ÖZET

TÜRKÇE DOKÜMANLARIN SINIFLANDIRILMASI

Rumeysa YILMAZ

Yüksek Lisans Tezi, Matematik Anabilim Dalı
Tez Danışmanı: Yrd. Doç. Dr. Rıfat AŞLIYAN
2013, 75 sayfa

İnternetin hızla gelişmesi elektronik ortamdaki bilgileri ve işlemleri arttırmıştır. Fakat, bu ortamlarda depolanan ve işlenen bilgilerin boyutunun artması aranan bilgiye erişmekte problemler çıkarmıştır. Bu doğrultuda, kullanıcıların istedikleri bilgiye daha doğru ve hızlı bir şekilde ulaşma ihtiyacı doğmuştur ve elektronik ortamdaki dokümanların sınıflandırılmasında yeni metotların geliştirilmesi çalışmaları devam etmektedir. Bu çalışmada, Türkçe metin içerikli web sitelerinden elde edilen dokümanların sınıflandırılması amaçlanmaktadır.

Dokümanlar, gövde tabanlı, sözcük tabanlı, hece tabanlı ve karakter tabanlı olmak üzere dört farklı kategoride ele alınmıştır. Gövde, sözcük, hece ve karakterler için n-gram analizleri yapılmıştır. Sistem K-En Yakın Komşu Modeli (K-NN), Çok Katmanlı Algılayıcı Modeli (MLP) ve Destek Vektör Makinesi (SVM) olmak üzere üç farklı yöntem ile eğitilmiş ve test edilmiştir. Çalışmanın gerçekleştirilmesinde eğitim ve test olmak üzere iki derlem oluşturulmuştur. Her biri internet ortamından temin edilen 75'er doküman içeren eğitim, ekonomi, kültür-sanat, otomobil, sağlık ve spor sınıfları ele alınmıştır. Bu dokümanlardan 25'er tane alınarak toplamda 150 doküman sistemin eğitilmesinde, 50'şer tane alınarak toplamda 300 doküman da sistemin test edilmesinde kullanılmıştır. Çalışmada sisteme verilen dokümanlar öncelikle önişlemden geçirilmiştir. Önişlemden geçirilen dokümanların frekansları ve olasılıkları hesaplandıktan sonra her bir sınıf için öznitelik vektör veritabanı oluşturulmuştur. Öznitelik vektör veritabanı oluşturulurken sözcüklerin dokümanlarda karşılaştırılmasında eşik değeri olarak 0,25, 0,50, 0,75 ve 0,90 değerleri kullanılmış. Eğitim setindeki dokümanlar sisteme verilmiş ve her bir sınıf için oluşturulan öznitelik vektör veritabanındaki sözcükler ile karşılaştırılarak dokümanın hangi sınıfa dahil olduğu belirlenmiştir. Daha sonra test setindeki dokümanlar sisteme verilmiş ve sistemin başarısı, kesinlik skoru, hassasiyet skoru, F-ölçüsü ve doğruluk değerlerine göre tespit edilmiştir.

Sonuç olarak en yüksek doğruluk başarı oranı SVM metodu kullanılarak sözcük 1-gramlarda %99,9 olarak bulunmuştur. F-ölçüsü değeri de %99,7 olmuştur.

Anahtar sözcükler: Doküman Sınıflandırma, K-En Yakın Komşu Modeli, Çok Katmanlı Algılayıcı Modeli, Destek Vektör Makinesi, n-gram.