

**KOCAELİ ÜNİVERSİTESİ**

Teknoloji Fakültesi  
Bilişim Sistemleri Mühendisliği Bölümü

**YAZILIM GELİŞTİRME LABORATUVARI DERSİ**

Proje: Ses Analizi ile Podcast Tür Tahmini

Hazırlayanlar:

Halit Mert Artun - 211307005

Latif Atmaca - 181307053

2024

**Ses Analizi ile Podcast Tür Tahmini**

**Giriş ve Proje Amacı**

Bu proje, günümüzün hızla artan dijital bilgi kaynaklarından biri olan podcast içeriklerinin analizini hedefleyen bir çalışma olarak tasarlanmıştır. Podcast'ler, özellikle eğitim amaçlı içerik üreticileri tarafından yaygın bir şekilde kullanılmaktadır. Eğitim içerikli podcast’lerin artan popülaritesi, bu içeriklerin analiz edilerek dinleyici davranışları, içerik trendleri ve ilgi alanlarının belirlenmesi gibi pek çok farklı alanda fayda sağlayabilir.

Proje, YouTube platformunda yer alan eğitim amaçlı podcast videolarının indirilmesi, bu videolardan elde edilen verilerin işlenmesi, analiz edilmesi ve görselleştirilmesi sürecini kapsamaktadır. Bu bağlamda, projenin amacı üç ana başlık altında toplanabilir:

1. **Veri Toplama**: YouTube Data API kullanılarak eğitim kategorisinde yer alan podcast videolarının indirilmesi. Bu adımda, API ile alınan verilerden yalnızca belirli filtrelemelere uygun olan içerikler seçilmektedir. Örneğin, belirli bir süre sınırının altındaki içerikler elenmiş ve yalnızca güncel içerikler projeye dahil edilmiştir.
2. **Veri İşleme**: İndirilen podcast açıklamalarının ve meta verilerinin metin işleme teknikleriyle analiz edilebilir forma getirilmesi. Bu aşamada, verilerin temizlenmesi, gereksiz bilgilerden arındırılması ve analiz için uygun hale getirilmesi sağlanmaktadır. Veri işleme sürecinde metin verilerinin tokenize edilmesi, noktalama işaretlerinin kaldırılması ve düşük bilgi değerine sahip olan yaygın kelimelerin (stopwords) filtrelenmesi gibi işlemler yapılmaktadır.
3. **Veri Analizi ve Görselleştirme**: İşlenen verilerin görselleştirilmesiyle podcast içeriklerinin analiz edilmesi. Görselleştirme adımında, yüksek boyutlu veriler, UMAP gibi boyut indirgeme algoritmaları kullanılarak iki boyutlu bir alana indirgenmekte ve farklı içerik türleri arasındaki ilişkiler analiz edilmektedir. Bu sayede, eğitim podcast’lerinin belirli bir konuya yoğunlaşma oranları, benzer içeriklerin kümelenme düzeyleri ve genel eğilimler gözlemlenebilmektedir.

Bu rapor, proje sürecinde karşılaşılan teknik sorunların nasıl çözüldüğünü, kullanılan teknolojilerin özelliklerini, veri toplama ve işleme aşamalarında yapılan işlemleri detaylandıracaktır. Proje, eğitim içeriklerinin dijital analizine dair bir temel oluşturmakta olup, ileride yapılacak daha ileri düzey analizler ve geniş çaplı araştırmalar için zemin hazırlamaktadır.

**Özet**  
 Bu proje, eğitim amaçlı podcast içeriklerini analiz etmek ve görselleştirmek için geliştirilmiş bir Python tabanlı veri işleme çalışmasıdır. Projede, YouTube üzerinde yer alan eğitim podcast’lerinin YouTube Data API kullanılarak otomatik olarak indirilmesi ve bu içeriklerin metin işleme teknikleriyle analiz edilmesi amaçlanmıştır.

Projenin ilk aşamasında, belirli anahtar kelimelere ve kriterlere göre veri toplanmış; API istekleri, eğitim içerikli podcast’lerin çeşitliliğini artırmak için optimize edilmiştir. Toplanan veri, tokenizasyon, stopwords filtreleme ve vektörleştirme gibi metin işleme adımlarıyla analiz edilebilir hale getirilmiştir. Elde edilen veriler, UMAP algoritması kullanılarak iki boyutlu alana indirgenmiş ve Matplotlib ile görselleştirilmiştir. Bu görselleştirmeler, eğitim içeriklerinin hangi konular etrafında kümelendiğini ve popüler eğilimleri anlamamıza yardımcı olmuştur.

Proje sürecinde karşılaşılan teknik zorluklar, Python sürüm güncellemeleri ve kütüphane uyumsuzlukları gibi sorunların çözülmesiyle aşılmıştır. Nihayetinde, proje, eğitim içeriklerinin dijital analizine dair bir temel oluşturmakta olup, podcast dinleyici davranışları ve içerik trendleri üzerinde daha ileri analizlerin yapılmasına olanak sağlamaktadır.

## Kullanılan Teknolojiler ve Kütüphaneler

Bu projede kullanılan başlıca teknolojiler ve kütüphaneler şunlardır:  
  
**1. Python:** Python, veri analizi ve işleme, görselleştirme ve makine öğrenimi alanlarında geniş kütüphane desteği sunar.  
**2. YouTube Data API:** YouTube üzerindeki videoların meta verilerine erişim ve veri çekme işlemleri için kullanılmıştır. API sayesinde, belirli anahtar kelimelere sahip eğitim içerikleri otomatik olarak alınabilmiştir.  
**3. Pyenv:** Python sürüm yönetimini sağlayan pyenv, farklı Python sürümlerini indirip kurmaya yardımcı olur. Projede Python 3.11.10 sürümü kullanılmak istenmiş ve pyenv ile uyumluluk sağlanmıştır.  
**4. NLTK (Natural Language Toolkit):** Doğal dil işleme süreçlerinde metinlerin parçalanması, stopwords filtrelenmesi ve tokenizasyon işlemleri için NLTK kullanılmıştır.  
**5. UMAP**: Yüksek boyutlu veriyi iki boyutlu bir alan üzerinde görselleştirmek için kullanılan UMAP, özellikle metin analitiğinde sıkça tercih edilir.  
**6. Matplotlib:** UMAP ile indirgenmiş verileri grafiksel olarak göstermek için Matplotlib kullanılmıştır.  
**7. Pandas:** Veri işleme ve analiz işlemleri için Pandas veri çerçevesi (DataFrame) yapısı kullanılmıştır.  
8. Venv: Python’un sanal ortam oluşturma aracı ile bağımlılıkların izole edilmesi sağlanmıştır.

## Sorun Tanımları ve Çözüm Yöntemleri

Proje sırasında karşılaşılan başlıca sorunlar ve bunlara yönelik çözüm yöntemleri:  
**1. Python Sürüm Güncellemesi:** Python 3.8.10’dan 3.11.10’a geçiş yapılmıştır. Pyenv ile yeni sürüm kurulmuş ve sanal ortam yeniden yapılandırılmıştır.  
**2. downloader.py Script Sorunu:** Script, YouTube API'de eğitim kategorisinde sürekli aynı podcast’i indiriyordu. nextPageToken ile sayfalama sağlanarak farklı içerikler çekildi ve indirilen videoların takip edilmesi için video ID’lerinin kaydedildiği bir sistem geliştirildi.  
**3. visualization\_script.py Script Hataları:** umap kütüphanesi import hatası verirken NLTK’nın bazı kaynaklarına erişim problemi yaşanmıştır. Gerekli kütüphanelerin doğru şekilde kurulması sağlanmış ve eksik kaynaklar indirildi.  
**4. Veri Toplama ve İşleme:** Eğitim podcast'lerinin etkin bir şekilde toplanıp işlenmesi için API istek parametreleri optimize edilmiştir. Örneğin, minimum süre ve yayın tarihi gibi filtreler eklenmiştir.

## Veri İşleme Adımları ve Detayları

Veri işleme sürecinde indirilen podcast’lerin açıklamaları metin ön işleme adımlarından geçmiştir. Bu işlemler aşağıdaki adımlardan oluşmaktadır:  
  
1. **Tokenizasyon:** Her açıklama metni kelime temelli parçalara ayrılmıştır.  
2. **Küçük Harfe Çevirme:** Büyük-küçük harf farklılıklarından kaynaklı hataları engellemek için metin küçük harfe dönüştürülmüştür.  
3. **Noktalama İşaretlerinin Kaldırılması:** Verinin temizlenmesi için noktalama işaretleri kaldırılmıştır.  
4. **Stopwords (Önemsiz Kelimeler) Filtrelemesi:** Türkçe ve İngilizce gibi dillerdeki yaygın ancak bilgi değeri düşük kelimeler, analizden çıkarılmıştır.  
5. **Vektörleştirme:** Metin verileri TF-IDF yöntemiyle sayısal vektör formuna dönüştürülerek analiz için hazırlanmıştır.

## Görselleştirme ve Analiz Süreci

Görselleştirme işlemi, indirilen ve işlenen podcast verilerinin analizinde önemli bir adımdır. Bu süreçte, verilerin iki boyutlu alan üzerinde görselleştirilmesi için UMAP yöntemi kullanılmıştır:  
  
**1. Boyut İndirgeme:** Podcast açıklama verileri yüksek boyutlu olduğundan UMAP kullanılarak iki boyuta indirgenmiştir. Böylece, metin verisinin farklı kümeleri arasındaki ilişkiler daha net analiz edilmiştir.  
**2. Scatter Plot:** İndirgenmiş veriler matplotlib kullanılarak scatter plot şeklinde çizilmiştir. Bu grafik, farklı podcast kategorilerinin birbirine olan benzerliklerini görselleştirmeyi sağlar.  
**3. Renk Kodlaması:** Farklı kategoriler için farklı renkler atanarak kategoriler arasında ayrım yapılmıştır. Bu analiz, eğitim amaçlı içeriklerde hangi tür içeriklerin popüler olduğunu görselleştirmeye olanak tanımıştır.

## Sonuç ve Öneriler

Projenin tüm adımları, Python sürüm güncellemesi, downloader.py ve visualization\_script.py ile ilgili sorunların çözülmesi ve verilerin işlenmesiyle başarıyla tamamlanmıştır.   
  
**Elde Edilen Çıktılar:**  
- Python sürüm güncellemesi ile uyumluluk sağlanmış, projede paket uyumsuzlukları giderilmiştir.  
- downloader.py scripti, daha çeşitli eğitim podcastlerini otomatik olarak indirebilecek şekilde optimize edilmiştir.  
- Görselleştirme adımlarıyla podcast içeriklerinin daha iyi analiz edilmesi sağlanmıştır.  
  
İleriye Yönelik Öneriler:  
**1. Makine Öğrenimi:** Eğitim içeriklerini analiz etmek için makine öğrenimi algoritmaları eklenebilir, özellikle içerik türlerini belirlemek için sınıflandırma yöntemleri kullanılabilir.  
**2. Veri Çeşitlendirmesi:** Farklı platformlardan veri çekmek ve analiz etmek için yeni API entegrasyonları yapılabilir.  
3. **Zaman Serisi Analizi:** Podcast içeriklerinin zamanla değişen popülerlik eğilimlerini analiz etmek için zaman serisi analizleri eklenebilir.

**Kaynakça**

1. **YouTube Data API Documentation**  
   Google Developers. (2024). *YouTube Data API Overview*. YouTube Data API v3.
2. **Python Documentation**  
   Python Software Foundation. (2024). *Python 3.11 Documentation*. Python Docs.
3. **Natural Language Toolkit (NLTK) Documentation**  
   Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.
4. **Pandas Documentation**  
   The Pandas Development Team. (2024). *Pandas Documentation*. Pandas Docs.
5. **Uniform Manifold Approximation and Projection (UMAP)**  
   McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv preprint arXiv:1802.03426.
6. **Matplotlib Documentation**  
   Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90-95.
7. **Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking**  
   Provost, F., & Fawcett, T. (2013). O'Reilly Media.