SENTETIK KISILIK ANALIZI VE NLP ILE HOBI TAHMINI

Yasin Yeşilyurt,
Abdullah Arda Gündoğdu

GIRIS

- Bu projede, Nvidia'nın Nemotron veri setini kullanarak, yapay zeka tarafından üretilen insan "personaları" ile "hobiler" arasında anlamsal bir bağ kurmayı amaçladık.
- Temel Zorluk: Veri setindeki hobi tanımları anlamsal olarak aynı olsa da metin olarak farklıydı (Örn: "building legos" vs "lego building").

AMACIMIZ

- Kirli ve tekrarlı metin verisini standartlaştırmak (Canonicalization).
- Bir vektör uzayından diğerine haritalama yapmak (Regression).

Veri Seti ve Engeller

Başlangıç noktamız 100.000 sentetik kişiliğe sahip Nvidia Nemotron veri setiydi.

Ancak proje ortasında Nvidia seti güncelleyip 100 Milyon veriye çıkardı ve yapıyı değiştirdi. 8 milyar parametreli embedding modellerinin işlem maliyeti nedeniyle, 93K'lık ilk versiyonu indirip deneyler için 3.000 adetlik temiz bir alt küme oluşturduk.

Veri analizinde (EDA), "city", "marital status" gibi hobiyle doğrudan ilgisi olmayan kolonları eledik

VERI ÖN ISLEME VE "SEMANTIC COMPRESSION"

Anlamsal Kümeleme (NCD)

Bu projenin en özgün kısımlarından biri burası. Basit bir vektör benzerliği, "guitar practice" ile "playing guitar"ı ayırmakta zorlanıyordu.

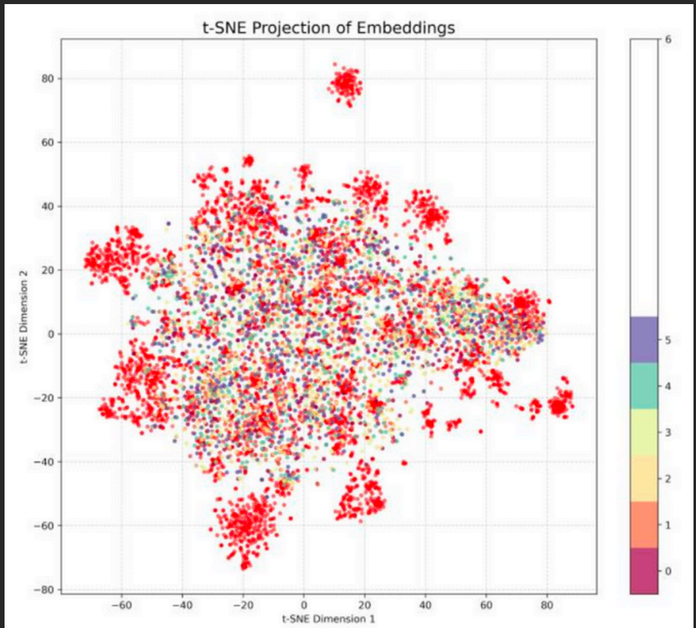
Bunu çözmek için Normalized Compression Distance (NCD) kullandık. Bu yöntem, kelime benzerliğinden ziyade bilginin sıkıştırılabilirliğine bakarak benzer hobileri kümeledi.

Kanonikleştirme (Canonicalization)

Böylece 12.000'den fazla benzersiz hobi metnini, yaklaşık 100 adet "Canonical" (standart) hobi etiketine indirgedik.

Anahtar kelime çıkarımı için YAKE algoritmasının, spaCy NER modellerinden daha başarılı olduğunu gördük.

EMBEDDING VE GÖRSELLESTIRME

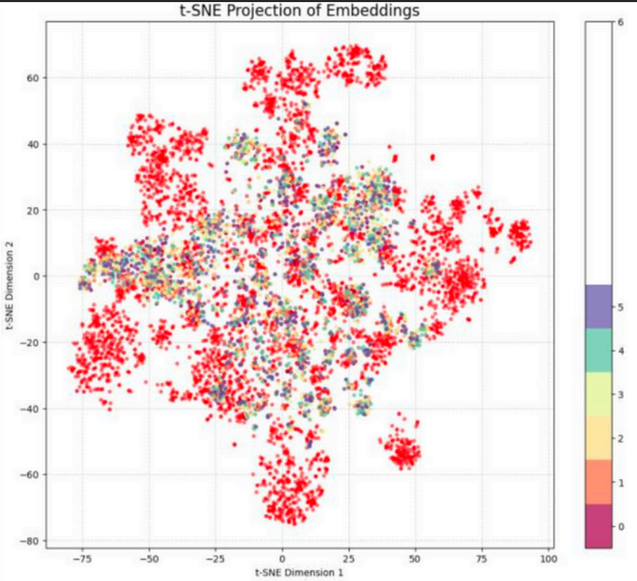


0.6B embedding with hobbies projected on to

Verileri vektör uzayına taşımak için Qwen 3 modelini kullandık (0.6B ve 8B varyantları).

t-SNE ile yaptığımız görselleştirmelerde, 0.6B modelinin persona ve hobileri tam ayrıştıramadığını, ancak 8B modelinin (ekrandaki grafik) anlamsal kümeleri çok daha net oluşturduğunu gördük.

EMBEDDING VE GÖRSELLEŞTİRME



Verileri vektör uzayına taşımak için Qwen 3 modelini kullandık (0.6B ve 8B varyantları).

t-SNE ile yaptığımız görselleştirmelerde, 0.6B modelinin persona ve hobileri tam ayrıştıramadığını, ancak 8B modelinin (ekrandaki grafik) anlamsal kümeleri çok daha net oluşturduğunu gördük.

REFERANS MODELLER VE ALTERNATIF YAKLAŞIMLAR

1. MatrixKNN (Baseline Yaklaşım)

Nedir: Herhangi bir eğitim gerektirmeyen, doğrudan vektör uzayındaki en yakın komşuyu (Cosine Similarity) bulan yöntem.

Sonuç: Hızlıdır ve ham embedding kalitesini ölçer ancak model lineer olmayan (non-linear) karmaşık ilişkileri öğrenemez. "Zero-shot" yeteneği ile sınırlıdır.

MatrixKNN'in hızını ve Cross-Encoder'ın öğrenme yeteneğini birleştirmek için Vektörden-Vektöre Regresyon (Vector-to-Vector Regression) yapan MLP modeline geçiş yaptık.

REFERANS MODELLER VE ALTERNATIF YAKLAŞIMLAR

2. Cross-Encoding (Semantic Reranking)

Nedir: Persona ve Hobi metnini tek bir girdi (pair) olarak alıp uygunluk skoru üreten daha ağır bir mimari.

Sonuç: Kelime eşleşmelerine (keyword overlap) aşırı odaklandı (Örneğin: IT çalışanına sürekli teknolojik hobi önermesi gibi). Nüansı yakalamakta zorlandı

REFERANS MODELLER VE ALTERNATIF YAKLAŞIMLAR

3. Çıkarım (Neden MLP?)

MatrixKNN'in hızını ve Cross-Encoder'ın öğrenme yeteneğini birleştirmek için Vektörden-Vektöre Regresyon (Vector-to-Vector Regression) yapan MLP modeline geçiş yaptık.

MODEL MIMARISI - MLP PROJECTOR

Problemi bir sınıflandırma (classification) değil, vektörden-vektöre regresyon (vector-to-vector regression) olarak kurguladık.

Geliştirdiğimiz "HobbyProjector" modeli, persona uzayından hobi uzayına doğrusal olmayan (non-linear) bir dönüşüm öğrenen bir Multilayer Perceptron (MLP) yapısıdır.

Eğitimde "Cosine Embedding Loss" kullandık çünkü amacımız vektörlerin yönünü (açısını) birbirine yaklaştırmaktı.

SONUÇLAR

Modelimiz 93 epoch sonunda 0.070 validasyon kaybına ulaştı.

Bu, tahmin edilen hobi vektörü ile gerçek hobi vektörü arasında 0.93 oranında bir kosinüs benzerliği olduğu anlamına geliyor. Yani model, kişiliği analiz edip doğru hobiyi vektör uzayındaki hedefi çok net bir şekilde işaret edebiliyor.

Ayrıca daha küçük (4096 hidden dim) modelin de benzer performans göstermesi, problemin çözümünün verimli olduğunu kanıtladı.

SONUÇLAR

Sonuç olarak, metin tabanlı öneri sistemlerinde vektör regresyonunun güçlü bir yöntem olduğunu kanıtladık.

Gelecekte veri seti ölçeğini artırmayı ve "bir kişiye birden fazla hobi" (one-to-many) önerisi yapabilecek bir yapı kurmayı hedefliyoruz.