Yildiz Technical University

Natural Language Processing

Finding the address information in the given document with RegEx

Author:
Toygar Tanyel (18011094)

October 31, 2021

Generalized RegEx

Following RegEx expression is the most generalized version even though it seems long, works well. That is because we can check different patterns may occur. The Python code and the EDA will be covering in the next pages.

Pattern

```
’(([\-ŞÖÜÇİĞA-Z0-9]{1,12}[a-z]*?\s*[âşöığüçiŞÖÜÇĞİA-Za-z]*?\s
[âşöığüçiŞÖÜĞÇİA-Za-z]*\s*(M\.).*(ÜMRANİYE))|([\-ŞÖÜÇİĞA-Z0-9]
{1,12}[a-z]*?\s*[âşöığüçiŞÖÜÇĞİA-Za-z]*?\s[âşöığüçiŞÖÜĞÇİA-Za-z]
*\s*(M\.|Mah\.|MAH\.|Mahallesi|MAHALLESİ).*(/\s?İSTANBUL|/\s?İstanbul))|
([\-ŞÖÜÇİĞA-Z0-9]{1,12}[a-z]*?\s*[âşöığüçiŞÖÜÇĞİA-Za-z]*?\s[âşöığüçiŞÖÜĞÇİA-Za-z]
*\s*((Cad\.|CAD\.|C\.).*(/\s?İSTANBUL|/\s?İstanbul)))|(([\-ŞÖÜÇİĞA-Z0-9]{1,12}[a-z]
*?\s*[âşöığüçiŞÖÜÇĞİA-Za-z]*?\s[âşöığüçiŞÖÜĞÇİA-Za-z]*\s*
(YER␣ALTI␣GEÇİDİ|SONDURAK|SON␣DURAK|METRO|DURAĞI|İskelesi).*
(/\s?İSTANBUL|/\s?İstanbul))))’
```

Accuracy

Expected Addresses: 35 Catched patterns: 35

Accuracy Rate: *Catched patterns/Expected Addresses = 1 Loss = 0%*

1 Exploratory Analysis of given document

What we will check before starting
- Abbreviations
- Long versions of abbreviations which are used in text

- Specific neighborhoods
- Specific main roads
- Specific streets that we should grab from text

- Specific districts ex.(X/ISTABUL)

- Patterns

We will use these pieces of information to compute loss (missed places) / expected (all) equation.

1.1 Exploring particular data

Abbreviations:
- " M. , Mah. , MAH. , Cad. , CAD(.)? , NO: , No: , Sok. , SOK(.)? , Dr. , C. , DOĞ. , ARS. , BLV. , APT. , K: "

Long versions of abbreviations which are used in text:
- " Caddesi , Mahallesi , Sokak , BULVARI , DURAK , BLOK , DAİRE , KAT "

!! This section includes mostly proper names, *MEHMET AKİF ERSOY* (MAH.) , *Abdi İpekçi* (Cad.) etc.

Specific neighborhoods:
- " Hürriyet Mah. , Yıldırım Mah. , Top Ağacı M. , Cumhuriyet Mahallesi , Yeni Mah. , Göztepe Mahallesi , YAVUZTÜRK MAH. , HASANPAŞA MAH. , ÖRNEK MAH. , GÜRSEL MAH. , ATATÜRK MAH. , MEHMET AKİF ERSOY MAH. , DERBENT MAH. , LEVAZIM MAH. , 19 MAYIS MAH. , YUNUS EMRE MAH. , ÇAMLIK MAH. , AŞAĞI DUDULLU MAH. , DENİZ KÖŞKLER MAH. , KARTALTEPE MAH. , KİRAZLIDERE MAH. , GÜRPINAR MAH. , FEYZULLAH MAHALLESİ , HIRKA-I SERİF MAH MAH. "

Specific main roads:
- "Abdi İpekçi Cad. , Dereboyu Caddesi , Sadık Ahmet C. , Ömer Besimpaşa Cad. , Tütüncü Mehmet Efendi Caddesi , Bağdat Cad. , PTT Evleri Bahçeköy Cad. , KENNEDY CAD. , KARADENİZ CAD. , FAHRETTİN KERİM GÖKAY CAD. , FİKRİ SÖN CAD. , 28 NİSAN CAD. , ALEMDAĞ CAD. , YILDIZ POSTA CAD. , ŞEMSETTİN GÜNALTAY CAD. , VEYSEL KARANİ CAD. , MUHSİN YAZICIOGLU CAD. , ALEMDAĞ CAD. , 29 EKİM CAD. , MARAŞAL FEVZİ ÇAKMAK CAD. , MİLLET CAD , ŞEHİT HIKMET ALP CAD , LÜLECİYEKTA CAD."

Specific streets:
- " Kemal Paşa Sok. , Rıdvanpaşa Sokak , KORU SOK , LADİN SOK. , AHMET KOCABIYIK SOK. , DİLEK SOK. , OYA SOK , LÜLECİ YEKTA SOK. "

Specific districts:
- " KARTAL/İSTANBUL , BAYRAMPAŞA/İSTANBUL , KÜÇÜKÇEKMECE / İSTANBUL , Kadıköy/İstanbul , Beykoz / İstanbul , Sarıyer/İstanbul , BAKIRKÖY/ İSTANBUL , FATİH/ İSTANBUL , ÜSKÜDAR/ İSTANBUL , BEYOĞLU/ İSTANBUL , KADIKÖY/ İSTANBUL , ESENYURT/ İSTANBUL , KAĞITHANE/ İSTANBUL , ÜMRANİYE/ İSTANBUL , BEŞİKTAŞ/ İSTANBUL , BEYKOZ/ İSTANBUL , SANCAKTEPE/ İSTANBUL , ÇEKMEKÖY/ İSTANBUL , AVCILAR/ İSTANBUL , BAHÇELİEVLER/ İSTANBUL , BEYLİKDÜZÜ/ İSTANBUL , MALTEPE/ İSTANBUL

1.2 Quick Explanation of Analysis

As we can see above, the text has lots of proper name, irregular and disordered words. To catch, we should gather them under common patterns as crowded as possible.

The neighborhoods, main roads and streets shows that specific abbreviations are distinctive features to get information without using *proper words as "MEHMET AKİF ERSOY"* However, we should not forget there are many possibilities for coming words which we cannot know is it right words for places before abbreviations, 1/2/3 words, needed to be consider.

On the other side, the most characteristic word is İstanbul and its kinds to catch addresses but that is not enough alone. ÜMRANİYE sentence breaks that pattern.

2 The Code & Output

2.1 Python Code

'X' is representing the pattern above, at first page.

Listing 1: Python Pseudo

```
import re

pattern = 'X'
line_num = 0
with open('Adresler-rev.txt', encoding='utf8') as f:
    for line in f:
        line_num += 1
        catched = re.findall(pattern, line)
        cleaned = [item[0] for item in catched]

        print("Line␣num␣{}:␣Catched␣pattern:␣{}".format(line_num, cleaned))
```

2.2 Looking at the addresses we catched from .txt file

Line num 1: Catched pattern: []

Line num 2: Catched pattern: [’Hürriyet Mah. Abdi İpekçi Cad. No:38/C KARTAL/İSTANBUL’]

Line num 3: Catched pattern: [’Yıldırım Mah. Kemal Paşa Sok. No:8 BAYRAMPAŞA/İSTANBUL’, ’Top Ağacı M. Dereboyu Caddesi N0:47 ÜMRANİYE’]

Line num 4: Catched pattern: [’Cumhuriyet Mahallesi Dr. Sadık Ahmet C. No:51/B KÜÇÜKÇEKMECE / İSTANBUL’]

Line num 5: Catched pattern: []

Line num 6: Catched pattern: [’Deniz Otobüsü İskelesi önü Rasimpaşa Kadıköy/İstanbul’]

Line num 7: Catched pattern: [’Yeni Mah. Dr. Ömer Besimpaşa Cad. No:64 Beykoz / İstanbul’]

Line num 8: Catched pattern: [’Göztepe Mahallesi Tütüncü Mehmet Efendi Caddesi Rıdvanpaşa Sokak Kadıköy/İstanbul’]

Line num 9: Catched pattern: [’Bağdat Cad. No:445 Kadıköy/İstanbul’]

Line num 10: Catched pattern: [’PTT Evleri Bahçeköy Cad. No:53 Sarıyer/İstanbul’]

Line num 11: Catched pattern: []

Line num 12: Catched pattern: []

Line num 13: Catched pattern: []

Line num 14: Catched pattern: [’YENİBOSNA METRO İSTASYONU BAKIRKÖY/ İSTANBUL’]

Line num 15: Catched pattern: [’KENNEDY CAD. SİRKECİ ARABALI VAPUR İSKELESİ FATİH/ İSTANBUL’]

Line num 16: Catched pattern: [’YAVUZTÜRK MAH. KARADENİZ CAD. NO:2 ÜSKÜDAR/ İSTANBUL’]

Line num 17: Catched pattern: [’HASANPAŞA MAH. FAHRETTİN KERİM GÖKAY CAD. KADIKÖY/ İSTANBUL’]

Line num 18: Catched pattern: []

Line num 19: Catched pattern: [’ÖRNEK MAH. DOĞ. ARS. BLV. FİKRİ SÖN CAD. A GİRİŞİ NO.215 9/2 ESENYURT/ İSTANBUL’]

Line num 20: Catched pattern: [’GÜRSEL MAH. 28 NİSAN CAD. NO:4/B KAĞITHANE/ İSTANBUL’]

Line num 21: Catched pattern: [’ATATÜRK MAH. ALEMDAĞ CAD. NO:61 ÜMRANİYE/ İSTANBUL’]

Line num 22: Catched pattern: [’YILDIZ POSTA CAD. TÜRK TELEKOM ÖNÜ GAZETE BAYİİ BEŞİKTAŞ/ İSTANBUL’]

Line num 23: Catched pattern: [’ORTAÇEŞME SONDURAK BEYKOZ/ İSTANBUL’]

Line num 24: Catched pattern: [’MEHMET AKİF ERSOY MAH. NATO YOLU BOSNA BULVARI NO: 115 ÜSKÜDAR/ İSTANBUL’]

Line num 25: Catched pattern: [’DERBENT MAH. DEREİCİ OTOBÜS SON DURAK SARIYER/ İSTANBUL’]

Line num 26: Catched pattern: [’LEVAZIM MAH. KORU SOK NO:7A BEŞİKTAŞ/ İSTANBUL’]

Line num 27: Catched pattern: [’19 MAYIS MAH. ŞEMSETTİN GÜNALTAY CAD. NO:144 KADIKÖY/ İSTANBUL’]

Line num 28: Catched pattern: [’YUNUS EMRE MAH. VEYSEL KARANİ CAD. NO:105 SANCAKTEPE/ İSTANBUL’]

Line num 29: Catched pattern: [’ÇAMLIK MAH. MUHSİN YAZICIOGLU CAD. NO:6/B ÇEKMEKÖY/ İSTANBUL’]

```
Line num 30: Catched pattern: ['AŞAĞI DUDULLU MAH. ALEMDAĞ CAD. NO:465/A ÜMRANİYE/ İSTANBUL']
Line num 31: Catched pattern: ['DENİZ KÖŞKLER MAH. ESKİ EDİRNE ASFALTI NUR APT. NO:2 AVCILAR/ İSTANBUL']
Line num 32: Catched pattern: ['KUYUMCU KENT ATÖLYE DURAĞI 29 EKİM CAD. LADİN SOK. K:1 NO:6 BAHÇELİEVLER/ İSTANBUL']
Line num 33: Catched pattern: ['FEVZİ ÇAKMAK MAH. AHMET KOCABIYIK SOK. ANADOLU APT. NO: 3-A KÜÇÜKÇEKMECE/ İSTANBUL']
Line num 34: Catched pattern: ['KARTALTEPE MAH. DİLEK SOK. NO:20 SEFAKÖY METROBÜS ALTGEÇİT KÜÇÜKÇEKMECE/ İSTANBUL']
Line num 35: Catched pattern: ['KİRAZLIDERE MAH. MARAŞAL FEVZİ ÇAKMAK CAD. OYA SOK NO:2/1 ÇEKMEKÖY/ İSTANBUL']
Line num 36: Catched pattern: ['KARTAL KÖPRÜSÜ ALTI METRO ÇIKIŞI ANKARA İSTİKAMETİ E5 YAN YOL KARTAL/ İSTANBUL']
Line num 37: Catched pattern: ['GÜRPINAR MAH. MİLLET CAD OKUTAN İŞ MERKEZİ NO:1 BEYLİKDÜZÜ/ İSTANBUL']
Line num 38: Catched pattern: ['FEYZULLAH MAHALLESİ ŞEHİT HIKMET ALP CAD ADATEPE SİTESİ B1 BLOK DÜKKAN 2 MALTEPE/ İSTANBUL']
Line num 39: Catched pattern: ['HIRKA-I SERİF MAH MAH. LÜLECİYEKTA CAD. LÜLECİ YEKTA SOK. NO : 1 DAİRE : 1 KAT : 1 FATİH/ İSTANBUL']
```

3 References

https://www.w3schools.com/python/python_regex.asp
https://regex101.com
https://docs.python.org/3/howto/regex.html
https://docs.python.org/3/library/re.html