



2020-2021 Güz Yarıyılı Doğal Dil İşlemeye Giriş

Konu : RegEx kullanarak verilen dosyadan adres bulma

Öğrenci Numarası : 16011049

Öğrenci adı : TARIK

Öğrenci soyadı : ÇARLI

RegExp Komut Satırı : `((.*MAH)\W*(.*CAD).*NO\D*(\d*)\W*(\w*)\W*(\w*))`

RegExp komutuna girdi olarak verilen dizgide toplam 6803 tane adres vardır. Yukarıda verilen regexp komutu bu adreslerin 3740 tanesini başarı ile bulup mahalle, cadde, numara, ilçe ve il olarak gruplandırmıştır. 3163 tane adres formata uygun olmadığı için bulunamamıştır. Bulamadığı örneklere baktığımızda ise cadde, mahalle veya numara bilgisi eksiktir.

REGULAR EXPRESSION

3740 matches, 1063369 steps (~1.46s)

/ ((.*MAH)\W*(.*CAD).*NO\D*(\d*)\W*(\w*)\W*(\w*)) /gmi

TEST STRING

YENIBOSNA METRO ISTASYONU BAKIRKÖY/ İSTANBUL
KENNEDY CAD. SIRKECI ARABALI VAPUR İSKELESİ FATİH/ İSTANBUL
YAVUZTURK MAH. KARADENİZ CAD. NO:2 USKUDAR/ İSTANBUL
HAMİDİYE MAH. ALPEREN SOK. NO:15/2 ÇEKMEKÖY/ İSTANBUL
UGUR MUMCU MAH. YUNUS EMRE CAD. NO:25 KARTAL/ İSTANBUL
BAĞLARBAŞI MAH. İNÖNÜ CAD NO:3 MALTEPE/ İSTANBUL
HASANPAŞA MAH. FAHRETTİN KERİM GÖKAY CAD. KADIKÖY/ İSTANBUL
P.T.T. EVLERİ BAĞCEKÖY CAD. NO: 53 SARIYER/ İSTANBUL
KARAKÖY YER ALTI GEÇİDİ NO:24 BEYOĞLU/ İSTANBUL
ÖRNEK MAH. DOĞ. ARS. BLV FİKRİ SON CAD. GİRİŞİ AGENA E NO. 215 9/2 ESENYURT/ İSTANBUL
GURSEL MAH.28 NİSAN CAD.NO:4/B KAGITHANE/ İSTANBUL
ATATÜRK MAH. ALEMDAG CAD. NO:61 ÜMRANIYE/ İSTANBUL
YILDIZ POSTA CAD. TÜRK TELEKOM ÖNÜ GAZETE BAYII BESİKTAS/ İSTANBUL
ARMAGAN EVLER MAH. ALEMDAG CAD. SITE ÖTOBUS DURAGI YANI ÜMRANIYE/ İSTANBUL
FETİHTEPE MAH. FATİH SULTAN CAD. NO:37/B BEYOĞLU/ İSTANBUL
BOZKURT MAH. KURTULUS CAD. NO:135/A SİSLİ/ İSTANBUL
PASABAĞCI MAH. BARBAROS CAD. NO:4/A BEYKÖZ/ İSTANBUL
YEŞİL PINAR MAH. SÜKRAN SOK. NO:36/B EYÜPSULTAN/ İSTANBUL
MUSTAFA KEMAL PAŞA CAD. AZİMKAR SOK.14/1 FATİH/ İSTANBUL
YENİ DOĞAN MAH. KISLA CAD. NUR SANAYİ SİTESİ NO: 69 BAYRAMPASA/ İSTANBUL
DENİZ KOSKİLER MAH. ESKİ LONDRA ASFALTI NO: 22/3 AVCILAR/ İSTANBUL
ORTAÇESME SONDURAK BEYKÖZ/ İSTANBUL
MEHMET AKİF ERSOY MAH. NATO YOLU BOSNA BULVARI NO: 115 USKUDAR/ İSTANBUL
DERBENT MAH. DEREİCİ ÖTOBUS SON DURAK SARIYER/ İSTANBUL
LEVAZİM MAH. KÖRÜ SOK NO:7A BESİKTAS/ İSTANBUL
CENGELKÖY MAH. CENGELKÖY CAD. NO:35/A USKUDAR/ İSTANBUL
BUYUKDERE CAD. NİMET ABLA CAMİ YANI SİSLİ/ İSTANBUL
RESİTİPAŞA MAH. POSTAYOLU CAD. NO :85 SARIYER/ İSTANBUL
RUMELİ HİSAR USTU BEBEK YOLU SOK SARIYER/ İSTANBUL
HASANPAŞA MAH. KURBAGALIDERE CAD. NO:23/A KADIKÖY/ İSTANBUL
SEYRANTEPE MAH. İ. KARAOĞLANOĞLU CAD. NO: 153/B KAGITHANE/ İSTANBUL
TAHTAKALE MAH. GAFFAR OKAN CAD. NO:10 AVCILAR/ İSTANBUL
BABA HASAN ALEMİ MAH. ATATÜRK BULVARI FATİH/ İSTANBUL
TOPKAPI TİCARET MERKEZİ KARSISI ÜSTGEÇİT ALTI ZEYTİNBURNU/ İSTANBUL

EXPLANATION

/ ((.*MAH)\W*(.*CAD).*NO\D*(\d*)\W*(\w*)\W*(\w*)) /gmi

1st Capturing Group ((.*MAH)\W*(.*CAD).*NO\D*(\d*)\W*(\w*)\W*(\w*))

2nd Capturing Group (.*MAH)

Quantifier — Matches any character (except for line terminators) *

Quantifier — Matches between zero and unlimited times, as many times as possible, giving back as needed (greedy)

MAH matches the characters MAH literally (case insensitive)

W matches any non-word character (equal to [^a-zA-Z0-9_])

Quantifier — Matches between zero and unlimited times, as many times as possible, giving back as needed (greedy)

3rd Capturing Group (.*CAD)

Quantifier — Matches any character (except for line terminators) *

Quantifier — Matches between zero and unlimited times, as many times as possible, giving back as needed (greedy)

MATCH INFORMATION

Match 1

Full match	106-158	YAVUZTURK MAH. KARADENİZ CAD. NO:2 USKUDAR/ İSTANBUL
Group 1.	106-158	YAVUZTURK MAH. KARADENİZ CAD. NO:2 USKUDAR/ İSTANBUL
Group 2.	106-119	YAVUZTURK MAH
Group 3.	121-134	KARADENİZ CAD
Group 4.	139-140	2
Group 5.	141-148	USKUDAR

QUICK REFERENCE

Search reference

All Tokens

Common Tokens

General Tokens

Anchors

Meta Sequences

Quantifiers

Group Constructs

Zero or more of a

One or more of a

Exactly 3 of a

3 or more of a

Between 3 and 6 of a

Start of string

End of string

A word boundary

Blank space boundaries

a*

a+

a{3}

a{3,}

a{3,6}

^

\$

\b

\s