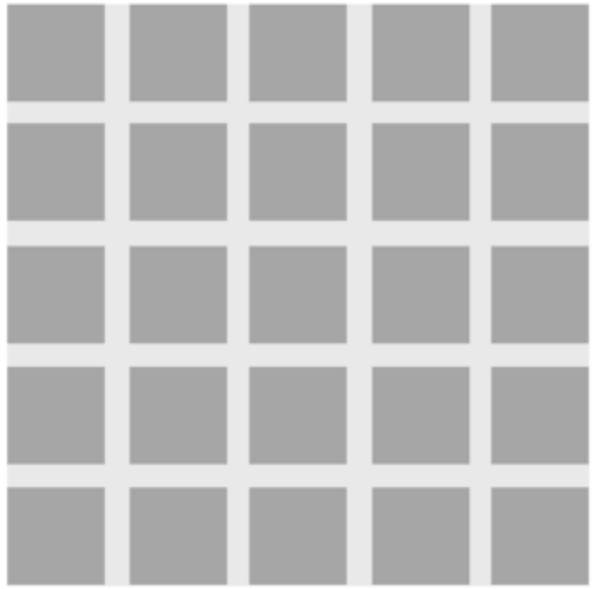
METİN MADENCİLİĞİ

Hazırlayan:

Dr. Ekin EKİNCİ

METIN MADENCILIĞINE GENEL BAKIŞ (1)*

Yapısal Veri	Yapısal Olmayan Veri
	
Veri tabanı, CRM, ERP	Metin, Ses, Video

Bugün üretilen tüm verilerin yüzde 80'inden fazlası yapısal olmayan veri olarak kabul edilmektedir.

METIN MADENCILIĞINE GENEL BAKIŞ (2)

İlan No	650143961
İlan Tarihi	11 Mayıs 2019
Marka	BMW
Seri	X5
Model	40e xDrive M Plus
Yıl	2017
Yakıt	Hybrid
Vites	Otomatik
KM	22.000
Kasa Tipi	SUV
Motor Gücü	313 hp
Motor Hacmi	1997 cc
Çekiş	4x4
Kapı	5
Renk	Gümüş Gri
Garanti	Evet
Plaka / Uyruk	Türkiye (TR) Plakalı
Kimden	Galeriden
Takas	Hayır
Durumu	İkinci El

VERİ

Yapısal Veri

Yarı Yapısal Veri

Yapısal Olmayan Veri

```
{
  "person":{
    "name": "Jennifer",
    "surname": "Green",
    "friends":[
      {
        "age":25,
        "isDeveloped":true
      },
      {
        "age":28,
        "isDeveloped":false
      }
      ]
  },
  "salary":1000
}
```

Hürriyet Lezizz @HurriyetLezizz · 8 sa.
Kavurma yapılışı en kolay ancak en lezzetli geleneksel lezzetlerimizden biri. İftarda yanında pilavla zevkle yiyebileceğiniz kuzu kavurma tarifimize bir göz atmaz mıydınız? hry.yt/sHi3Z

METIN MADENCILIĞINE GENEL BAKIŞ (3)*

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

METIN MADENCILIĞINE GENEL BAKIŞ (4)

- Yapısal ve yapısal olmayan milyarlarca içeriği biz kullanıcılarına sunan Web, günümüzün önemli veri kaynaklarından birisi haline gelmiştir.

- Sunulan içerik her geçen gün büyümektedir.

- İçeriğin %80'i dokümanlar şeklinde organize edilmiştir: haberler, forumlar, e-mailler, haber grupları, sosyal medya, ...

METIN MADENCILIĞINE GENEL BAKIŞ (5)

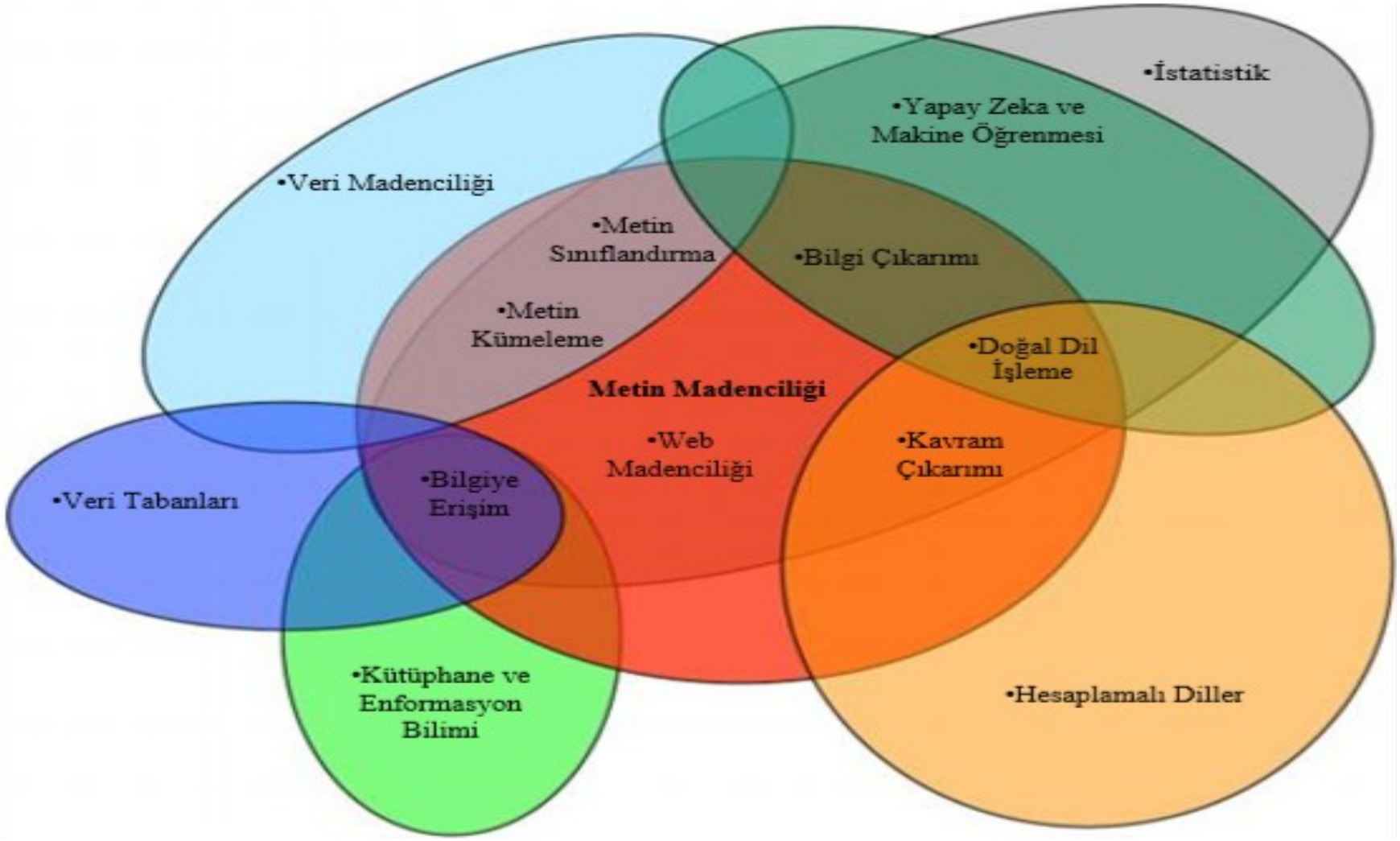
- Sunulan içerik her geçen gün büyümekte, bu içerikten **istenilen bilginin otomatik bir şekilde çıkartılması ve çıkartılan bilginin**:
  - **Organize edilme,**
  - **Analiz edilme ve**
  - **Anlaşılması adımında ise metin madenciliğine ihtiyaç duyulmaktadır.**

METIN MADENCILIĞI NEDIR? (1)

- Her geçen gün artan veri miktarı bu verileri yönetmeyi ve içerisinden önemli olan ancak keşfedilmemiş bilgiyi çıkarmayı gerekli hale getirmiş ve metin madenciliği kavramı ortaya çıkmıştır.

- Eldeki dokümanlardan belli bir amaç çerçevesinde önceden bilinmeyen ancak potansiyel olarak faydalı bilginin çıkarılması şeklinde tanımlanmaktadır (Visa, 2001).

METIN MADENCILIĞI NEDIR? (2)*



METIN MADENCILIĞI BILEŞENLERI (1)

- **Bilgi Edinme (Information Retrieval):** Büyük koleksiyonlardan (genellikle bilgisayarlarda saklanan) belli bir amaca yönelik bir bilgi ihtiyacını karşılayan, yapılandırılmamış nitelikte bir materyalin (genellikle belgeler) elde edilmesidir (*).

- **Doküman Kümeleme (Document Clustering):** Büyük miktardaki doküman koleksiyonunu her birinin bir konuyu temsil ettiği az sayıdaki anlamlı kümelere dağıtma görevidir.

- **Doküman Sınıflandırma (Document Classification):** Dokümanlarını önceden tanımlı bir ya da daha fazla sınıfa atama görevidir.

METIN MADENCILIĞI BILEŞENLERI (2)

- **Web Madenciliği (Web Mining):** Veri madenciliğinin alt dallarından biri olan web madenciliği webden elde edilen verilerden bilginin çıkartılmasını amaçlar.

- **Bilgi Çıkarımı (Information Extraction):** Yapılandırılmamış dokümanlardan yapılandırılmış bilginin çıkartılması görevidir.

- **Doğal Dil İşleme (Natural Language Processing):** doğal dil üzerine inceleme, çözümleme, yorumlama, bilgi çıkarma, üretme yapan bilgisayar sitemi şeklinde tanımlanmaktadır (Oğuzlar, 2011).

METIN MADENCILIĞI BILEŞENLERI (3)

- **Kavram Çıkarımı (Concept Extraction):** Kelimelerin ve öbeklerin anlamsal olarak benzer gruplar altında öbeklenmesi görevidir.

METIN MADENCILIĞI UYGULAMA ALANLARI

- Konu çıkarımı
- Duygu analizi
- Soru cevaplama sistemleri
- Yazar analizi
- Doküman özetleme
- Haberlerin sınıflandırılması
- Spam filtreleme,…

METIN MADENCILIĞI ADIMLARI

METNIN ELDE EDILMESI

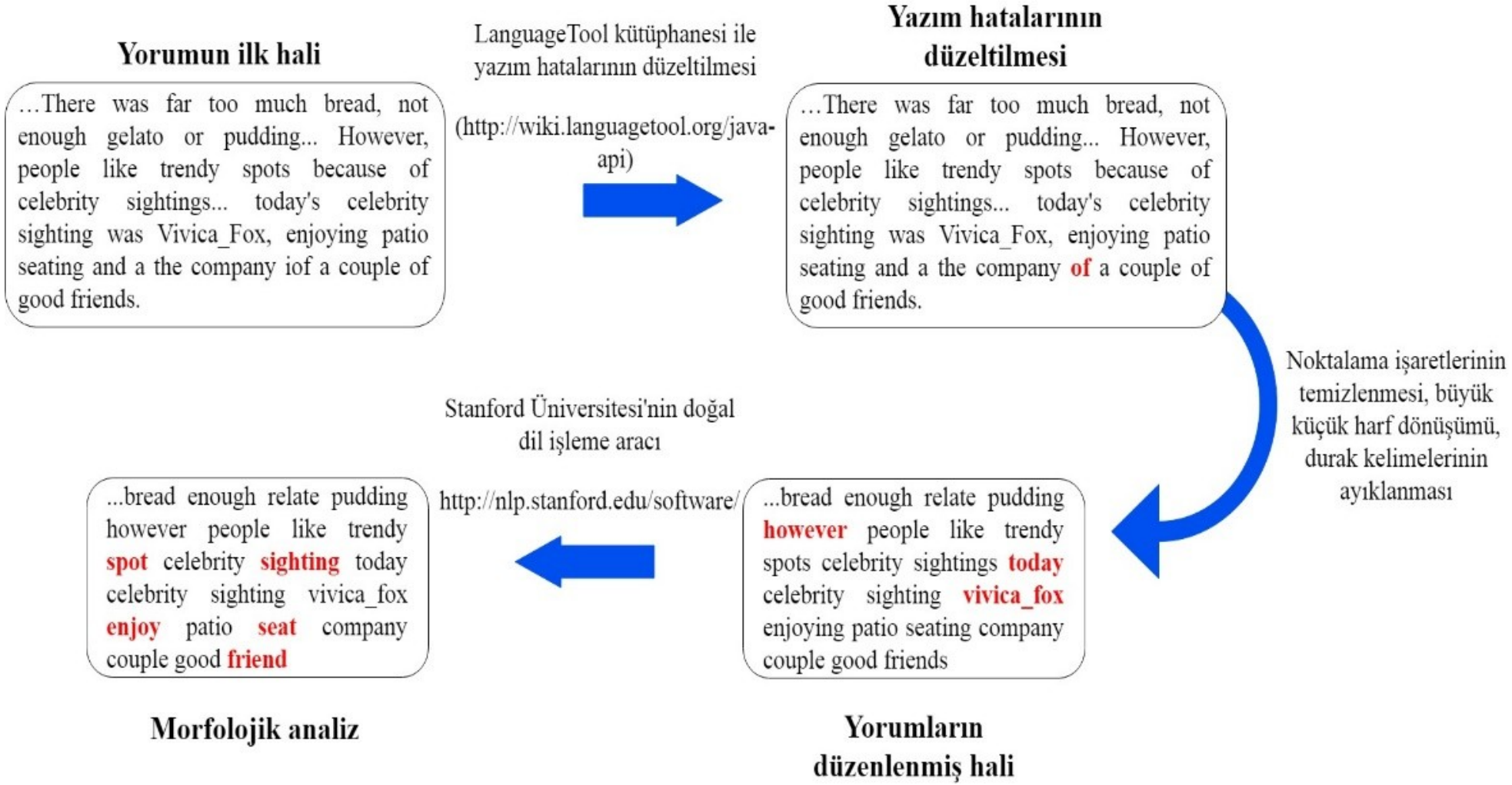
- Metin madenciliği adımlarını gerçekleştirebilmemiz için ilk olarak amaca yönelik bir veri kümesinin elde edilmesi gerekmektedir.
  - Hazır veri kümelerini kullanabiliriz:
    - UCI Machine Learnnig Repository
    - Kaggle
    - Kemik Doğal Dil İşleme Grubu
  - Kendi veri kümemizi kendimiz oluşturabiliriz:
    - Web Crawler ile

METIN ÖNIŞLEME

- Metinler üzerinde yapılacak önişleme çalışılacak amaca göre farklılıklar göstermekle birlikte temel önişleme adımları;
  - noktalama işaretleri, sayı ve özel karakterlerin eldeki metinlerden çıkartılması,
  - büyük küçük harf duyarlı olmamasından ötürü büyük harflerin küçük harflere dönüştürülmesi,
  - metni meydana getiren ve çok sık tekrarlanan ancak doküman için önemli olmayan durak kelimelerinin eldeki metinlerden ayıklanması,
  - yazım hatalarının düzeltilmesi (normalizasyon),
  - POS tagging
  - gövdelemenin gerçeklenmesi şeklinde sıralanmaktadır.

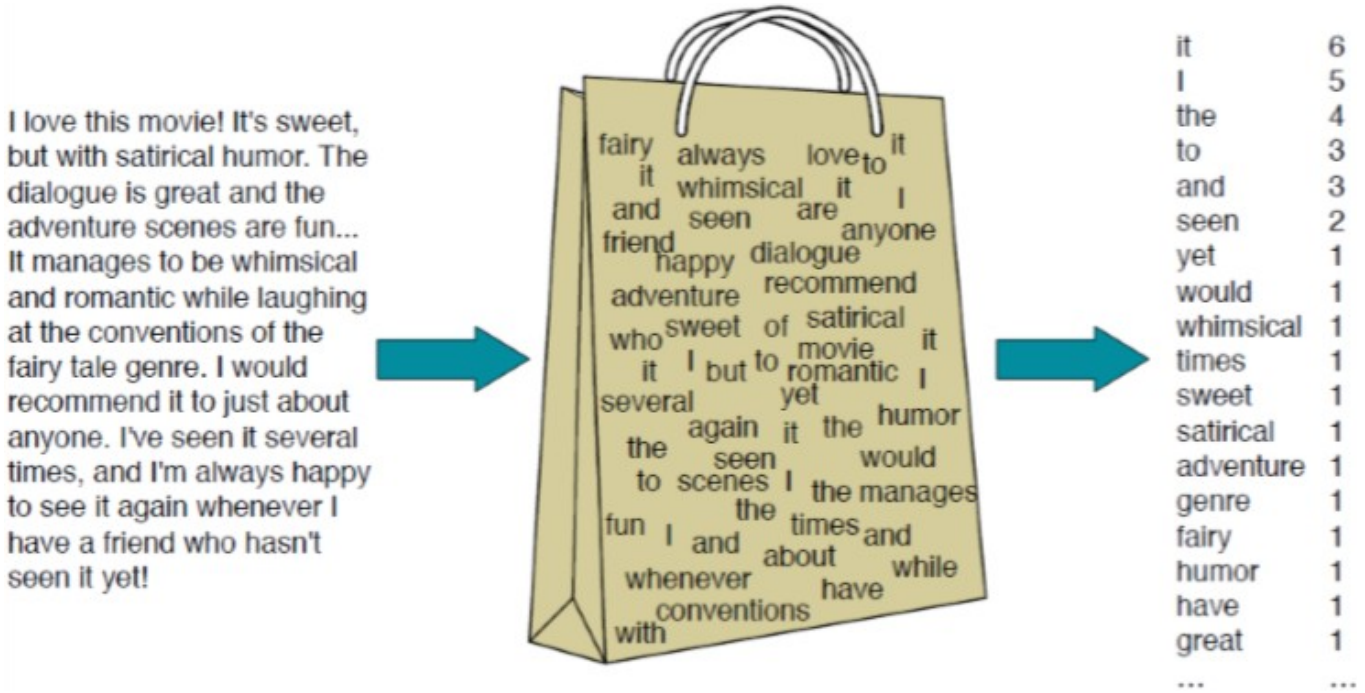
- Bu önişleme adımları doğal dil işleme sürecini oluşturmaktadır.

METIN ÖNİŞLEME



METIN DÖNÜŞÜMÜ (1)

- **Kelime torbası:** Bir dokümanın tipik temsilidir. Kelimeler frekansları ile temsil edilmektedir, kelimelerin doküman içerisindeki konumu göz ardı edilmektedir.
- Kelimelerin ağırlıklarının hesaplanması gerekmektedir.
- Ağırlık hesabı ise kelimenin ilgili sınıfta geçme sıklığı şeklinde hesaplanır.



METIN DÖNÜŞÜMÜ (2)

- **Vektör Uzayı Modeli:** Dokümanların ortak bir uzayda vektörler olarak gösterilmesi, vektör uzay modeli olarak ifade edilmektedir.
- Bu modelde dokümanlar ağırlık vektörü olarak temsil edilmektedir.
- Terim ağırlıkları Tf (term frequency) ya da Tf-Idf (term frequency-invert document frequency) şemalarına göre hesaplanmaktadır.

METIN DÖNÜŞÜMÜ (3)

- **Terim frekansı:** Bir terimin ilgili dokümanda kaç kere geçtiğini temsil etmektedir.
- i. terimin j. dokümandaki frekansı ile temsil edilmektedir.

	Doc 1	Doc 2	...	Doc *n*
Term(s) 1	12	2	...	1
Term(s) 2	0	1	...	0
...	...	...	...	
Term(s) *n*	0	6	...	3

METIN DÖNÜŞÜMÜ (4)

- **Ters metin frekansı:** Eğer bir terim çok fazla sayıda dokümanda bulunuyorsa muhtemelen bu terim gerçekleştirilecek görev için önemli değildir.
  - N toplam doküman sayısı iken df_i i. terimin geçtiği toplam doküman sayısıdır.

$$idf_i = log \frac{N}{df_i}$$

METIN DÖNÜŞÜMÜ (5)

- i. terimin j. dokümandaki ağırlığı $w_{i,j}$ ile temsil edilmektedir.

$$w_{i,j} = tf_{i,j} * idf_i$$

$$\begin{pmatrix} & D_1 & D_2 & \ldots. & D_t \\ T_1 & w_{11} & w_{21} & \ldots & w_{t1} \\ T_2 & w_{12} & w_{22} & \ldots & w_{t2} \\ : & : & : & & : \\ : & : & : & & : \\ T_n & w_{1n} & w_{2n} & \ldots & w_{tn} \end{pmatrix}$$

METIN DÖNÜŞÜMÜ (5)

- **Cosinüs Benzerliği:** Cosinüs benzerliği, iki vektör arasındaki açının cosinüsünü ölçer.
- İki dokümanın ağırlık vektörleri üzerinden benzerliklerini ölçmek için cosünüs benzerliğinden yararlanılmaktadır.

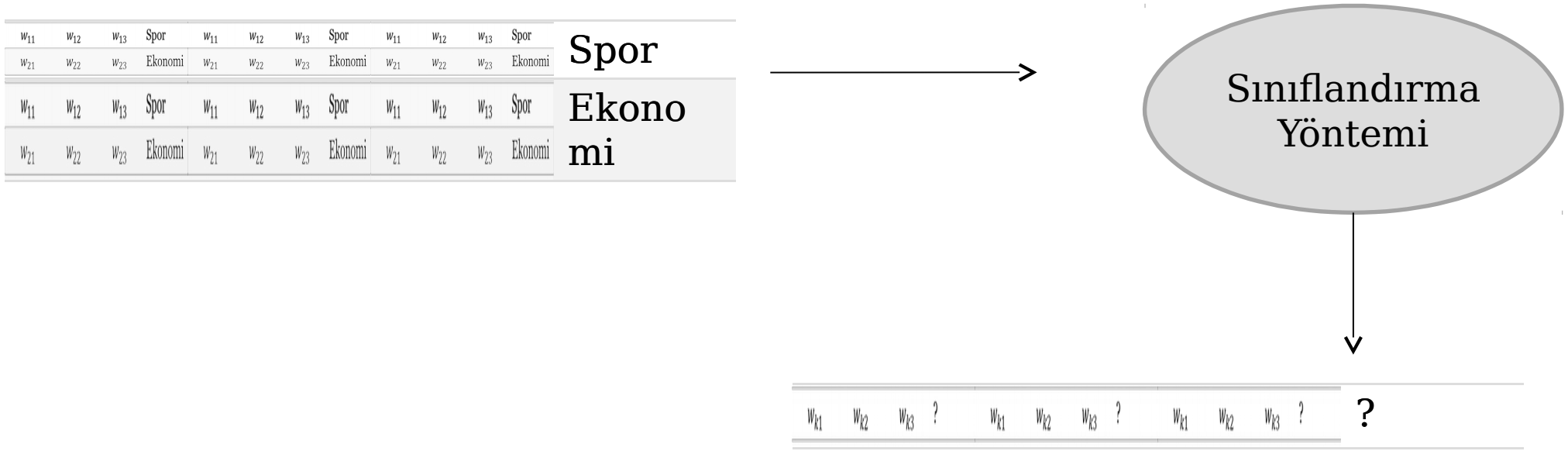
$$\cos(d_j, d_k) = \frac{\langle d_j . d_k \rangle}{\|d_j\| \|d_k\|} = \frac{\sum_{i=1}^{V} w_{i,j} \times w_{i,k}}{\sqrt{\sum_{i=1}^{V} w_{i,j}^2} \sqrt{\sum_{i=1}^{V} w_{i,k}^2}}$$

ÖZELLIK SEÇIMI

- Model oluşturmada kullanılacak önemli özelliklerin bir alt kümesini seçme işlemidir.
- Gereksiz özellikler veri madenciliği görevi için herhangi bir katkı sağlamamaktadır.
- Ki-kare istatistiği, tekil değer ayrışımı, doküman frekansı için bir eşik değeri belirleme,…

VERI MADENCILIĞI (1)

- **Sınıflandırma:** Dokümanın önceden tanımlanmış sınıflardan birine atanmasıdır.



VERI MADENCILIĞI (2)

- K-en yakın komşu
- Destek vektör makineleri
- Naive Bayes
- Yapay Sinir Ağları
- Karar Ağaçları…

DEĞERLENDIRME (1)

- Doğruluk, doğru sınıflandırılan kayıtların sayısının yanlış sınıflandırılan kayıtların sayısına oranı olarak tanımlanmaktadır.
- Doğruluk, diğer bir adıyla sınıflandırıcının doğru tahmin oranıdır. Kesinlik (p); gerçek sınıfı ve tahmin edilen sınıfı 1 olan kayıtların, tahmin edilen sınıfı 1 olan kayıtlara oranı şeklinde tanımlanmaktadır.
- Duyarlılık (r), gerçek sınıfı ve tahmin edilen sınıfı 1 olan kayıtların gerçek sınıfı 1 olan kayıtlara oranıdır.
- F-ölçümü kesinlik ve duyarlılık ölçümlerinin harmonik ortalaması alınarak bulunmaktadır.

DEĞERLENDIRME (2)

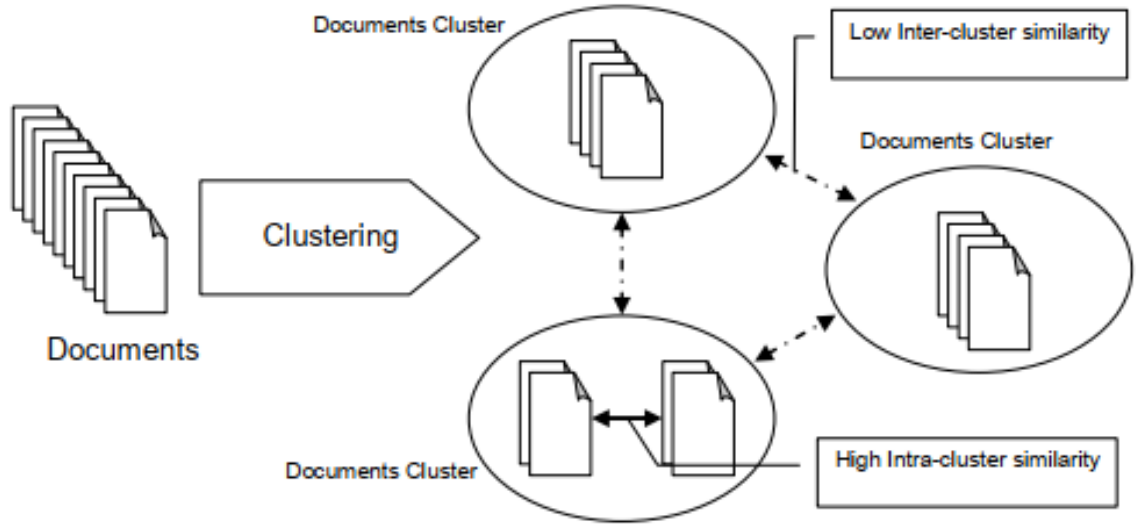
	TAHMİN EDİLEN SINIF		
GERÇEK SINIF		Sınıf=1	Sınıf=0
	Sınıf=1	a (Doğru Pozitif)	b (Yanlış Negatif)
	Sınıf=0	c (Yanlış Pozitif)	d (Doğru Negatif)

$$Doğruluk = \frac{a+d}{a+b+c+d} \qquad Duyarlılık = \frac{a}{a+b}$$

$$Kesinlik = \frac{a}{a+c} \qquad F-\text{ölçümü} = \frac{2 \times Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık}$$

VERI MADENCILIĞI (3)

- **Kümeleme:** Doküman koleksiyonunda yer alan dokümanları kümeler altında gruplar. Küme içi benzerlik maksimum iken, kümeler arası benzerlik minimum olmalıdır.
- Sınıflandırmanın aksine kümeleme yapılacak veri kümesindeki dokümanlarının sınıf etiketi bulunmamaktadır.



*

VERI MADENCILIĞI (4)

- K-means
- Hiyerarşik kümeleme,…

DEĞERLENDIRME (3)

- Kümeler içi benzerlik maksimum, kümeler arası benzerlik minimum olması gerekmektedir. Temel değerlendirme ölçütü bu kuraldır.

PYTHON KÜTÜPHANELERI

- **NLTK (Natural Language Toolkit) :** Önişleme adımlarının gerçekleştirilmesini sağlayan kütüphanedir.
- **Spacy:** NLTK ile aynı görevleri gerçekleştirmektedir.
- **Scikit-learn:** Makine öğrenmesi yöntemlerini sunan kütüphanedir. Ayrıca metin önişleme görevlerinin de yerine getirilmesini sağlamaktadır.
- **Gensim:** Konu modelleri, vektör uzayı modellerini sunan kütüphanedir.