



Ana Sayfa Dersler Makine Öğrenimi

Konu Modelleme Nedir? Örneklerle Bir Giriş

Konu modellemeyle yapılandırılmış verilerden elde edilen içgörülerin kilidini açın. Temel kavramları, LSA ve LDA gibi teknikleri, pratik örnekleri ve daha fazlasını keşfedin.

İçindekiler Eki 2023 · 13 dakika okuma



Kurtis Pykes

Veri Bilimi ve Yapay Zeka Blogcusu | Yapay Zeka ve Veri Bilimi Alanı nda En İyi 1000 Orta Düzey Yazar

KONULAR

Makine öğrenme

Analitiklerin amacı verilerden içgörü elde etmektir. Geleneksel olarak bu tür veriler yapılandırılmıştı, yani verimli erişim için standartlaştırılmış bir formattadır. Dünya değiştiğinde ve dijitalleştiğinde, üretilen verilerin çoğu yapılandırılmamış durumda, yani önceden tanımlanmış bir veri modeli yok.

Gartner'a göre yapılandırılmış veriler, tüm yeni kurumsal verilerin %80-90'ını temsil ediyor. Üstelik yapılandırılmış verilere göre üç kat daha hızlı büyüyor. Sonuç olarak, analitik uzmanlarını yeni kümelerinden ilgili bilgileri elde etmek için yeni teknikler kullanması gerekiyor.

Kullanılan tekniklerden biri de metin madenciliği alanından konu modellemedir. Bu makalenin geri kalanında şunları ele alacağız:

- Konu modelleme nedir?
- Temel kavramlar
- İki popüler konu modelleme tekniği
- Uygulamalı bir örnek
- Konu modelleme kullanımı örnekleri
- Konu modellemenin diğer tekniklerden farkı

Konu Modelleme Nedir?

Konu modelleme, bir metin külliyyatı tarafından tasvir edilen gizli anlamsal kalıpları keşfetmek ve onun içinde var olan konuları otomatik olarak belirlemek için sıklıkla kullanılan bir yaklaşımdır.

Yani, [denetimsiz makine öğreniminden](#) yararlanan bir tür istatistiksel modellemedir bir metin gövdesi içindeki benzer sözcük kümelerini veya gruplarını analiz etmek ve tanımlamak.

Örneğin, bir belgenin içeriğinin bunun bir fatura, şikayet veya sözleşme olduğunu ima edip etmediğini belirlemek için bir konu modelleme algoritması kullanılabilir.

İş dünyasında konu modellemenin rolü

Bazı kaynaklara göre ortalama bir kişi saniyede 1,7 MB'tan fazla dijital veri üretiyor. Bu sayı, günde 2,5 kentilyon bayttan fazla veri anlamına geliyor ve bunların %80-90'ı yapılandırılmamış.

Bir işletmenin, her bir yapısal olmayan veri parçasını incelemek ve bunları temel konuya göre bölümlere ayırmak için tek bir kişiyi çalıştırdığı bir senaryo düşünün. Bu imkansız bir görev olurdu.

Tamamlanması önemli miktarda zaman alır ve son derece sıkıcı olur, ayrıca İnsanlar doğal olarak önyargılı olduğundan ve hataya daha yatkın olduğundan çok daha fazla risk söz konusudur. makineler.

Çözüm konu modellemedir.

Konu modellemeyle verilerden içgörüler daha hızlı ve muhtemelen daha iyi elde edilebilir. Bu teknik, konuları anlaşılabilir bir yapıya da birleştirerek işletmelerin ne olduğunu hızlıca anlayın.

Örneğin müşterilerin en büyük zorluklarını anlamak isteyen bir işletme, Bu bilgiyi yapılandırmamızı veriler aracılığıyla öğrenmek için konu modellemeyi kullanır.

Kısa sırasıyla konu modelleme işletmelere şu konularda yardımcı olur:

- Yapılandırılmış İmamı ş metinsel veriler üzerinde gerçek zamanlı analiz gerçekleştirme
- Yapılandırılmış İmamı ş verilerden geniş ölçekte bilgi edinme
- Biçimi ne olursa olsun, verilerle ilgili tutarlı bir anlayış oluşturun.

Konu Modellemenin Temel Kavramları

Veri profesyonellerinin hızlı bir şekilde analiz etmelerini ve tanımlamalarını sağlayan konu modellemeyi oluşturduk belirli bir ölçekte bir metin gövdesi içindeki benzer sözcük kümeleri veya grupları.

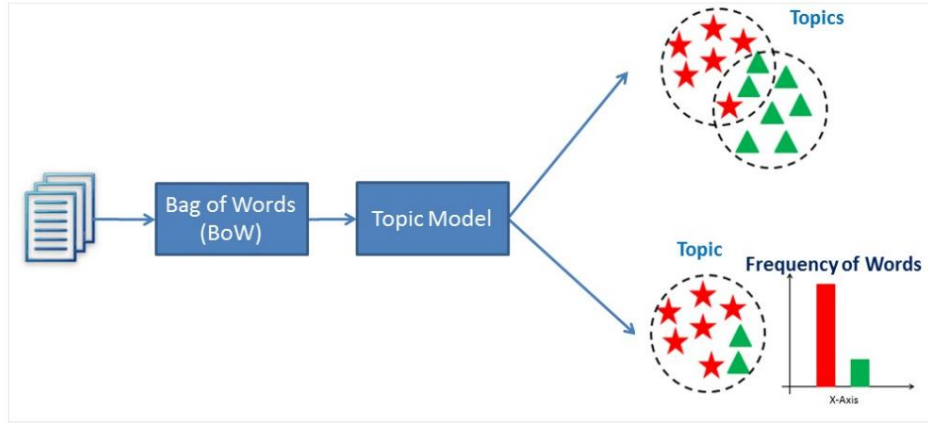
Peki konular nelerdir ve konu modelleme nasıl çalışır?

Konular nedir ve konu modelleri nasıl çalışır?

Konular, bir metin topluluğunun (büyük bir grup) gizli açılımalarıdır. Sezgisel olarak, belgeler Belirli bir konuyla ilgili olarak belirli kelimeleri daha sık üretme olasılıkları daha yüksektir.

Örneğin, "köpek" ve "kemik" kelimelerinin belgelerde görünme olasılığı daha yüksektir köpeklerle ilgili olarak ise "kedi" ve "miyav" kelimelerinin belgelerde bulunma olasılığı daha yüksektir kedilerle ilgili. Sonuç olarak, konu modeli belgeleri tarayacak ve üretecektir. benzer kelimelerin kümeleri.

Temel olarak konu modelleri, sözcükleri çıkararak ve benzer olanları konu başlıkları halinde gruplandırarak çalışır. konu kümeleri oluşturun.



Konu modellemenin nasıl çalıştığına dair görselleştirme

Konu Modelleme Tekniklerini Keşfetmek

İki popüler konu modelleme tekniği Gizli Semantik Analiz (LSA) ve Gizli Dirichlet Tahsisi (LDA). Amaçları, tarafından tasvir edilen gizli anlamsal kalıpları keşfetmedir. metin verileri aynıdır ancak bunu nasıl elde ettikleri farklıdır.

Gizli Anlamsal Analiz (LSA)

[Gizli Anlamsal Analiz \(LSA\)](#) doğal bir [dil işlemedir](#) analiz etmek için kullanılan teknik belgeler ve içerdikleri terimler arasındaki ilişkiler. Yöntem ilk olarak

1988 tarihli bir makalede tanımlanır ["Gizli Semantik Analizi Kullanmak"](#) ile [Erişimi İyileştirin](#) [Metinsel Bilgiler](#) ve bugün hala bir koleksiyondan yapılandırılmış İmamı ş veriler oluşturmak için kullanılıyor. yapılandırılmış İmamı ş metin.

Yani LSA, benzer anlamlara sahip kelimelerin benzer belgelerde görüneceğini varsayar. öyle bu nedenle, her satırın belge başına kelime sayısı içerdiği bir matris oluşturarak benzersiz bir kelimeyi temsil eder ve sütunlar her belgeyi temsil eder ve ardından Tekil bir sözcük kullanılarak. Benzerliği korurken satır sayısı azaltmak için Değer Ayrıştırma (SVD) Sütunlar arasındaki yapı. SVD, verileri korurken basitleştiren matematiksel bir yöntemdir. önemli özellikler. Burada kelimeler ve kelimeler arasındaki ilişkileri sürdürmek için kullanılır. belgeler.

Belgeler arasındaki benzerliği belirlemek için kosinüs benzerliği kullanılır. Bu bir ölçü iki vektör arasındaki açıyı kosinüsünü hesaplayan bu örnekte, belgeler. 1'e yakının bir değer, belgelerin içindeki kelimelere göre çok benzer olduğu anlamına gelir. 0'a yakının bir değer ise oldukça farklı oldukları anlamına gelir.

Gizli Dirichlet Tahsisi (LDA)

Gizli Dirichlet Tahsisi (LDA) ilk olarak 2000 yılında başlıklı bir makalede önerildi. “Çıkarım Çokluokuslu genotip verilerini kullanarak popülasyon yapısını belirlemesi. Makale ağırlıklı olarak odaklandı genetik farklılıklarla ilgilenen genetiğin bir alt alanı olan popülasyon genetiği hakkında Popülasyonların içinde ve arasında. Üç yıl sonra [Gizli Dirichlet Tahsisi](#) uygulandı makine öğrenme.

Makalenin yazarları tekniği şu şekilde tanımlamaktadır: “metin ve diğerleri için üretken bir model ayrı k verilerin toplanması. Dolayısıyla LDA doğal bir dil tekniği olarak tanımlanabilir. Bir belgenin ait olduğu konuları, içinde yer alan kelimelere göre tanımlamak için kullanılır.

Daha spesifik olarak, LDA bir Bayes ağıdır, yani üretken bir istatistiksel modeldir belgelerin konuları n belirlenmesine yardımcı olan sözcüklerden oluştuğunu varsayar. Böylece, Belgeler, belgedeki her bir kelimeye bir konu atayarak bir konu listesiyle eşlenir. farklı konular. Bu model, bir belgede geçen kelimelerin sırasını göz ardı eder ve onları bir torba kelime gibi.

LSA ve LDA

Gizli Semantik Analiz (LSA) ve Gizli Dirichlet Tahsisi (LDA) doğaldır bir koleksiyondan yapılandırılmış verileri oluşturmak için kullanılan dil işleme teknikleri yapılandırılmış metin.

Ancak LSA, boyutsallığı azaltmak için Tekil Değer Ayrıştırma (SVD) yararlanır. terim-belge matrisi ve benzer anlamlara sahip kelimelerin olduğu varsayımı na dayanmaktadır. benzer belgelerde görünecektir. Metnin daha düşük boyutlu bir temsili oluşturarak, model, kelimeler arasındaki temel ilişkileri yakalayarak ne kadar benzer olduğunu belirleyebilir. iki belge var.

Buna karşılık LDA, Bayesian çıkarımını kullanan üretken bir olasılıksal modeldir. Bir metin külliyyatındaki temel konular. Her belgenin bir kombinasyonu olduğunu varsayar az sayıda gizli konu vardır ve her kelime belirli bir konu tarafından üretilir.

Sonuçta LSA kelimeler arasındaki temel ilişkileri keşfetmeye çalışır. LDA, bir metin bütününe altı yatan konuları keşfetmeyi amaçlamaktadır. Her ikisi de olması na rağmen Metnin vektör temsili oluşturmak için kullanılan teknikler, farklı temeller oluşturur. varsayım lar.

Konu Modellemenin Pratik Uygulaması

Bu tekniklerin nasıl çalıştığını görelim. Bu [DataCamp Çalma Alanı](#) nı kullanarak birlikte takip etmek kod.

Veri Hazırlama

İhtiyacı mız olan ilk şey veridir.

Konu modelleme için kullandığımız veriler basitçe bir metin koleksiyonu olan derlem adı verilir.

İşte internetteki gerçekleri kullanarak oluşturduğum küçük bir derleme:

```
# Örnek belgeler oluşturma
doc_1 = "Dünyadaki suyun yüzde 96,5'i okyanusları mızdadır ve 71'i kaplar.

doc_2 = "Hayatını üçte biri uykuyla geçiriyor. Her gece 7-9 saat uyumak

doc_3 = "Yeni doğmuş bir bebeğin yüzde 78'i sudur. Yetişkinlerin ise yüzde 55-60'ı sudur. Su
```



```
doc_4 = "Bir öğrenci lisedeyken 264,4 saat uykusuz kaldı .
```

```
doc_5 = "Suyun her üç halini de deneyimliyoruz: katı buz, sıvı su ve gaz
```

```
# Derlem derlemi oluştur
= [doc_1, doc_2, doc_3, doc_4, doc_5]
```

[Kodu aç](#)[OpenAI](#)

Ön işleme

Bir sonraki adı m metni temizlemektir:

```
# Kod kaynağı : https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-to
```



```
dizeyi içe aktar nltk
nltk.download
('stopwords') nltk.download('wordnet')
nltk.download('omw-1.4') nltk.corpus'tan
engellenen kelimeleri içe aktar
nltk.stem.wordnet'ten içe aktar WordNetLemmatizer
```

```
# engellenen sözcükleri, noktalama işaretlerini kaldı rı n ve derlemi normalleştirin stop =
set(stopwords.words('english')) include = set(string.punctuation)
lemma = WordNetLemmatizer()
```

```
def temiz(belge):
    stop_free = " ".join([i for i in doc.lower().split() if i stop'ta değilse ])
    punc_free = "" .join ( ch hariç tutulmazsa stop_free'deki ch için ch ) normalleştirilmiş =
    " ".join(lemma.lemmatize(word) for word in punc_free.split())
    normalleştirilmiş dönüş
```

```
clean_corpus = [ derlemdeki belge için clean(doc).split() ]
```

[Kodu aç](#)[OpenAI](#)

Yukarı daki kodda şunları yapı yoruz:

1. Gerekli kütüphaneleri içe aktardı k ve engellenecek sözcükleri ve **wordnet'i** indirdik
2. İngilizce engellenen kelimeleri tanı mladı
3. Hariç tutmak istediğimiz noktalama işaretlerini örneklendirdik
4. Wordnet lemmatizer örneğini oluşturduk
5. Engellenen sözcükleri ve noktalama işaretlerini kaldı rmak ve sözcükleri lemmatize etmek için bir işlev oluşturuldu belgeler.
6. Derlemdeki her belgeye temizleme işlevi uygulandı .

Ancak bu yine de hazı r olduğumuz anlamı na gelmiyor.

Bu verileri bir LDA veya LSA modeline girdi olarak kullanabilmemiz için, bunun bir terim-belge matrisine dönüştürülmesi gerekir. Bir terim-belge matrisi, bir dizi belgenin ve bunları n içerdiği terimlerin yalrı zca matematiksel bir temsildir.


Her belgede her terimin geçtiği yerlerin sayısı lması ve ardı ndan analiz için kullanı labilecek bir değerler matrisi oluşturmak üzere sayı ları n normalleştirilmesiyle oluşturulur.

Bunu Python'da yapmak için **Gensim'den** yararlanacağız z. kütüphane.

```
gensim import corpora'dan

# Belge-terim matrisi sözlüğü oluşturma =
corpora.Dictionary(clean_corpus) doc_term_matrix = [dictionary.doc2bow(doc)
for doc in clean_corpus]
```



 Kodu aç OpenAI

Artık modellerimizi yerleştirebiliriz.

Modelleme

LSA'da kullanacağımız ilk model:

`gensim.models`'den `LsiModel`'i içe aktarıyoruz




```
# LSA modeli
lsa = LsiModel(doc_term_matrix, num_topics=3, id2word = sözlük)

# LSA modeli
print(lsa.print_topics(num_topics=3, num_words=3))

"""

[(0, '0,555*"su" + 0,489*"yüzde" + 0,239*"gezegen"), (1, '0,361*"uyku" +
0,215*"saat" + 0,215*"hareketsiz"), (2, '-0,562*"su" + 0,231*"yağmur" +
0,231*"gezegen")]

"""
```

 Kodu aç OpenAI

Bu, konuların (her satır) ayrı konu terimleri (terimler) ve ağırlıklarıyla birlikte çıktıyı verir.

LDA ile deneyelim:

`gensim.models`'den `LdaModel`'i içe aktarıyoruz




```
# LDA modeli
lda = LdaModel(doc_term_matrix, num_topics=3, id2word = sözlük)

# Sonuçlar
print(lda.print_topics(num_topics=3, num_words=3))

"""

[(0, '0,071*"su" + 0,025*"durum" + 0,025*"üç"), (1, '0,030*"hareketsiz"
+ 0,028*"saat" + 0,026*"uyuyor"), (2, '0,073*"yüzde" + 0,069*"su" +
0,031*"yağmur")]

"""
```

 Kodu aç OpenAI

Konu Modelleme Ne İçin Kullanılır?

Konu modelleme, manuel ve tekrarlanan görevleri ortadan kaldırarak süreçleri basit bir şekilde kolayca ve ucuz bir şekilde hızlandırabilir. İşte birkaç örnek:

Destek biletlerinin etiketlenmesi

Konu modelleme, müşteri hizmetleri personelinin, temel sorunları belirlemek ve tekrar tekrar meydana gelen sorunları belirlemek için destek sorgularını analiz etmesine yardımcı olmak için kullanılabilir. Bu verilere dayanarak, daha bilgilendirici self-servis içerik oluşturabilir veya müşterilere doğrudan yardımcı olabilir.

Müşteri deneyimini geliştirmek

Konuşmaları en uygun ekibe yönlendirilecek şekilde etiketlemek için konu modelleme kullanılabilir. Örneğin, "fiyatlandıırma", "abonelik", "yenileme" vb. kelimeleri içeren bir görüşme, destek için doğrudan muhasebe departmanına gönderilebilir.

Konu Modelleme ve Diğer Teknikler

Konu modelleme ve kümeleme

Konu modelleme, bir belge koleksiyonunda var olan gizli konuları keşfetmek için kullanılır.

Bu, belgelerde görünen kelime ve ifadelerdeki kalıpları tanımlamayı ve bunları benzerliklerine göre konulara göre gruplandırmayı içerir.

Buna karşılık kümeleme, benzer nesneleri benzerlik ölçüsüne göre gruplamak için kullanılan bir tekniktir. Bu tür yöntemler, benzer veri noktalarını bir arada gruplayarak verilerdeki kalıpları ve yapıyı keşfetmek için kullanılır.

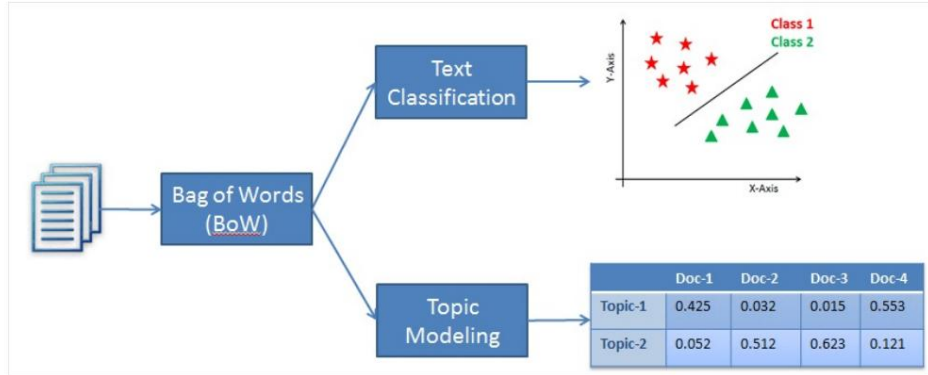
Her iki yaklaşım da metin verilerindeki kalıpları ortaya çıkarabilirse de farklı hedeflere sahiptir. Konu modelleme, bir belge koleksiyonundaki gizli konuları belirlenmesiyle ilgilenirken, kümeleme, benzer veri noktalarını bir arada gruplandırılmasıyla ilgilidir.

Konu modelleme ve metin sınıflandırması

Metin sınıflandırma, [doğal dil işleme](#) yoluyla teknik, denetimli öğrenme kategorisine girer. Yani, metin sınıflandırması önceden tanımlanmış kategorileri veya belirli bir metin parçasını etiketlemek için kullanılır. Modelin bu başarıya ulaşabilmesi için, yeni, görünmeyen metin örnekleri üzerinde tahminlerde bulunmadan önce, etiketlenmiş bir veri kümesinden öğrenmesi gerekiyor.

Öte yandan konu modelleme, bir metin belgeleri koleksiyonunda altta yatan konuları bulmak için kullanılan denetimsiz bir öğrenme tekniğidir. Bu, etiketli bir veri kümesinden öğrenmesinin gerekmediği anlamına gelir.

Bu nedenle, iki yöntem arasındaki fark, metin sınıflandırmanın metne önceden tanımlanmış etiketler atamak için kullanılması, oysa konu modellemenin bir belge koleksiyonundaki temel konuları keşfetmesidir.



Bir sınıflandırma örneği

konu modelleme

Çözüm

Konu modelleme, yapılandırılmış verilerden oluşan bir koleksiyondan yapılandırılmış veriler oluşturmak için kullanılan popüler bir doğal dil işleme tekniğidir. Başka bir deyişle bu teknik, işletmelerin bir metin kütüphanesi tarafından tasvir edilen gizli anlamsal kalıpları öğrenmesine ve içinde yer alan konuları otomatik olarak tanımlamasına olanak tanır.

İki popüler konu modelleme yaklaşımı LSA ve LDA'dır. Her ikisi de metin verilerindeki gizli kalıpları keşfetmeye çalışır ancak amaçları na ulaşmak için farklı varsayımlarda bulunurlar.

LSA, benzer anlamlara sahip kelimelerin benzer belgelerde görüneceğini varsayarken, LDA, belgelerin konuları nı belirlenmesine yardımcı olan kelimelerden oluştuğunu varsayar.

Bu derste konu modellemenin temel kavramlarını, pratik uygulamasını ve konu modellemenin metin sınıflandırma ve kümeleme gibi diğer tekniklerden nasıl farklı olduğunu ele aldık. Öğrenmenize devam etmek için diğer bazıları mızza göz atın

kaynaklar:

- [R'de Doğal Dil İşleme'ye Giriş](#)
- [R'de Metin Analizine Giriş](#)
- [Python kullanarak Gizli Anlamsal Analiz](#)



YAZAR

Kurtis Pykes



KONULAR

Makine öğrenme

Konu Modelleme Yolculuğunuza Bugün Başlayın!

KURS

Python'da Doğal Dil İşleme'ye Giriş

4 saat 111.7K

Python'u kullanarak temel doğal dil işleme tekniklerini ve bunları n gerçek dünyadaki metin verilerinden öngörüler elde etmek için nasıl uygulanacağı nı öğrenin.

Aynı niteliklere bakın →

Kursu Başlat

Daha fazla gör →

İlgili



MLOps Dünyası'nda Gezinmek
Sertifikaları

Adel Nehme



2024'te Makine Öğrenimi Nasıl Öğrenilir?

Adel Nehme



Yazar ve Şef Guy Kawasaki ile Dikkat
Çekici Olmak...

Richie Pamuk

Daha fazla gör →

DataCamp for Mobile ile veri becerilerinizi geliştirin

Mobil kursları mız ve günlük 5 dakikalık kodlama zorlukları mızla hareket halindeyken ilerleyin.



ÖĞRENMEK

Python'u öğrenin

R'yi öğrenin

Yapay Zekayı öğrenin

SQL öğrenin

Power BI'ı öğrenin

Tableau'yu öğrenin

Veri Mühendisliğini Öğrenin

Değerlendirmeler

Kariyer Yolları

Beceri Parçaları

Dersler

Veri Bilimi Yol Haritası

VERİ DERSLERİ

Python Kursları

R Kursları

SQL Kursları

Power BI Kursları

Tablo Kursları

Azure Kursları

Elektronik Tablo Kursları

Yapay Zeka Kursları

Veri Analizi Kursları

Veri Görselleştirme Kursları

Makine Öğrenimi Kursları

Veri Mühendisliği Kursları

Olası İlişkiler ve İstatistik Kursları

ÇALIŞMA ALANI

Başlamak

Şablonlar

Entegrasyonlar

Dokümantasyon

SERTİFİKA

Sertifikalar

Veri Bilimcisi

Veri Analisti

Veri Mühendisi

Veri Uzmanları'nı İşe Alın

KAYNAKLAR

Kaynak Merkezi

Yaklaşan Etkinlikler

Blog

Birlikte Kodlama

Öğreticiler

Açık kaynak

RBelgeler

Kurs Editörü

DataCamp for Business ile Demo Rezervasyonu Yapın

Veri Portföyü

Portföy Skor Tablosu

PLANLAR

Fiyatlandırma

İş için

Üniversiteler için

İndirimler, Promosyonlar ve Satışlar

DataCamp Bağlıları

DESTEK

Yardımcı Merkezi

İştirak Olun

HAKKINDA

Hakkımızda

Öğrenci Hikayeleri

Kariyer

Eğitmen Olun

Basmak

Liderlik

Bize Ulaştın

DataCamp İspanyolca



Gizlilik Politikası | Çerez Bildirimi | Kişisel Bilgilerimi Satma Erişilebilirlik Güvenlik Kullanım Koşulları

© 2024 DataCamp, Inc. Tüm Hakları Saklıdır.