İnönü Üniversitesi
Bilgisayar Mühendisliği Bölümü
Bitirme Projesi (2022-2023 GÜZ)

Proje başlığı	Trend Aramaların Sınıflandırılması
Öğrenci(ler)	Civan BOZAN - Hafta 8
Proje özeti	**Yöntem:** Apache Spark veri işleme çerçevesi ile arama motorlarında sıklıkla aratılan ifadelerin, Spark Streaming özelliği ile gerçek zamanlı olarak işlenmesi ve Spark MLlib makine öğrenmesi kütüphanesi kullanılarak bu verilerin sınıflandırılması (kategorilere ayrılması) sağlanıp uygun formatta (json, xml..) gerçek zamanlı olarak sunulması ve geçmişe yönelik analizlerin gerçekleştirilebilmesi açısından bu verilerin Apache HBase veritabanı ile tutulması ve ilgili analizlerin bir web sitesi aracılığıyla servis edilmesi. **Amaç:** Kullanıcı eğilimlerinin tespit edilerek, ilgili kategoriler hakkında içerik üreten yazarlara ilham kaynağı olmak.
Tamamlanması gereken özellikler	1. Spark ve İlgili kütüphanelerin incelenmesi. 2. HBase incelenmesi. 3. HBase ve Spark entegrasyonu 4. Arama motorlarından sıklıkla aratılan ifadelerin elde edilmesi. 5. Elde edilen ifadelerin spark ile gerçek zamanlı olarak işlenmesi. 6. Verilerin kategorilere ayrılması ve uygun formata dönüştürülmesi. CRISP-DM süreci a. İşin Anlaşılması b. Verinin Anlaşılması c. Verinin Hazırlanması d. Modelleme e. Değerlendirme f. Konuşlandırma 7. Spark ile işlenmiş tüm verilerin HBase'de tutulması. 8. Geçmiş verilerin web sitesi üzerinden sunulması.
Sonradan eklenmesi gereken özellikler	1. Verilerin bir API aracılığıyla sunulması. 2. Abonelik özelliği ile sadece ilgili kategorilere ait aramaların sunulduğu bir web sitesi.
Bu hafta gerçekleştirilenler	1. Spark kümesi oluşturulması için gerekli hazırlıklar üzerinde çalışıldı. (Sanal makine, hadoop bileşenleri..) 2. CRISP−DM sürecinin ilk 3 aşaması gerçekleştirildi. ● **İşin Anlaşılması** ● **Verinin Anlaşılması** ● **Verinin Hazırlanması**
Kaynaklar	1. **Veri seti:** https://www.kaggle.com/datasets/savasy/multiclass-classification-data-for-turkish-tc32