Graf Tabanlı Metin Özetleme Projesi

Kocaeli Üniversitesi Bilgisayar Mühendislik Bölümü Yazılım Laboratuvarı
II Proje 3

200201137 Marah Alasi — 190201140 Mohamed Helwa — 200201147
Muhammad Abdan SYAKURA

26.05.2023

Özet

Bu proje, gelişmiş bir metin özetleme sürecini desteklemek amacıyla oluşturulan bir masaüstü uygulamanın geliştirilmesini içerir. Metin özetleme, bilgiyi yoğunlaştırma ve kullanıcıların karmaşık metinlerden hızlı bir şekilde anlamlı bilgi almasına yardımcı olma gibi hayati avantajlara sahiptir. Bu özelleştirilmiş uygulama, metinlerin daha anlaşılabilir ve öz bir biçimde özetlenmesine imkan sağlar. Projenin ana bileşeni, çeşitli dil işleme tekniklerinin ve modern algoritma temelli yaklaşımların kullanıldığı bir anlamsal ilişkili graf tabanlı özetleme yöntemidir. Bu yaklaşım, cümleler arası ilişkileri daha etkin bir şekilde anlamlandırmak ve böylece daha kaliteli metin özetleri oluşturmak için kullanılmaktadır.

---

1 Giriş

Dijital çağda bilgiye hızlı ve etkili bir şekilde ulaşmanın önemi tartışılmaz bir gerçek. Bilgi miktarının hızla artması, özellikle geniş kapsamlı metinlerin değerlendirilmesi ve anlaşılmasını zorlaştırıyor. Bu nedenle, kullanıcıların metinlerden çıkarılabilecek temel bilgilere daha hızlı bir şekilde ulaşabilmelerini sağlayacak etkin metin özetleme yöntemlerine olan ihtiyaç artmaktadır. Bu çalışmanın amacı, bu ihtiyaca yanıt vermek ve metin özetleme süreçlerini geliştirmektir.

Projemiz, karmaşık metinlerin anlamlı ve öz bir biçimde özetlenmesini sağlayan bir masaüstü uygulamanın geliştirilmesi üze-

rine odaklanmaktadır. Bu, çeşitli dil işleme tekniklerini ve modern algoritmaları içeren bir anlamsal ilişkili graf tabanlı özetleme yöntemi kullanılarak gerçekleştirilmektedir. Bu metodoloji, metinlerin derinlemesine analizini sağlar ve cümleler arası ilişkileri görselleştirmek için bir grafik yapısına dökülmesine olanak tanır. Bu yapı, metni özetlemek için en uygun cümlelerin belirlenmesine yardımcı olur. Bu sayede, metinlerden çıkarılabilecek en önemli bilgilerin etkin bir şekilde sunulması amaçlanmaktadır.

Bu proje, teknolojinin dil işleme yeteneklerini daha da ileriye taşımayı ve metin özetleme süreçlerini daha etkin ve kullanıcı dostu hale getirmeyi hedeflemektedir.

2 Yöntem

Aşağıdaki akış diyagramı ve yalancı kod, projenin genel yapısını ve işleyişini göstermektedir:

Masaüstü Arayüzü Geliştirilmesi ve Graf Yapısının Oluşturulması Bu aşamada Python'un etkileşimli masaüstü uygulama geliştirme araçlarından biri olan PyQt5 kullanıldı. Graf oluşturma ve görselleştirme işlemleri için ise NetworkX ve Matplotlib kütüphaneleri tercih edildi. Bu kütüphaneler sayesinde hem graf verisi oluşturulabildi, hem de bu veri kullanıcı dostu bir görsel formatla sunulabildi.

Cümleler Arası Anlamsal İlişkinin Kurulması Ön işlem adımlarında NLTK kütüphanesi kullanıldı ve Tokenization, Stemming, Stop-word Elimination, Punctuation gibi işlemler gerçekleştirildi.

- 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.

Anlamsal ilişkileri belirlemek için Word Embedding ve BERT modelleri kullanıldı. Bu modellemeler sonucunda her bir cümlenin vektör temsili elde edildi ve bu vektörler üzerinden cümleler arasındaki benzerlik kosinüs benzerliği metodu ile ölçüldü.

Cümle Skoru Hesaplama Algoritması Her bir cümle için bir skor hesaplandı. Skor hesaplama algoritması:

- (P1): özel isim kontrolü
- (P2): numerik veri kontrolü
- (P3): benzerlik threshold'unu geçen node'ların bulunması
- (P4): başlıktaki kelimelerin kontrolü
- (P5): TF-IDF değerinin hesaplanması

parametreleri içermektedir.

Metin Özetleme Algoritması Metni özetlemek için cümle seçerek özetleme yöntemi kullanıldı. Bu yöntem, skorları kullanarak önemli cümleleri belirler ve özeti oluşturur. Bu özet, masaüstü arayüzde son kullanıcıya sunulur.

Özetlemenin Başarısının ROUGE Skoru İle Hesaplanması Son aşamada, özetleme başarısı ROUGE skorlaması ile ölçüldü.

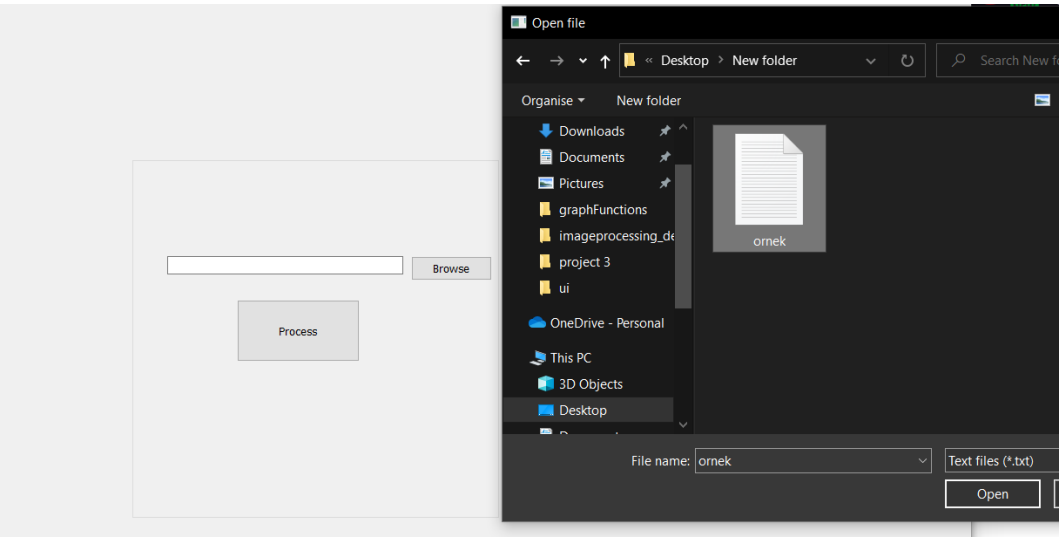
Bu skorlama, oluşturulan özeti, dokümanın gerçek özeti ile karşılaştırarak bir başarı oranı sunar.

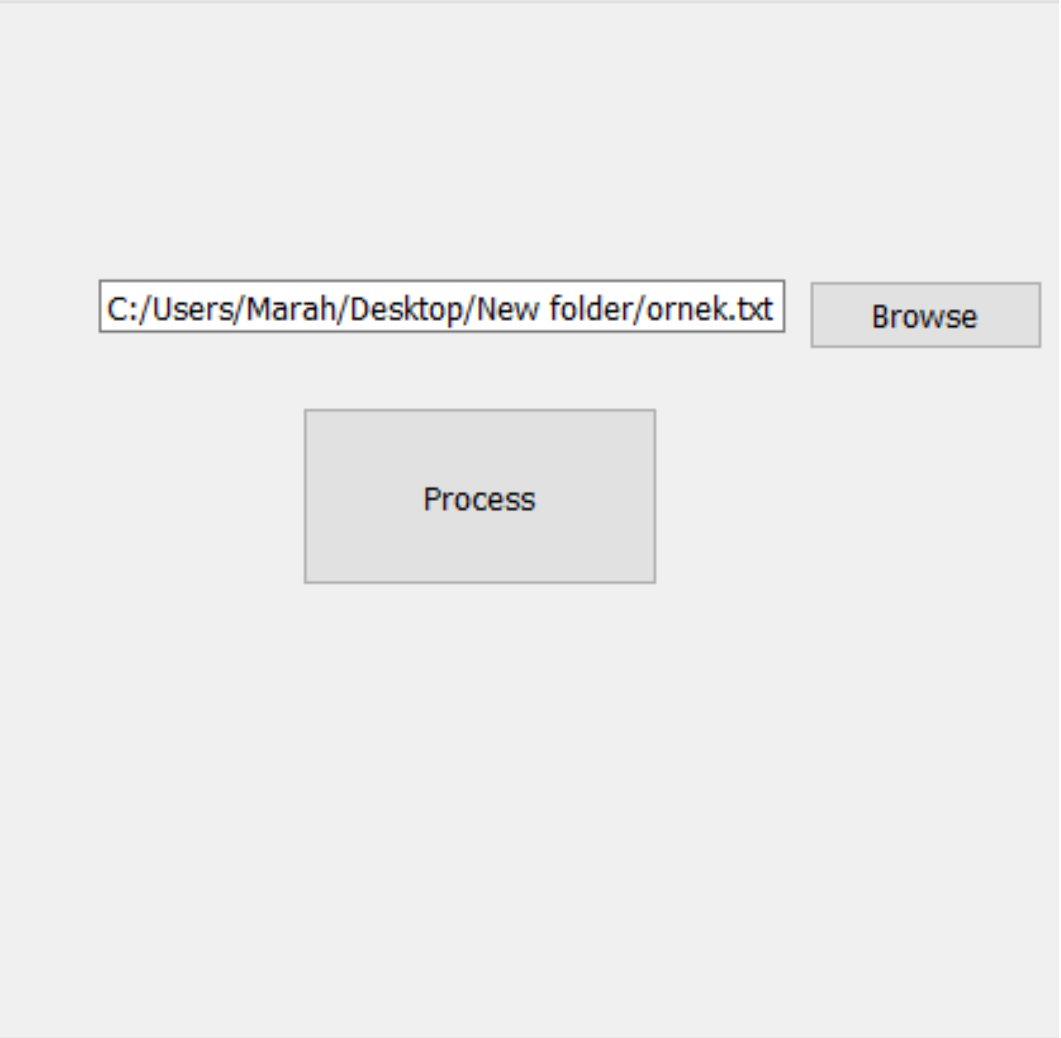
3 Sonuç

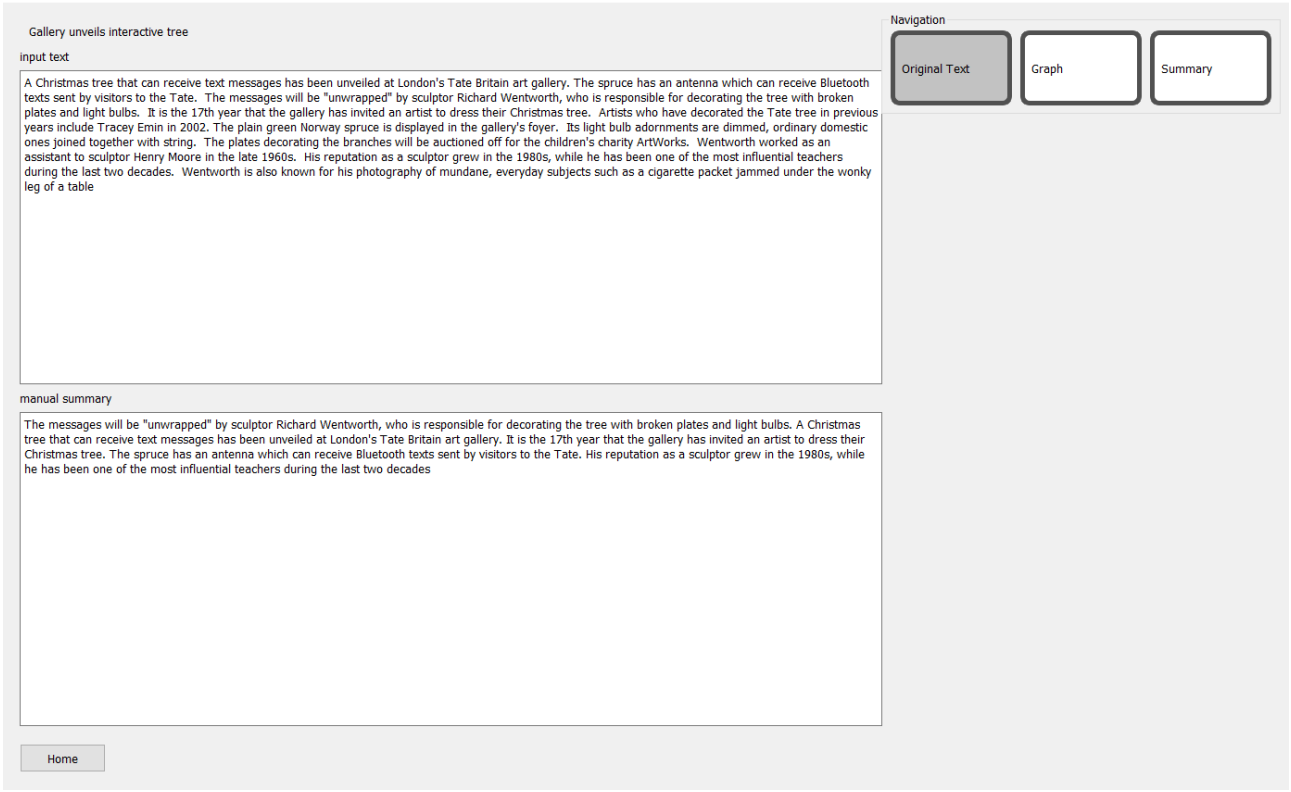
Bu proje, karmaşık bir dokümanın özetini çıkarma yeteneği olan bir masaüstü uygulamanın geliştirilmesini içerir. Bu, metin analizi ve doğal dil işleme tekniklerinin birleştirilmesiyle gerçekleştirilmiştir. Projedeki başarı, daha etkili bir metin özeti çıkarılması ve bu sayede kullanıcıların zaman ve kaynak tasarrufu sağlamasına yardımcı olmuştur.

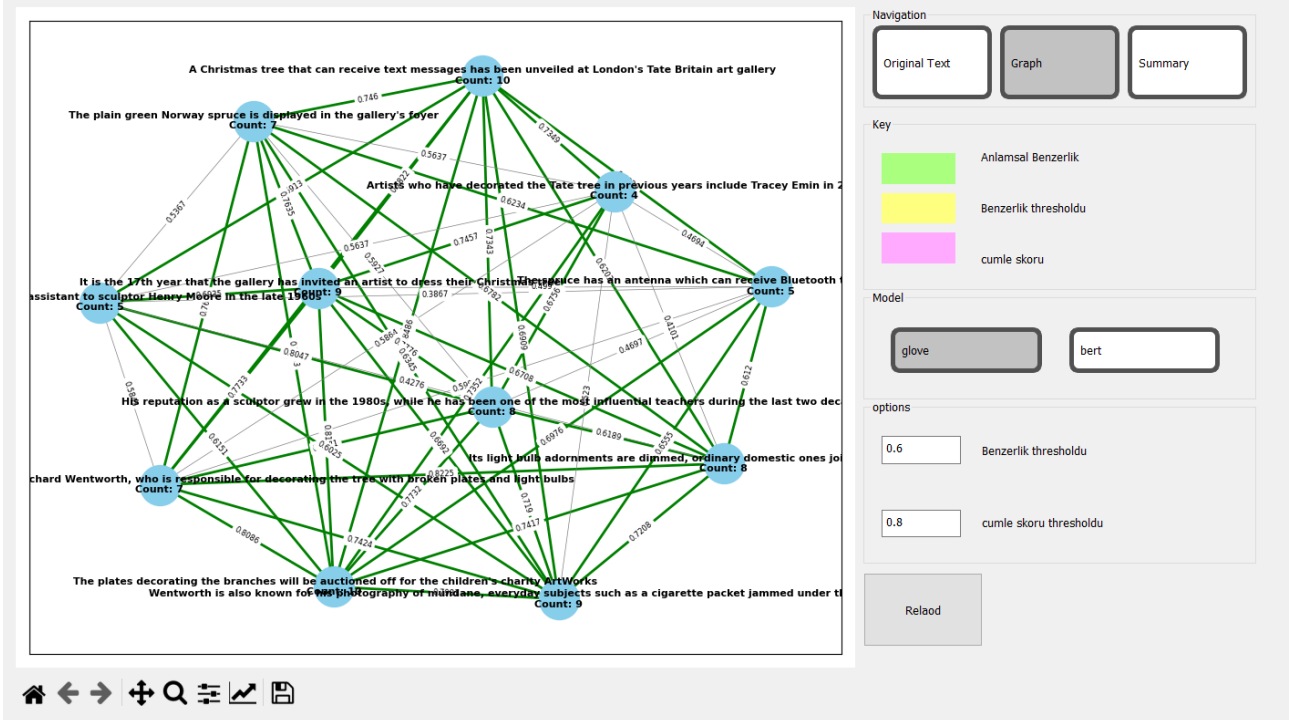
4 Kaynakça

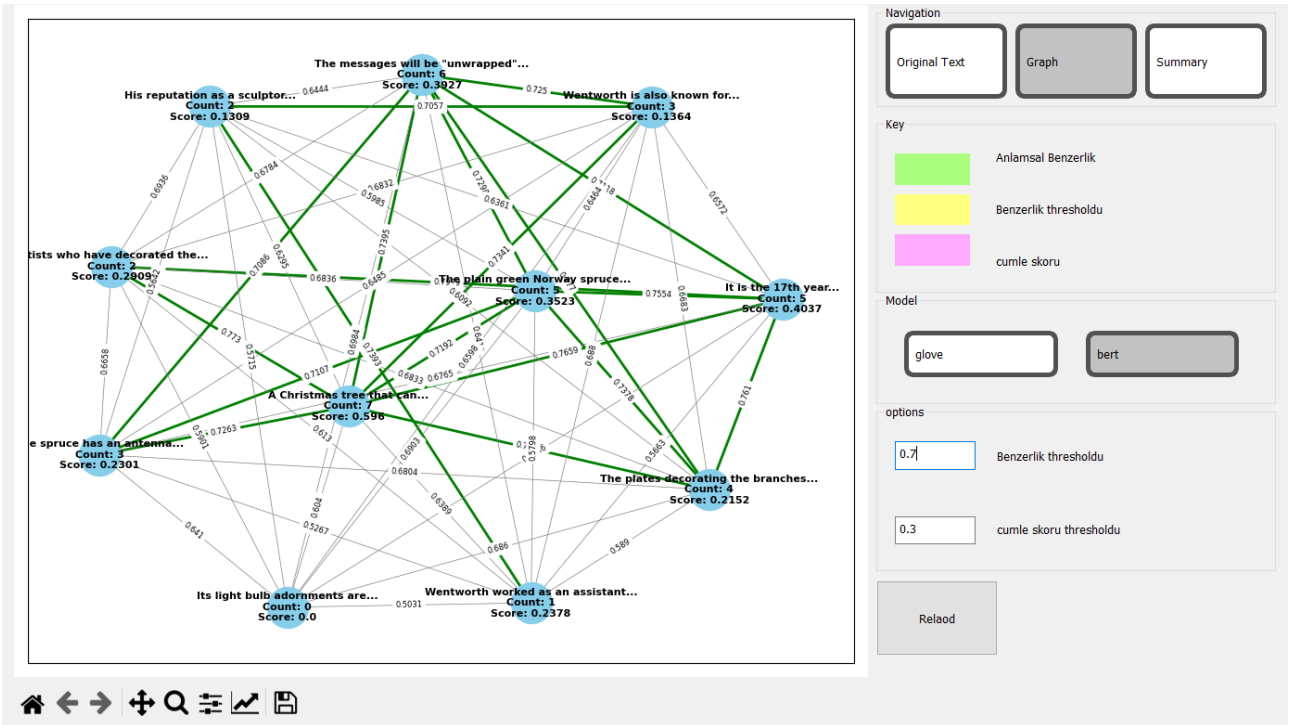
- https://www.kaggle.com/code/awadhi123/text-preprocessing-using-nltk
- https://machinelearningknowledge.ai/11-techniques-of-text-preprocessing-using-nltk-in-python/
- https://towardsdatascience.com/stemming-lemmatization-what-ba782b7c0bd8
- https://www.kaggle.com/code/adepvenugopal/nlp-text-similarity-using-glove-embedding/notebook
- https://turbolab.in/better-word-embeddings-using-glove/
- https://www.sbert.net/docs/usage/semantic_textual_similarity.html

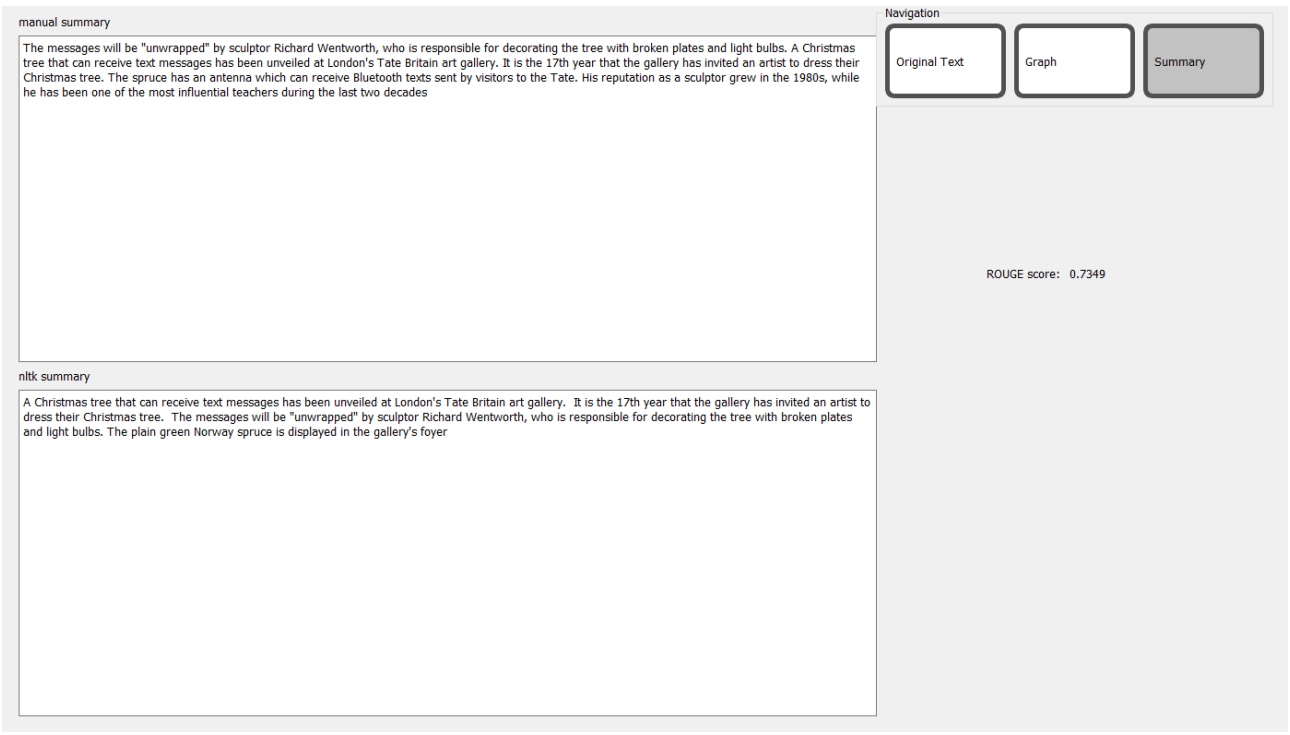

Browse
Process
Open file
Desktop
New folder
Search New f
Organise
New folder
Downloads
Documents
Pictures
graphFunctions
imageprocessing_de
project 3
ui
OneDrive - Personal
This PC
3D Objects
Desktop
ornek
File name:
ornek
Text files (*.txt)
Open


C:/Users/Marah/Desktop/New folder/ornek.txt
Browse
Process


Gallery unveils interactive tree
input text
A Christmas tree that can receive text messages has been unveiled at London's Tate Britain art gallery. The spruce has an antenna which can receive Bluetooth texts sent by visitors to the Tate. The messages will be "unwrapped" by sculptor Richard Wentworth, who is responsible for decorating the tree with broken plates and light bulbs. It is the 17th year that the gallery has invited an artist to dress their Christmas tree. Artists who have decorated the Tate tree in previous years include Tracey Emin in 2002. The plain green Norway spruce is displayed in the gallery's foyer. Its light bulb adornments are dimmed, ordinary domestic ones joined together with string. The plates decorating the branches will be auctioned off for the children's charity ArtWorks. Wentworth worked as an assistant to sculptor Henry Moore in the late 1960s. His reputation as a sculptor grew in the 1980s, while he has been one of the most influential teachers during the last two decades. Wentworth is also known for his photography of mundane, everyday subjects such as a cigarette packet jammed under the wonky leg of a table
manual summary
The messages will be "unwrapped" by sculptor Richard Wentworth, who is responsible for decorating the tree with broken plates and light bulbs. A Christmas tree that can receive text messages has been unveiled at London's Tate Britain art gallery. It is the 17th year that the gallery has invited an artist to dress their Christmas tree. The spruce has an antenna which can receive Bluetooth texts sent by visitors to the Tate. His reputation as a sculptor grew in the 1980s, while he has been one of the most influential teachers during the last two decades
Home
Navigation
Original Text
Graph
Summary


A Christmas tree that can receive text messages has been unveiled at London's Tate Britain art gallery
Count: 10
The plain green Norway spruce is displayed in the gallery's foyer
Count: 7
Artists who have decorated the Tate tree in previous years include Tracey Emin in 2
Count: 4
Count: 5
Count: 8
Count: 9
Navigation
Original Text
Graph
Summary
Key
Anlamsal Benzerlik
Benzerlik thresholdu
cumle skoru
Model
glove
bert
options
0.6
Benzerlik thresholdu
0.8
cumle skoru thresholdu
Relaod


The messages will be "unwrapped"...
Count: 6
Score: 0.3927
His reputation as a sculptor...
Count: 2
Score: 0.1309
Wentworth is also known for...
Count: 3
Score: 0.1364
tists who have decorated the...
Count: 2
Score: 0.2809
The plain green Norway spruce...
Count: 3
Score: 0.3523
It is the 17th year...
Count: 5
Score: 0.4037
A Christmas tree that can...
Count: 7
Score: 0.596
e spruce has an antenna...
Count: 3
Score: 0.2301
The plates decorating the branches...
Count: 4
Score: 0.2152
Its light bulb adornments are...
Count: 0
Score: 0.0
Wentworth worked as an assistant...
Count: 1
Score: 0.2378
Navigation
Original Text
Graph
Summary
Key
Anlamsal Benzerlik
Benzerlik thresholdu
cumle skoru
Model
glove
bert
options
0.7
Benzerlik thresholdu
0.3
cumle skoru thresholdu
Relaod


manual summary
The messages will be "unwrapped" by sculptor Richard Wentworth, who is responsible for decorating the tree with broken plates and light bulbs. A Christmas tree that can receive text messages has been unveiled at London's Tate Britain art gallery. It is the 17th year that the gallery has invited an artist to dress their Christmas tree. The spruce has an antenna which can receive Bluetooth texts sent by visitors to the Tate. His reputation as a sculptor grew in the 1980s, while he has been one of the most influential teachers during the last two decades
nltk summary
A Christmas tree that can receive text messages has been unveiled at London's Tate Britain art gallery. It is the 17th year that the gallery has invited an artist to dress their Christmas tree. The messages will be "unwrapped" by sculptor Richard Wentworth, who is responsible for decorating the tree with broken plates and light bulbs. The plain green Norway spruce is displayed in the gallery's foyer
Navigation
Original Text
Graph
Summary
ROUGE score: 0.7349

Algorithm 1

```
function COMPUTE_SENTENCE_SIMILARITY_GRAPH
    G ← nx.Graph()
    for sentence in sentences do
        G.add_node(sentence)
    end for
    n_sentences ← len(sentences)
    similarity_matrix ← np.zeros((n_sentences, n_sentences))
    for i in range(n_sentences) do
        for j in range(n_sentences) do
            if i = j then
                similarity_matrix[i, j] ← 1
            else if i < j then
                sentence1 ← sentences_i
                sentence2 ← sentences_j
                if method = "bert" then
                    if "bert_embedding" not in G.nodes[sentence1] then
                        G.nodes[sentence1]["bert_embedding"] = sentence_vector_bert(
                        sentence1, model, tokenizer).reshape
                    end if
                    if "bert_embedding" not in G.nodes[sentence2] then
                        G.nodes[sentence2]["bert_embedding"] ← sentence_vector_bert(
                        sentence2, model, tokenizer).reshape
                        similarity ← round(1 - cosine(G.nodes[sentence1]["bert_embedding"],
                        G.nodes[sentence2]["bert_embedding"]), 4)
                    end if
                else if method = "glove" then
                    if "glove_embedding" not in G.nodes[sentence1] then
                        G.nodes[sentence1]["glove_embedding"]← sentence_vector_glove(
                        sentence1, model)
                    end if
                    if "glove_embedding" not in G.nodes[sentence2] then
                        G.nodes[sentence2]["glove_embedding"] = sentence_vector_glove(
                        sentence2, model)
                    end if
                    similarity ← round(1 - cosine(G.nodes[sentence1]["glove_embedding"],
                    G.nodes[sentence2]["glove_embedding"]), 4)
                    similarity_matrix[i, j] ← similarity_matrix[j, i] = similarity
                    G.add_edge(sentence1, sentence2, similarity←similarity)
                end if
            end if
        end for
    end for
    return similarity_matrix, G
end function
```