Bölüm 16: NLP'de Ustalaşmak için Adım Adım Kılavuz - LSA Kullanarak Konu Modelleme

GELİŞMİŞ NLP METNİ

Bu makale Veri Bilimi Blogathon'unun bir parçası olarak yayınlandı

giriiş

Bu makale Doğal Dil İşleme (NLP) üzerine devam eden bir blog serisinin parçasıdır. Önceki makalede, Negatif Olmayan Matris Faktorizasyon adlı temel Konu Modelleme tekniğini tamamladık. Şimdi bu bölümün devamında Gizli Semantik Analiz adı verilen başka bir Konu modelleme tekniği üzerinde tartışmamıza başlayacağız.

Bu nedenle, bu makalede, Gizli Semantik Analiz (LSA) adı verilen Konu Modelleme tekniğine derinlemesine dalacağız ve bu tekniğin, herhangi bir NLP Sorunu ifadesi üzerinde çalışırken çok yararlı bir şey haline gelen bu gizli konuları nasıl ortaya çıkardığını göreceğiz.

Bu, Doğal Dil İşleme İçin Adım Adım Kılavuz hakkındaki blog serisinin 16. kısmıdır.

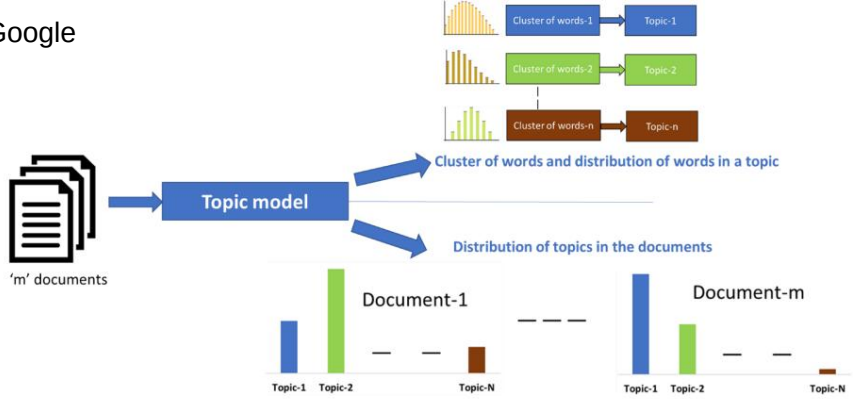
İçindekiler

Konu Modellemenin Özeti

Tüm konu modelleme algoritmalarının dayandığı temel varsayımlar:

- Her belge birden fazla konudan oluşur ve
- Her konu bir kelime koleksiyonundan oluşur.

Başka bir deyişle konu modelleme algoritmaları, belgemizin anlambiliminin aslında metin materyalini gördükten sonra doğrudan gözlemlemediğimiz bazı gizli veya "gizli" değişkenler tarafından yönetildiği fikri etrafında inşa edilmiştir.



Resim Kaynağı: Google Görseller

Sonuç olarak, belgemizin ve derlemimizin anlamını şekillendiren bu gizli değişkenleri ortaya çıkarmak için konu modelleme algoritmalarına ihtiyacımız var. Bu blog yazısının ilerleyen bölümlerinde, farklı konu modellerinin bu gizli konuları nasıl ortaya çıkardığına dair bir anlayış geliştireceğiz. Ancak bu makalede öncelikle LSA tekniğini tartışacağız ve daha sonra ilerledikçe LDA, pLSA vb. gibi farklı Konu modelleme tekniklerini de tartışacağız.

Neden Gizli Semantik Analize ihtiyacımız var?

İlk makaleden itibaren bahsettiğimiz gibi, tamamen doğal dillerin, bir makinenin yakalaması oldukça zor olan kendine has incelik ve nüanslara sahip olduğunu ve hatta bazen bunların biz insanlar tarafından bile yanlış anlaşıldığını tartışmıştık! Bu, aynı anlama gelen farklı sözcükleri ve ayrıca yazılışları aynı olan ancak farklı anlamlar veren sözcükleri de içerir.

Örneğin aşağıdaki iki cümleyi düşünün:

Cümleler: Premchand'ın son romanını oldukça beğendim. Yeni bir pazarlamaya geçmek istiyorlar
kampanya.

İlk cümlede 'roman' kelimesi bir kitabı temsil ederken, ikinci cümlede yeni veya taze anlamına gelir.

Biz bir insan olarak bu iki kelimenin ardındaki bağlamı anlayabildiğimiz için bu iki kelimeyi kolaylıkla ayırt edebiliriz. Ancak makineler kelimelerin kullanıldığı bağlamı anlayamadıkları için bu kavramı yakalayamayacaklardır. Gizli Semantiğin rolü buradadır

Analiz (LSA) devreye giriyor!

LSA, konu adı verilen gizli veya gizli kavramları yakalamak için kelimelerin etrafındaki bağlamdan yararlanmaya çalışır.

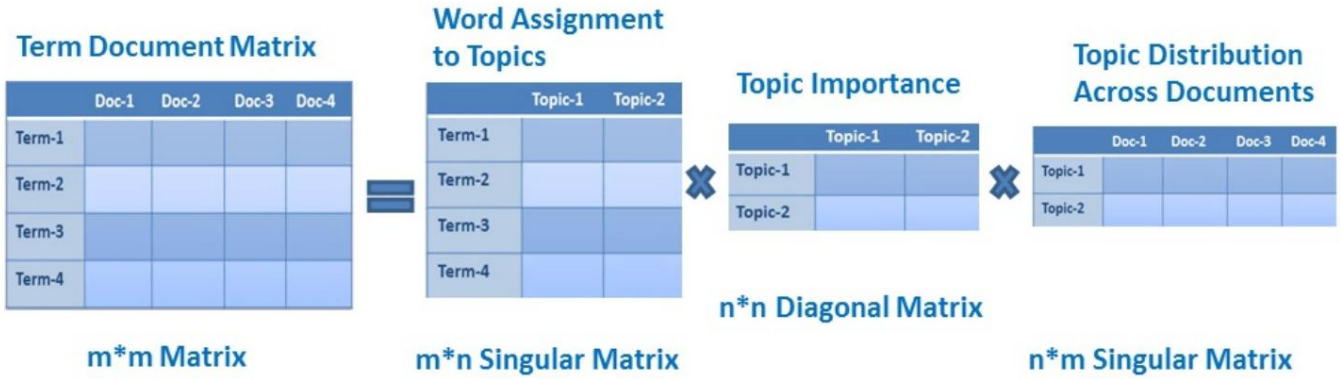
Yani, eğer kelimeleri basitçe belgelerle eşleştirirsek, bu bizim için gerçekten yararlı olmayacaktır. Yani asıl ihtiyacımız olan, kelimelerin ardındaki gizli kavramları veya konuları ortaya çıkarmaktır. LSA, bu makalede tartışacağımız bu gizli konuları bulmak için kullanılabilecek tekniklerden biridir.

Gizli Semantik Analiz Nedir?

Gizli Semantik Analiz anlamına gelen LSA, konu modellemede kullanılan temel tekniklerden biridir. Temel fikir, belgelerden ve terimlerden oluşan bir matris almak ve onu iki ayrı matrise ayırmaya çalışmaktır:

- Belge-konu matrisi
- Bir konu-terim matrisi.

Bu nedenle, gizli konular için LSA'nın öğrenilmesi, Tekil değer ayrıştırması kullanılarak belge-terim matrisi üzerinde matris ayrıştırılmasını içerir. Tipik olarak boyut küçültme veya gürültü azaltma tekniği olarak kullanılır.



Resim Kaynağı: Google Görseller

LSA Uygulanırken İlgili Adımlar

Diyelim ki, toplam n sayıda benzersiz kelimeden oluşan m sayıda metin belgemiz var, yani kelime dağarcığınızdaki kelime sayısı. Amacımız belgelerdeki tüm metin verilerinden k konu çıkarmaktır. Kullanıcının konu sayısını (k) belirtmesi gerekir.

Aşama 1

İlk adım , her satırın bir belgeyi temsil ettiği ve her sütunun bazı puanlara sahip bir kelimeyi temsil ettiği mxn şeklinde bir belge-terim matrisi oluşturmaktır .

Resim Kaynağı: Google Görseller

Puanlar nasıl hesaplanabilir?

Puanları hesaplamak için aşağıdaki gibi çeşitli yöntemler kullanabiliriz:

1. Her sütun girişini, j'inci kelimenin i'inci belgede kaç kez göründüğünün ham sayımıyla doldurabiliriz. Ancak pratikte ham sayımlar pek işe yaramıyor çünkü belgedeki her kelimenin önemi dikkate alınmıyor.

2. Tipik olarak LSA modelleri, belge terim matrisindeki ham sayımları bir TF-IDF puanıyla değiştirir.

TF-IDF veya Terim frekansı-Ters belge sıklığı, aşağıdaki formül yardımıyla i belgesindeki j terimine bir ağırlık atar:

Resim Kaynağı: Google Görseller

Sezgisel olarak, yukarıdaki formülden, bir terimin belgede sık sık ama bütün metinde nadiren geçmesi durumunda büyük bir ağırlık verilmiştir.

Adım 2

Belge-terim matrisi A'yı hesapladıktan sonra, belgelerde veya derlemde mevcut olan gizli veya gizli konularımız hakkında düşünmeye başlayabiliriz. Burada belge terimi matrisinin olduğunu gözlemleyebiliriz

- A çok seyrektir,
- A gürültülüdür,
- A, birçok boyutunda gereksizdir.

Sonuç olarak, kelimeler ve belgeler arasındaki ilişkileri yakalayan birkaç gizli konuyu bulmak için belge terim matrisi A'da boyut azaltma işlemi yapmak istiyoruz.

Daha sonra yukarıdaki belge terim matrisinin boyutlarını istenilen konu sayısını belirten k boyutuna indirgememiz gerekir.

Boyutsallık Azaltma işlemini nasıl yapabiliriz?

Bu boyutluluk azaltımı, kesik Tekil Değer ayrıştırması kullanılarak yapılabilir.

Tekil Değer Ayrışımı (SVD)

SVD'nin amacı en değerli bilgiyi bulmak ve aynı şeyi temsil etmek için daha düşük t boyutunu kullanmaktır. Tekil Değer Ayrışımı, herhangi bir A matrisini 3 ayrı matrisin çarpımına ayıran bir doğrusal cebir tekniğidir. Sinyal işleme, psikoloji, sosyoloji, iklim ve atmosfer bilimi, istatistik ve astronomi gibi birçok alanda birçok yararlı uygulamaya sahiptir. Matrisin üç farklı matrise ayrıştırılması:

- Ortogonal sütun matrisi, (V)
- Ortogonal satır matrisi, (U)
- Bir tekil matris. (S)

M=U*S*V

burada S, A'nın tekil değerleri olarak köşegen elemanlara sahip bir köşegen matristir.

Kritik olarak, kesik SVD'nin yaptığı şey, yalnızca en büyük t tekil değerleri seçerek ve U ve V matrisinin yalnızca ilk t sütunlarını tutarak boyutluluğu azaltmaktır. Burada t, seçebileceğimiz ve sayısını yansıtacak şekilde ayarlayabileceğimiz bir hiperparametredir. çıkarmak istediğimiz konular.

Sezgisel olarak bunu, dönüşmüş halimizdeki yalnızca en önemli boyutları korumak olarak düşünebiliriz.
uzay.

Bu durumda U ℝ(mk) matrisi belge konu matrisimiz olur ve V ℝ(nk) matrisi terim konu matrisimiz olur. Her iki matriste de U ve V sütunları t konularımızdan birine karşılık gelir.

- U matrisi: Bu matristeki satırlar, konular cinsinden ifade edilen belge vektörlerini temsil eder.
- V matrisi: Bu matriste satırlar, konular cinsinden ifade edilen terim vektörlerini temsil eder.

SVD'yi uyguladıktan sonra ne elde ederiz?

Uk matrisinin (belge-konu matrisi) her satırı karşılık gelen matrisin vektör temsilidir.
belge. Burada bu vektörlerin uzunluğu istediğimiz konu sayısı olan k'ye eşittir ve verilerimizdeki terimlerin vektör temsili Vk matrisinde (terim-konu matrisi) bulunabilir.

Yani SVD bize verilerimizdeki her belge ve terim için vektör temsillerini döndürür. Her vektörün uzunluğu k olacaktır. Bu vektörlerin önemli bir kullanımı, kosinüs benzerliği metriğinin yardımıyla benzer kelimeleri ve benzer belgeleri bulabilmemizdir.

Belge ve Terim Vektörlerinin Uygulanması

Artık bu belge vektörleri ve terim vektörlerinin yardımıyla, değerlendirilecek kosinüs benzerliği gibi bazı ölçümleri kolayca hesaplayabiliriz:

- Farklı belgelerin benzerliği.
- Farklı kelimelerin benzerliği.
- Arama sorgumuzla en alakalı pasajları almak istediğimizde bilgi erişiminde faydalı olacak terimlerin veya sorguların ve belgelerin benzerliği.

Önceki Bilginizi Test Edin

1. Eğitim verilerinde her biri 100 cümle içeren 10 belgemiz olduğunu varsayalım. Her cümle en fazla 20 kelimeden oluşur. Bu verileri kullanarak her cümleyi sınıflandıran bir model oluşturmamız gerekiyor. Aşağıdakilerden hangisi X_train'in boyutunu temsil eder?

- (1000, 20)
- (12000, 1)
- (10, 100, 20)
- (100, 10, 20)

2. Aşağıdaki algoritmalardan hangisi, bir belge koleksiyonunda sık kullanılan kelimelerin ağırlığını azaltırken, pek kullanılmayan kelimelerin ağırlıklarını artırır?

- Dönem Sıklığı (TF)
- Ters Belge Sıklığı (IDF)
- Kelime çantası (YAY)
- Yukarıdakilerin hiçbiri

3. Python'daki hangi modül düzenli ifadeleri destekler?

- tekrar
- normal ifade
- piregx
- Yukarıdakilerin hiçbiri

LSA'nın Avantajları ve Dezavantajları

Gizli Semantik Analiz çok faydalı olabilir ancak sınırlamaları da vardır. LSA'nın her iki tarafını da anlamak önemlidir; böylece ondan ne zaman yararlanacağınız ve ne zaman başka bir şeyi deneyeceğiniz konusunda bir fikriniz olur.

LSA'nın Avantajları

1. Verimlidir ve uygulanması kolaydır.

2. Ayrıca düz vektör uzayı modeline kıyasla çok daha iyi, iyi sonuçlar verir.

3. Yalnızca belge terimi matris ayrıştırmasını içerdiğinden, mevcut diğer konu modelleme algoritmalarıyla karşılaştırıldığında daha hızlıdır.

LSA'nın dezavantajları

1. Doğrusal bir model olduğundan doğrusal olmayan bağımlılıklara sahip veri kümelerinde iyi sonuç vermeyebilir.

2. LSA, belgelerdeki terimlerin Gauss dağılımını varsayar; bu, tüm problemler için doğru olmayabilir.

3. LSA, hesaplama açısından yoğun olan ve yeni veriler ortaya çıktıkça güncellenmesi zor olan SVD'yi içerir.

4. Yorumlanabilir yerleştirmelerin olmaması (konuların ne olduğunu bilmiyoruz ve bileşenler keyfi olarak olumlu/olumsuz olabilir)

5. Doğru sonuçlar elde etmek için gerçekten çok sayıda belgeye ve kelime dağarcığına ihtiyaç var

6. Daha az verimli temsil sağlar

Optimum Konu Sayısının Belirlenmesi

Verilen derlem metnindeki optimum konu sayısını belirlemek o kadar kolay bir iş değildir, bazen bu çok zorlayıcı bir iş haline gelebilir. Ancak problem tanımına göre optimum konu sayısını belirlemek için aşağıdaki seçenekleri deneyebiliriz:

1. İlk yöntem, her konuyu ayrı bir küme olarak ele alıp, bir kümenin etkinliğini Silhouette katsayısı yardımıyla bulmaktır.

2. Konu tutarlılığı ölçüsü, konu sayısını belirlemek için gerçekçi bir ölçüdür.

Konu modellerini değerlendirmek için Konu Tutarlılığı yaygın olarak kullanılan bir ölçümdür. Gizli değişken modellerini kullanır. Oluşturulan her konunun bir kelime listesi vardır. Konu tutarlılığında, bir konudaki mevcut kelimelerin ikili kelime benzerliği puanlarının ortalamasını veya ortancasını bulacağız.

Sonuç: Konu tutarlılık puanının yüksek değerini alırsak model iyi bir konu modeli olarak değerlendirilecektir.

LSA uygulamaları

LSA, Gizli Semantik İndeksleme (LSI) ve Boyut Azaltma algoritmalarının öncüsü olarak kabul edilir.

1. LSA, boyutsallığın azaltılması için kullanılabilir. Herhangi bir bağlamı veya bilgiyi kaybetmeden vektör boyutunu büyük ölçüde milyonlardan binlere düşürebiliriz. Sonuç olarak hesaplamayı azaltır

güç ve hesaplamayı gerçekleştirmek için harcanan zaman.

2. LSA arama motorlarında kullanılabilir. Latent Semantic Indexing (LSI), LSA'dan geliştirilen vektör yardımıyla verilen arama sorgusuyla eşleşen dokümanların bulunmasından dolayı LSA'yı temel alarak geliştirilmiş bir algoritmadır.

3. LSA aynı zamanda belge kümeleme için de kullanılabilir. Gördüğümüz gibi LSA, atanan konuya göre her belgeye konu atamaktadır, böylece belgeleri kümelendirebiliriz.

Kendi Kendine Öğrenme: Çözülmüş bir LSA örneğini görmek istiyorsanız aşağıdakileri okuyun:

Çözülmüş Örnek için Okuyun

Bu, Doğal Dil İşleme ile ilgili Blog Dizimizin 16. Kısmını sonlandırıyor!

Diğer Blog Yazılarım

Daha önceki blog yazılarıma da göz atabilirsiniz.

Önceki Veri Bilimi Blog gönderileri.

LinkedIn

İşte Linkedin profilim benimle bağlantı kurmak istersen. Sizinle bağlantı kurmaktan mutluluk duyacağım.

E-posta

Sorularınız için bana Gmail'den e-posta gönderebilirsiniz .

Son Notlar

Okuduğunuz için teşekkürler!

Umarım makaleyi beğenmişsinizdir. Beğendiyseniz arkadaşlarınızla da paylaşın. Bahsetmediğiniz bir şey mi var veya düşüncelerinizi paylaşmak mı istiyorsunuz? Aşağıya yorum yapmaktan çekinmeyin, size geri döneceğim.

Makale URL'si - https://www.analyticsvidhya.com/blog/2021/06/part-16-step-by-step-guide-to-master-nlp-topic-modelling-using-lsa/

chirag676