
PROJE TESLİM TARİHİ: 21.05.2018

KARAKTER N-GRAM TABANLI DOKÜMAN SINIFLANDIRMA
Naive Bayes Algoritması öğrenci tarafından
kodlanılacaktır. Algoritmaya ait internetten alınan
kod parçacıkları kesinlikle kopya olarak
değerlendirilecektir.

- Proje masaüstü uygulaması olarak geliştirilecektir. Nesneye yönelik programlama ile geliştirmeniz istenmektedir. Kullanılabilecek diller: C++, C#, Python ve Java.
- Proje kaynak kodları (source code) CD içerisinde raporla beraber teslim edilecektir.
- Proje ile alakalı sorularınız için fidan.kaya@kocaeli.edu.tr ve ekin.ekinci@kocaeli.edu.tr adreslerine ulaşabilirsiniz.

Bu projenin amacı 1150 haberden oluşan veri kümesinin n-gram tabanlı olarak Naive Bayes sınıflandırma yöntemi ile sınıflandırılmasıdır.

Verilenler: 5 sınıftan (ekonomi, magazin, sağlık, siyasi, spor) oluşan 1150 habere ait veri kümesi. 1150 Haber isimli veri kümesi <http://www.kemik.yildiz.edu.tr/?id=28> adresinden elde edilecektir.

İstenler:

- Veri kümesi üzerinde temel ön işleme adımları uygulanır.
- Veri kümesindeki karakter 2-gram ve 3-gramlar frekansları ile birlikte eldeki haberlerden çıkartılır. Frekans değeri 50'den büyük olan n-gramlar sınıflandırmada kullanılacak özellikler olarak belirlenir.
- “bugün hava” cümlesindeki karakter 2-gramlar şu şekildedir: bu, ug, gü, ün, n_, _h, ha, av, va. Boşluk karakteri “_” ile temsil edilmektedir.
- Veri kümesinin rastgele %25'i test kümesi geri kalan %75'i ise eğitim kümesi olarak belirlenir.
- Eğitim kümesi Naive Bayes sınıflandırma yöntemi ile eğitilip elde edilen model test kümesine uygulanır ve test kümesindeki her haberin sınıf etiketi belirlenir (ekonomi, magazin, sağlık, siyasi, spor).

- Test kümesi için elde edilen sınıf etiketleri ile bu kümedeki her haberin gerçek etiketi karşılaştırılır ve precision, recall ve F-measure değerleri hesaplanır.

Proje Değerlendirme Yönergesi Açıklamalar

1. **PROJE RAPORU:** Her proje sonunda teslim edilmesi gereken projenizi her yönüyle açıklayacağınız bir dokümandır.

1.1. Problem Tanımı: Proje kapsamında sizden çözüm bulmanız beklenen problem ile ilgili açıklama yapılması gerekmektedir. Burada amaç projenin ne kadar anlaşıldığını test etmektir.

1.2. Yapılan Araştırmalar: Proje geliştirilmesi aşamasında karşılaşılan sorunlara nasıl çözümler bulunduğu ve bu konularda yapılan araştırmalar açıklanmalıdır.

1.3. Tasarım

1.3.1. Akış şeması: Proje içerisinde yer alan algoritma ve işlemleri şekilsel olarak ifade edecek şema oluşturmanız beklenmektedir.

1.3.2. Yazılım mimarisi: Projenin kodlanması aşamasında kullanılacak kod yapısı ve geliştirme aşamalarını gösteren bir yapı hazırlanması beklenmektedir.

1.3.3. Veri tabanı diyagramı: Projeye ait ER diyagramının oluşturulması beklenmektedir.

1.4. Genel Yapı: Projenizi genel yapısı bakımından her yönüyle özetlemeniz gerekmektedir.

1.5. Referanslar: Proje geliştirilirken ve araştırma aşamasında faydalanan kaynaklar rapor dokümanının en altında listelenmeli ve doküman içerisinde de ilgili yerlerde indekslenmelidir.

Referans formatı aşağıda verilen örneklerle uygun olmalıdır.

Kitap, çok yazarlı

Larson, G. W., Ellis, D. C., & Rivers, P. C. (1984). Essentials of chemical dependency counseling. New York: Columbia University Press.

Report from a private organization (author & publisher same)

National League for Nursing. (1990). Self-study report for community health organizations (Pub. No. 21-2329). New York: Author.

Unpublished master's thesis

Paulosky, K. A. (1997). Knowledge and attitudes of pain and activities of nurse administrators. Unpublished master's thesis, Northern Michigan University, Marquette, Michigan.

Article in a journal (continuous pagination throughout volume)

Burke, R. J., Shearer, D., & Deszca, E. (1984). Correlates of burnout phases among police officers. *Group and Organizational Studies*, 9, 451-466.

Article in a Popular Magazine

Caloyianis, N. (1998, September). Greenland sharks. *National Geographic*, 194, 60-71.

Web Site

http://en.wikipedia.org/wiki/Neural_network (Access date: 07.10.2013)

PROJELER İKİ KİŞİLİK GRUPLAR HALİNDE YAPILACAKTIR!

Proje gruplarının oluşturulacağı link:

<https://docs.google.com/spreadsheets/d/1Mhi7fneF8BPZbuBRXmj5RCFOzSSuSoHvFxBYLcOynPQ/edit?usp=sharing>