Word2Vec

One-hot encoding kelimeleri tanımlamak için kullanılan yöntemlerden biridir. Vektörün boyu elimizdeki kelimelerin çeşitliliği kadardır. Vektörde sadece ilgili kelimenin bulunduğu yerin değeri 1, diğer tüm kelimelerin değeri 0'dır.

→ Skip-gram

Skip-gramda girdi hedef kelime iken çıktılar ise hedef olarak verilen kelimenin etrafındaki kelimelerdir. Girdiler ile çıktıları olasılıksal olarak birbirine benzeterek anlamsal olarak en uygun şekilde temsil etmek hedeflenmektedir.

Mimari CBOW'a benzer, ancak mevcut kelimeyi bağlama göre tahmin etmek yerine, aynı cümledeki başka bir kelimeye göre bir kelimenin sınıflandırmasını maksimize etmeye çalışır. Daha doğrusu, her mevcut kelimeyi sürekli projeksiyon katmanına sahip bir log-lineer sınıflandırıcıya girdi olarak kullanırız ve mevcut kelimeden önce ve sonra belirli bir aralıktaki kelimeleri tahmin ederiz.

→ CBOW

CBOW (Continuous Bag of Words) skip-gram'a çok benzer bir yaklaşımdır. Aralarındaki tek fark girişlerle çıkışların yer değişmiş olmasıdır. Buradaki fikir bir kelimenin etrafındaki kelimeler verildiğinde hangi kelimenin bu kelimeler içinde görülme olasılığının en yüksek olduğunu bilmek istemesidir.

Doğrusal olmayan gizli katmanın kaldırıldığı ve yansıtma katmanının tüm kelimeler için paylaşıldığı ileri beslemeli NNLM'ye benzer (yalnızca projeksiyon matrisi değil); böylece, tüm kelimeler aynı konuma yansıtılır (vektörlerinin ortalaması alınır).

→ Skip-gram vs. CBOW

CBOW modelleri genel olarak daha küçük küçük datasetlerde daha iyi çalışırken Skip-gram modeller daha büyük datasetlerde daha iyi çalışmaktadır. CBOW daha az hesaplama gücü gerektirirken, Skip-Gram daha fazla hesaplama gücü gerektirir.

FastText

"FastText" 2016 yılında Facebook tarafından geliştirilmiş Word2Vec'in bir uzantısıdır. Tek tek kelimeleri yapay sinir ağına girdi olarak vermek yerine kelimeleri birkaç harf bazlı "n-gram" halinde parçalar. Örneğin elma sözcüğü için tri gram: elm, lma'dır. N-gram ifadesinde yer alan n tekrar derecesini ifade etmektedir. Yani kaçar kaçar böleceğimizi buradaki n ifadesi sağlar.Bir kelime veya harften ne kadar olduğunu anlamamızı sağlar. Elma'nın word vektörü tüm bu n-gram vektörlerinin toplamıdır.

Her bir kelimenin n-gram karakterlerden oluşan bir paket olarak temsil edildiği atlama programı modeline dayalı yeni bir yaklaşımdır. Bir vektör temsili, her karakter n-gram ile ilişkilidir; kelimeler bu temsillerin toplamı olarak temsil edilmektedir. Yöntem hızlıdır, modelleri büyük bir kurum üzerinde hızlı bir şekilde eğitmeye ve eğitim verilerinde görünmeyen kelimeler için kelime temsillerini hesaplamaya izin verir.

GloVe: Global Vectors for Word Representation

Unsupervised algoritmaların temelinde verilerin istatistikleri yer almaktadır. Skip-gram, CBOW gibi modeller anlamsal bilgileri yakalar ama birlikte kullanılma istatistiklerini kullanmazlar. Matris

ayrıştırma yöntemleri bu istatistikleri kullanmasına rağmen anlamsal ilişkileri yakalayamamaktadırlar. Bu tarz modellerde anlamsallık yoktur. Pennington ve diğerleri tarafından önerilen “GloVe” modeli olasılık istatistiklerinden yararlanarak yeni bir objektif fonksiyon oluşturarak bu problemi çözmeyi amaçlamaktadır.

Word2Vec vs. GloVe vs. FastText

Bu üç algoritma kelime düğünlerinin nasıl hesaplanacağına dair üç genel fikri temsil ediyorlar:

Word2Vec, metinleri bir sinir ağı için eğitim verisi olarak alır. Ortaya çıkan embedding, kelimelerin benzer bağlamlarda görünüp görünmediğini yakalar.

GloVe, tüm külliyatta kelime birlikteliklerine odaklanır. Yerleştirmeleri, iki kelimenin bir arada görünme olasılıkları ile ilgilidir.

FastText, kelime bölümlerini de dikkate alarak Word2Vec'i geliştirir. Bu numara, daha küçük veri kümeleri üzerinde yerleştirmelerin eğitimini ve bilinmeyen kelimelere genellemeyi sağlar.

Son yıllarda daha fazla gelişme, her ikisi de dağıtım hipotezine veya aynı bağlamlarda geçen kelimelerin genellikle benzer anlamlara sahip olduğuna dair gözlemlere dayanan Word2Vec ve GloVe da olmuştur. FastText çerçevesi kısa süre sonra, nadir bulunan veya eğitim külliyatında yer almayan kelimeler için iyi vektör temsilleri oluşturmaya izin verdi.

Referanslar

[1] Mikolov, Tomas; etal. (2013). "Efficient Estimation of Word Representations in Vector Space".arXiv:1301.3781[cs.CL].

[2] Enriching Word Vectors with Subword Information,Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov, 2016

[3] Bag of Tricks for Efficient Text Classification,Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, 2016

[4] https://medium.com/codable/word2vec-fasttext-glove-d4402fa8cce0

[5] https://towardsdatascience.com/the-three-main-branches-of-word-embeddings-7b90fa36dfb9

[6] https://www.alpha-quantum.com/blog/word-embeddings/introduction-to-word-embeddings-word2vec-glove-fasttext-and-elmo/