Hesaplamalı Anlambilim Dersi Proje Konuları

Projenizi aşağıdaki konularda ya da kendi önereceğiniz yeni bir konuda yapabilirsiniz.

Projede ne yapmayı planladığınızı (kullanacağınız ya da oluşturacağınız veri kümesini, sistemin sayısal başarısını nasıl ölçeceğinizi) bir paragraf halinde açıklayıp dersin yürütücüsüne 5 Mayıs 2025 saat 09:30'a kadar "hesaplamalı anlambilim proje önerisi" başlığı ile email (mfatihamasyali@gmail.com) atınız. Birden fazla öneride de bulunabilirsiniz. Onay aldığınızda projenize başlayabilirsiniz.

Onay almadan yapılan projeler geçerli olmayacaktır.

Projede herkes tek başına çalışacaktır.

Teslim edilecekler: Kod, sunum, rapor (siu 2025 - IEEE formatında)

Teslim şekli: online.yildiz.edu.tr

Son teslim tarihi: Dersin final tarihi, final saati

Kullanabileceğiniz örnek derlemler, veri kümeleri, modeller:

https://drive.google.com/open?id=1mTeSCsf9NaMUL2oD2B09GpD8aOkcMVCO

https://huggingface.co/datasets/merve/turkish_instructions

https://huggingface.co/spaces/malhajar/OpenLLMTurkishLeaderboard

https://huggingface.co/ytu-ce-cosmos

Projeleriniz final günü sınıfta sunulmalıdır. Sunum yapmayanlar en fazla %30 not alabilirler.

Proje konuları:

1- Ortak bir sözlüğe ve mimariye sahip word2vec modellerinin ağırlık bazında birleştirilmesi: 100 MB'lik Türkçe bir derlem alın. 100 MB ile eğitilen modeli, 50 50 lik 2 modelin birleştirilmesi ile elde edilen modeli, 25 25 25 25 lik 4 modelin birleştirilmesi ile elde edilen modeli, 50 ve 25 lik lerin herbirini en az 3 Türkçe veri kümesini lojistik regresyon ile sınıflandırmada karşılaştırın.

2- Kelime ve token tabanlı word2vec modellerinin karşılaştırılması: 100 MB'lik Türkçe bir derlem alın. en az 3 tokenizer (1'i Türkçeye özel olmalı) kullanın. kelime tabanlı ve 3 token tabanlı 4 modeli en az 3 Türkçe veri kümesini lojistik regresyon ile sınıflandırmada karşılaştırın

3- Token temsili path leri: N layerdan oluşan bir LLMde her layerda her token için oluşan temsiller X boyutlu olsun. Bu durumda her token için X boyutlu uzayda N adet noktadan oluşan bir path oluşur. Eğitilmiş bir llm üzerinde oluşan bu path leri kullanarak kendi seçeceğiniz bir uygulama (sınıflandırma, üretim, benzerlik skoru vb.) gerçekleştirin.

4- Logit path leri: LLM ler bir sonraki token ı tahmin ederken softmax öncesi X boyutlu logit üretir. her bir token üretimi sonrası bu logit değerleri oluşur. N adet token üretildiğinde X boyutlu uzyda N adet noktadan oluşan bir path oluşur. Eğitilmiş bir llm üzerinde oluşan bu path leri kullanarak kendi seçeceğiniz bir uygulama (sınıflandırma, üretim, benzerlik skoru vb.) gerçekleştirin.

5- Attention matrisleri: LLM lere M token dan oluşan bir metin verildiğinde her layerda her bir head için N*N lik bir attention matrisi oluşur. Eğitilmiş bir llm üzerinde oluşan bu matrisleri kullanarak kendi seçeceğiniz bir uygulama (sınıflandırma, üretim, benzerlik skoru vb.) gerçekleştirin.

6- Çözebilir mi?: https://huggingface.co/datasets/ytu-ce-cosmos/gsm8k_tr veri kümesinde soru ve cevap bulunmaktadır. İçeriklerinden bağımsız olarak, sorunun türünü ve cevabın yöntemini ifade eden metinler üretin. Üretimi prompt optimizasyonu ile çok başarılı modellere yaptırın. 4 metnin https://huggingface.co/ytu-ce-cosmos/turkish-e5-large ile embedding lerini alın.

https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 ve https://huggingface.co/google/gemma-2-9b-it modellerine soruları verip çözüp çözemediklerine göre etiketleyin (2 sınıf).

(4 metin*2 model) 8 embedding veri kümesini %50 eğitim %50 test olarak ayırın.

lojistik regresyon ile sınıflandırma başarılarını ölçün. metin türüne ve modele göre sonuçları yorumlayın.

Veri kümelerini t-sne ile 2 boyuta indirgeyin. sınıflarına göre renklendirin.

7- Bağlamda öğrenme: https://huggingface.co/datasets/ytu-ce-cosmos/gsm8k_tr veri kümesinde 5-shot ile çözümde soruya en benzeyen örnekleri ve rasgele seçilmişleri kullanmanın etkisi ölçülecektir. Benzerlik için veri kümesindeki sorudan içeriğinden bağımsız olarak, sorunun türünü vb. ifade eden 3 adet metin üretin. Üretimi prompt optimizasyonu ile çok başarılı modellere yaptırın. 4 metnin https://huggingface.co/ytu-ce-cosmos/turkish-e5-large ile embedding lerini alın. Bu embedding lere göre benzerlikleri bulun. Soru cevaplarını https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 ile alın.

8- Soruları bozalım: https://huggingface.co/datasets/ytu-ce-cosmos/gsm8k_tr veri kümesinden https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 ve https://huggingface.co/google/gemma-2-9b-it modellerinin ikisinin de çözebildiği 50 soru seçiniz.

Sorular üzerinde harf / kelime üzerinde silme / ekleme / yer değiştirme / rasgele seçilenle değiştirme işlemleriyle adım adım değişiklik yapınız. Sorulara doğru cevaplar alınabildiği sürece değişikliklere devam ediniz. Soruların bozulmadan önce değişim türlerinin ve değişim sayılarının etkisini inceleyiniz.

9- Benzerleri çözebilir mi?: https://huggingface.co/datasets/ytu-ce-cosmos/gsm8k_tr veri kümesinden https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 modelinin çözebildiği 50 ve çözemediği 50 soru seçiniz. Başarılı bir llm ile her soru için en az 5 farklı prompt la her prompt tan 5 er benzer soru (toplam 25 adet) üretiniz.

Üretilen soruların https://huggingface.co/ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 tarafından çözülüp çözülemediğini belirleyiniz.

Orijinal ve üretilen soruların https://huggingface.co/ytu-ce-cosmos/turkish-e5-large ile embedding lerini alın. t-sne ile 2 boyuta indirgeyip, çözülüp çözülememeye göre renklendirerek görselleştirin. Sonuçları prompt türüne vb. göre yorumlayın.