İstanbul Teknik Üniversitesi Elektronik ve Haberleşme Mühendisliği Bölümü

Isaret Isleme için Makine Öğrenmesi Donem Odevi

2022 Sonbahar

Ogrenci: Giray Arat

Ogretmen: Bilge Gunsel

Proje Adi: Sosyal Medya'dan Alınan Verileri Kullanarak Olay Tanımlama

Proje Amaci

Sosyal medya'dan alınan verilerin yakinsaklik vektörlerinin oluşturulması, öbeklenmesi ve sonrasinda kelime ogrenme yapılmasi ile "olay tanımlama" yapmak.

Proje Referanslari

Proje makalesi DOI:
https://doi.org/10.1007/s10994-021-05988-7
Proje gerceklemesi içeren GitHub depositorysi:
https://github.com/HHansi/Embed2Detect
Projede kullanılan sosyal medya verisinin işlenmiş halini içeren GitHub depositorysi:
https://github.com/HHansi/Twitter-Event-Data-2019

Projede Kullanilacak Giris Verisi

https://github.com/HHansi/Twitter-Event-Data-2019/blob/master/BrexitVote/ids_7.30-17.30.txt

Projede Yapılacaklar

Genel bilgi

Kelimeler arasinda syntax ile alakali ve semantic benzerlikler vardir. Syntax benzemeler sadece ayni kelime yapisina sahip olmakla alakalıdır. Semantic yani anlamsal benzerlikler ise ayni

anlama gelen farkli soyleyislerin kullanilmasidir. Makale bunun icin verdigi örnekte kendi ürettikleri veriden iki tweet arasinda ayni anlama gelen fakat farkli syntax'e sahip iki deyişi secmis:

hashtags. In addition, different word phrases such as *impact assessments* and *economic analysis* were used to mention the same subject discussed in them. In such cases, semantics are needed to understand the relationships between terms to extract valuable information.

Bu sebeple verilerin semantic yakınsaklıklar kurularak işlenmesi gerekiyor. Bunun için kullanılabilecek birçok "Word2vec" algoritmasi var. Word2vec algoritmalari, "text" yani metini vektorlere çevirmek üzere geliştirilmiş bazi algoritmalarin adi. Syntax uzerine semantic benzerlikleri de yakalayabileceği için seçilen Skip-gram modeli kullaniliyor makalede. Skip-gram modeli bir denklem.

3.1.1 Skip-gram model

Skip-gram model is a log-linear classifier which is composed by a 3-layer neural network with the objective to predict context/surrounding words of a centre word given a sequence of training words $w_1, w_2, ...w_n$ (Mikolov et al. 2013b). More formally, it focuses on maximizing the average log probability of context words $w_{k+j} | -m \leq j \leq m, j \neq 0$ of the centre word w_k by following the objective function in Eq. 1. The length of the training context is represented by m.

$$j = \frac{1}{n} \sum_{k=1}^{n} \sum_{-m<j<m, j\neq 0} \log p(w_{k+j}|w_k) \quad (1)$$

Kelime gömülmesini (Word Embedding) öğrenme

Sonraki adim Skip-gram modeli kullanarak alınan ham veriden sayisal vektörel veri oluşturmak. Benim anladigim kadariyla bu adimda sadece ham veriden sayisal vektorler olusturma islemi yapiliyor. Fakat bunu yapan fonksiyonun "learn word embeddings" diye bir basligi var. Neden sadece "word embedding" degil de "learn word embeddings" yaziyor bilemiyorum.

Olay Pencerelerini (Event Windows) öğrenme

Twitter'dan alinan veriler kullanilarak olay pencereleri belirleme. Sadece belirli token üzerinden filtrelenmis ve baska on-isleme geçirmemiş veriyi alarak ilk önce sayısal değerlere sahip vektörleri oluşturmuşlardı. Simdi kodun kalan kisminda bu vektörleri kullanarak olay pencerelerini belirleyecekler.

Olay pencereleri

Bu makine öğrenmesi uygulamasinin amaci olan olay tanimlamayi yapmak için vektörel verinin icerisinde bazi olay pencereleri belirlemek gerekiyor. Olay penceresinin ne oldugunu anlamak icin makaledeki ornegi tekrar edebiliriz. Makalede kullanılan veri dizilerinden bir tanesi bir futbol maçıyla ilgili. Diger veri seti ise Brexit ile ilgili.

Bir veri seti politika ile ilgili iken diğeri sporla ilgili. Makaleyi yazan doktora ogrencisinin beraber calistigi profesörün bununla ilgili makalesinde gosteriyormus ki, politika ile ilgili verilerin olay penceresi daha uzun ve sporla ilgili verilerin olay penceresi daha kısa olur. Bu demektir ki sporla ilgili olaylar olduğunda kisa bir zaman dilimi içinde cok sik ve cok fazla Tweet atilir. Buna kıyasla politikayla ilgili Tweetler daha uzun bir zaman dilimine yayilir. O zaman Brexit'le ilgili olabilecek olaylari tanımlayabilmek için daha uzun olay pencereleri tanimlamak gerekir. Futbol maçıyla ilgili olabilecek olaylar bir kisinin gol atmasi, birisinin ceza karti gormesi olabilir. Bu olaylari tanımlayabilmek için daha kisa olay pencereleri tanimlamak gerekir.

Kod'un yaptigi: (Cluster) Obeklerin Arasindaki Degisimleri Hesaplamak

Kod'un bu bolumunde olay pencerelerini tanimlayacak islemler yapilir. Onceki adimlarda "word embeddings" uretilmisti. Bu word embeddingsdeki semantic ve syntax ile ilgili degisimlere bakarak, obekler arasindaki degisimler hesaplanir.
Clustering icin hierarchical agglomerative clustering (HAC) secilmis. Iki cluster arasindaki mesafe icin bir sonraki denklemi kullanmislar:

$$D(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{w_p \in C_i} \sum_{w_q \in C_j} d(w_p, w_q),$$

Kelime öbeklerinde tokenler vardi. Tokenler Tweetlerde geçen bir kaç kelime için geçerli, o kelimenin benzer kelimelerinin toplandigi kelimeler. Makale mesela futbol maçından örnek veriyor. Futbol maçıyla ilgili Tweetlerde "goalll" ve "goallll" kelimeleri geçiyorsa, bu kelimeler tek bir token altında toplanıyor. Bu iki kelime de "goalll" sayılıyor.
Her zaman penceresi için bu tokenler arasındaki değişimleri içeren bir matrix kuruluyor. Bu matrislerin tokenleri arasında karar vermek icin Dendogram seviyesinde benzerlik kullaniliyor.

$$DL\ Similarity_{(w_i, w_j)} = \frac{dl_{(w_i, w_j)}}{max(dl_{r \to x} : x \in L) + 1}$$

Kelime Damıtma (Word Extraction) yapma

Önceki adımda olay pencereleri tanımlanmıştı. Eğer bir olay penceresi, içinde olay gerçekleşmiş bir olay penceresi olarak belirlenirse, "event word extractor", yani olay kelime damıtıcısı, olayla ilgili kelimeleri o pencereden damıtıyor.
Olaylar, metin dağarcığında değişimler yaratiyor. Yani sosyal medyada akan, yaşayan metin havuzunda, olayların gerçekleşmesiyle beraber ani değişimler oluyor. Bu sebeple onceki olay penceresine göre öbek farkları taşıyan kelimeler, olay kelimeleri olarak seçiliyor. Obek farklari yani "cluster changes" gosteren kelimeler.
Yukarida gosterilen $DL\ Similarity_{(w_i, w_j)}$ degeri sifirdan farkli olan çiftlere ait olan kelimelerin hepsi, geçici "temporal" obekler değişimine sahip kelimeler sayılıyor.

Benim anladığım kadarıyla olay tanımlama bu şekilde gerçekleşiyor. Makine öğrenmesi algoritmaları olay penceresi ve sonrasında olay kelimeleri belirlemede kullanılmış oluyor.