

**FAQ**



## **Soru Cevaplama**

### **Question Answering - QA**



Doç.Dr.Banu Diri

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## **İÇERİK**

- Soru Cevaplama Nedir?
- TREC Yarışması
- Önceki Sistemler
- Tasarlanan Sistem
- Sistemin Akışı
- Sistemin Değerlendirilmesi
- Sonuç ve Öneriler

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



# Soru Cevaplama Sistemleri

Arama Motorları

=

Site Arama Motorları

Soru Cevaplama Sistemleri

=

Cevap Arama Motorları

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

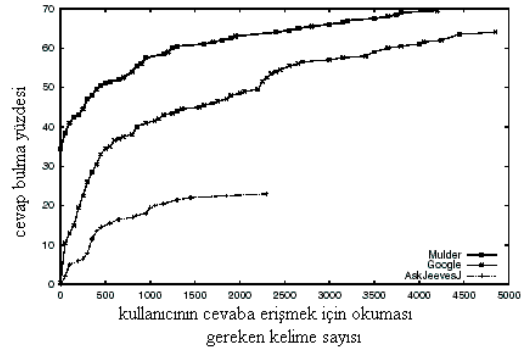


## Soru Cevaplama Sistemlerinin Avantajları

- Doğal Dil ile Çalışması
- Cevabı İçeren Sitelerin Listesi Yerine Cevabın Kendisini vermesi

Sonuç olarak;

Daha az kullanıcı emeği



YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## TREC (Text REtrieval Conference )Konferansları

- Büyük Dokümanlar (1 milyon gazete makalesi) İçinde Cevap Arayan Sistemlerin Yarıştığı Bir Konferans
- Her yıl yaklaşık 30 sistem katılıyor
- Her sisteme aynı sorular sorulup puan veriliyor

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Yarışmanın Kuralları

- Cevaplar dokümanlardan elde edilmeli
- Sistemlerden sıralanmış 5 adet cevaptan oluşan bir liste isteniyor



YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sistemlerin Puanlanması (Mean Reciprocal Rank)

- Sistemlerin verdiği sıralı cevap listesinden kaçının doğru cevabı içerdiğine göre sistemler puanlanır.

Sistemin  $i$ . sırada verdiği doğru cevap için aldığı puan

$$RR_i = \frac{1}{r_i}$$

Sistemin  $n$  adet sorudan aldığı puanların ortalaması (MRR)

$$MRR = \frac{1}{n} \sum_{i=1}^n RR_i$$

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## TREC Soru ve Cevap Örnekleri

Soru 1	How far is it from Denver to Aspen?
Cevap 1	The Aspen/Snowmass area is about 200 miles southwest of Denver.
Soru 2	What country is Modesto, California in?
Cevap 2	Modesto in Stanislaus Country.
Soru 3	Who was Galileo?
Cevap 3	Galileo Galileo, the astronomer.
Soru 4	What is an atom?
Cevap 4	Atoms, long considered to be the smallest units of matter.
Soru 5	What was the name of the first Russian astronaut to do a spacewalk?
Cevap 5	Aleksei A. Leonov.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## TREC ve Internet Veritabanları Arasındaki Farklar

	TREC	INTERNET
Doküman sayısı	<1,000,000	<3,000,000,000
Doküman türü	Gazete makaleleri	Her tür
İşlem	Internet'ten bağımsız	Internet'e bağımlı
Cevap formatı	50-250 bytes	Her türlü
Yanlış bilgi	Nadir	Çok fazla
Doğru cevabı içeren doküman sayısı	Az	Çok fazla

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Önceki Sistemler

- WEBCLOPEDIA
- QuASM (Question Answering System)

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



# WEBCLOPEDIA

- Southern California Üniversitesi,1999
- Soru Türü Kütüphanesi
- Cevap Şablonlarının Otomatik Olarak Toplanması
- Tanımlar için WORDNET

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Otomatik Şablon Toplama-1

- Örnek soru-cevap ikilileri alınmış (Mozart-1756),(Gandhi-1869) ve Altavista'ya sorgu olarak gönderilmiştir.
- İlk 1000 sonuç sayfasındaki bu ikililerin beraber geçtiği cevap cümlelerinde en uzun eşleşen parçalar bulunmuştur. Örneğin sonuç sayfalarından yandaki gibi cümleler elde edilmiştir.
- Cevap cümlelerindeki en uzun eşleşen parçadaki soru kelimesi;<Name>, cevap kelimesi ise <Answer> ile etiketlenmiştir.

Mozart (1756-1791).

Mozart was born on January 27, 1756

Mozart was born in 1756 in Salzburg, Austria

Mozart was born in Salzburg, Austria, Jan. 27, 1756, the son of Leopold Mozart

Mozart, born in 1756

Mozart was born there in 1756

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Otomatik Şablon Toplama-2

- Doğum yılını soran yukarıdaki ikilliler için
    - 1.00 <NAME> ( <Answer> - )
    - 0.87 <NAME> was born on < Answer > , in
    - 0.87 <NAME> was born on < Answer >
    - 0.86 <NAME> was born on < Answer > in
    - 0.86 <NAME> was born on < Answer >
- şeklinde birçok şablon elde edilmiştir.
- Şablonların başlarındaki ağırlık değerleri ise şablonun geçtiği doğru cevabı içeren cümle sayısının, şablonun geçtiği cevap cümlesi sayısına bölümünden elde edilmiştir.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## QuASM

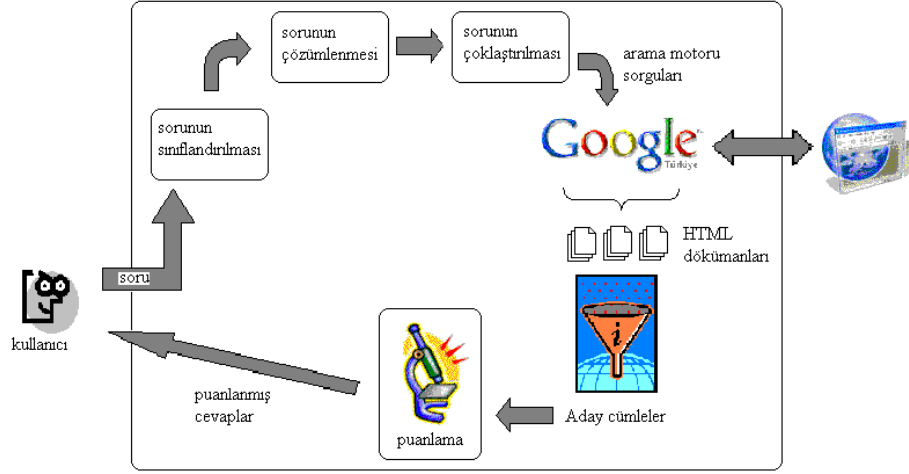
- Amerika'daki istatistiksel veritabanlarını sorgulamak için geliştirilmiş bir sistemdir (Pinto, 2000). Sistem [www.fedstats.gov](http://www.fedstats.gov) adresindeki yaklaşık 170,000 web sitesinde arama yapmaktadır.
- Kullanıcının girdiği soru incelenerek türü belirlenir. Bu işlem Naive Bayes metodu kullanılarak yapılmıştır. Soru ve türlerinden oluşan bir veritabanında kelimelerin hangi tür sorularda bulunduklarının olasılıkları ( $P(w_i|C)$ ) önceden çıkarılmıştır. Kullanıcı sorusunun hangi sınıfa ait olduğu ise aşağıdaki formülle bulunmuştur.

$$P(Q|C) = P(w_1|C) * P(w_2|C) * ... * P(w_n|C)$$

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Baybilmiş Soru Cevaplama Sisteminin Mimarisi\*



\* Hemen hemen tüm QA sistemleri benzer bir mimariye sahiptir.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sistemin Akışı

- Sorunun Tipinin Belirlenmesi
- Kullanıcı Sorusunun Çözülmesi
- Kullanıcı Sorusunun Çoklaştırılması
- Arama Motorundan Sonuçların Alınması
- Sonuç Sayfalarından Aday Cümlelerin Seçilmesi
- Aday Cümlelerin Puanlanması

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ





## Sorunun Tipinin Belirlenmesi

- Sorgudaki soru kelimeleri incelenmektedir. Elle oluşturulan bir tablo kullanılmıştır.

sorukelime	sorutürü	silinecekkelime
yüzde kaç	oran	kaç
yıl nedir?	yıl	nedir?
yıl kaçtır?	yıl	kaçtır?
yapıtaşları nelerdir?	bileşen	nelerdir?
tarih?	zaman	tarih?
tarih nedir?	zaman	tarih nedir?
tarifi nedir?	nedir	nedir?
şehri hangisidir?	il	hangisidir?
sebebi nedir?	neden	nedir?
sayısı nedir?	sayı	nedir?
nüfusu nedir?	nüfus	nedir?
nüfusu ne kadardır?	nüfus	ne kadardır?
nüfusu kaçtır?	nüfus	kaçtır?

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Kullanıcı Sorusunun Çözümlemesi

- Sistemde, ilk olarak kelimenin olası tüm gövdeleri bulunmaktadır. Örneğin; “ağacı” kelimesi için “ağ”, “ağa” ve “ağaç” kelimeleri gövde olarak bulunmaktadır.
- Çekim eklerinin varlığı araştırılmaktadır. Eğer gövde isim soylu ise kelimenin içerdiği isim çekim ekleri, şayet fiil soylu ise çekim ekleri ve fiilimsi ekleri araştırılır.
- Türkçe’deki eklerin birbirini içermesinden dolayı (örnek: -m, -miz gibi), bu işlemler sonucunda birden fazla muhtemel çözüm bulunmaktadır. Bu çözümlerden bir kısmı, kelimenin tamamını içermediğinden dolayı zaten elenir. Geriye kalan olası çözümler içerisinde doğru çözümü bulurken, kelimenin cümle içerisindeki durumuna bakılmaktadır. Örneğin, “evi” kelimesi için “ev” gövde olarak ve “-i” eki ise ya hal eki ya da iyelik eki olarak bulunur. Eğer cümlede “Ayşe’nin evi” benzeri bir tamlama varsa bunun iyelik eki olduğu anlaşılır. “Şu evi beğendim” cümlesinde ise “-i” ekinin hal eki olduğu anlaşılır.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Çözüm Örnekleri

Cümle	Orijinal kelime	Eklere ayrılmış kelime	Gövde ve eklerin tanımları	Kelimenin görevi
13 Aralık'ta okullar açıldı.	13	13	i[ɫ]	gagun
	Aralık'ta	aralık*da	i[ɫ]*bul	gagay
	okullar	okul*lar	i[ɫ]*y	
	açıldı	açıl*dı	f[ɫ]*dga	
29 Mayıs 1453'te okullar açıldı.	29	29	i[ɫ]	ttgum
	Mayıs	mayıs	i[ɫ]	ttay
	1453'te	1453*te	i[ɫ]*bhazyr	ttıyl
	okullar	okul*lar	i[ɫ]*y	
29 km yürüdüm.	açıldı	açıl*dı	f[ɫ]*dga	
	29	29	i[ɫ]	sayı
	km	km	i[ɫ]	uz
	yürüdüm	yürü*dü*üm	f[ɫ]*dga*setl	
29 km'lik yolu 3 saate aldık.	29	29	i[ɫ]	sayı
	km'lik	km*lık	i[ɫ]*bhazyr	uz
	yolu	yol*lu	i[ɫ]*bli	
	3	3	i[ɫ]	sayı
	saate	saat*te	i[ɫ]*yon	zaman
	aldık	aldı*ık	f[ɫ]*dga*seçl	
Ali üniversiteyi 3 yılda bitirdi.	Ali	ali	ö[ɫ]	özelisim
	üniversiteyi	üniversite*yi	i[ɫ]*dh*bli	
	3	3	i[ɫ]	sayı
	yılda	yıl*da	i[ɫ]*bul	zaman
Ankara ilini gördüm.	bitirdi	bitir*di	f[ɫ]*dga	
	Ankara	ankara	ö[ɫ]	özelisim
	ilini	il*ını*	i[ɫ]*at3*dh*bli	yer
Bir fil 2500 kilodur.	gördüm	gür*dü*üm	f[ɫ]*dga*setl	
	Bir	bir	st[st]	
	fil	fil	i[ɫ]	
	2500	2500	i[ɫ]	sayı
	kilodur	kilo*dur	i[ɫ]*f_gaz	ag

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Kelimelerin etiketlemesinde kullanılan anahtar kelimelere örnekler

[yer]	[zaman]	[ag]	[uz]
ili	dakika	gram	metre
barajı	saat	kilo	km
dağı	gün	kg	kilometre
gölü	hafta	kilogram	santim
şehri	ay	ton	cm
yöresi	yıl	litre	santimetre
bölgesi	asır		mil
kenti			

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Kullanıcı Sorusunun Çoklaştırılması

- Kullanıcının sorduğu sorudan iki adet arama sorgusu elde edilmiştir.
  - Soru kelimeleri atılarak elde edilen
  - Çekim ekleri de atılarak elde edilen

Kullanıcının Sorusu	1. sorgu Soru kelimeleri atılmış hali	2. Sorgu Eklerinden arındırılmış hali
Asya ile Avrupa'yı ayıran boğazın adı nedir?	Asya ile Avrupa ayıran boğazın adı	Asya ile Avrupa ayır boğaz ad
enzimlerin vücuttaki görevi nedir?	enzimlerin vücuttaki görevi	enzim vücut görev
Deli Dana hastalığının belirtileri nelerdir?	Deli Dana hastalığının belirtileri	Deli Dana hastalık belirti
dünyanın en uzun boylu insanı kimdir?	dünyanın en uzun boylu insanı	dünya en uzun boy insan
su hangi atomlardan oluşur?	su atomlardan oluşur	su atom oluş

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Google

Google'da Ara: Öğrenme algoritmaları - Microsoft Internet Explorer

Dosya Düzen Görünüm Sık Kullanılanlar Araçlar Yardım

Geri İleri Duraklatma Ara Sık Kullanılanlar Ortam

Adres http://www.google.com.tr/search?hl=tr&ie=UTF-8&oe=UTF-8&q=%C3%96%C4%9Frenme+algoritmalar

Google™ Gelişmiş Arama Tercihler Dil Araçları Arama İpuçları

Öğrenme algoritmaları Google'da Ara

Web'de Ara Web'de Ara Türkçe sayfalarda ara

Web Grafikler Gruplar Dizin

Web Öğrenme algoritmaları için arandı.

[Ethem Alpaydın'a ödül](#)  
... Ödül alma nedenim Yapay **Öğrenme**, özellikle de yapay sinir ağlarındaki **öğrenme algoritmaları** konusunda yaptığım çalışmalar. ...  
[www.boun.edu.tr/~pubre/news/focak2003/6.htm](#) - 7k - [Önbellek](#) - [Benzer sayfalar](#)

[Derslerin Listesi](#)  
... Sınırlanmamış optimizasyon ve **öğrenme algoritmaları**. Lineer, quadratik programlama ve lineer türetilen problemler için sinir ağları. ...  
[www.fbe.ktu.edu.tr/~anabilimdallari/bilgisayar/bilgisidet.html](#) - 24k - [Önbellek](#) - [Benzer sayfalar](#)

[İstanbul Tabip Odası](#)  
... Bölümü D Bilgisayar mühendisliği alanında yapay **öğrenme** ve özellikle yapay sinir ağlarındaki **öğrenme algoritmaları** konusundaki uluslararası ...  
[www.istabip.org.tr/gene/tubitak2002.asp](#) - 18k - [Önbellek](#) - [Benzer sayfalar](#)

[BASKENT ÜNİVERSİTESİ BİLGİSAYAR MÜHENDİSLİĞİ ...](#)  
... Sinirsel çağrışım yardımcı bellekler, şekil sınıflayıcılar, fonksiyon yaklaşımları ve **öğrenme algoritmaları**. ...  
[www.cs.baskent.edu.tr/fylderslericenk.html](#) - 23k - [Önbellek](#) - [Benzer sayfalar](#)

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



- En büyük içerik:  
8,083,324,652 web sayfası

- En hızlı arama
- En iyi puanlandırma tekniği
- Sayfa özetleri
- Önbelleğe sahip

## Arama Motorundan Sonuçların Alınması

- Cevap cümlelerinin elde edilebileceği iki kaynak bulunmaktadır.
  - Arama motorunun sonuç sayfasındaki sayfa özetlerinden
  - Sonuç sayfasındaki bağlantıların gösterdiği sayfalardan
- Her iki yaklaşımın avantaj ve dezavantajları aşağıda verilmiştir.

Cevabın nerede aranacağı	Avantajları	Dezavantajları
Sadece sayfa özetlerinde arama	hızlı	Cümleler kesik ve anlamsız dolayısıyla daha az doğru sonuçlar (%43.9)
Sayfaların kendilerinde arama	Cümleler tam ve anlamlı ve dolayısıyla daha doğru cevaplar (%43.9)+%16.6	Çok yavaş

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Google'ın Önbelleği

- Sayfaların kendilerinden cevap üretilirken sayfaların orijinal hallerinin yerine Google'ın önbelleğindeki halleri kullanılmıştır. Bunun iki önemli avantajı vardır.
  - Sayfalardaki bilgilerin indirilmesi daha az zaman almaktadır
  - Sayfanın arama yapılırkenki (belki değiştirilmiş ya da yayından kaldırılmış) içeriği yerine Google'da indekslendiği hali kullanılarak daha doğru bilgiye erişim sağlanmaktadır.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sonuç sayfalarından aday cümlelerin seçilmesi

- Bir cümlemin cevap cümlelerinden olabilmesi için aşağıdaki şartları sağlaması gerekmektedir.
  - İçinde soru kelimelerinin sayısının en az yarısı kadar soru kelimesi içermeli
  - Sorgudaki özel isim olarak etiketlenmiş en son kelimeyi içermeli
  - Soru türü, cevapta sayısal bir değer beklenen türdence (yıl, mesafe, oran, ağırlık vb.) içinde sayı bulundurmalıdır
- Puanlama işlemi zaman alıcı bir işlem olduğundan, yukarıdaki filtreleme sayesinde sistemin cevap verme süresi kısaltılmıştır

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Aday cümlelerin puanlanması

- Filtreden geçebilen cümleler 3 farklı şekilde puanlanmaktadır:
  - Eşleştirme puanı
  - Sıralama puanı
  - Şablon puanı

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Eşleştirme puanının belirlenmesi

- Cümlelerin içerdiği sorgu kelimesi sayısı ile bulunur
- Ör: “Boğaz Köprüsü’nün yüksekliği ne kadardır?” sorusu için

Cevap cümlesi	Eşleştirme puanı
Boğaz Köprüsü’nün yüksekliğinin yaklaşık 3 katına eşdeğerdir .	3
64 metre yüksekliğindeki Boğaz Köprüsü’yle, Saipem 7000 .	3
Boğaz Köprüsü’nün denizden yüksekliği ise 60 metre.	3
Mesela Boğaziçi Köprüsü’nün deniz seviyesinden yüksekliği sadece 64 .	3
boyu 165 metre olan Birinci Boğaz Köprüsü’nün .	2

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sıralama puanının hesaplanması-1

- Cümle içindeki sorgu kelimelerinin birbirine yakınlığı, cevabın o cümlede olma olasılığını arttırmaktadır. Bu yüzden sorgudaki kelimelerin, cümledeki sorgu kelimelerinin sayısına ve yakınlığına bağlı olarak bir formül yardımıyla 0-10 arasında bir sıralama puanı hesaplanır.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sıralama puanının hesaplanması-2

$$SP = \left\lfloor \frac{10 * C_k * (C_k - 1)}{Tp * S_k} \right\rfloor$$

Parametreler:

- $C_k$  cümlelerin içerdği sorgu kelimesi sayısı
- $S_k$  sorgudaki kelime sayısı
- $Tp$  cümledeki sorgu kelimelerinin birbirlerine kelime bazındaki uzaklıklarının toplamı

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sıralama puanının hesaplanması-3

- Ör: “A B C D” sorgu cümlesi için  
“A E F G H B C” cevap cümlesinin sıralama puanının hesaplanması:  
Cevap cümlesindeki sorgu kelimelerinin cümledeki kaçınıcı kelime oldukları bulunur. (1 – 6 – 7)  
 $Tp$  ifadesi; bu dizideki her rakamın bir sonrakinden farkı bulunduktan sonra, bu farkların mutlak değerleri toplanarak elde edilir.  
 $Tp = |(1-6)| + |(6-7)| = 6$   
Örneğin sıralama puanı  
 $SP = \lfloor 10 * 3 * (3-1) / (6 * 4) \rfloor = 2.5$

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sıralama puanının hesaplanması-4

- Ör: “Boğaz Köprüsü’nün yüksekliği ne kadardır? ” sorusu için

Cevap cümlesi	Sıralama puanı
64 metre yüksekliğindeki Boğaz Köprüsü’yle, Saipem 7000 arasında yaklaşık 8 metre mesafe kalacak.	10
Fatih Sultan Mehmet Köprüsü’nün güvenli geçiş yüksekliği 64 metre olduğu için bu köprünün altından geçişte bir sorun görülüyor.	2
Gemi, denizden 64 metre yüksekliğinde olan Boğaziçi Köprüsü’nün altından yaklaşık 10 dakikada geçişini tamamladı.	6

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Şablon puanının belirlenmesi -1

- Kullanıcının girdiği soru türüne ait şablonlar veritabanından çekildikten sonra cümlede bu şablonların olup olmadığı araştırılır. Eğer varsa o var olan şablona ait ağırlık değeri cümlenin şablon puanı olarak belirlenir.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ





## Şablon puanının belirlenmesi -2

- Örnek olarak “plastik neden elde edilir? ” sorusunun türü “bileşen” olarak belirlendikten sonra; şablonlar veritabanından “bileşen” sorusuna cevap olabilecek şablonlar seçilir.

Şablonlar		
sorutipi	Cevaptipi	ağırlık
bileşen	orj=ana%maddesi	5
bileşen	orj=elde%edil	6
bileşen	orj=hammadde	5
bileşen	orj=içerir	5
bileşen	orj=içinde	5
bileşen	orj=kullanılarak	5
bileşen	orj=yapılır	7
bileşen	tnm=ayr#orj=elde%edil	9
bileşen	tnm=ayr#orj=oluş	7
bileşen	tnm=ayr#orj=üretilir	8
bileşen	tnm=ayr#orj=yapılır	9

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Şablonlar Veritabanının Oluşturulması-1

- Şablonlar iki farklı yolla elde edilmiştir:
  - “Nedir” şablonlarının oluşturulması
  - “Nedir haricindeki” şablonların oluşturulması

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Şablonlar Veritabanının Oluşturulması-2

- “Nedir” şablonlarının oluşturulması:
  - “Nedir” soruları bir kelimenin tanımını sorduğu için Internet'teki bir sözlükten yararlanılmıştır. Yaklaşık 4000 Türkçe kelimeden oluşan listedeki kelimeler Türk Dil Kurumu'nun Internet sitesindeki sözlüğe (<http://www.tdk.gov.tr/tdksozluk/sozara.htm>) gönderilip sonuç sayfalarındaki kelime anlamları bir dosyaya kaydedilmiştir.
  - Bir sözlükteki en çok tekrar eden kelimeler Türkçe'deki tanım ifadelerini içereceğinden; kelime anlamlarını içeren dosyadaki kelimelerden en çok tekrarlanan 50 kelime (Türkçe'de en sık tekrarlanan kelimeler çıkarıldıktan sonra) “nedir sorularının” şablonları olarak belirlenmiştir.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Şablonlar Veritabanının Oluşturulması-3

“Nedir” sorularının şablonları

orj=:	orj=gösteren	orj=olarak%bilin
orj=ad	orj=herhangi	orj=olarak%tanım
orj=ad%verilir	orj=için%	orj=olmak
orj=alan	orj=için%kullan	orj=olmayan
orj=alınan	orj=ifade%eder	orj=oluşan
orj=amacıyla	orj=ilgili	orj=oluşturan
orj=anlamında	orj=ismidir	orj=saglamak
orj=araç	orj=kalan	orj=saglayan
orj=aracı	orj=karşılık	orj=tanım
orj=bakımından	orj=kimse	orj=tanımlamak%için%kullan
orj=biçimde	orj=kısaltılmış%halidir	orj=tarafından
orj=biçiminde	orj=kısaltma	orj=türlü
orj=bulunan	orj=kısaltması	orj=veren
orj=denir	orj=kişi	orj=verilen
orj=dir%.	orj=konulan	orj=verilen%addır
orj=duruma	orj=kullanılan	orj=yani
orj=eden	orj=kullanılır	orj=yapan
orj=edilen	orj=madde	orj=yapıl
orj=etmek	orj=olan	orj=yarayan
orj=genellikle	orj=olarak	orj=yaşayan

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Şablonlar Veritabanının Oluşturulması-4

- Nedir haricindeki soruların şablonlarının oluşturulması
  - Her soru türü için soru cevap çiftleri belirlenmiş ve Internet'te soru ve cevabın birlikte geçtiği cümleler bulunmuştur. Bu cümleler incelenerek, her soru türü için sıkça kullanılan kelimeler ve yapılar elle belirlenmiştir.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Şablonlar Veritabanının Oluşturulması-5

Zaman şablonlarının elde edilmesinde kullanılan cümlelerden örnekler

X (Y -	ERZURUM KONGRESİ (23 Temmuz -
Y X	23 TEMMUZ ERZURUM KONGRESİ
X Y de	Erzurum Kongresi 23 Temmuz 1919'da
Y günü ... X	4 Eylül 1919 günü toplanan Sivas Kongresi
Y de ... X	4 Eylül 1919'da toplanan Sivas Kongresi'nin
X, Y günü	Sivas Kongresi, 4 Eylül 1919 günü
Y tarihinde ... X	17 Ağustos 1999 tarihinde saat 03.02'te meydana gelen
Y X (Y)	13 Kasım, Napolyon Viyana'ya girdi (1805).
Y tarihli X	3 Kasım 1839 tarihli Gülhane Hattı Hümayunu'nun
Y: X	19 Temmuz: Paris'te metro hizmete girdi.
Y yılında ... X	25 Haziran 1950 yılında çıkan Kore Savaşı

X' ler soruyu, Y' ler ise cevabı temsil etmektedir.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Cevapların kullanıcıya verilmesi

- 3 farklı şekilde puanlandırılan cümlelerin aldıkları puanlar toplanarak o cümlenin puanı elde edilmiştir.
- Puanlanmış cümlelerden, birbirinden farklı olan ve en yüksek puanı almış ilk 5 cümle kullanıcıya cevap olarak iletilmektedir.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Uygulama Örnekleri-1

soru	Cevaplar
TBMM ne zaman açıldı?	1 İste hemen sonra da Ankara'da 23 Nisan 1920 tarihinde TBMM açıldı.
	2 Sonuçta 23 Nisan 1920 Tarihinde Ankara'da TBMM açıldı ve Türk .
	3 23 Nisan 1920'de geniş bir katılımla TBMM açıldı.
	4 Teklif, bugün (24 Şubat 2000) koalisyon partileri genel başkanlarının düzenledikleri zirveden hemen sonra TBMM'de imzaya açıldı.
	5 1 Kasım 1927, TBMM'nin üçüncü dönemi açıldı.
Gaziantep'in kurtuluş günü ne zamandır?	1 Şöyle ki, bugün, 25 Aralık 1999, Gaziantep İlimizin kurtuluş günüdür.
	2 25 Aralık'ta Gaziantep'in kurtuluş gününde açacağımız sergiye gelecek insanlardan böyle bir grup çıkarabileceğimizi ve gelecek sene beş .
	3 -
	4 -
	5 -
Tarkan ne zaman doğdu?	1 Tarkan Süalp 1968 yılında İstanbul'da doğdu.
	2 TARKAN GÖZÜBÜYÜK Tarkan Gözübüyük Bass Gitar & Vokal 1970 yılında Erzurum'da doğdu.
	3 Tarkan Gözübüyük: Bass, Vocal 1970 yılında Erzurum'da doğdu.
	4 -
	5 -

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Uygulama Örnekleri-2

Türkiye İnternet ile ne zaman tanıştı?	1	12 Nisan 1993'de Ankara-Washington 64K'lık bağlantısı ile Türkiye İnternet'le tanıştı.
	2	Türkiye İnternet ile 12 nisan 1993'de tanıştı ve İnternetimiz 6.
	3	12 Nisan 1993'te Ankara-Washington arasında kiralık hatla kurulan bağlantı ile Türkiye internetle tanıştı.
	4	Türkiye İnternet'le 12 nisan 1993'de ODTÜ'den gerçekleştirilen Ankara-Washington bağlantısı ile tanıştı.
	5	-
Boğaz Köprüsü'nün yüksekliği ne kadardır?	1	Boğaz Köprüsü'nün yüksekliğinin yaklaşık 3 katına eşdeğerdir .
	2	64 metre yüksekliğindeki Boğaz Köprüsü'yle, Saipem 7000 .
	3	Boğaz Köprüsü'nün denizden yüksekliği ise 60 metre.
	4	Mesela Boğaziçi Köprüsü'nün deniz seviyesinden yüksekliği sadece 64 .
	5	boyu 165 metre olan Birinci Boğaz Köprüsü'nün .
Ay'ın Dünya'dan uzaklığı ne kadardır?	1	Pembe-mavi benekli bu uydunun Neptün'e uzaklığı, yaklaşık olarak Dünya-Ay arası uzaklığı kadardır (354.200 Km.
	2	Ay'ın Yer yüzeyine uzaklığı 350000 km ile 400000 km .
	3	Ay'ın Dünya'ya olan ortalama uzaklığı 384000 km dir.
	4	Yer çevresinde yaklaşık olarak her 48 saatte bir dönmekte ve bu hareketi sırasında kimi zaman Yer'e Ay uzaklığı kadar yaklaşmakta, kimi zaman da Ay .
	5	Bu uzaklık, Ay-Dünya arasındaki uzaklığın (6) katıdır.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## İnternet'teki bilginin güvenilirliği

‘Atatürk öldüğünde kaç yaşındaydı?’ sorusu için  
cevap:

15 yıllık Cumhurbaşkanlığı  
olan Atatürk, öldüğünde 54 yaşındaydı.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sistemin Değerlendirilmesi-1

- Soru cevaplama sistemlerinin değerlendirilmesi oldukça güç bir işlemdir. TREC yarışmalarında bile sistemlerin değerlendirilmesi elle yapılmaktadır.
- Birbirinden farklı sorularla test edilen iki sistemin birbirleriyle karşılaştırılması pek bilimsel bir yöntem değildir.
- Bu yüzden geliştirilen sistemin diğer sistemlerle karşılaştırılabilmesi için aynı sorularla olmasa bile aynı tür sorularla test edilebilmesi için öncelikle TREC-9 ve TREC-10 soruları yarışmanın sitesinden toplanmış ve daha sonra Türkçe'ye çevrilmiştir.
- Türkçe web sitelerinde cevabı olması mümkün gözükmeyen bazı özel sorular elendikten sonra 4 kişiye bu soruları başka türlü nasıl sorabilecekleri sorulmuştur.
- Kişilerden gelen sorular birleştirilerek 524 adet test sorusu elde edilmiştir.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sistemin Değerlendirilmesi-2

- Cevapları sadece Google'ın sonuç sayfasında arayan sisteme, 524 soru verilmiş ve cevapları aldıkları puanlarla birlikte bir veritabanına kaydedilmiştir. Sistemin verdiği cevaplardan oluşan veritabanını incelemek için bir sonraki slide'ta görülen arabirim tasarlanmıştır.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sistemin Değerlendirilmesi-3

**Cevapların Değerlendirilmesi**

Kriter =  Soru Göster

Soru No = 20030600\_1720-45\_15 Soru Türü = Zaman

Soru = Amerika bağımsızlık günü

**CEVAPLAR :**

Eşleme	Cevap - 1
Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde * 4'ün Today (Bugün) adı programında Amerika'nın bağımsızlık Eylül 2001 Çarşamba Bİlindiyi üzere 11 Eylül 2001 Salı günü Amerika'da çok Amerika'nın Bağımsızlık Günü'nden itibaren; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günleri	<input type="radio"/> 0 <input type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5

**Eşleme - Sıralama**

Cevap - 2	
4'ün Today (Bugün) adı programında Amerika'nın bağımsızlık Eylül 2001 Çarşamba Bİlindiyi üzere 11 Eylül 2001 Salı günü Amerika'da çok Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde * Amerika'nın Bağımsızlık Günü'nden itibaren; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günleri	<input type="radio"/> 0 <input type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5

**Eşleme - Şablon**

Cevap - 3	
Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde * Eylül 2001 Çarşamba Bİlindiyi üzere 11 Eylül 2001 Salı günü Amerika'da çok Amerika'nın Bağımsızlık Günü'nden itibaren; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günleri 4'ün Today (Bugün) adı programında Amerika'nın bağımsızlık	<input type="radio"/> 0 <input type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5

**Sıralama - Şablon**

Cevap - 4	
Eylül 2001 Çarşamba Bİlindiyi üzere 11 Eylül 2001 Salı günü Amerika'da çok Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde * 4'ün Today (Bugün) adı programında Amerika'nın bağımsızlık Amerika'nın Bağımsızlık Günü'nden itibaren; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günleri	<input type="radio"/> 0 <input type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5

**Eşleme - Sıralama - Şablon**

Cevap - 5	
Eylül 2001 Çarşamba Bİlindiyi üzere 11 Eylül 2001 Salı günü Amerika'da çok Hürriyet, Amerika'daki Bağımsızlık günü kutlamalarına ilişkin 4 Temmuz tarihli haberinde * 4'ün Today (Bugün) adı programında Amerika'nın bağımsızlık Amerika'nın Bağımsızlık Günü'nden itibaren; 4 Temmuz, oğlum Berent'in doğumgünü :) Geçtiğimiz günleri	<input type="radio"/> 0 <input type="radio"/> 3 <input type="radio"/> 1 <input type="radio"/> 4 <input type="radio"/> 2 <input type="radio"/> 5

Soru Puanı Kaydet

## Sistemin Değerlendirilmesi-4

- Sistemin verdiği cevaplar 5 farklı puanlama tekniğine göre sıralanmıştır
  - 1. eşleme
  - 2. eşleme+sıralama
  - 3. eşleme+şablon
  - 4. sıralama+şablon
  - 5. eşleme+sıralama+şablon
- Doğru cevabın kaçınıcı sırada olduğu bir tabloya kaydedilmiştir.



## Sistemin Değerlendirilmesi-5

Soru ve doğru cevap sayılarının soru türlerine göre dağılımını gösteren tablodan bir kesit

Soru türü	Soru Sayısı	Eşleme	Eşleme+ Sıralama	Eşleme+ Şablon	Sıralama+ Şablon	Eşleme+ Sıralama+ Şablon
neden	6	2	1	3	1	1
nedir	117	33	33	40	39	40
nelerdir	13	2	2	2	2	2
nerededir	20	12	12	13	13	13
nesne	11	2	2	2	2	2
nüfus	10	7	6	6	7	5
olumzaman	3	3	3	3	3	3
oran	11	8	7	8	7	7
para	4	2	1	2	2	2
renk	7	1	1	1	1	1
sayı	26	15	15	14	14	14
şehir	3	1	1	1	1	1
sembol	1	0	1	1	1	1
süre	7	1	1	1	1	1
süreyl	1	1	1	1	1	1
yıl	11	3	3	3	3	3
zaman	50	22	21	21	22	22
toplam	524	224	210	230	218	220
		42.70%	40.00%	43.90%	41.60%	41.98%

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sistemin Değerlendirilmesi-6

Soru türlerine göre MRR puanlarının dağılımını gösteren tablodan bir kesit

Soru türü	Soru Sayısı	Eşleme	Eşleme+ Sıralama	Eşleme+ Şablon	Sıralama+ Şablon	Eşleme+ Sıralama+ Şablon
neden	6	0.25	0.083333333	0.305555556	0.083333333	0.083333333
nedir	117	0.178917379	0.197578348	0.212393162	0.235327635	0.242735043
nelerdir	13	0.115384615	0.115384615	0.115384615	0.115384615	0.115384615
nerededir	20	0.268333333	0.276666667	0.435	0.405833333	0.433333333
nesne	11	0.181818182	0.090909091	0.181818182	0.075757576	0.090909091
nüfus	10	0.483333333	0.375	0.458333333	0.475	0.35
olumzaman	3	0.777777778	0.777777778	0.777777778	0.777777778	0.777777778
oran	11	0.545454545	0.590909091	0.545454545	0.454545455	0.5
para	4	0.375	0.25	0.375	0.333333333	0.333333333
renk	7	0.028571429	0.028571429	0.028571429	0.028571429	0.028571429
Sayı	26	0.509615385	0.477564103	0.490384615	0.448717949	0.467948718
şehir	3	0.333333333	0.333333333	0.333333333	0.333333333	0.333333333
sembol	1	0	1	0.333333333	1	1
Süre	7	0.142857143	0.071428571	0.142857143	0.071428571	0.071428571
Süreyl	1	1	1	0.5	0.5	0.5
Yıl	11	0.151515152	0.2	0.227272727	0.212121212	0.212121212
Zaman	50	0.362333333	0.361666667	0.373333333	0.385	0.395

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ





## Sonuç-1

- En fazla soruya doğru cevap veren puanlama türü  
– eşleme+şablon (%43.9)
- En yüksek MRR puanına sahip puanlama türü  
– eşleme+sıralama+şablon (0.313)

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Sonuç-2

TREC yarışmasının en iyileri (MRR puanları)

	TREC-8	TREC-9	TREC-10
En iyi	0.66	0.58	0.68
En iyi ikinci	0.56	0.32	0.59

Sistem örnekleri (MRR puanları)

Sistem	MRR
Quantum	0.223
Shapaqa	0.545
Webslopedia	0.435
QuASM	0.253
AskMSR	0.507

Geliştirilen Sistemin Puanlama türlerine göre MRR puanları

	Soru Sayısı	Eşleme	Eşleme+ Sıralama	Eşleme+ Şablon	Sıralama+ Şablon	Eşleme+ Sıralama+ Şablon
MRR	524	0,298	0,289	0,308	0,301	<b>0,313</b>

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Gelecek Çalışmalar İçin Öneriler

- Sistemde kullanılan **şablonlar veritabanı** daha da geliştirildiğinde sistemin performansının artacağı kuşkusuzdur.
- Sisteme **yeni kelime etiketlerinin (Name Entity)** eklenmesi, şablonların ve filtrelerin dolayısıyla tüm sistemin performansını arttıracaktır.
- Aday cevapların puanlanmasında ve kullanıcı sorusunun arama motoru sorgularına dönüştürülmesinde, **eşanlamlı kelimelerin ya da yapıların** kullanılması sistemin başarısına en büyük katkıyı sağlayacaktır.
- **Google yerine Türkçe için tasarlanmış olan arama motorları** içerik olarak daha da zenginleştiği takdirde başarıya olumlu bir katkısı olacaktır. Örneğin Google'da 'tuz' ifadesi aratıldığında içinde 'tuzlu' olan sayfalar döndürülmemektedir. Bu durum Türkçe için tasarlanmış arama motorlarında söz konusu değildir. Örneğin www.arama.com'da bulunan arama motoru içinde 'tuzsuz', 'tuzlama', 'tuzlanıp' kelimelerini içeren sayfaları da döndürmektedir. Sistemin bu tür arama motorlarında kullanılabilmesi için sadece arama sorgusundaki kelimelerin çekim eklerinden arındırılmış hallerini göndermek yeterli olacaktır.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



## Aşamalı Sorgulama

- Soru: *What is the occupation of Bill Clinton's wife?*
  - Cevabı içeren doküman yoksa
- Aşamalı sorgu: soruyu ikiye böl
  - Önce *Who is Bill Clinton's wife?* = X'i bul.
  - Sonra *What is the occupation of X?* 'i ara.

YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



- Kabul edilebilirlik

- *İstanbul Ankara arası 12 metredir.*
- *Haliç Köprüsünün yüksekliği 23 cm'dir.*

- Wordnet:

- Cevap türü şehir olan bir sorunun olası cevaplarını Wordnet'ten kontrol edebiliriz. WordNet: Helsinki, Tampere, Luanda... şehir isimleri olarak yer almaktalar.



## Kaynaklar

- Bir Soru Cevaplama Sistemi: Baybılmış, M.Fatih Amasyalı, Banu Diri, Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 2005
- Open-Domain Question Answering, Eric Nyberg
- Questioning and Answering, Alison Huettnner
- Processing of large document collections, Helena Ahonen-Myka
- An Introduction to Information Retrieval and Question Answering , Jimmy Lin



Son ☺



YILDIZ TEKNİK ÜNİVERSİTESİ  
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

