, BOW) 99
4.2. TF-IDF (Term Frequency - Inverse Document Frequency) 102
4.3. One-Hot Encoding Yaklaşımı 105
4.4. Eş oluşum matrisleri (Co-occurrence Matrices) 108
4.5. N-Gram 112
4.6. Açı Örüntüler 114
4.7. Bir Boyutlu Yerel İkili Örüntüler 116
4.8. Motif Örüntüler 118

BÖLÜM 5 **METİN GÖSELLEŞTİRME 121**
5.1. Kelime Bulutu (Word Cloud), ... 121
5.2. Bar Chart (Çubuk Grafiği) 126
5.3. Heatmap (Isı Haritası) 128
5.4. Topic Modeling (Konu Modellemesi Görselleştirmesi) 130
5.5. Network Graphs (Ağ Grafikleri): 133
5.6. Word Tree (Kelime Ağacı): 135
5.7. N-Gram Analiz ve Görselleştirme 139

BÖLÜM 6 **MAKİNE ÖĞRENMESİ VE METİN MADENCİLİĞİ SINIFLANDIRMA UYGULAMALARI 143**
6.1. Veri Setleri 143
6.2. N-Gram ile Duygu Tespiti 145
6.3. Kelime Çantası (BOW) ile Duygu Tespiti 148
6.4. TF-IDF Öznitelikler ile Duygu Tespiti 151
6.5. One-Hot Encoding Yaklaşımı ile Spam E-Posta Tespiti 153
6.6. Eş oluşum matrisleri (Co-occurrence Matrices) ile Duygu Tespiti 156
6.7. Açı Örüntüler ile Spam E-Posta Tespiti 159
6.8. 1B-YİÖ ile Spam E-Posta Tespiti 163
6.9. Motif Örüntüler ile Spam E-Posta Tespiti 168

BÖLÜM 7 **KONU MODELLEME (TOPİK MODELLEME) 175**
7.1. Latent Dirichlet Allocation (LDA) 175

7.2. Latent Semantic Analysis (LSA) 178
7.3. Non-Negative Matrix Factorization (NMF) 180
7.4. Hierarchical Dirichlet Process (HDP) 183
7.5. Correlated Topic Model (CTM) 186
7.6. Biterm Topic Model (BTM) 189
7.7. BERTopic ile Topik Modelleme 192

BÖLÜM 8 **DERİN ÖĞRENME METOTLARI İLE METİN SINIFLANDIRMA 197**
8.1. LSTM (Long Short-Term Memory) 198
8.2. GRU (Gated Recurrent Unit) 200
8.4. Arıza Veri Seti 205
8.5. LSTM ile Metin Sınıflandırma 205
8.6. GRU (Gated Recurrent Unit) ile Metin Sınıflandırma 209
8.7. 1D-CNN (Bir Boyutlu Konvolüsyonel Sinir Ağı) ile Metin Sınıflandırma 212

BÖLÜM 9 **EŞ OLUŞUM AĞLARI İLE METİN ANALİZİ 217**
9.1. Eş Oluşum Ağları Nedir? 218
9.2. Eş Oluşum Ağları ve Metin Madenciliği 222
9.3. Eş Oluşum Ağları için Veri Seti 223
9.4. Kelime Bazlı Eş Oluşum ağlarının Oluşturulması 227

BÖLÜM 10 **ANAHTAR KELİME ÇIKARIM YAKLAŞIMLARI 231**
10.1. RAKE (Rapid Automatic Keyword Extraction) 231
10.2. TextRank Yaklaşımı 235
10.3. YAKE! (Yet Another Keyword Extractor) Yaklaşımı 237
10.4. TF-IDF Tabanlı Anahtar Kelime Çıkarımı 240
10.5. KPMiner Yaklaşımı 243
10.6. Multipartite Rank Algoritması 247

BÖLÜM 11 **BÜYÜK DİL MODELLERİ VE PYTHON UYGULAMALARI 251**
11.1. Büyük Dil Modelleri Nedir? 251
11.2. Büyük Dil Modellerinin Gelişim Süreci 252
11.3. Doğal Dil İşleme (NLP) ve Büyük Dil Modelleri 252
11.4. Dil Modellerinin Tarihçesi 253
11.5. İstatistiksel Dil Modelleri ve Derin Öğrenme Tabanlı Modeller 253
11.6. Transformer Mimarisine Giriş (BERT, GPT, vs.) 254
11.7. Dönüştürücülerin (Transformers) Çalışma Prensipleri 254
11.8. Büyük Dil Modellerinin Uygulamaları 258
11.9. Büyük Dil Modellerinin Python Uygulamaları 261

KAYNAKLAR

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Hofmann, M., & Chisholm, A. (Eds.). (2016). *Text mining and visualization: Case studies using open-source tools*. CRC Press.

Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. " O'Reilly Media, Inc.".

Sarkar, D. (2016). *Text analytics with python* (Vol. 2). New York, NY, USA:: Apress.

Sarkar, D. (2019). *Text analytics with Python: a practitioner's guide to natural language processing* (pp. 1-674). Bangalore: Apress.

Sri, M. (2021). *Practical Natural Language Processing With Python: With Case Studies from Industries Using Text Data at Scale*. Apress.

Singh, J. (2023). *Natural Language Processing in the Real World: Text Processing, Analytics, and Classification*. Chapman and Hall/CRC.

Surdeanu, M., & Valenzuela-Escárcega, M. A. (2024). *Deep learning for natural language processing: a gentle introduction*. Cambridge University Press.