KOCAELİ ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ YAZILIM

LAB. 2- 3. Proje

GRAF TABANLI METİN ÖZETLEME

Seda Nur Ekici-200201050

Begüm Erva Şahin-200201020

• Özet

Bu rapor Yazılım Laboratuvarı 2 Dersinin 3.projesini açıklamak ve sunumunu gerçekleştirmek amacıyla oluşturulmuştur. Raporda projenin tanımı, isterleri, yapım aşaması kullanılan araç ve yöntemler, kod parçacıkları vb. Bulunmaktadır. Proje aşamasında yararlanılan kaynaklar raporun son bölümünde bulunmaktadır.

• Giriş

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.

• İlerleyiş ve Yöntem

1. Başlamadan Önce

Bu projeye başlamadan önce proje dokümanında verilen bilgileri derleyip kullanabileceğimiz metodları araştırdık.

Kütüphaneden ve internet üzerinden yararlanabileceğimi kaynakları araştırdık ve Python ile yapmaya karar verdik.

2. Başlangıç

Projede masaüstü uygulama geliştirmemiz gerekmektedir. Masaüstü uygulamada ilk olarak doküman yükleme işlemi gerçekleştirilecektir. Ardından yüklenen dokümandaki cümleleri

graf yapısı haline getirmemiz ve bu graf yapısını görselleştirmemiz 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 bizim geliştirmemiz istenmektedir. Özet metni arayüzde sunmamız beklenmektedir. Sonuç olarak bize verilen bir metnin özetini bu yöntem ile çıkarmamız ve gerçek özet ile benzerliğini "ROUGE" skorlaması ile ölçmemiz istenmektedir.

3. İLERLEYİŞ

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 geliştirmemiz beklenmektedir. Arayüz aşağıdaki isterleri içermektedir: 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.

Dokümandaki cümleleri graf yapısına dönüştürmek için NetworkX kullandık. Bu Python programlama dili için açık kaynaklı bir graf kütüphanesidir. Düğümler ve kenarlar gibi grafik elemanlarını temsil etmek için birden fazla graf sınıfı sağlar

Cümlelere NLTK kütüphanesi kullanılarak aşağıdaki ön işleme adımları uygulanmıştır: 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.

İki cümle arasındaki anlamsal ilişkiyi kurmak için Word Embedding kullanılmıştır. 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. 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 sırasında aşağıdaki parametreler oluşturulmuştur: 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. Yukarıdaki parametrelerin hepsini kullanarak cümle skorlamak için bir algoritma geliştirmemiz beklenmektedir. Algoritma sonucunda her bir node un skoru oluşmalıdır. Önemli cümleler üzerinden gidilerek özet çıkarılacaktır. Cümle seçerek özetleme kullanılmıştır. Cümle seçerek özetleme: Burada amaç metin içerisindeki önemli cümleleri puanlandırma yöntemleri kullanarak, istatistiksel metotlar ve sezgisel yaklaşımlar ile cümle seçmektir.

Algoritma sonucu oluşan Özet ile metnin gerçek özeti arasındaki benzerliği ROUGE skoru ile hesaplamamız istenmiştir. "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 özetlememiz ve yine size verilecek gerçek özet ile karşılaştırmamız istenmektedir.

• *Son Söz*

Bu proje bize çok fazla bilgi birikimi sağladı ve farklı bir bakış

açısı kazandırdı. Projeyi yaparken çok fazla araştırma yapıp eksiklerimizi kısa zamanda tamamladık ve projede bizden istenenleri elimizden gelen en iyi şekilde yapmaya çalıştık.

• Kaynakça

https://miracozturk.com/python-kutuphaneleri-ve-ozellikleri/

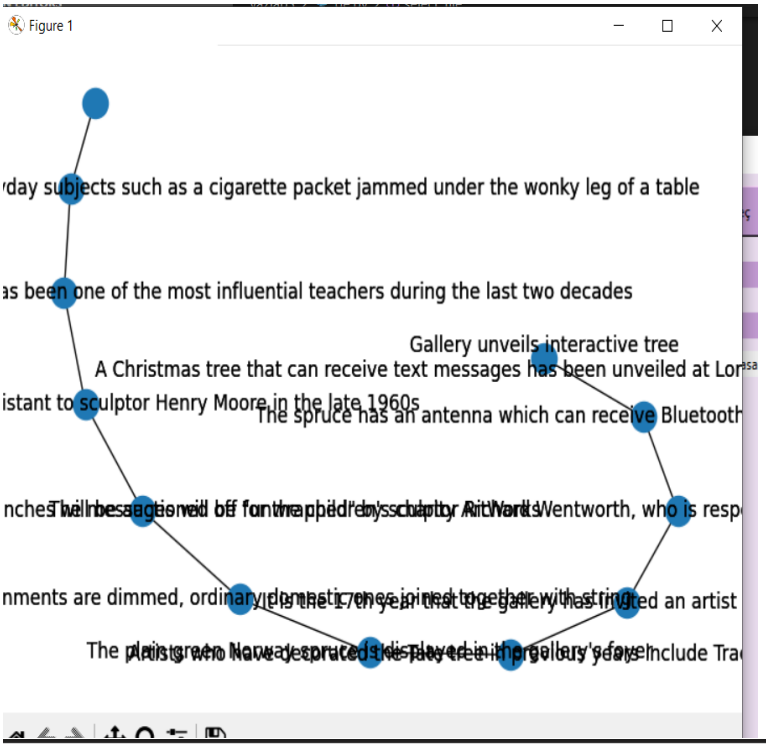
https://www.w3schools.com/python/numpy/numpy_intro.asp

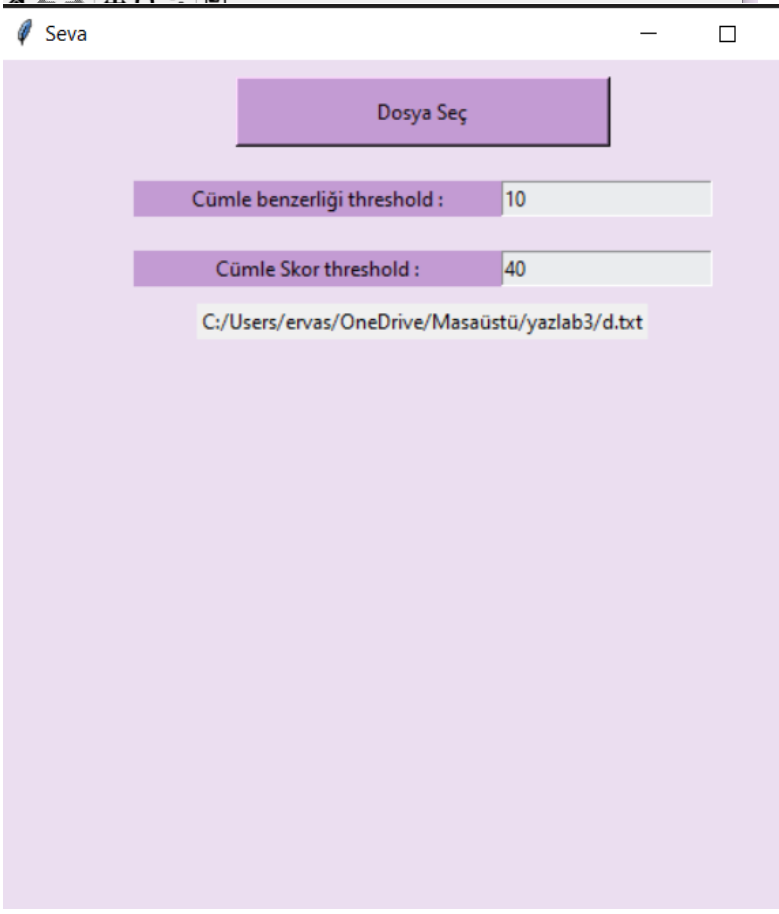
https://medium.com/@ilyaskaraca/python-gui-programlama-tkinter-d63a99b43179

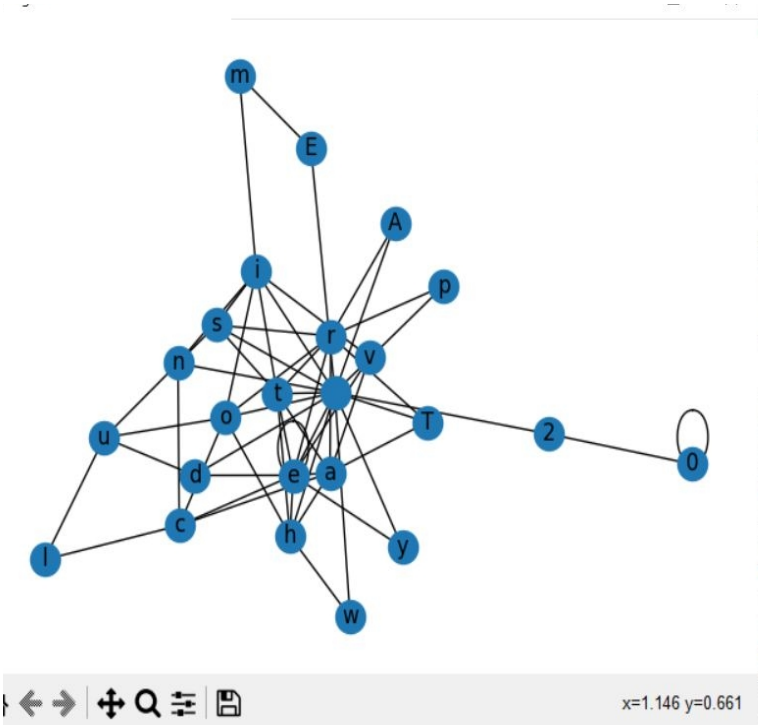
https://dergipark.org.tr/en/download/article-file/819661

https://learn.microsoft.com/tr-tr/azure/architecture/guide/ai/query-based-summarization

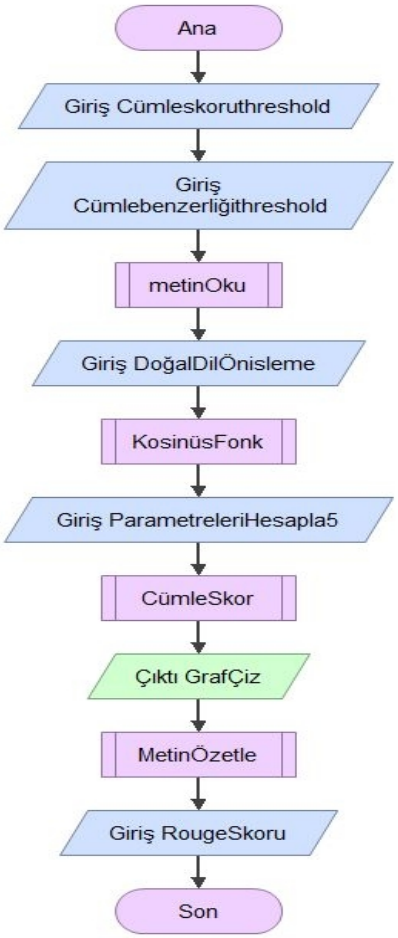
• Arayüz fotoğrafları







• *Akış Şeması*


Ana
Giriş Cümleskoruthreshold
Giriş Cümlebenzerliğithreshold
metinOku
Giriş DoğalDilÖnisleme
KosinüsFonk
Giriş ParametreleriHesapla5
CümleSkor
Çıktı GrafÇiz
MetinÖzetle
Giriş RougeSkoru
Son