KOCAELİ ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
BLM306 YAZILIM LAB. II
PROJE 3

GRAF TABANLI METİN ÖZETLEME PROJESİ

Proje İlan Tarihi: 27 Nisan 2023
Proje Teslim Tarihi: 24 Mayıs 2023
Sunum Tarihleri: 25/26 Mayıs 2023

Bu projede verilen bir dokümandaki cümlelerin graf yapısına dönüştürülmesi ve bu graf modelinin görselleştirilmesi istenmektedir. Ardından graf üzerindeki düğümler ile özet oluşturan bir algoritma oluşturulması beklenmektedir.

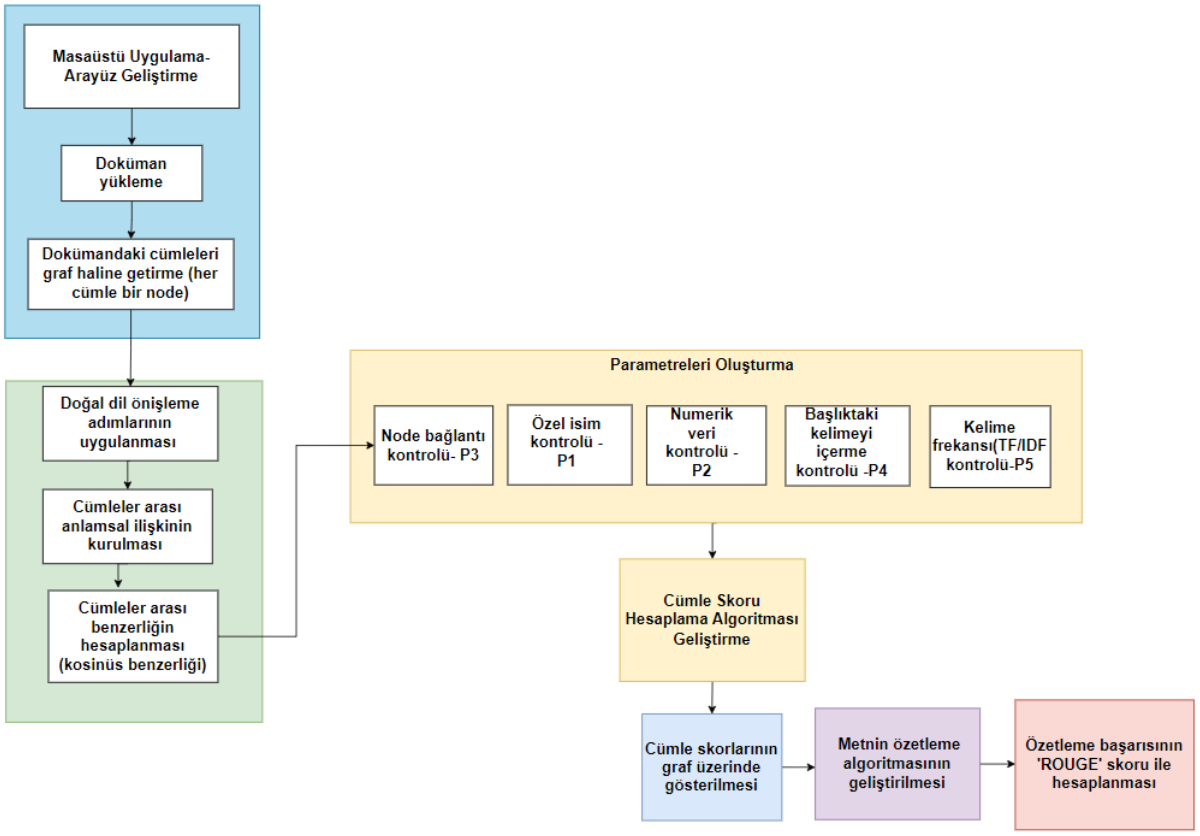
Amaç: Proje gerçekleştirimi ile öğrencilerin veri yapıları bilgisinin pekiştirilmesi ve problem çözme becerisinin gelişimi amaçlamaktadır.

Programlama Dili: Proje C++, C#, Java veya Python dili kullanılarak gerçekleştirilecektir.

Projede aşağıdaki isterleri yerine getirmeniz beklenmektedir.

İSTERLER:
Projede masaüstü uygulama geliştirmeniz gerekmektedir. Masaüstü uygulamada ilk olarak doküman yükleme işlemi gerçekleştirilecektir. Ardından yüklenen dokümandaki cümleleri graf yapısı haline getirmeniz ve bu graf yapısını görselleştirmeniz beklenmektedir. Bu grafta her bir cümle bir düğümü temsil edecektir. Cümleler arasındaki anlamsal ilişki kurulmalı, cümleler skorlanmalıdır. Belirli parametreleri kullanarak cümle skorunun hesaplama algoritmasını ve cümle skorlarına göre metin özeti çıkarma algoritmalarını sizin geliştirmeniz istenmektedir. Özet metni arayüzde sunmanız beklenmektedir. Sonuç olarak size verilen bir metnin özetini bu yöntem ile çıkarmanız ve gerçek özet ile benzerliğini "*ROUGE*" skorlaması ile ölçmeniz istenmektedir.

Şekil 1'de projenin akış diyagramı sunulmuş, ayrıntılar aşağıda açıklanmıştır.



Şekil 1. *Proje akış diyagramı*

Projede temel amaç; cümleleri graf yapısına çevirip Cümle Seçerek Özetleme (Extractive Summarization) gerçekleştirmektir. Graf yapısına çevirerek cümlelerin metindeki anlamsal ilişkilerini görselleştirmek ve bu ilişkileri kullanarak önemli cümleleri belirlemek amaçlanmaktadır.

Masaüstü Arayüzü Geliştirilmesi ve Graf Yapısının Oluşturulması

- Masaüstü arayüzü geliştirmeniz beklenmektedir. Arayüz aşağıdaki isterleri içermelidir:
    - Kullanıcının doküman yükleyebileceği bir alan,
    - Dokümanın graf halinde görüntüleneceği bir alan,
    - Cümle benzerliği için threshold seçilebilecek bir araç,
    - Cümle skorunun belirlenmesi için threshold seçilebilecek bir araç.
    - Cümle benzerliği algoritmasına alternatif oluştururursanız bunun arayüzden seçilebilmesini sağlayan bir araç.
- Dokümandaki cümleleri graf yapısına dönüştürmek için hazır bazı veritabanları, kütüphaneler veya API kullanabilirsiniz. Bunlardan bazıları;
    - Neo4j: Grafik veritabanı yönetim sistemi olarak bilinir ve grafik yapısını kullanarak verileri depolar ve işler.
    - NetworkX: Python programlama dili için açık kaynaklı bir graf kütüphanesidir. Düğümler ve kenarlar gibi grafik elemanlarını temsil etmek için birden fazla graf sınıfı sağlar.

- Graph-tool: C++ programlama dili için açık kaynaklı bir graf kütüphanesidir. hızlı ve büyük veri kümeleri için daha uygun olabilir.
- Gephi: Java programlama dili için açık kaynaklı bir grafik analiz aracıdır. Grafiklerin görselleştirilmesine ve analiz edilmesine yardımcı olan bir dizi araç sağlar
- igraph: R, Python ve C/C++ için açık kaynaklı bir graf kütüphanesidir. Düğümler ve kenarlar gibi grafik elemanlarını temsil etmek için birden fazla graf sınıfı sağlar. Ayrıca, grafikleri manipüle etmek ve farklı ölçütlere göre analiz etmek için bir dizi fonksiyon sunar.

Cümleler Arası Anlamsal İlişkinin Kurulması

- Cümlelere NLTK kütüphanesi kullanılarak aşağıdaki ön işleme adımları uygulanmalıdır:
  - Tokenization: Bir metnin küçük parçalara ayrılmasıdır.
  - Stemming: Kelimelerin kökünün bulunması işlemidir.
  - Stop-word Elimination: Bir metindeki gereksiz sözcükleri çıkarma işlemidir. Stop word'ler, genellikle yaygın olarak kullanılan, ancak metnin anlamını belirlemekte önemli bir rol oynamayan kelime ve ifadelerdir.
  - Punctuation: Cümledeki noktalama işaretlerinin kaldırılmasıdır.

  NOT: Cümle skoru hesaplama adımında yapılması gereken özel isim içerme ve nümerik veri içerme adımları bu ön işlemlerden önce gerçekleştirilmelidir.

- İki cümle arasındaki anlamsal ilişkiyi kurmak için aşağıdaki yöntemlerden en az biri kullanılmalıdır (İki yöntemin de kullanılması durumunda ek puan verilecektir):
  - Word Embedding: Kelime düzeyindeki anlamsal ilişkileri yakalamak için kullanılan bir makine öğrenimi tekniğidir. Cümleleri temsil etmek için word embedding kullanıldığında, her kelime; vektörleri ile temsil edilir ve cümle vektörü, içerdikleri kelime vektörlerinin toplamıdır. Bu şekilde, cümlelerin anlamsal ilişkileri vektör uzayında ölçülebilir hale gelir.
  - BERT: Özellikle doğal dil işleme (NLP) alanında kullanılan bir derin öğrenme modelidir. BERT, bir cümleyi tamamen anlamak ve cümleyi oluşturan kelimelerin birbirleriyle olan ilişkilerini anlamak için kullanılabilir. BERT, önceden eğitilmiş bir modeldir ve büyük bir metin korpusunda önceden eğitilir. Bu sayede, dildeki örüntüleri ve anlamsal ilişkileri öğrenir ve genelleştirir.
- Benzerliği ölçmek için "kosinüs benzerliği" yöntemini uygulamalısınız. Kosinüs benzerliği, iki vektör arasındaki benzerliği ölçmek için kullanıldığı gibi, iki cümle arasındaki benzerliği de ölçmek için kullanılabilir.

Cümle Skoru Hesaplama Algoritmasının Geliştirilmesi

- Cümle Skoru Hesaplama sırasında aşağıdaki parametreleri oluşturmalısınız:
  - Cümle özel isim kontrolü (P1)

    Cümledeki özel isim sayısı / Cümlenin uzunluğu

- Cümlede numerik veri olup olmadığının kontrolü (P2)

  Cümledeki numerik veri sayısı / Cümlenin uzunluğu
- Cümle benzerliği threshold'unu geçen node'ların bulunması (P3)

  Tresholdu geçen nodeların bağlantı sayısı / Toplam bağlantı sayısı
- Cümlede başlıktaki kelimelerin olup olmadığının kontrolü (P4)

  Cümledeki başlıkta geçen kelime sayısı / Cümlenin uzunluğu
- Her kelimenin TF-IDF değerinin hesaplanması (P5). Buna göre dokümandaki toplam kelime sayısının yüzde 10'u 'tema kelimeler' olarak belirlenmelidir.

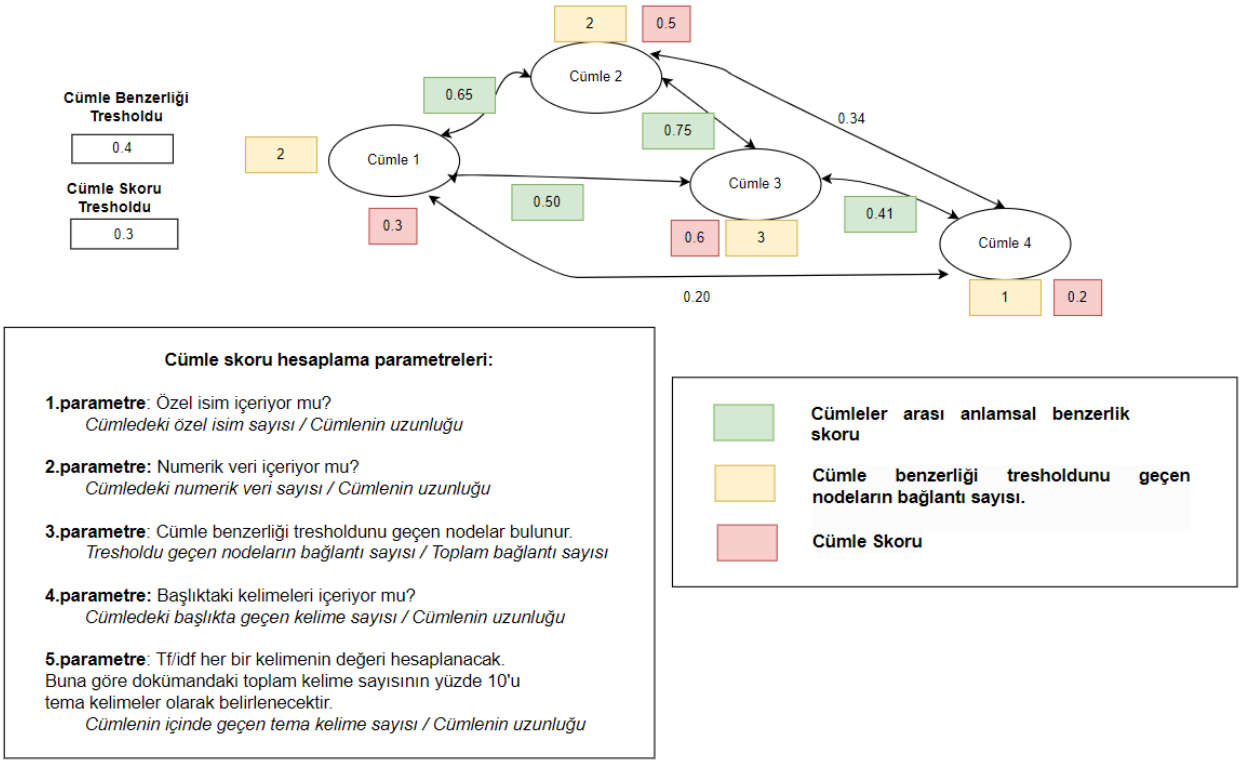
  Cümlenin içinde geçen tema kelime sayısı / Cümlenin uzunluğu

NOT: "TF-IDF, bir metinde belirli bir kelimenin ne kadar önemli olduğunu belirlemek için kullanılan bir istatistiksel yöntemdir. Bu, belirli bir kelimenin ne kadar yaygın olduğunu ve ne kadar nadir olduğunu dikkate alır.
Aşağıdaki formül kullanılarak TF-IDF değeri hesaplanabilir:

$$TF\text{-}IDF = TF \times IDF$$

Daha fazla bilgi için: https://mdurmuss.github.io/tf-idf-nedir/

Yukarıdaki parametrelerin hepsini kullanarak cümle skorlamak için bir algoritma geliştirmeniz beklenmektedir. Algoritma sonucunda her bir node un skoru oluşmalıdır. Şekil 2'de örnek bir graf yapısı gösterilmektedir.



Şekil 2. *Örnek graf yapısı*

Skorlara Göre Metin Özetleme Algoritmasının Geliştirilmesi

- Önemli cümleler üzerinden gidilerek özet çıkarılacaktır. Özet çıkarmada kullanılan bazı yöntemler şunlardır;
    - Cümle seçerek özetleme: Burada amaç metin içerisindeki önemli cümleleri puanlandırma yöntemleri kullanarak, istatistiksel metotlar ve sezgisel yaklaşımlar ile cümle seçmektir.
    - Yorumlayarak özetleme : Bu tip özetlemedeki amaç metin içerisindeki cümlelerin kısaltılmasıdır.

Projede cümle seçerek özetleme yapılmalıdır, yani "***var olan cümle yapısı bozulmadan cümleler seçilerek çıkarılıp özet elde edilecektir***". Oluşan node skorlarına göre node seçip bunlar ile özet oluşturacak bir metin özetleme algoritması geliştirmeniz beklenmektedir. Algoritmanızda metin özetlenirken hangi cümlelerin hangi sıra ile seçileceğini, cümle skorlarını kullanarak sizin belirlemeniz gerekmektedir. Oluşturulan özet arayüzde gösterilmelidir.

Özetleme Başarısının ROUGE Skoru ile Hesaplanması

- Algoritma sonucu oluşan Özet ile metnin gerçek özeti arasındaki benzerliği ROUGE skoru ile hesaplamalısınız."ROUGE" skoru, iki metnin benzerliğini ölçmek için kullanılır. Bu benzerlik genellikle referans metinde bulunan kelimelerin özetlenmiş metinde de bulunup bulunmadığına dayanır. Size verilen bir dokümanı özetlemeniz ve yine size verilecek gerçek özet ile karşılaştırmanız istenmektedir.

ÖDEV TESLİMİ

- Proje raporu Yazlab rapor formatında ve en az 4 sayfa uzunluğunda olmalıdır. Rapor; akış diyagramı veya yalancı kod içermeli, özet, giriş, yöntem, deneysel sonuçlar, sonuç ve kaynakça bölümünden oluşmalıdır.
- Dersin takibi projenin teslimi dâhil edestek.kocaeli.edu.tr sistemi üzerinden yapılacaktır. edestek.kocaeli.edu.tr sitesinde belirtilen tarihten sonra teslim edilen projeler kabul edilmeyecektir.
- Proje ile ilgili sorular edestek.kocaeli.edu.tr sitesindeki forum üzerinden Arş. Gör. Gamze Korkmaz Erdem veya Arş. Gör. Ayşe Gül Eker'e sorulabilir.
- Sunum tarihleri daha sonra duyurulacaktır.
- Sunum sırasında;
    - Algoritma, geliştirdiğiniz kodun çeşitli kısımlarının ne amaçla yazıldığı ve geliştirme ortamı hakkında sorular sorulabilir.
    - Kullandığınız algoritmaları birkaç cümle ile açıklayabilecek yeterlilikte olmanız beklenmektedir.
    - Kullandığınız herhangi bir satır kodu açıklamanız istenebilir.

Projenin tanıtım toplantısı 3 Mayıs Çarşamba günü saat 11:30'ta bölüm duyurularında ve e-destekte duyurulacak toplantı linki üzerinden online yapılacaktır.

Proje grupları en fazla 3 kişiden oluşmalıdır. Proje grup bilgileri e-destekte paylaşılacak link üzerinden en geç 6 Mayıs Cumartesi gününe kadar girilmelidir. Bu tarihten sonra gruplarda herhangi bir değişiklik yapılamayacaktır.

Sunumlar belirtilen tarihlerde yüzyüze alınacaktır.