Step by Step Guide to Master NLP – Topic Modelling using LSA Makalesinin Özeti

Hazırlayan

Fırat Kaan Bitmez

Öğrenci Numarası

23281855

Dersin Hocası

Asst.Prof.Dr. İsmail İşeri

Giriş

Bu Raporda Doğal Dil işleme ve Topic Modelling (konu modellemesi) içeresindeki LSA yöntemini ve LSA'nın Doğal dil işlemedeki problemlerinki önemi ,avantajları ,dezavantajları uygulama adımları ve kullanım alanları inceleyen makalenin kısa Türkçe özeti ele alınmıştır.

Step by Step Guide to Master NLP – Topic Modelling using LSA Makalesinin İçeriği

1. Konu Modellemenin Özeti
2. Neden Latent Semantic Analysis (LSA) gerekir?
3. Latent Semantic Analysis (LSA) Nedir?
4. LSA Uygulanırken İzlenen Adımlar
5. LSA'nın Avantajları ve Dezavantajları
6. Optimal Konu Sayısını Nasıl Seçilir?
7. LSA'nın Uygulama Alanları

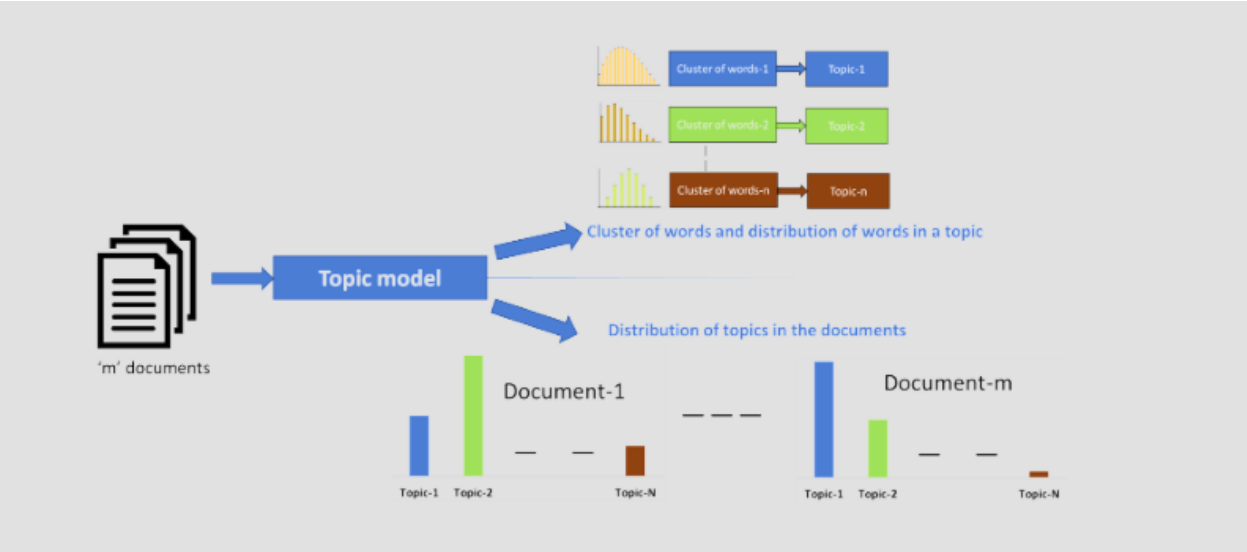
Biz burada Konu modelleme Nedir ve LSA Ne için kullanılır gibi Konularının üzerinde duracağız.

Konu Modelleme (Topic Modelling)

Tüm konu modelleme algoritmalarının temel varsayımları şunlardır:

Her belge birden fazla konuyu içerir,

Her konu bir kelime koleksiyonundan oluşur. Başka bir deyişle, konu modelleme algoritmaları, belgelerimizin anlamının aslında metinsel malzemeyi gördükten sonra doğrudan gözlemlemediğimiz bazı gizli veya "görünmez" değişkenler tarafından yönetildiği fikri etrafında inşa edilmiştir.



LSA Neden Gereklidir?

Tüm doğal dillerin kendi incelikleri ve nüansları vardır ve bunlar bilgisayarlar tarafından yakalanması oldukça zordur ve bazen biz insanlar tarafından bile yanlış anlaşılabilir! Bu, aynı şeyi ifade eden farklı kelimeleri ve aynı yazılışa sahip ancak farklı anlamlar veren kelimeleri içerir.

Örneğin, aşağıdaki iki cümleyi düşünün:

Cümleler:

I liked the last novel of Premchand quite a lot.
They would like to go for a novel marketing campaign.

İlk cümlede "**novel**" kelimesi bir kitabı temsil ederken, ikinci cümlede yeni veya taze

anlamına gelir.

Biz insanlar, bu iki kelimenin arkasındaki bağlamı anlayabileceğimiz için bu iki kelime arasında kolayca ayrım yapabiliriz. Ancak makineler, kelimelerin nasıl kullanıldığı bağlamı anlayamadıkları için bu kavramı yakalayamazlar. İşte bu noktada, Latent Semantic Analysis (LSA) kendini gösterir.

LSA, kelimelerin etrafındaki bağlamı kullanarak, konular olarak adlandırılan gizli veya görünmez kavramları yakalamaya çalışır.

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

5. Daha az verimli temsil sağlar.

Optimal Konu Sayısının Belirlenmesi

Verilen metin korpusunda optimal konu sayısını belirlemek kolay bir iş değildir, bazen bu çok zorlu bir görev haline gelir. Ancak, problem ifadesine göre, optimum konu sayısını belirlemek için aşağıdaki seçenekleri deneyebiliriz:

LSA'nın Uygulama Alanları

LSA, Latent Semantic Indexing (LSI) ve Boyut Azaltma algoritmalarının öncüsü olarak kabul edilir.

Singular Value Decomposition (SVD)

SVD bir matrisin üç ayrı matrise çözülmesini sağlayan bir tekniktir. Bu matrisler şunlardır:

Ortogonal sütun matrisi (V)

Ortogonal satır matrisi (U)

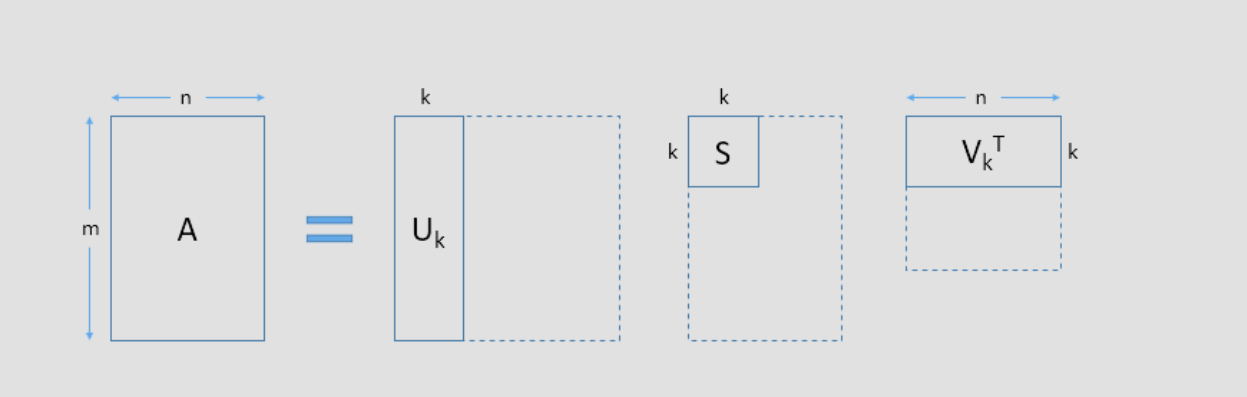
Bir tekil matris (S)

M=U*S*V

$$A = USV^T$$

Matris A, M=USV şeklinde ifade edilir,

Burada S, A'nın tekil değerlerini içeren diyagonal bir matristir. Truncated SVD ise, yalnızca t en büyük tekil değeri seçerek boyut azaltma işlemidir. Bu durumda, dönüştürülmüş uzayımızdaki en önemli boyutları koruduğumuzu düşünebiliriz. Matris U, belge-konu matrisi olarak kullanılırken, matris V ise terim-konu matrisi olarak kullanılır. Bu matrislerdeki sütunlar t konulara karşılık gelir. Bu vektör temsilleri, benzerlik ölçümlerinde kullanılarak farklı belgelerin, terimlerin veya sorguların benzerliklerinin değerlendirilmesine yardımcı olur.



Sonuç

İncelemiş olduğumuz bu makalede, doğal dil işleme alanında konu modellemesi ve özellikle Latent Semantic Analysis (LSA) üzerine makaleyi inceledik ve baz öğrenimler kazandık. Konu modellemesinin temel prensiplerini ve LSA'nın nasıl çalıştığını anladık. LSA'nın avantajları ve dezavantajlarına göz attık ve optimal konu sayısını belirlemenin zorluklarını tartıştık. Ayrıca, LSA'nın çeşitli uygulama alanlarını keşfettik.

Sonuç olarak, **Step by Step Guide to Master NLP – Topic Modelling using LSA** makalesi, LSA'nın temel kavramlarını anlamanıza ve doğal dil işleme projelerinizde bu teknikten nasıl yararlanabileceğinize dair bir kılavuz sağlamıştır. Bu bilgileri uygulamada kullanarak, metin verilerinden anlamlı bilgiler çıkarmak ve çeşitli endüstrilerde uygulamak için güçlü bir temel oluşturabilirsiniz.

Kaynaklar

https://www.datacamp.com/tutorial/what-is-topic-modeling

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