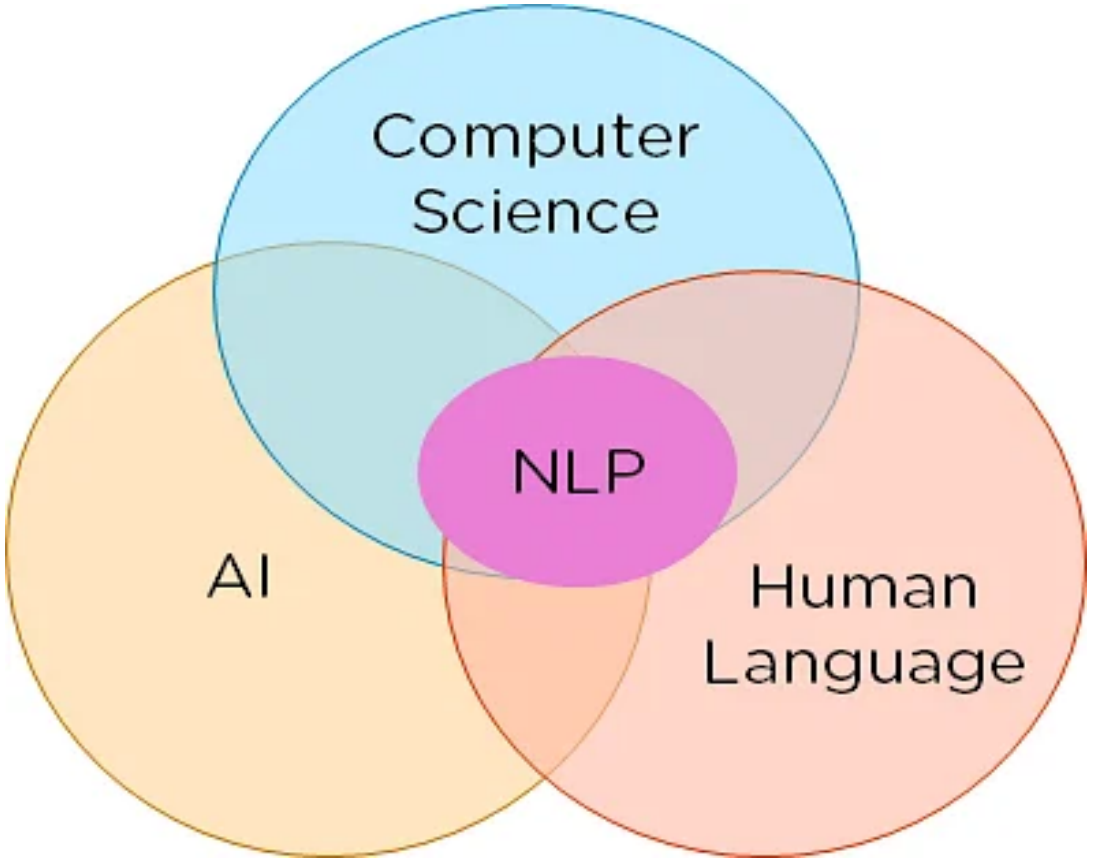
Doğal Dil İşlemeye Giriş

2024 GÜZ

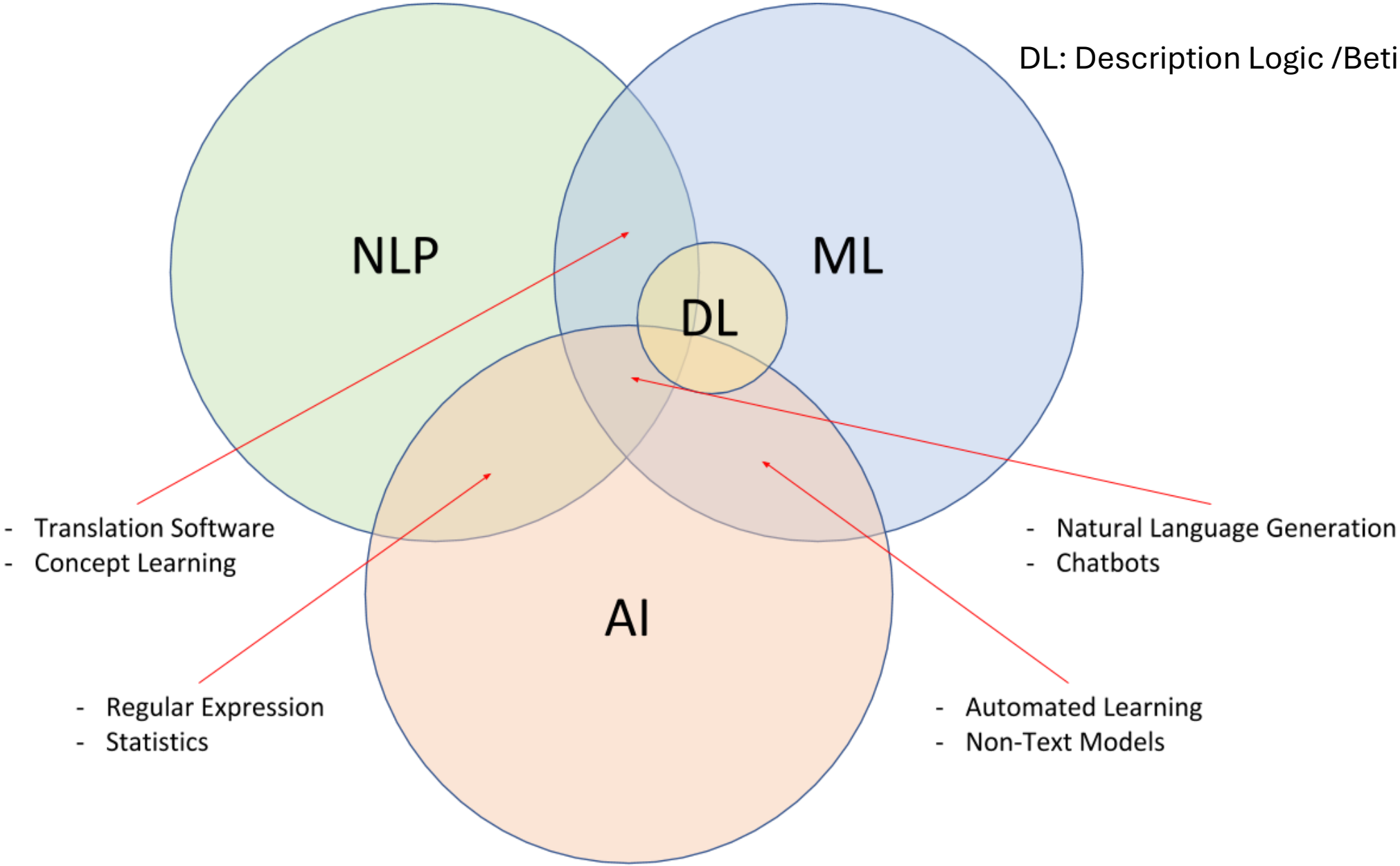
DDİ nedir?

- DDİ, dilbilim (linguistics) bilgisayar bilimleri ve yapay zekanın (YZ) alt disiplini olan bir çalışma alanıdır.
- DDİ, konuşma dili ile bilgisayarları anlamak ve onlarla iletişim kurabilmek için *makine öğrenmesini* kullanır.
- DDİ, bilgisayarlarla doğal diller (konuşma dili) arasındaki etkileşimle ilgilenilir.
- Doğal dili anlamak (understand) ve dilden bilgiyi çıkarmak (information extraction) için makineler nasıl programlanır? sorusuna cevap verilir.
- Uygulama örnekleri: Arama motorları, metin madenciliği, makine çevirisi, diyalog sistemleri, duygu analizi …

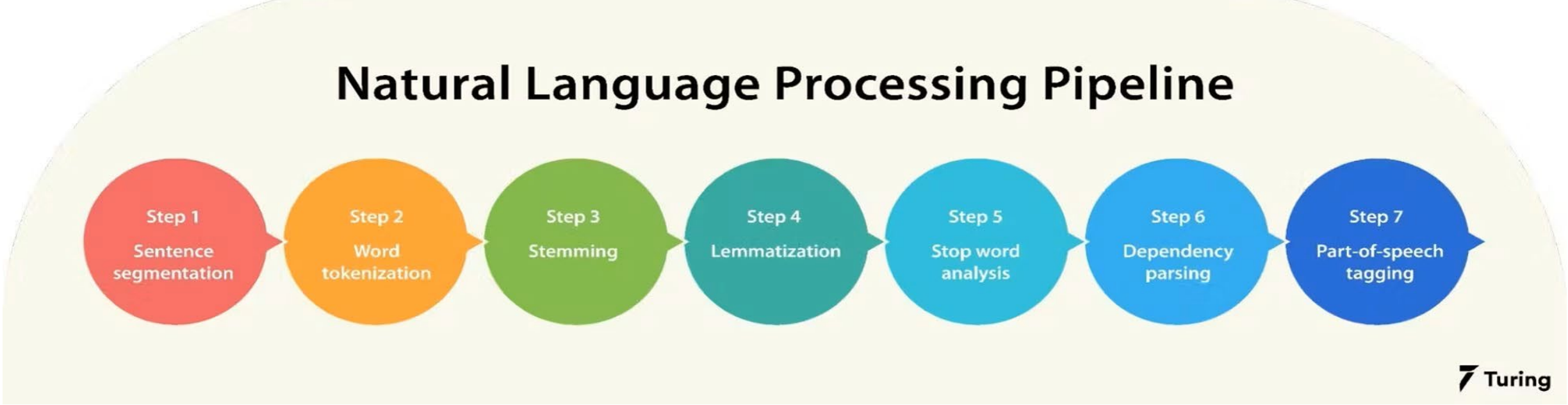
DDİ ‘nin Diğer Çalışma Alanları ile İlişkisi



DL: Description Logic /Betimleme Mantığı



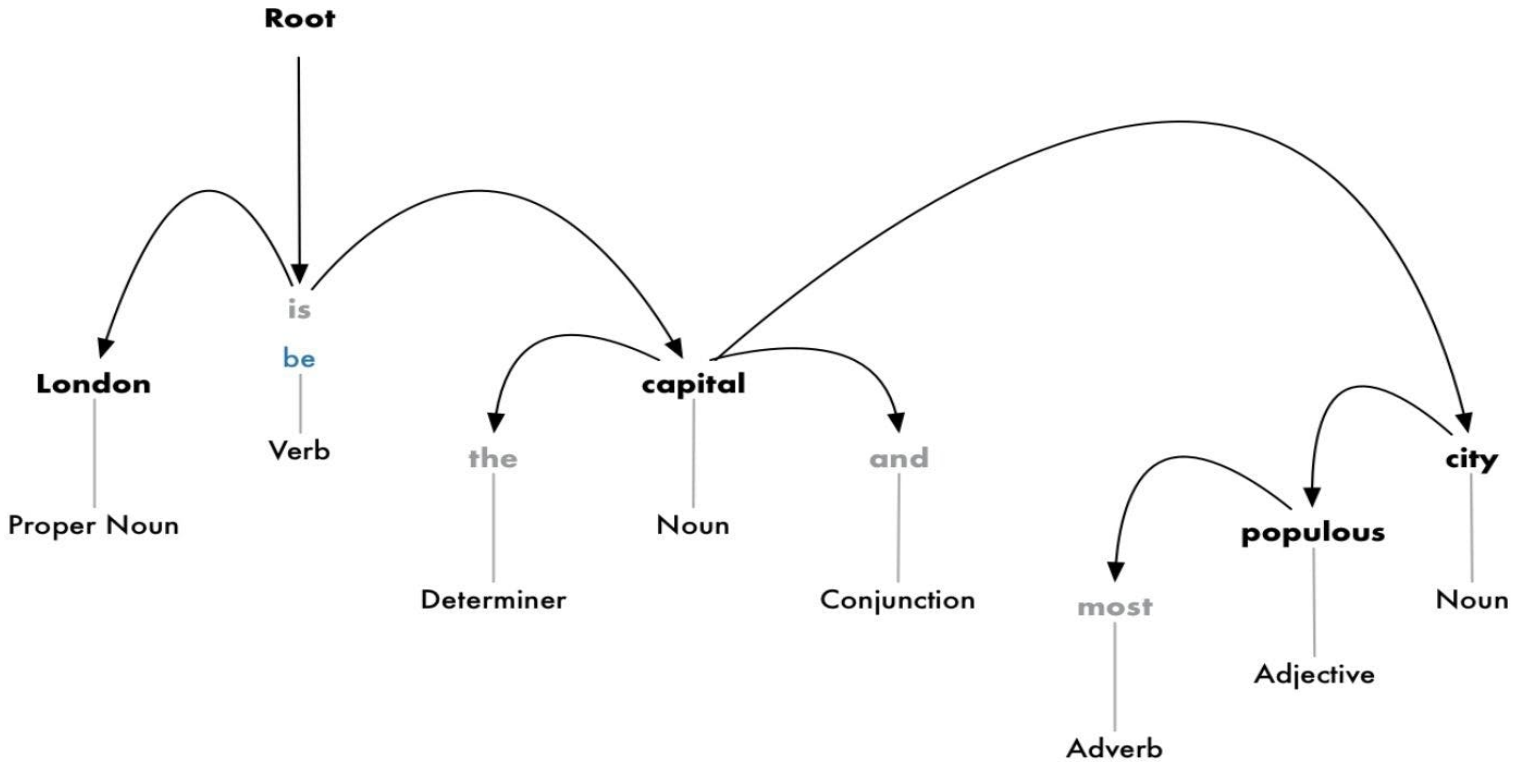
Regular Expressions / Düzgün İfadeler



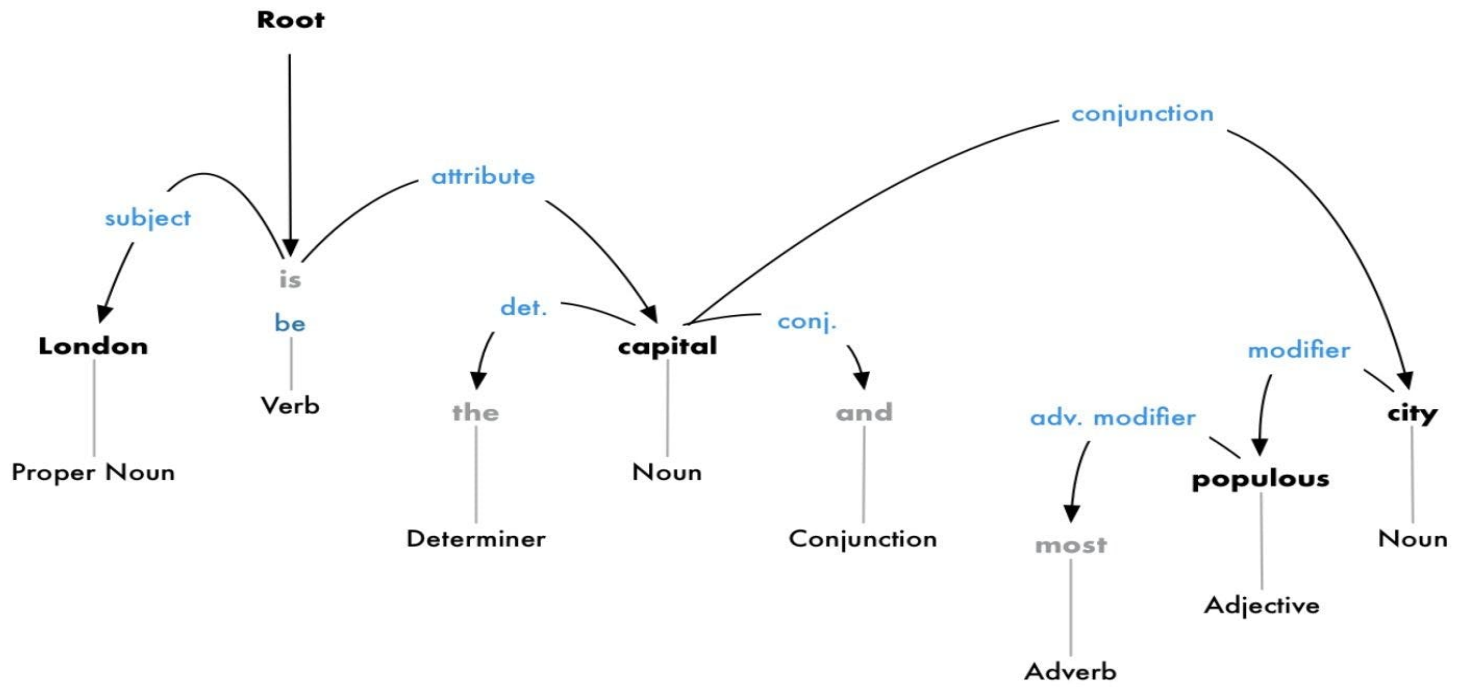
Adım 1: 1. “London is the capital and most populous city of England and the United Kingdom.”
2.“Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia.”
3. “It was founded by the Romans, who named it Londinium.”
Adım 2: “London”, “is”, “the”, “capital”, “and”, “most”, “populous”, “city”, “of”, “England”, “and”, “the”, “United”, “Kingdom”, “.”
Adım 3: *Köklendirme:* Kelimede bulunan ön ek ve/veya son ek çıkarılır. Süreç sonunda kelimenin sadece köküne inilir. Örneğin “boşluk” kelimesi kök hali “boş” kelimesine dönüşür. Cümlenin ifade ettiğinineyi anlamak için, konuşma bölümlerini analiz eder. Metnin ön işleme adımlarından biridir (preprocessing).
Adım 4: *Lemmatizasyon:* çekim eklerini kaldırır ve kelimeyi / lemmayı döndürür. Köklendirmeden farkı *lemma* gerçek, yani anlamı olan bir sözcüktür. 'playing' ve 'plays’ sözcükleri, 'play’ lemmasının farklı formlarıdır.
Adım 5: Sık kullanılan ve anlam ifade etmeyen sözcükler kaldırılır. Cümlede : is", "a", "the", "and".

Adım 6: Bağımlılığın Ayrıştırması /Dependency Parsing : Cümledeki sözcüklerin birbirleri ile ilişkisini gösterir. Ağaç yapısında oluşturulur.

Bağımlılığı bulmak için ana (parent) sözcük olarak bir kelime atanır. Bu, cümlenin eylemi kök (root) düğüm olarak davranır.



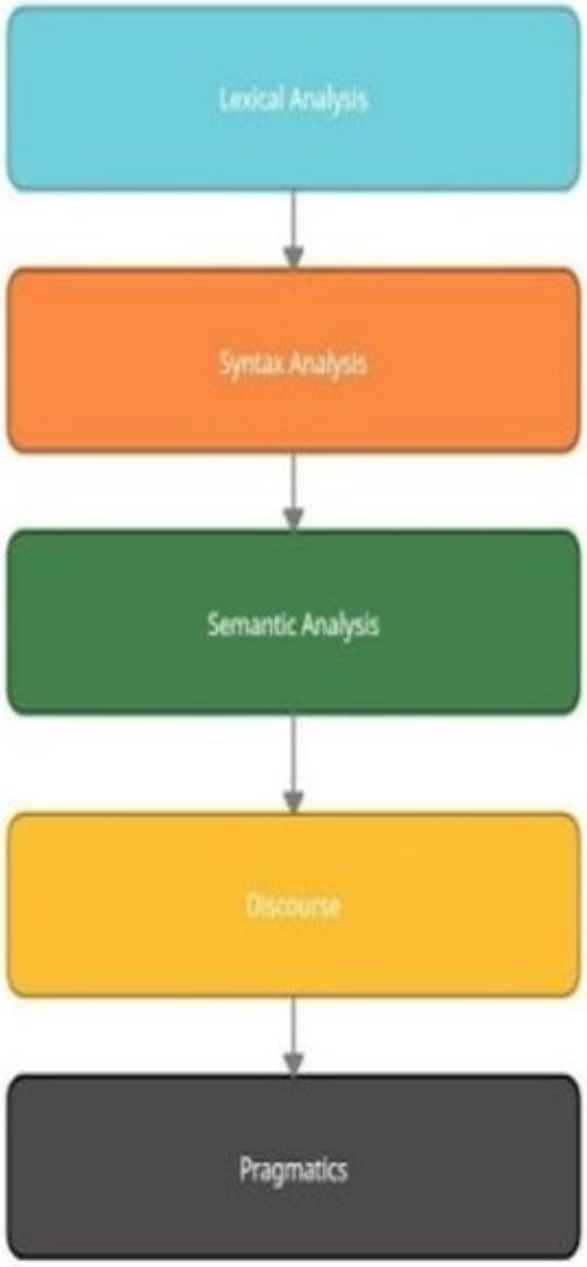
Adım 7: POS Etiketleme / Cümlenin öğelerine ayrılması: Eylemleri, zarfları, isimleri, sıfatları içerir ve sözcüğün anlamı cümle içerisinde dilbilgisi kurallarına uygun olarak belirtilir.



Özetle: DDİ, makinelerle insan dilini yorumlar. Yani, DDİ 'nin amacı metinden anlam çıkarmaktır

Kaynak: How Does Natural Language Processing Function in AI?
https://www.turing.com/kb/natural-language-processing-function-in-ai

DDİ ‘de Geleneksel Akışın Aşamaları



Kaynak:https://byteiota.com/stages-of-nlp/

1. Biçimbirimsel/Morphological Analysis -Sözcüksel /Lexical Analysis

❖ Morfolojik /Sözcüksel Analiz , metin tek tek sözcükler düzeyinde düzenlenir. Sözcüğün en küçük birimidir ve biçimbirimleri (morpheme) belirler. *irrationally* sözcüğü , *ir* (önek), *rational* (kök) ve *-ly* (sonek) olarak parçalanır. Sözcüksel Analiz, anlam birimler arasındaki ilişkiyi bulur ve sözcüğü kök biçimine dönüştürür. Sözcük POS (Part of Speech) olarak atanır. Dilin sözlüğünü oluşturulması hedeflenir. Sözcüğün olası türü (POS) atanır.

2. Sözdizimsel Analiz

❖ Metin parçasının dilbilgisi yapısının doğruluğu hedeflenir. Bunun için cümlenin öğelerine ayrıştırılması gerekir. Önceki adımda olası POS ğretildiğinden, cümle yapısına göre POS etiketleri atanır

Doğru: Sun rises in the east **Yanlış:** Rise in sun the east.

3. Anlamsal Analiz

«Elma bir muz yedi» cümlesi sözdizimsel olarak doğru olsa da, elmalar yiyemediği için mantıklı değildir. Anlamsal analiz incelene cümlede anlam arar. Ayrıca kelimeler ifadeler halinde birleştirilir. “Robert Hill” birlikte değerlendirilir. Robert ve Hill olarak incelenmez.

4. Söylem

Önceki cümlenin incelenen cümle üzerindeki etkisiyle ilgilenir. Metinde, “Jack parlak bir öğrencidir. O, zamanının çoğunu kütüphanede geçirir.” Burada söylem, “o”yu “Jack”e atıfta bulunmak için atar.

5. Pragmatik (faydacı) Metin önceki bir cümlenin ele alınan cümle üzerindeki etkisiyle ilgilenir. Verilen bir cümle, "Işıkları kapatın", ışıkları kapatma emri ya da isteğidir.

DDİ Uygulama Alanları

- **Makine Çevirisi** / *Machine Translation:* Bir metin ya da konuşmanın bir dilden diğerine çevrilmesidir
- **Bilgi Erişimi, Bilgi Geri Kazanımı** /I*nformation Retrieval*: Büyük veri kümelerinde ilgili bilgiyi bulmak, genellikle arama motorlarında kullanılır.
- **Bilgiyi Seçip Çıkarma** /*Information Extraction*: Yapılandırılmamış (unstructured) metinden yapılandırılmış (structured) bilgiyi otomatik olarak çıkarma.
- **Metin Madenciliği** /Text Mining Metinden analitik yöntemlerle nitelikli bilgiler türetilmesidir
- **Duygu Analizi** */ Sentiment Analysis* : Bir dizi kelimeye ait duygusal tonu belirleyerek ifade edilen tutumları, görüşleri ve duyguları anlamak.
- **Konuşma Tanıma** */ Speech Recognition:* Konuşma dilini metne dönüştürmek
- **Chatbots ve Sanal Asistanlar** İnsanlarla doğal dilde konuşabilen sistemler geliştirmilmesidir.

DDİ ile günümüzde ne yapılır?

- Yazım denetimi (spell check), çeviri (translation), sosyal medya izleme araçları gibi spesifik işlerin otomatikleştirilmesi gerçekleştirilir
- İnsan dilinin ***kural tabanlı modellemesi*** olan *hesaplamalı dilbilimin*, istatistiksel modelleme, makine öğrenimi (ML) ve derin öğrenmeyle birleştirilmesiyle bilgisayarların ve dijital cihazların metin, konuşma ve videoyu *tanımasını*, anlamasını ve metin, konuşma ve video *üretmesini* sağlar.

Hesaplamalı Dilbilim -HD - Computational Linguistics

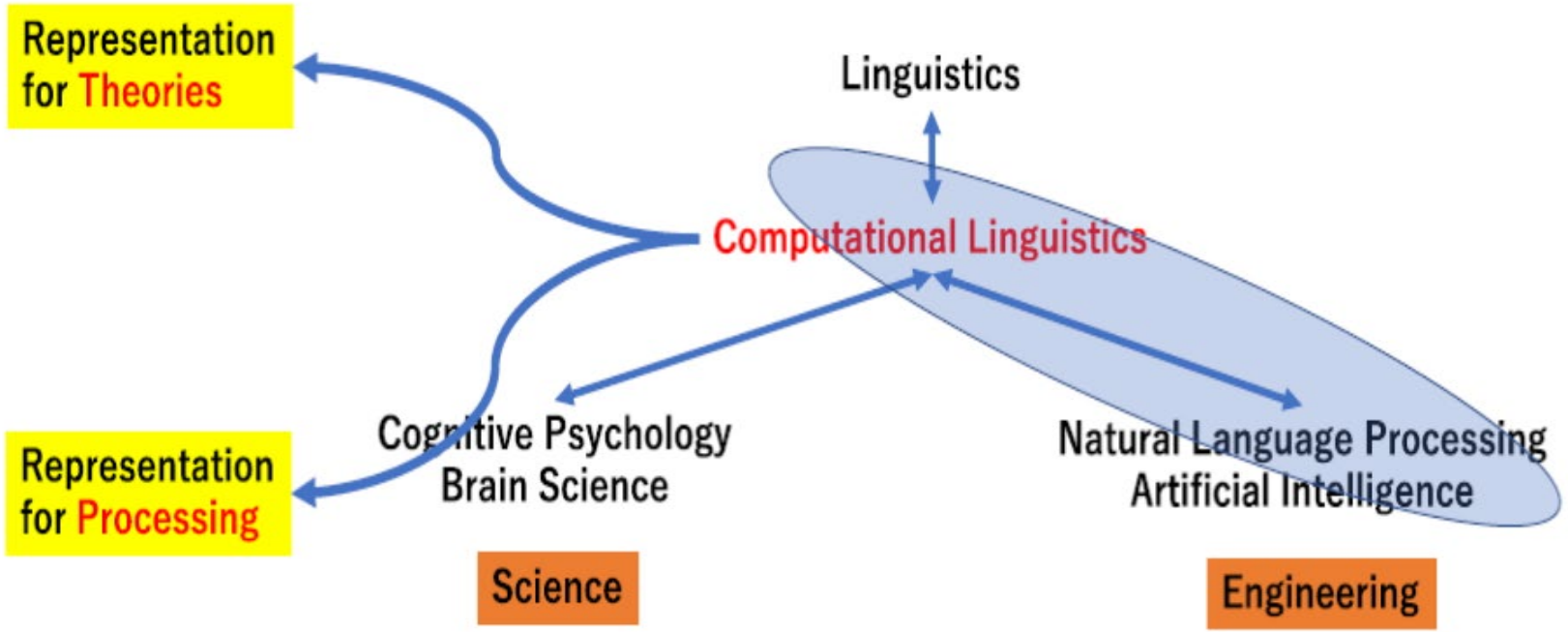
- Dilin (yazılı ve sözlü) analizi ve anlamada bilgisayar bilimini / algoritmaların kullanımını uygulayan disiplinler arası bir alandır.
- Dili hesaplamalı bir bakış açısıyla anlayabilmek üzere dilbilim, bilgisayar bilimi, yapay zeka (YZ), mühendislik, sinirbilim ve antropoloji birlikte incelenir.
- Bilgisayarın yazılı ya da sözlü olarak dili anlaması sağlandığında, yazılım ve makinelerle etkileşimi kolaylaşacaktır.
- HD uygulama alanları, müşteri hizmetleri, bilimsel araştırmalar, YZ araçları gibi çeşitli alternatiflerle sağlanır.

HD Niçin Önemlidir?

- Günümüzde insanlar yapılacak işlerini daha verimli tamamlamak üzere araçlar geliştirmek amacıyla teknolojiyi kullanmaktadır.
- HD ‘nin ilk uygulamaları, Çince gibi dilleri bilgisayar kullanarak İngilizceye çevirmek üzere ortaya çıkmıştır.
- Bugün ise bir chatbot ile bir ürünü iade etmeye çalışılmak veya iPhone'larda Siri yardımıyla hızlıca bilginin bulunması gibi müşteri hizmetlerini desteklemektedir.
- Hesaplamalı dilbilim, müşterilerin ne sorduğunu çözme ve yapay zekanın sürece dahil olması ile verilere dayanarak sorularına doğru yanıtlar vermesini sağlama sürecidir.

Hesaplamalı Dilbilim ve DDİ

- HD, makinelerin dilleri anlama, öğrenme veya çıktı oluşturma ile ilgili olarak sistem ve kavramlar üzerinde hesaplanabilirliğe odaklandığı ifade edilmişti.
- DDİ, bir bilgisayar programının insan dilinin yazıldığı veya konuşulduğu gibi anlamasını sağlayan işlemlerin uygulaması olarak tanımlanmıştı.
- Metin madenciliği, bilgi çıkarımı, makine çevirisi gibi uygulamalar HD ve DDİ birlikteliği ile çözümlenir.



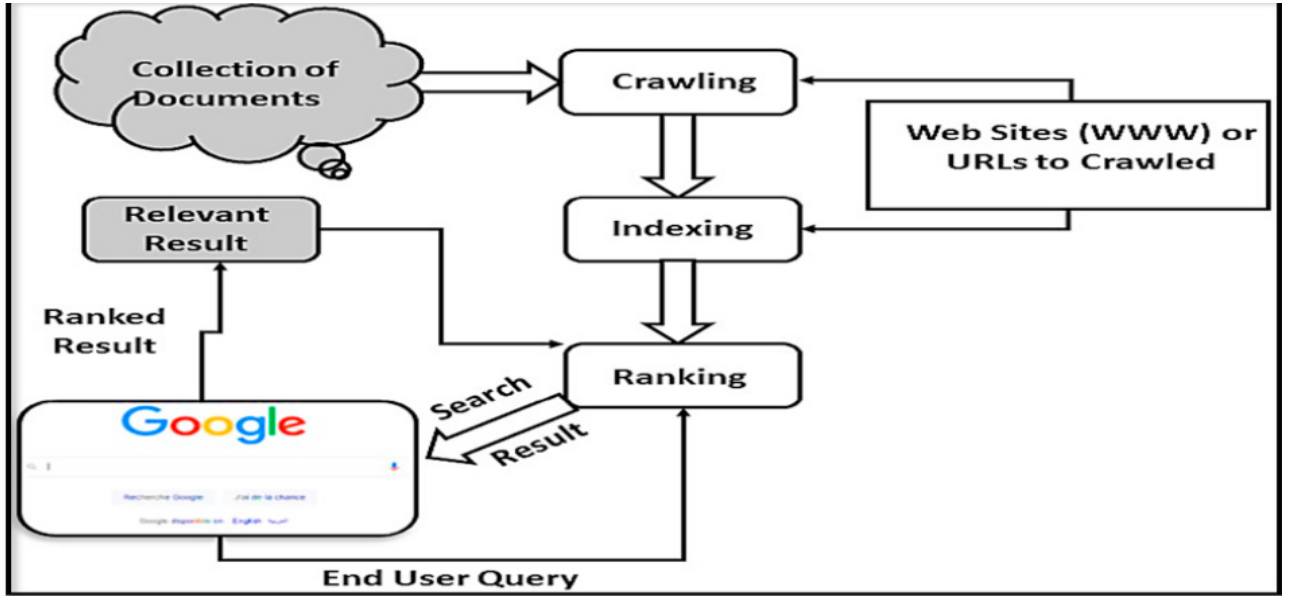
Dilin İşlendiği Uygulamalar

1970 li yıllar Soru Cevap Sistemleri
i) Dil ve Özbilgi (knowledge)
ii) Çıkarım odaklı dil yapıları
1980 ler ve 1990 ların ortaları Makine Çevirisi
i) İçeriklerin Çevirisi
ii) (Sığ) semantik yapıların transferleri
iii) Sözcükleştirilmiş (lexicalized) kurallar
1990 lar - 2010 Çözümleme (Parsing)
i)Farklı dilbilgisi formları arasındaki hesaplanabilirliğin denkliği
ii) Dilbilgisi formları odaklı çözümleme
2000 –ler Biyoloji alanında yapısal –tabanlı metin madenciliği
i)Dil ve Alan Özbilgisi
ii) Dilbilimsel yapıları kullanarak IE

Bilgiye Erişim - Information Retriewal (IR)

Günümüzde IR; web tarama /crawling veya web örümcekleme / spidering olarak adlandırılan otomatikleştirilmiş veri madenciliği yapılarak gerçekleştirilen *web kazıması* (scraping) işlemidir.

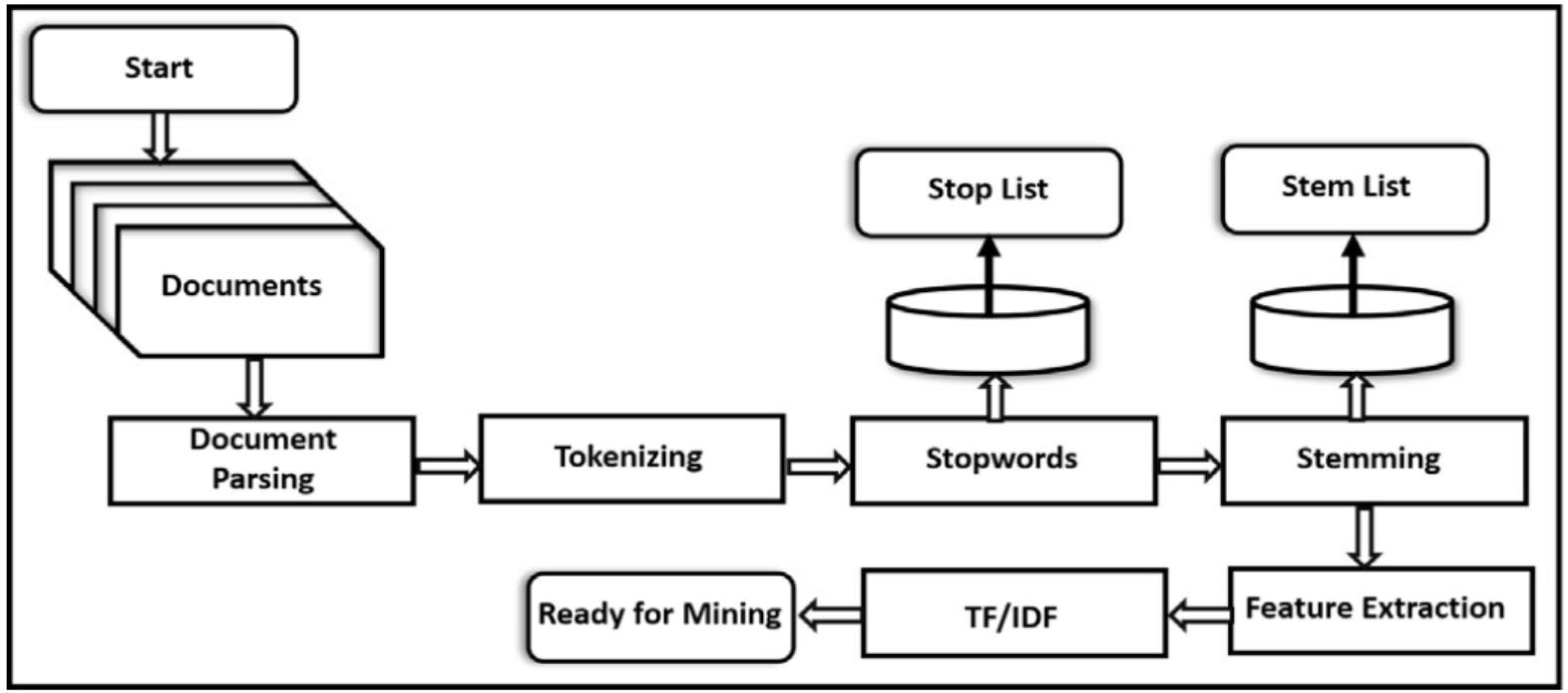
Bir web sayfasına ait koleksiyonunun bir program çerçevesinde incelenmesi, veri çıkarılması ve analiz işlemlerinin gerçekleştirildiği süreçtir.



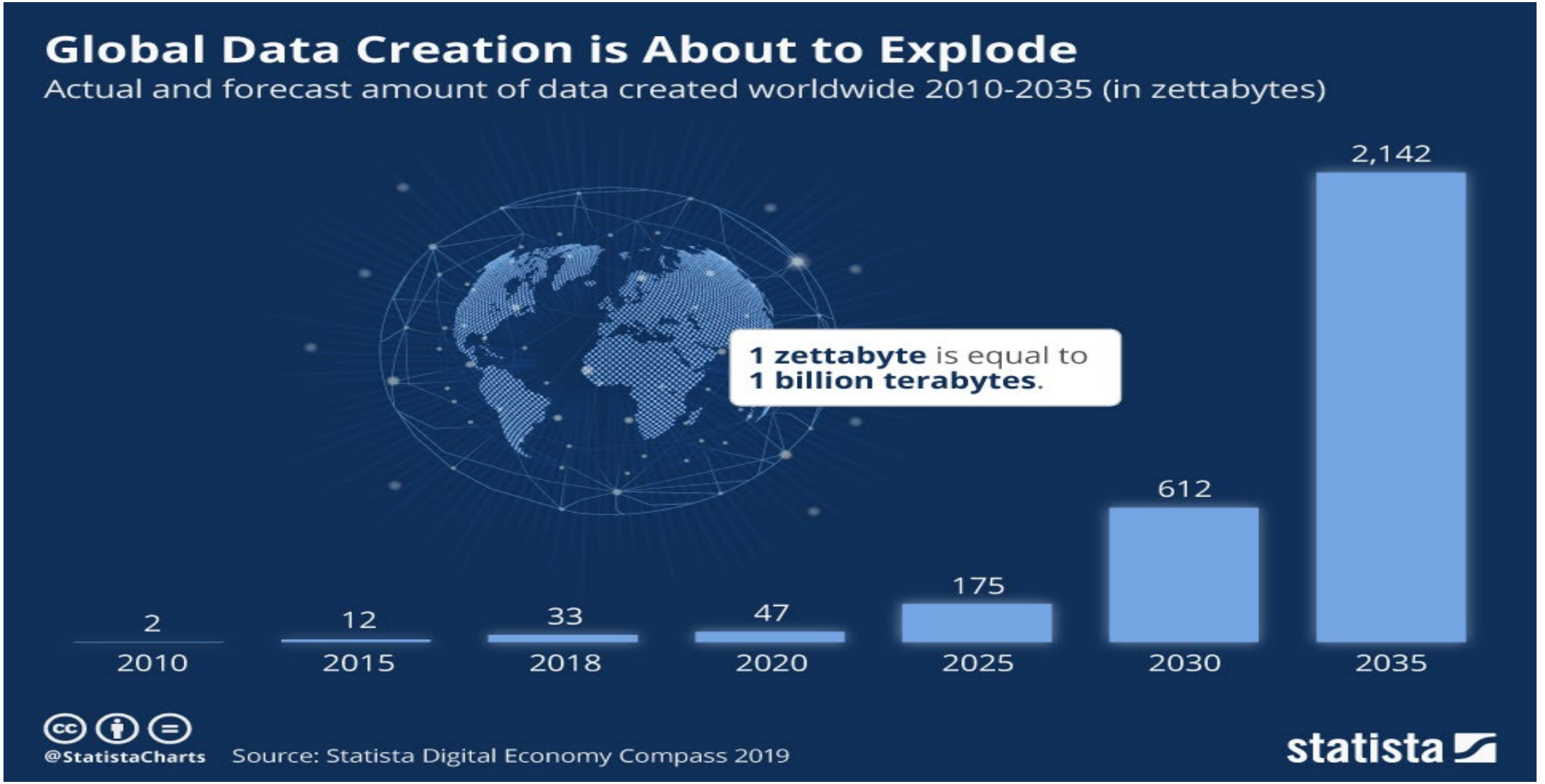
IR Bileşenleri

Veri kazıma (scraping) ile, dokümanlardan oluşan bir koleksiyondan veri (data) çıkartılır ve metni simgeleyen bir büyük bir *derlem* (corpus) ya da *niceliksel* (quantitative) *veri* elde edilir.

IR’de Önişleme Teknikleri



Niçin Önemli?



IR sürecinde Veri Kazıması Nasıl Gerçekleşir?

- Scraper, web sitelerindeki verileri toplamak için tasarlanmış programdır.
- Belirlenen URL'lerden istenen sayfaları getirir ve bu sayfalardaki HTML kodlarını analiz eder.
- Analizler sonucunda, kullanıcıların belirlemiş olduğu veri yapılarını çıkararak işlenebilir hale getirir.
- HTML' e ait belli etiketlerin içeriği veya tabloların içerisindeki veriler bu yöntemle elde edilmiş olur.

Bilgiyi Seçip Çıkarma /Information Extraction –IE

- Yapılandırılmamış bilgi yapılandırılmış data olarak elde edildiğinde IE gerçekleşir.
- Böylece ilişkisel bir veri tabanı elde edilerek sonraki işlemlere geçilebilir.
- IE gerçekleştirilmesi için, metnin içerisinden «named entity recognition» (NER) denilen isimlendirilmiş varlıklar elde edilir.
- Yapılan metinde adlandırılmış olan bir varlığın tanımlandığı her adlandırılmış varlığın (*named entity*) bulunması ve türünün etiketlenmesi «Named Entity Recognition» için ilk adımdır.

«Named Entity – Proper Names» nasıl belirlenir?

Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lowercost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as[LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

13 named entities: 5 organizasyon, 4 lokasyon, 2 zaman /times, 1 kişi /person, 1 para /money

Elde edilen *adlandırılmış varlıklardan* / named entities, dil işlenir ve pek çok uygulamada yararlanılabilir.

Bu yaklaşım, soru /cevap sistemleri, Wikipedia gibi özbilgi (knowledge) kaynaklarının işlenmesinde kullanılmaktadır.

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Golden Gate Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Veri Biliminin HD’ deki Rolü

- Veri bilimi, dili işleyen veya dili üreten ürünler oluşturmada yapılandırılmamış formatta büyük miktarda yazılı metni analiz eder.
- Bir chatbot'un veya uygulamanın yüksek kaliteli hizmet vermesini sağlar
- Mühendisler geliştirilen sistemin yönergelerini tanımlamak üzere hesaplamalı modeller kullanabilir.

HD Yaklaşımları

- 1950 li yıllardan beri HD üzerinde çalışmalar yapılmaktadır.
- ***Gelişimsel yaklaşım:*** Bir çocuğun zamanla bir dili öğrenmesi gibi, gelişimsel yaklaşım da bir dil edinme stratejisini simüle eder. Algoritmalar, dilbilgisi içermeyen istatistiksel bir yaklaşım oluşturmak üzere programlanır.

- ***Yapısal yaklaşım:*** Bu yaklaşım daha teoriktir ve dilin altında yatan yapıları daha iyi anlamak için dilin büyük örnekleri HD modelleri aracılığıyla çalıştırılır.

Gerçek Dünyada HD Örnekleri

Makine Çevirisi : YZ kullanarak bir dilden diğer dile çeviridir. Örnek «Google Translate»

Chatbots:. Müşteri hizmetleri bot için, konuşma dili ve yazılı dil aracılığıyla insan konuşmasını simüle eden yazılım programlarıdır. Amazon gibi şirketler, telefon ve e-postaya ek canlı sohbet olanağı sunar.

Bilgi Çıkarımı - Knowledge extraction: Yapısal olmayan ya da yapısal metin kaynaklarından özbilginin oluşturulması Örnek: Wikipedia,

Rastgele editörlerin ürünüdür ve açık bir bilgi çıkarıcının *precision* ve *recall* değerleri eğitilir.

Doğal Dil Arayüzü : Siri ve Alexa gibi araçlardır. Bu araçlar, insanın konuşma dilini kullanarak cihazların işletim sistemleriyle etkileşime girmesini sağlar.

Duygu Analizi - Sentiment analysis: Metindeki veya konuşulan dildeki duygusal tonu belirleyen bir DDİ aktivitesidir.

HD Yaklaşımları

Üretim yaklaşımı: Üretim yaklaşımı, metin üretmek için bir HD algoritması kullanır ve bu, metin tabanlı veya konuşma tabanlı etkileşimli yaklaşımlara ayrılabilir.

Metin tabanlı etkileşimli yaklaşım: Bu, bir insan tarafından yazılan metnin algoritmik bir yanıt oluşturmak için kullanıldığı üretim yaklaşımına girer. Bilgisayar daha sonra desenleri tanıyabilir ve kullanıcı girdisine ve anahtar kelimelere dayalı bir yanıt üretebilir.

Konuşma tabanlı etkileşimli yaklaşım: Metin tabanlı yaklaşıma benzer şekilde, bu yaklaşımda ses dalgaları ve desenler için konuşma girdilerini taramak üzere algoritmalar kullanılır.

Anlama yaklaşımı: Bu yaklaşımla, DDİ motoru basit kurallar kullanarak yazılı komutları doğal olarak yorumlayacak şekilde programlanır.

DDİ ‘nin Bileşenleri

- DDA, DDİ nin bir alt alanıdır.
- Dilin içeriğini, anlamının ve amacının anlaşılmasına odaklanılır.