SORGU TABANLI METİN ÖZETLEME

1. GİRİŞ

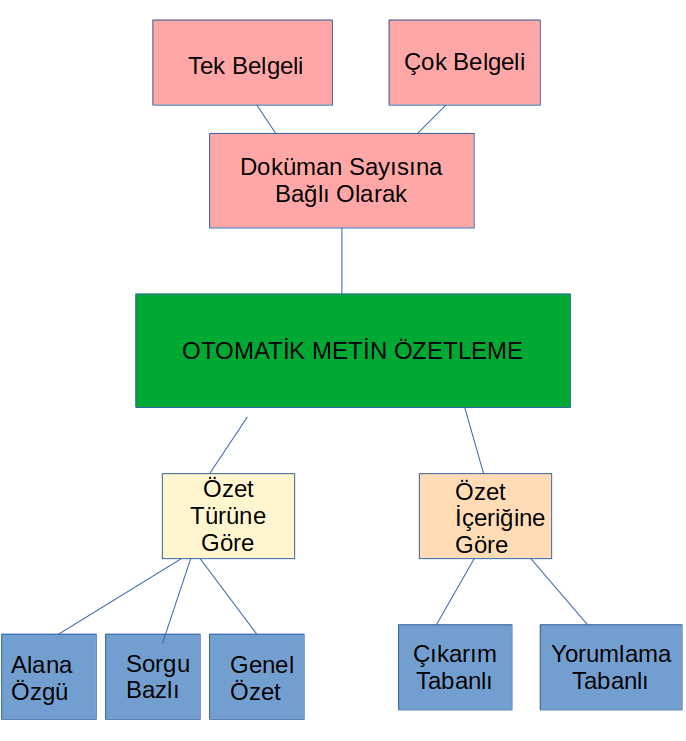
Metin özetleme; farklı kaynaklardan elde edilmiş metinlerin kullanılan algoritmaya göre otomatik olarak özetlenmesi, metnin içindeki önemli bilgileri elde etmek için metni kısaltma işlemidir. Bilgi işlem sistemlerinin hızla değişimi ve bu kaynaklardan elde edilen verinin logaritmik olarak artması sonucunda elde edilen veriler devasa büyüklüklere gelmiştir. Elde edilen verilerden anlamlı bilgilerin çıkarılması amacıyla birçok model, algoritma vs. gibi yöntemler geliştirilmiştir. Araştırmacılar için Yapay Zekâ (YZ)’nın bir alt kolu olan DDİ (Doğal Dil İşleme) alanı, içerisinde güncel ve dinamik konuları barındıran heyecan verici bir çalışma alanıdır. Aynı zamanda YZ’nın en zorlu problemlerini barındıran alanlarından bir tanesidir. Bu durum DDİ’nin araştırmacıların ilgi odağı olmasını sağlamaktadır. DDİ problemlerinin çözülebilmesi için; Derin Öğrenme (DÖ), Makine Öğrenmesi (MÖ), İstatiksel Analiz (İA), Kural Tabanlı Yaklaşımlar (KTY) gibi birçok yöntem ve metot kullanılmaktadır.

İnternetin gelişmesiyle birlikte, bilgi her zamankinden daha hızlı yayılıyor. İnternette yayılan bilgiler arasında, bilgilerin çok büyük bir kısmı metin olarak paylaşılmaktadır. Yararlı bilgilerin özetlenmesi ve tanımlanması için Metin Özetleme (MÖ) yöntemleri internet kullanıcılarına fayda sağlayabilir (Yu, 2022). Bu çalışma ile bir MÖ yöntemi olan Sorgu Tabanlı Metin Özetleme yöntemini inceleyeceğiz.

1. OTOMATİK METİN ÖZETLEME

İnternetin ve teknolojinin hızla gelişmesiyle elde edilen verilerin katlanarak arttığını belirtmiştik. Bu verilerin büyük çoğunluğunu ise metin verileri oluşturmaktadır. Çok büyük belgelerin insanlar tarafından okunması, yorumlanması, özetlenmesi uzun zaman alması nedeniyle metin özetleme çalışmaları sürekli olarak güncel kalan araştırma konusudur. Metin özetlemenin temel amacı kaynak doküman içerisindeki önemli bilgileri elde ederek, kaynak dokümandan daha küçük boyutta doküman elde etmektir. Joshi otomatik metin özetleme çalışmasını, büyük hacimdeki dokümanların daha küçük hacimli bir türünü elde etmek için gerçekleştirilen Doğal Dil İşleme dalı olarak tanımlanmıştır (Joshi et al., 2019).

Otomatik metin özetlemeyi amaçları ve yöntemlerine göre kategorilere ayıracak olursak Şekil-1’i inceleyebiliriz.



**Şekil 1.** Otomatik Metin Özetleme Türleri

Bu makale ile sorgu tabanlı metin özetleme yöntemine yakından bakacağız. Sorgu tabanlı metin özetlemede insanlar tarafından algoritmaya iletilen sorgular yönünde metin özetleme yapılmaktadır. Bu sorgular kalıplar halinde önceden hazırlanmış sorgular olabileceği gibi, kullanıcılar tarafından özel olarak da hazırlanan sorgular olabilir.

1. SEMANTİK AĞ’LAR HAKKINDA

İyi bir sorgu tabanlı metin özetleme sağduyu bilgisine dayanmaktadır. Semantik ağ, birbirine bağlı düğümlerin ve yolların bulunduğu bir ağdaki kavramlar arasındaki anlamsal ilişkileri temsil eder. Nesne kavramı, düğümleri temsil eder ve yol, iki kavram ve nesne arasındaki farklı ilişkileri verir. Semantik ağda farklı kavramlarla farklı ilişkiler mevcuttur. Semantik ağı doğal dil işlemeyle bütünleştirerek insan benzeri sağduyu bilgisini elde edebiliriz. Semantik ağ, sözcüklerden çok kavramlara daha fazla önem verilen daha zengin anlamsal ilişkiler içerir. Semantik ağ, iddiaları ve kavramları içerir. İddia kenar ile temsil edilebilir. Bazen birden fazla kenar aynı iddiayı belirtir. Her bir kenarın iki kavram arasındaki ilişkiyi temsil ettiği bir hiper grafik olarak temsil edilir. Bu kavramlar, yarı yapılandırılmış İngilizce parçalardır ve sorgulama genişletme ve anlamsal benzerlik ölçümü için kullanılabilir. Sorguya göre ilgili cümleleri bulmak amacıyla, giriş metni ile sorgu arasındaki anlamsal ilişki puanını bulmak için farklı ilişkilerin kullanıldığı Word-Net kullanılır. WordNet, iki içerik sözcüğü arasında bir dizi ilişkinin bulunduğu bir sözlüktür (WordNet: A Lexical Database for English George A. Miller, n.d.). Anlamsal ilişkililik ölçüsü, iki kelimenin ne kadar ilişkili olduğunu tanımlar. İçerik sözcükleri (isim, fiil, zarf ve sıfat) daha alakalı bilgi taşıdıkları için ilişkililik puanını bulmak için kullanılır. Aralarında herhangi bir ilişki varsa, iki kelime anlamsal olarak ilişkili olarak söylenebilir (Rahman & Borah, 2020).

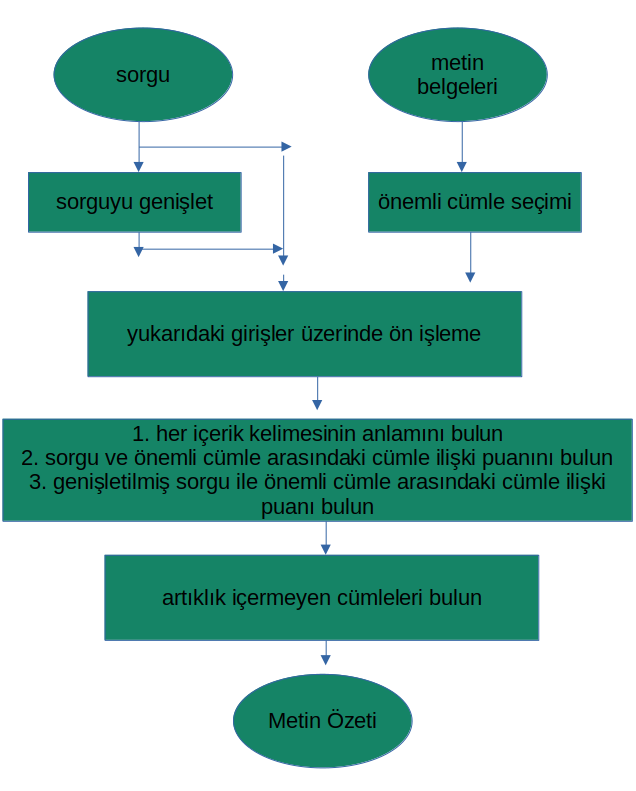
1. SORGU TABANLI METİN ÖZETLEME

Büyük miktardaki metin verisinden yararlı bilgileri özetlemek ve belirlemek kullanıcılar için zorlu bir iş olabilir. Sorgu tabanlı metin özetleme, kullanıcı tanımlı sorgu bilgilerini kılavuz olarak kullanan ve otomatik metin özetleme sürecine müdahale eden bir yöntemdir. Sorgu tabanlı, aynı zamanda konu- tabanlı, kullanıcı odaklı veya sorgu odaklı metin özetleme, kullanıcılar tarafından sağlanan sorgu bilgilerinden yararlanan bir metin özetleme görevi olarak kabul edilebilir. Genel metin özetleme görevleri için amaç, verilen belgelerin genel bir özetini elde etmektir. Sorguya verilen yanıt, hem belgelerdeki sorguyla ilgili önemli noktaları kapsayan hem de belgeleri sıkıştıran metinleri çıkarır veya oluşturur. Bu anlamda, metin özetleme üç alt sınıfa sahip daha yüksek bir hiyerarşi olarak kabul edilebilir: genel metin özetleme, gerçek metin özetleme (soru-cevaplama olarak da bilinir) ve sorgu tabanlı metin özetleme (Yu, 2022b). Ancak bazen soru cevaplama, bir tür sorgu tabanlı özetleme olarak da kabul edilebilir.

Metin özetleme açısından sorgu tabanlı metin özetleme, sorgu tabanlı çıkarımsal ve soyutlayıcı metin özetleme olarak ikiye ayrılabilir. Kullanıcı tarafından sağlanan sorgu bilgileri göz önüne alındığında, sorgu tabanlı metin özetleme, kelime sorgusu ve cümle sorgusu olarak ikiye ayrılabilir (Yu, 2022b). Ayrıca, bu konudaki bazı çalışmalar tek belge özetlemeye odaklanırken, bazı çalışmalar çoklu belge özetlemeye odaklanmaktadır (Rahman & Borah, 2020).

Sorgu tabanlı metin özetlemede özetlenen metni sorguya göre buluruz. Girdiler, sorgu ve metin belgeleridir ve çıktı, özetlenmiş metindir. Özetleme, sorgu cümlesiyle anlamsal olarak ilişkili olan cümleleri seçer. Anlamsal ilişki puanı, bir sorgu cümlesi ile bir giriş metni cümlesi arasında hesaplanır. Girilen her metin cümlesinin puanını bulmak için, sorgu ile her cümle için anlamsal ilişkililik puanı hesaplanır. Bunun için yalnızca sorguda bulunan içerik sözcükleri dikkate alınır. Özet için cümleler seçilirken aynı bilginin birden fazla cümle ile aktarılma olasılığı yüksektir. Gereksiz tümceleri ortadan kaldırmak, özet boyutunu azaltmak için metin özetlemede önemli bir görevdir. Sorgu ve giriş metni cümlesi arasındaki anlamsal ilişki puanını bulmadan önce, giriş metni belgesinden önemli cümleleri çıkarılır. Aslında, önemli cümleleri ayıklamak, özet için zengin bilgi bulmaya yardımcı olur. Önemli cümleler, daha önemli bilgiler taşıyan cümlelerdir. Sorgu ve giriş metni kelimeleri arasındaki anlamsal ilişki puanını bulurken, sorgu kelimeleri veya cümleleri ile ek kelimeler veya kelime öbekleri eklenir. Bu, sorgu genişletme olarak bilinir. Sorgu genişletmede girdi, sorgu sözcükleri veya tümcecikleridir ve çıktı, ilgili sözcükler veya tümceciklerdir. Genişleyen sorgu sözcükleri, kullanıcının ihtiyacına göre daha yararlı ve temel cümlelerin çıkarılmasına yardımcı olur. Anlamsal ilişki puanını bulmadan önce, kelimenin cümlenin belirli bağlamına uygun olacak doğru anlamını bulunur. Böylece, kelime anlamı belirsizliği giderme, cümlenin bağlamına göre uygun anlamı bulmak için buraya gelir. (Rahman & Borah, 2020).

Sorgu tabanlı çoklu belge özetlemenin genel diyagramı Şekil 2 de verilmiştir.



**Şekil 2.** Sorgu Tabanlı Çoklu Belge Özetlemenin Diyagramı

Sözcük anlam ayrımını kullanarak sorgu tabanlı metin özetlemenin iyileştirilmesi konusunda semantik ağlar kullanıldığından dolayı, semantik ağ’lar konusuna yüzeysel olarak değindik. Biz sorgu tabanlı metin özetleme görevimizde embedding matrix’leri kullanacağız.

1. EMBEDDİNG MATRİX

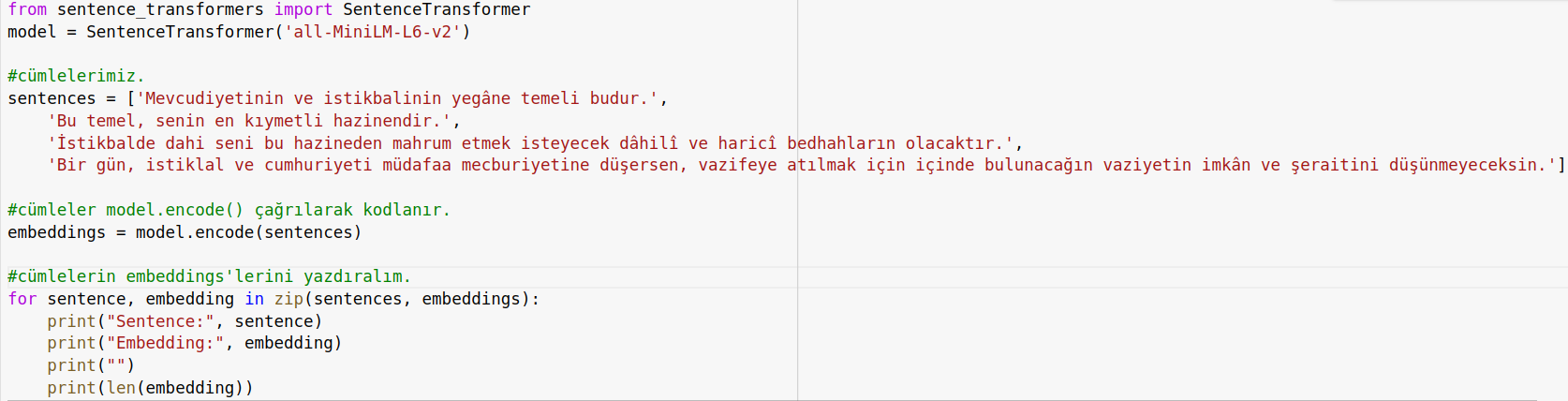
DDİ’yi insan dilinin bilgisayar sistemleri tarafından anlaşılmasını sağlamak amacıyla yapılan çalışmaları içeren YZ’nın bir alt dalı olduğunu belirtmiştik. DDİ tıpkı Konvolüsyonel Sinir Ağları (KSA) gibi özellik çıkartmayı hedefler. Nasıl KSA’lar resimlerdeki önemli özellikleri çıkartmayı hedefliyorsa DDİ modelleri de cümlelerden özellik çıkartmaya çalışır.

Embedding matrix kelimelerin tek tek gerçek sayısal değer içeren vektörler olarak ifade edilmesidir. Yani bizim derin öğrenme problemlerimizin çözümü için string değil int veri tiplerine ihtiyacımız vardı. İşte embedding matrix tam olarak kelimeleri sayısal vektörlere dönüştürüyor. Ayrıca kelime uzayında birbiri ile aynı anlamı taşıyan kelimelerin birbirine yakın ve aynı yönde olduğunu söylüyor. Bu durumda bir cümlenin anlam bütünlüğünü bozmadan anlamamız için embedding matrix’ler bizim için çok önemlidir. Esasen KSA ‘nda amacımız filtreleri öğrenmek ise, DDİ projelerinde amacımız embedding matrix’ini öğrenmektir desek yanlış olmayacaktır. Embedding matrix’leri kendimizde DDİ ağımızı eğiterek elde edebiliriz. Ancak State Of The Art denilen başarışı kanıtlanmış ağların embedding matrix’lerini kullanmak bizim için çok daha faydalı olacaktır.

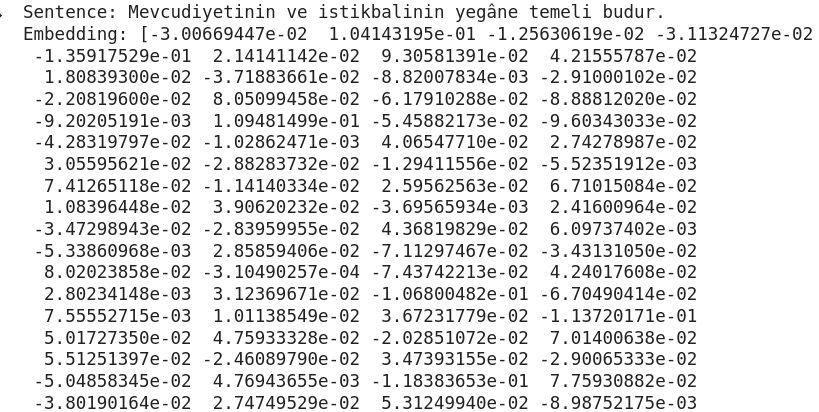
1. DENEYLER

Sorgu tabanlı dokümantasyon özetlemenin nasıl yapılacağına dair çok basit bir örnek yapacağız. Bu uygulama kullanıcıdan bir sorgu almak ve ana belgemizde bu sorgunun yanıtını en iyi temsil eden grupları ya da cümleleri bulmayı hedefleyeceğiz. Bunun için Sentence Transformers’da bulunan soru/cevap alma özelliğini kullanacağız. Sentence Transformers içinde bulunan modeller 1 milyardan fazla eğitim çifti ile eğitilmiştir. Bu çalışmamızda nq-distilbert-base-v1 adında önceden eğitilmiş bir cümle dönüştürücü modeli kullanacağız. Bu model cümleleri 768 (bu düzleştirilmiş bir vektördür, aslında her cümle için 384\*384’lük bir embeddings matrix elde ederiz) boyutlu yoğun bir vektör uzayına eşler. Kümeleme veya semantik arama gibi görevler için kullanılabilir.

Şekil 3 de verdiğimiz kod incelendiğinde SentenceTransformer içerisinden indirdiğimiz eğitimli modelimiz, input olarak verdiğimiz cümle için tamamen düzleştirilmiş 768 boyutlu bir vektör elde edecektir.

**Şekil 3.** Girilen Cümleden Embeddings Matrix’in Oluşturulması

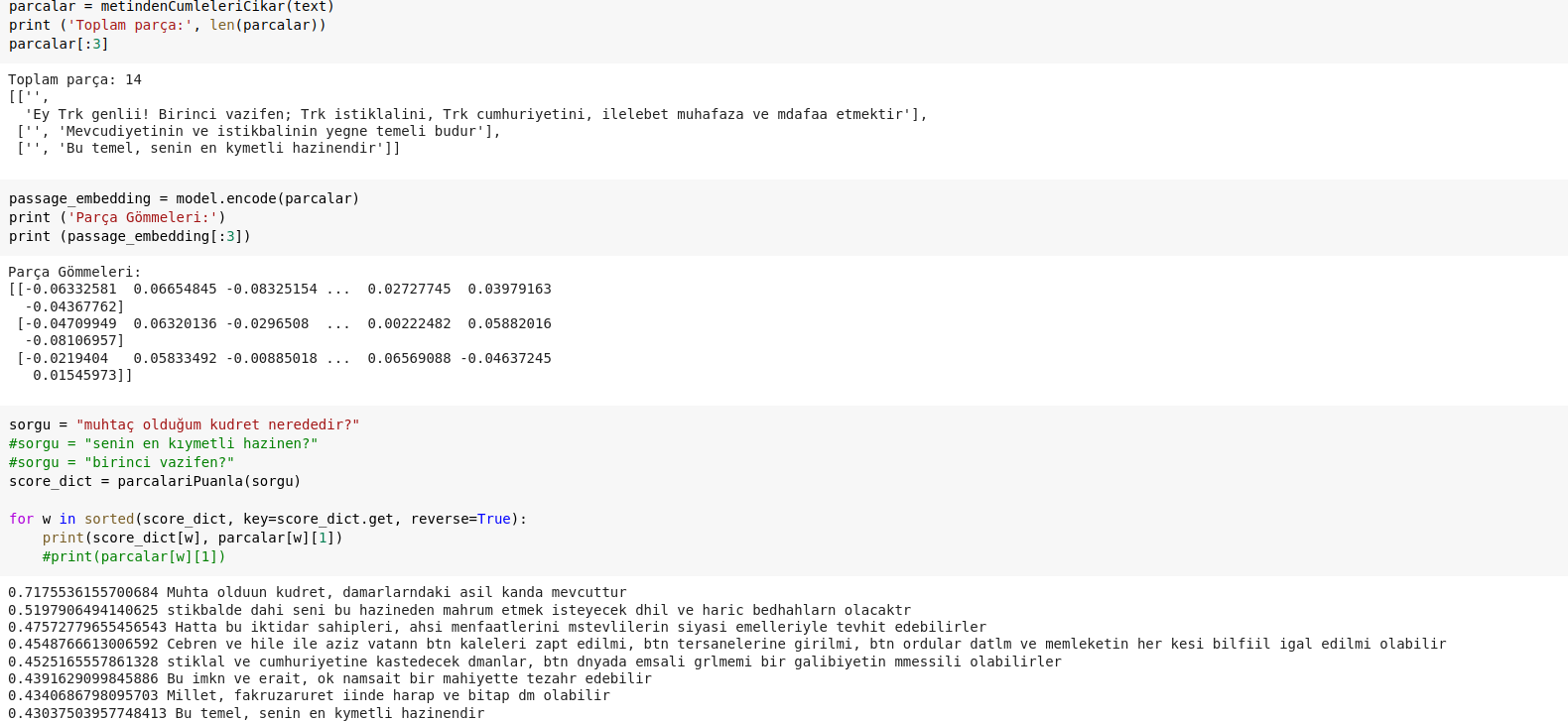
İnput olarak verdiğimiz cümleden elde edilen embeddings matrix’imizin sonuçları ve uzunlukları ise Şekil 4 de gösterilmiştir.



**Şekil 4.** Girilen Cümlenin 384\*384 Boyutundaki Embeddings Matrix

Elimizde embeddings matrix’imiz olduğuna göre bundan sonraki işimiz çok kolay. İlk olarak verilen cümleleri embedding matris olarak ürettikten sonra, artık kullanıcının sormuş olduğu sorguyu embeddings matrix’e çevirip elimizdeki özetlenecek metnin embeddings matrix’leri ile karşılaştırarak en uygun cümle ya da paragraf grubu kullanıcıya çıktı olarak vermemiz yeterli olacaktır. Deneyle sonuçları Şekil 5 ve Şekil 6 da görebiliriz.

**Şekil 5.** Özetlenecek Metnin Cümlelere Ayrılması ve Embeddings Matrix’in Oluşturulması

**Şekil 6.** Sorgu Cümlesi ve Çıktı

1. SONUÇ

Bu makelemizde otomatik metin özetlemenin bir alt dalı olan sorgu tabanlı metin özetleme konusuna yüzeysel olarak değindik. Sorgu tabanlı metin özetlemeyi önceden eğitilmiş modellein embeddings matrix’ini kullanarak nasıl kolayca tasarlayacağımızı gördük. Sorgu tabanlı metin özetlemenin daha iyi ve daha doğru sonuçları vermesi adına semantik ağın modellere eklenmesi gerektiğini gördük. Bir sonraki çalışma ile tamamen türkçe kelimelerden oluşan bir embeddings matrix oluşturularak yeni bir çalışma yapılabilir ya da semantik ağlar türkçe metinler üzerinde kullanılarak sonuçlar incelenebilir.

1. KAYNAKLAR

Joshi, A., Fidalgo, E., Alegre, E., & Fernández-Robles, L. (2019). SummCoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. *Expert Systems with Applications*, *129*, 200–215. <https://doi.org/10.1016/j.eswa.2019.03.045>

Rahman, N., & Borah, B. (2020). Improvement of query-based text summarization using word sense disambiguation. *Complex and Intelligent Systems*, *6*(1), 75–85. <https://doi.org/10.1007/s40747-019-0115-2>

*WordNet: A Lexical Database for English George A. Miller*. (n.d.).

Yu, H. (2022a). *Survey of Query-based Text Summarization*. <http://arxiv.org/abs/2211.11548>

Yu, H. (2022b). *Survey of Query-based Text Summarization*. <http://arxiv.org/abs/2211.11548>