Bilgisayar Mühendisliğinde Özel Konular

Dr. Fahrettin Horasan

METİN MADENCİLİĞİ

METİN MADENCİLİĞİ

- Metin madenciliğinde veri kaynağı olarak yapılandırılmamış veri olarak nitelendirilen metin dokümanları işlenmektedir.
- Yapılandırılmamış veri; bilgisayar tarafından anlamlandırılmayan, direk işleme tabi olabilecek şekilde veri yapısına sahip olmayan metin ya da sayısal ifadelerin bir arada sunulduğu organize edilmemiş veriler bütünüdür.
- Yapılandırılmış veri: bilgisayar tarafından tanınan, işleme direk tabi olabilen, veri yapısı belirlenmiş kategorik ya da sayısal değerler içeren veriler olarak adlandırılmaktadır

METİN MADENCİLİĞİ

- Metin madenciliği, veri kaynağı olarak yapılandırılmamış veri olan metinsel dokümanları işleyerek yeni bilgiler keşfedilmesini sağlayan veri madenciliği alanıdır.
  - ❖Örneğin; metinlerin benzerliği, sınıflandırılması, özetlenmesi, temsilci kelimelerinin oluşturulması, metinlerden duygu analizi, metinlerden yazan kişinin tespiti, metin içeriğine bağlı öneri sistemleri, soru-cevap sistemleri gibi birçok çalışma alanı mevcut ve gelişmeleri devam etmektedir.

METİN MADENCİLİĞİ

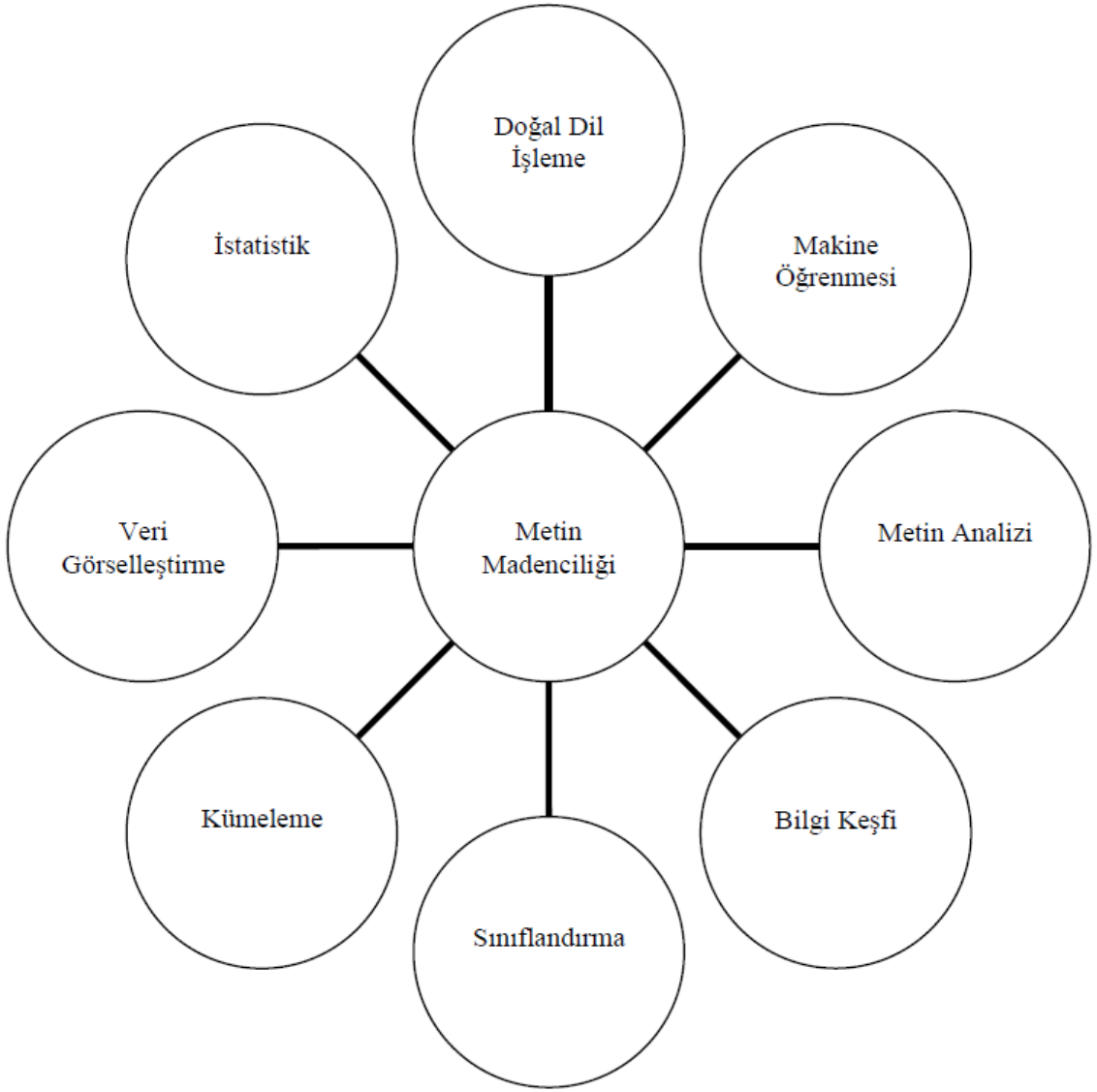
- Metin madenciliği metin analizi teknolojisine dayanarak ortaya çıkan işlevler bütünüdür. Çok kapsamlı olmayan bir çalışma da bile otomatik olarak metinlerin önişleme sürecinden geçirilmesi, işlenmesi ve kullanıma hazır hale getirilmesi önemli bir çaba gerektirmektedir.

METİN MADENCİLİĞİ

- Metin madenciliği veri madenciliğinin bir alt kolu gibi görünse de her ikisini de birbirinden ayıran önemli farklar vardır. Veri madenciliği sayısal ya da kategorik olarak yapılandırılmış verileri inceleyip bilinmeyen ilişkileri çıkarmayı amaçlarken metin madenciliği yapılandırılmamış metin türündeki verileri inceleyerek farkında olmadığımız bilgilere ulaşmayı amaçlamaktadır.

METİN MADENCİLİĞİ

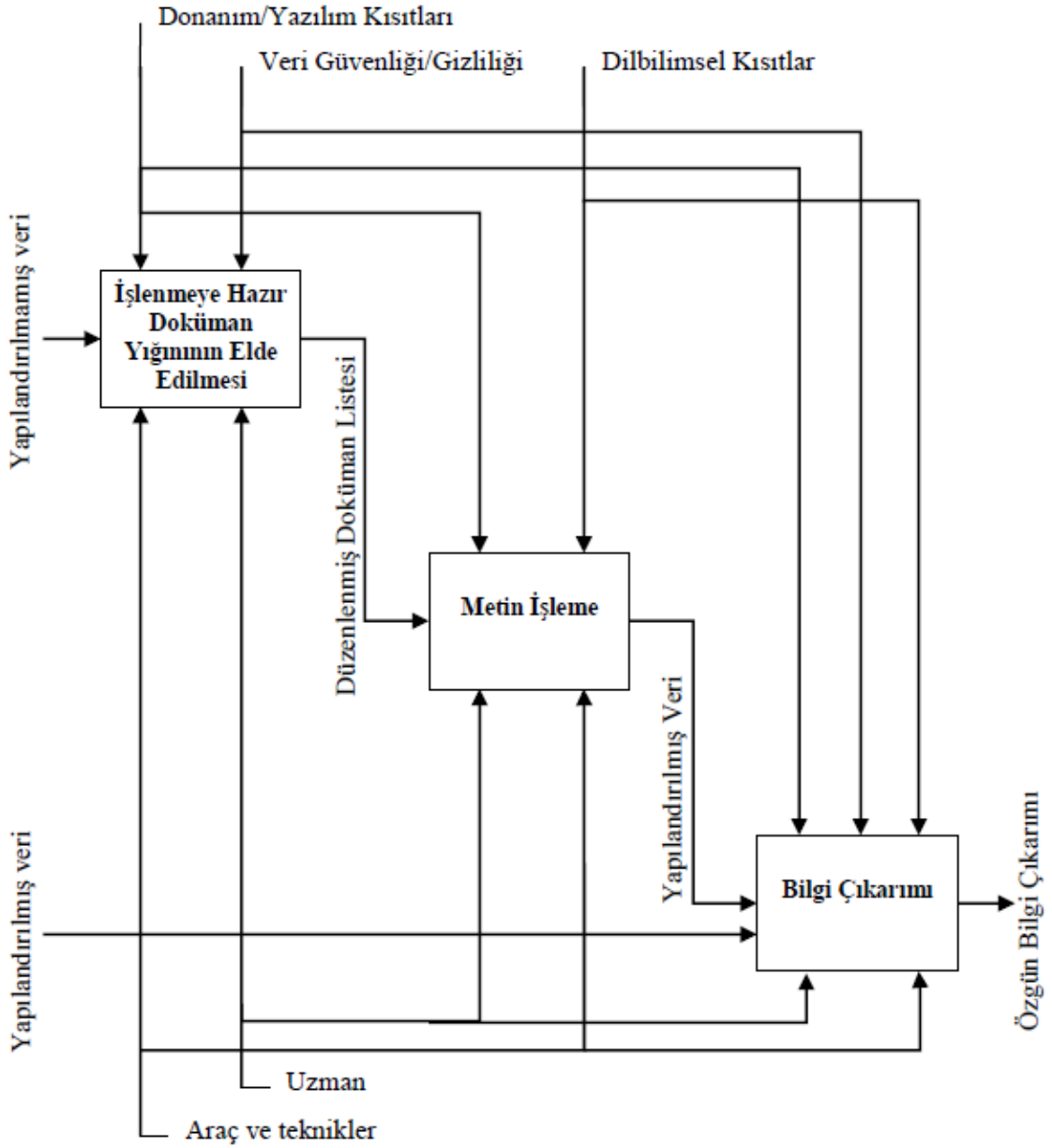
- Metin madenciliği makine öğrenmesi, istatistik, bilgi keşfi, doğal dil işleme, metin analizi, sınıflandırma, kümeleme, veri görselleştirme gibi birçok alanla birlikte anılmaktadır (Şekil 1-Bir sonraki slaytta).
- Ayrıca veri madenciliği yöntemlerinde geliştirilen yeni algoritmaların da bu alanda kullanılması mümkün olabilmektedir.



Şekil 1: Metin madenciliğin diğer disiplinlerle ilişkisi

METİN MADENCİLİĞİ

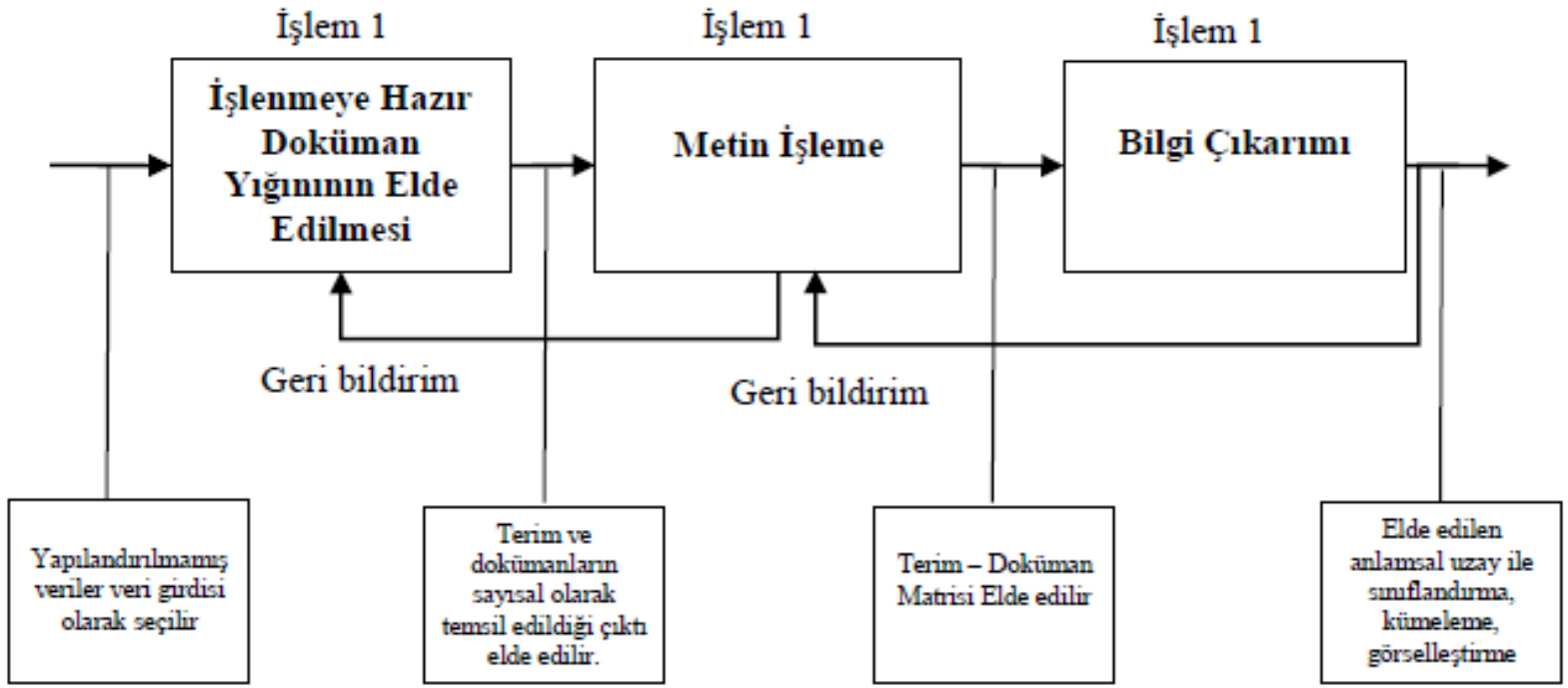
- Metin madenciliği süreci, işlenmesi gereken verilerin toplandığı ve işleme hazır hale getirildiği aşama, verilerin işlenerek yapılandırılmış verilerin elde edildiği aşama ve bunların ardından gelen bilgi çıkarım aşaması olarak üç aşamada incelenir.
- Şekil 2'de gösterildiği gibi yapılandırılmamış veriden bilginin keşfine doğru uzanan bu sürece tercih edilen yazılım ile ilgili kısıtlar, donanımsal kısıtlar, dilbilimsel kısıtlar da dâhil olmaktadır. Ayrıca bu sürecin öncesi ve sonrasındaki tüm aşamada kişisel verilerin gizliliği ve güvenliğinin sağlanmasına dikkat edilmelidir.



Şekil 2: Metin Madenciliği ve Paydaşları

METİN MADENCİLİĞİ

- Metin madenciliği sürecinde (Şekil 3.) ilk olarak çalışılacak doküman yığını ya da korpus olarak adlandırılan veri kümesinin oluşturulması gerekmektedir. Doküman yığınları metin dosyalarının bir arada tutulduğu bir dizin ya da metin türünde verilerin tutulduğu veri tabanı olarak ele alınabilir.



Şekil 3: Metin madenciliği süreci

METİN MADENCİLİĞİ

- Metinsel dokümanlar bilişim sistemleri ile anlamsal olarak işlenebilmesi ve sonuç olarak çıkarımlarda bulunması, metinler üzerinde anahtar kelimelerin sorgulanması ile ya da metne ait kelimelerin bir araya gelerek oluşturduğu anlamsal yapıya dikkat ederek mümkün olmaktadır.
- Metinde yer alan kelimelerin tamamının metin içindeki öneminin yanı sıra diğer dokümanlarla olan ilişkilerine de dikkat edilmektedir.

METİN MADENCİLİĞİ

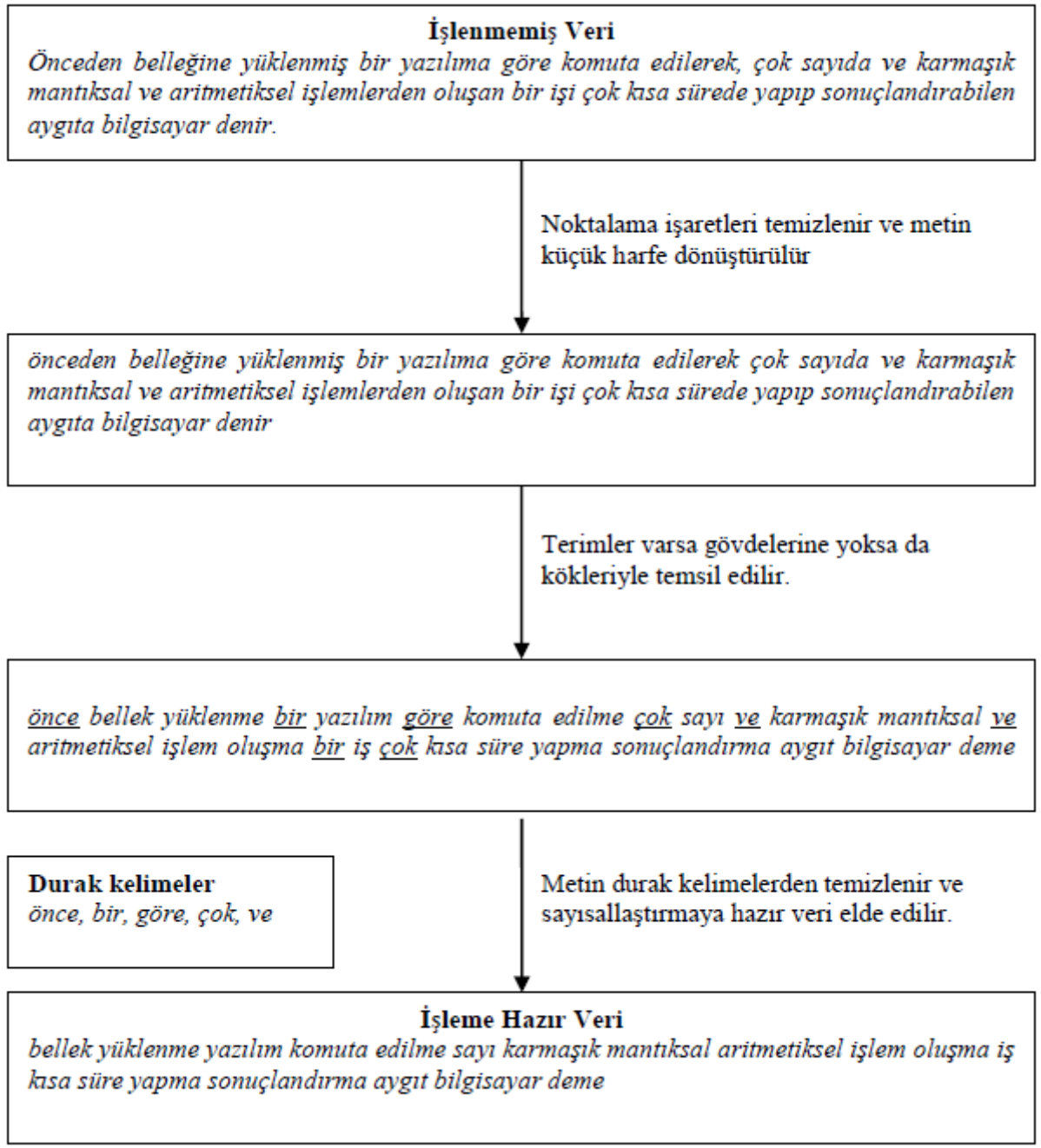
- Metin içindeki her bir terim dilin yapısına göre standart bir şekilde temsil edilmelidir.
- Bu amaçla dilin özellikleri dikkate alınarak elde edilen doküman yığınındaki her bir terim kök ya da gövdelerine dikkat edilerek ön işleme sürecinden geçirilerek işlenmelidir.
- Ayrıca metin içerisinde çok geçmekle birlikte tek başına anlamı olmayan kelime grubu olarak tanımlanan durak kelimelerinin işleme dâhil olması engellenmelidir.

METİN MADENCİLİĞİ

- İşlem öncesi yapılan veri temizleme aşamasında ilk olarak veri yığınında sadece terim ve doküman şeklinde verilerin işlenebilmesi amacı ile noktalama işaretlerinin temizlenmesi sağlanır.
- Daha sonra da hedef olarak seçilen verilerin anlamı bozabilen gürültü niteliği taşıyan anlama olumlu ya da olumsuz katkısı olmayıp işlem süresini uzatan durak kelime olarak adlandırılan bağlaç, edat zamir gibi tek başına anlamı olmayan kelimelerin çıkarılması gerekmektedir.
- Böylece dokümanların içerisinde sayıca daha çok bulunan ancak anlamı etkilemeyen bu kelimelerin çıkarılmasıyla yapılacak işlem sürecinin zaman ve kaynak yönünden maliyeti azalır.
- Bunun yanında sadece anlamı daha çok etkileyen kelimelerin işleme dâhil olmasını sağlayarak bilgiye erişim performansının da artmasına neden olur.

Tablo 1: Örnek bir terimin kök ya da gövdesine ayrıştırılması

Kelime	Kök	Gövde	İşlenen kısım
kitapçıdan	kitap	kitapçı	kitapçı
kitaplık	kitap	kitaplık	kitaplık
kitabeden	kitap	kitabe	kitabe
kitapçılık	kitap	kitapçılık	kitapçılık
kitaplaştırma	kitap	kitaplaştırma	kitaplaştırma
kitabı	kitap	-	kitap
kitaptan	kitap	-	kitap
kitap	kitap	-	kitap


İşlenmemiş Veri
Önceden belleğine yüklenmiş bir yazılıma göre komuta edilerek, çok sayıda ve karmaşık mantıksal ve aritmetiksel işlemlerden oluşan bir işi çok kısa sürede yapıp sonuçlandırabilen aygıta bilgisayar denir.
Noktalama işaretleri temizlenir ve metin küçük harfe dönüştürülür
önceden belleğine yüklenmiş bir yazılıma göre komuta edilerek çok sayıda ve karmaşık mantıksal ve aritmetiksel işlemlerden oluşan bir işi çok kısa sürede yapıp sonuçlandırabilen aygıta bilgisayar denir
Terimler varsa gövdelerine yoksa da kökleriyle temsil edilir.
önce bellek yüklenme bir yazılım göre komuta edilme çok sayı ve karmaşık mantıksal ve aritmetiksel işlem oluşma bir iş çok kısa süre yapma sonuçlandırma aygıt bilgisayar deme
Durak kelimeler
önce, bir, göre, çok, ve
Metin durak kelimelerden temizlenir ve sayısallaştırmaya hazır veri elde edilir.
İşleme Hazır Veri
bellek yüklenme yazılım komuta edilme sayı karmaşık mantıksal aritmetiksel işlem oluşma iş kısa süre yapma sonuçlandırma aygıt bilgisayar deme

METİN MADENCİLİĞİ

- Doküman yığınları içerisindeki bütün kelimeler ve bu kelimelerin bulunduğu her bir dokümanın temsil edildiği terim doküman matrisi kelimelerin dokümanlardaki ağırlığını dikkate alarak elde edilir.
- Tablo 2’de örnek bir terim doküman matrisini görebilirsiniz. Metin madenciliğinde genellikle terim sayısı doküman sayısından büyük olduğu görülmektedir.
- Ancak işlenmekte olan doküman sayısının katlanarak arttığı durumlarda terim sayısı doküman sayısından az olmaktadır. Bu duruma arama motorlarındaki terim ve doküman sayıları örnek verilebilir.

METİN MADENCİLİĞİ

Tablo 2: Örnek bir Terim Doküman Matrisi

Dokümanlar / Terimler	Doküman 1	Doküman 1	Doküman 3	…	Doküman n
Terim 1	1,2	0	0,78	…	0
Terim 2	0,05	0	0		0,1
Terim 3	0	1,1	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮
Terim m	0	0	2,1	…	0

METİN MADENCİLİĞİ

- Metin madenciliğinde istatistiksel yöntemlerle metin içerisinden anahtar kelimelerin belirlenmesi ya da sadece terim doküman matrisindeki frekans sayıları tek başına yeterli değildir.
- Terim doküman matrisi vasıtası ile doküman yığını için bir anlamsal uzay elde edilerek metin içerisindeki anlamı etkilemeyen ya da anlamı bozan bileşenlerin göz ardı edildiği analiz işlemleri yerine getirilir.
- Bu anlamsal uzay sayesinde metin içerisindeki anlamsal kalıplar dikkate alınarak metne ait ilginç ya da belirli amaçlar için kullanılacak yararlı bilgiyi temsil eden verileri sunulur.

- Örnekler
  - Metin özetleme
  - Metin benzerliği
  - Metin sınıflandırma
  - Metin üretme
  - Sahtekarlık tesbiti
  - Yazar tespiti
  - Metne konu başlığı bulma
  - Metinden anahtar kelime çıkarma
  - Soru cevaplama
  - Chatbot uygulamarı
  - İçerik tabanlı tavsiye sistemleri
  - Duygu analizi
  - …….

- Teşekkürler.