MAKİNE ÖĞRENMESİ PROJE DETAYLARI

Amaç:

- Paylaşılan Twitter veri kümelerinde sınıflandırma problemini makine öğrenme yöntemleriyle ele alarak duygu analizi yapmak

Veri kümeleri:

- "MACHINE LEARNING PROJECT TEAM SHEET" dokümanında da her öğrenciye atandığı üzere Twitter veri kümeleri Kemik Doğal Dil işleme grubunun paylaştığı Türkçe tweetlerden oluşmaktadır.
- 3bin ve 17bin tweet içeren 2 farklı veri kümesi bulunmaktadır. Bunlardan 17bin tweet içeren veri kümesinin etiket bilgisi .csv dosyalarında yer almaktadır. Sizden beklenen test ve train olarak ayrılan bu veri kümelerinin tek bir .csv dosyasında birleştirilmesidir. 3bin tweet içeren veri kümesinin etiket bilgisi her cümle için arff dosyasında toplanmıştır. Dolayısıyla dosyanızı hangi formatta kullacaksanız yazacağınız ufak bir scriptle ister csv ister txt dosyanızda her bir cümlenin etiket bilgisinin de olduğu tek bir doküman haline getirmeniz beklenmektedir.
- Veri kümeleriyle alakalı her bilgiyi içeriği, etiket bilgisi vb her detayı raporunuzda paylaşmanız beklenmektedir. Dosyaları tek bir yerde toplamak için yaptığınız tüm işlemleri raporunuzda özellikle belirtmelisiniz.
- Veri kümeleriyle ilgili detaylar "okubeni" dosyasında yer almakta olup veri kümesine referans vermek için sunulan yayınlardan (SIU 2013 ve ASYU 2018 konferansındaki yayınlar) veri kümelerinin detaylarını toplayıp raporunuza kendi cümlelerinizle eklemeniz beklenmektedir.

Eğitim Kümesi Yüzdeleri:

- Veri kümenizi sınıflandırıcınıza input olarak vermeden önce eğitim ve test kümesi olarak 2' ye bölmeniz gerekmektedir. Bu aşamada 3 farklı eğitim kümesi yüzdeleri için sınıflandırıcılarınızı çalıştırmanız beklenmektedir Bunlar %80 , %50 ve %30 eğitim yüzdeleridir.
- Dolayısıyla her bir eğitim yüzdesi için sınıflandırıcılar farklı sonuçlar üretecektir.

Yöntemler:

- Sınıflandırma için gerekli olan makine öğrenmesi teknikleri k-nearest neigbour, naive Bayes, support vector machine, decision trees, multilayer perceptronlar, convolutional neural networkler, recurrent neural networkler, long-short term memory networkler, Word2Vec, GloVe, FastText dir.
- Bu teknikler kullanılarak size atanan veri kümelerini sınıflandırmaya çalışacaksınız. Hazırlayacağınız raporun içeriğinde, sizlere "MACHINE LEARNING PROJECT TEAM SHEET.pdf" dokümanından atanan sınıflandırıcıların nasıl çalıştığını detaylı bir şekilde anlatmanız ve varsa ilgili sınıflandırıcıya dair şekil /formül sunmanız beklenmektedir.
- Projenizin sunumunda sınıflandırıcının nasıl çalıştığına dair sorularla karşılaşacağınızdan modelinizi anlamanızda fayda vardır.

Sonuçlar:

- Sınıflandırıcılardan elde edeceğiniz sonuçları accuracy (doğruluk) metriğiyle raporunuzda tablolaştırarak sunmanız beklenmektedir.

Örneğin, 3bin tweet içeren veri kümesi için

	Sınıflandırıcılar			
Eğitim Yüzdesi	Knn(k=1)	Knn(k=3)	Knn(k=5)	DT
Ts80	0.78	Xxx	Xxx	Xxx
Ts50	0.71	Xxx	Xxx	Xxx
Ts30	0.62	Xxx	Xxx	Xxx

- Sınıflandırıcısı **k-NN** olanlar belirtilen k parametrelerine göre modellerini ayrı oluşturup sonuçlarını da tablodaki gibi sunmalılar.
- Sınıflandırıcısı **SVM** olanlar farklı kernellardaki (linear, polynomial,rbf) sonuçları da tablodaki gibi sunmalılar.
- Sınıflandırıcısı **NB** olanlar farklı NB modelleri (Gaussian NB, multinomial NB, Bernoulli NB) için sonuçları da tablodaki gibi sunmalılar.
- Sınıflandırıcısı **MLP, RNN, CNN, LSTM, Word2Vec, GloVe, FastText** olanlar modelin aldığı tüm parametrelerle (window size, epoch, vb..) oynayarak en iyi sonucu elde edene kadar modellerini iyileştirmeliler. Her çalışmadaki deney sonuçlarını tabloya işlemeliller. (Burada çok uç değerlerle çalışmamak adına internette sıklıkla denenen parametre değerlerini almanızda fayda vardır. Mesela window size 3,5,7 gibi)

Teslim: Proje teslim tarihi **10.06.2020 23:59** a kadardır. Projelerinizi e-destek üzerinden açılacak olan platforma yüklemeniz beklenmektedir. Proje **sunumlarını ders günü ve saatinde 11.06.2020** tarihinde almaya başlayacağım. Sisteme raporlarıyla birlikte yüklenmeyen projelerin sunumu alınmayacaktır.

Not: Kafanıza takılan sorular için ZOOM dersi yapılması planlanmaktadır. Sorularınıza hızlı yanıt almak için zeynep.kilimci@kocaeli.edu.tr adresine mail atmanız yeterli olacaktır.

Kolaylıklar dilerim.

Zeynep.