Ders Proje Konuları için Örnek Projeler

Doğal Dil İşleme Dersi kapsamında gerçekleştirilecek olan projeler aşağıdaki başlıklardan oluşabilir. Sizlerin de getireceği öneriler olabilir. Öncesinde tarafıma bilgi verilip, onay alınması gereklidir. Birçok proje için veri seti ihtiyacı çok önemlidir. İngilizce veri seti bulmak Türkçe'ye göre çok daha kolaydır. Ancak hiç yok değildir. Belki sizlerin desteği ile yavaş yavaş oluşturabiliriz.

Doküman Sınıflandırma / Text Classification

- Spam maillerin tespiti
- Yazar Tanıma
- Doküman Türünü Bulma
- Dokümanın yazarını Bulma
- Duygu Analizi
- Kişilerin Duygu Durumlarının Tespiti

Konu başlıkları için

- makine öğrenmesi veya
- derin öğrenme modelleri (word2vec, fastText, doc2vec, BERT, Alberto, vs.)

kullanarak bir uygulama geliştirmek.

Etiketsiz Dokümanların Kümelenmesi / Text Clustering

- TR / EN etiketsiz gazete yazılarının farklı kümeleme algoritmaları ile arşivlemesi (KMeans, Fuzzy- C Means, Dbscan, BIRCH) algoritmalarından biri seçilerek detaylı inceleme yapılabilir
- Topic Detection –Latent Drichlet Allocation Alg. Kullanılması

Yarı Eğiticili Yöntemler / Semi Supervised Methods

- **YATSI**

Bilgi Getirimi / Information Retrieval

- Bir sorgu kelimesi ile yapılan arama sonucunda gelen sayfaları sorgu kelimesinin anlamına göre ayırarak getirme. Örnek: Sorgu kelimesi "bayram" olsun. Gelen sayfaların "isminde bayram olanlar", "milli ve dini bayramlar ile ilgili sayfalar", "bayram ismi verilen bir müzik grubu", vs.

Bilgi Çıkarımı / Information Extraction

- Seçilecek 2 veya 3 farklı kitap, gayrimenkul, market ürünleri gibi farklı sayfalardan aradığımız ürünleri belirlenen belli kriterlere sahip olanları bulup çıkarma
- Text olarak verilmiş yapısal olmayan özgeçmiş bilgileri yer alan bir dokümandan önceden belirlenmiş bilgilerin varsa çıkarılması

Soru ve Cevap / Question Answering

- Önceden belirlenmiş alanlarda sorulmuş olan soruların cevaplarını internetten bumak

Otomatik Etiketleme / Auomatic Tagging

- Seçilen bir alan üzerinde otomatik etiketleme yaparak (Türkçe için) etiketli veri seti oluşturmak. Knime yardımcı araç olarak kullanılabilir. Özellikle varlık isimlerinin tespitinde. (kişi, yer, organizasyon, tarih, para, özel varlık)

Aklıma yeni öneriler geldikçe ekleyip dosyanın son halini paylaşacağım sizinle. Bunun dışında sizlerin önerilerine de açığım. Ancak önce onayımın alınması gerekmektedir.

NLTK-Natural Language Toolkit
Zemperek

Reuters-21578
The 20 newsgroups dataset
ModApte split for top 10 categories in Reuters-21578 from publication: Semi-supervised text categorization
NRC Emotion Lexicon

Banu Diri