Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Arvutiteaduste instituut

Anne Ott

Medistiinitekstidest info eraldamine EstNLTK abil

Projekt aines eesti keele töötlus Pythonis

Juhendaja: Sven Laur

Tartu 2021

Sisukord

Sissejuhatus

Käesoleva töö eesmärk on eraldada meditsiinitekstidest patsientide analüüside andmeid. Patsientide epikriiside andmetes on olemas tekstiline väli, mis sisaldab pool struktureeritud andmeid patsiendi kohta. Selline tekstiväli ehk veerg on olemas erinevates tabelites (näiteks anamneesi, protseduuride, allergiate jne tabelites). Sagedasti on sattunud sinna tekstilisse välja patsiendi analüüsi tulemused (näiteks vere-, uriinianalüüsid), mis sinna tegelikult ei kuulu. Vaja on need analüüside tulemusi kirjeldavad osad tekstidest eraldada.

Analüüsi tekstid on tavaliselt tekstis enam-vähem tabeli formaadis. Üldistatult võib öelda, et tabeleid on seitsmes erinevas formaadis, edaspidi on toodud ka mitmeid näited nende kohta. Igat analüüsi teksti tabelit saab enamjaolt tuvastada esimese ja viimase rea järgi, sest need omavad sarnaseid struktuure. Töö eesmärk on võtta ette kõik tekstivälja tekstid ja leida üles nendes olevad analüüsid ning eraldada need eraldi tabelisse.

Esimeses peatükis antakse ülevaade kasutatud andmetest. Teises peatükis keskendutakse kuupäeva formaadile. Kolmas kuni kümnes peatükk tutvustavad erinevat tüüpi tekstides leiduvaid analüüside tabeleid, nende eraldamiseks kasutatavaid regulaaravaldisi ja näiteid. Üheteistkümnendas peatükis käsitletakse alguse ja lõpu regulaaravaldiste ühendamist, kaheteistkümnendas tekstide prioriteete, kolmeteistkümnendas saadud tulemusi ja neljateistkümnendas peatükkis on kokkuvõte.

1 Andmed

Antud projektis on kasutatud Eesti Geenivaramu andmeid, mida on kokku umbes 37 miljonit rida. Kasutatud on 10 erinevat tabelit ja 16 erinevat veergu kõigi nende tabelite peale kokku. Tabelite suurused varieeruvad 1000 reast kuni 4 miljoni reani. Tekste otsime järgnevatest tabelitest: *Surgery, Surgery Entry, Summary, Procedures, Procedures entry, Objective finding, Death, Anamnesis, Allergy, Allergy entry.*

Erinevat tüüpi analüüsi tabelite leidmiseks tuli käsitsi läbi vaadata sadu erinevaid tekstivälja väärtusi, et leida mustreid, mille järgi regulaaravaldisi kirjutada. Seejärel tuli leitud mustrid kategoriseerida ja igale kategooriale luua oma *RegexTagger*. Kokkuvõttes ilmnes 7 erinevat tabeli tüüpi, mõni väga kindla struktuuriga, teine aga väga varieeruva struktuuriga.

Kuna antud töös on tegu meditsiini andmetega, siis ei tohi neid andmeid avalikustada. Seega kõigis näidetes on andmete sisu (kuupäevad, mõõtmiste tulemused jne) muudetud ja säilitatud on vaid teksti struktuur. Samuti on raportiga kaasa pandud koodis vaid mõned näited.

2 Kuupäevad

Enne tekstide juurde asumist tuleb tähelepanu pöörata kuupäevadele. Kuupäevad on ühed põhilistest komponentidest, mis analüüsitekstides läbivalt esinevad.

Seega on selle jaoks eraldi regulaaravaldis, mis on välja toodud järgmistel ridadel.

```
substitutions = {
    "YEAR": r"(?P<YEAR>((19[0-9]{2})|(20[0-9]{2})|([0-9]{2})))",
    "MONTH": r"(?P<MONTH>(0[1-9]|1[0-2]))",
    "DAY": r"(?P<DAY>(0[1-9]|[12][0-9]|3[01]))",
}
date_regex = r"{DAY}\.\s*{MONTH}(\.\s*{YEAR}\s*)?".format(**
    substitutions)
```

Kuupäevad, mida regulaaravaldised tuvastavad on näiteks ”21.02.2020”, ”21. 02. 2020” ja ”21.02”.

Koodi real 2 on kirjeldatud lubatud aasta formaat - kui aasta on neljakohaline, siis tohib ta alata kas 19 või 20-ga ja teised kaks numbrid võivad olla 0-9ni (1900, 1901, ..., 1999, 2000, 2001,..., 2099). Kui aasta on kahekohaline, siis võivad mõlemad numbrid olla 0-9ni (00, 01, ..., 99).

Real 3 on kirjeldatud lubatud kuu formaat - kuu võib alata kas 0 või 1-ga. Kui kuu algab 0-ga, siis sellele võivad järgneda 1-9 (01,02,...,09 ehk jaanuar-september). Kui kuu algab 1-ga, siis sellele võivad järgneda 0-2 (10,11,12 ehk oktoober-detsember).

Real 4 on kirjeldatud lubatud päeva formaati - päev võib alata numbritega 0,1,2,3. Kui päev algab 0-ga, siis sellele võib järgneda ükskõik milline number (leiab üles 01, 02, ..., 09). Kui päev algab 1 või 2-ga siis samamoodi võib järgneda ükskõik milline number (11, 12, ..., 19, 21, 22, ..., 29). Kui päev algab 3-ga, siis sellele võib järgneda kas 0 või 1 (30, 31).

Real 6 panen kokku aasta, kuu ja päeva, mis on eraldatud üksteisest punktiga ja valikuliste (*optional*) tühikutega. Samuti on aasta valikuline.

3 Tagger teksti tüübile 1

3.1 Kirjeldus

Esimest tüüpi tekstid on kõige lihtsama ja stabiilseima struktuuriga. Stabiilse all pean silmas, et ei esine väga palju erinevaid variatsioone. Sellist tüüpi tekste sai 37 miljonist reast eraldatud 33 181 (7.2% kõikidest eraldatud tekstidest). Kusjuures ühest tekstist võib eraldada mitu tabelit. *RegexTaggeris* kasutatavad alguse ja lõpu tuvastamise regulaaravaldised näevad välja järgnevad

```
start_regex = date_regex + ".*?\s*(Kliiniline veri|Uriin)\s*"
end_regex = "(RBC.*;|Erikaal.*;)"
```

Tüübi 1 analüüside tekstid on alati kaherealised, esimesel real on kuupäev ja analüüsi nimi, teisel real parameetrite nimed ja väärtused (näide 1). Analüüsi nimeks on alati kas ”Kliiniline veri” või ”Uriin”. Seega otsib alguse regulaaravaldis *start_regex* rida, mis

1. algab kuupäevaga
2. ".*?" tähistab ükskõik millist teksti, mis tüüp 1 puhul on tavaliselt numbrite kombinatsioon või sõne "line veri" (näited 2 ja 3)
3. lõppeb analüüsi nimega, kas "Kliiniline veri" või "Uriin"

Lõpu regulaaravaldis (*end_regex*) otsib lõpurida, mis algab sõndedega "RBC" või "Erikaal". ".*?" võtab kogu teksti kuni rea lõpuni, kus tuvastati "RBC" või "Erikaal".

```
09.08.2011  Kliiniline veri
RBC 3.9 ; HCT 38.0; MCV 98 ; HGB 128; MCH  49; MCHC 290; PLT  192; MPV
     9.4 ; WBC 5.8 ; SR   0; Gly     ; CRV     ; ANONYM; keppt.  0;
    segmt.  0; eos.    0; basof.  0; lümf.   0;
```

Näide 1: Analüüsi tekst tüüp 1

```
26.02.2016 61209 URIIN

Erikaal: 2.080; Atset.: -; ANONYM; ANONYM; Urobil.: -; Bilir: -; ANONYM
    ; Reakts.: 4.9     ; Leukots.: -       ; bakterid: -   ;
```

Näide 2: Analüüsi tekst tüüp 1

```
15.10.2009 line veri Kliiniline veri

RBC 6.00; HCT 65  ; MCV 86 ; HGB 127; MCH  50; MCHC 112; PLT  410; MPV
         ; WBC 7.1 ; SR   0; Gly     ; CRV     ; <ANONYM id="1" type="
    per" morph="\_S\_ sg n"/>:; keppt.  0; segmt.  0; eos.    0; basof.
      0; lümf.   0;
```

Näide 3: Analüüsi tekst tüüp 1

3.2 Ootamatud tulemused

Lisaks standardsetele tekstidele, leidis *tagger* ka muid analüüsi teksti tabeleid. Sellised tekstid on pigem erandlikud ja seega regulaaravaldist kirjutades polnud ma nendest teadlik. Siiski on see positiivne üllatus, sest kogu töö eesmärk on eraldada võimalikult palju tekste. Mõned näited:

```
28.08.2017: Telefonikontakt: ema soovib teada analüüsi vastuseid:
    Kliiniline veri:  WBC 5.1 x10*9/L;  RBC 7.02 x10 12/L;  HGB 127 g/L
    ;  HCT% 49.0 %;  MCV 65.0 fL;  MCH 51.8 pg;  MCHC 299 g/L;  PLT 327
     x 10 9/L; valem: <ANONYM id="2" type="per" morph="_H_ sg n"/>%
    58.6 %;  MO% 1.4 %;  GR% 87.0 %;  TSH 71.2 mU/l;
```

Näide 4: Analüüsi tekst tüüp 1

```
22.12.2013: Patsient tegi vereanalüüsid: *Kliiniline veri --> WBC 6.7
    x10*9/L;  RBC 6.91 x10 12/L;  HGB 131 g/L;  HCT% 35.2 %;  MCV 83.2
    fL;  MCH 24.5 pg; MCHC 401 g/L; PLT 271 x 10 9/L; valem: <ANONYM id
    ="2" type="per" morph="_H_ sg n"/> 1.9    109/l; MO 1.7 x 10 9/L;
    GR 6.4 10x 9/L; sete 31 mm/h. Biokeemia:  CRP <7 mg/L;
```

Näide 5: Analüüsi tekst tüüp 1

3.3 Vale negatiivsed

Siin on toodud välja tekstid, mis jäävad eraldamata, kuigi on väga sarnased tüüp 1 struktuurile ja võiksid saada eraldatud. Nimelt on probleem selles, et andmed pole kahel real vaid rohkematel (näited 6,7,8). Hinnanguliselt jääb selle probleemi tõttu eraldamata väga väike osa tekste ehk umbes 50 teksti.

Näited:

```
15.05.2019    66
215 - Kliiniline veri
...
```

Näide 6: Analüüsi tekst tüüp 1

```
29.06.2016    - Kliiniline veri
R
BC 5.63; HCT 42.8; MCV 890; HGB 171; MCH 41; MCHC
...
```

Näide 7: Analüüsi tekst tüüp 1

```
2014 URIIN URIIN
URIIN

27.08.2013 10:25   27.08.2013 11:05   Uriini
ribaanalüüs  U-Strip  16309
           Erikaa
...
```

Näide 8: Analüüsi tekst tüüp 1

4 Tagger teksti tüübile 2

4.1 Kirjeldus

Tekstid tüübiga 2 on esinemissageduse poolest teisel kohal, eraldatud sai 129 000 teksti, mis on 28% kõikidest eraldatud tekstidest. *Taggeri* regulaaravaldised näevad välja järgnevad

```
substitutions = {
    "AN": r"(ANALÜÜSI(DE)?\s*TELLIMUS\s*nr:\s*\d*\s*)",
    "MAT": r"(MATERJAL:\s*.*?\s*)",
    "VAS": r"(VASTUSED:.*?)",
    "MARK": r"(Märkus:(.*?\s*){0,2}\s*)",
}

start_regex = "({AN}{MAT}{MARK}?{VAS}|{AN}{MARK}?{VAS}|{AN}{MAT}|{MAT}{
    VAS}|{VAS}|{AN})".format(**substitutions)
end_regex = "(" + date_regex + ")|"  + "((\)|\w+)\s*$)"
```

Nagu alguse regulaaravaldisest paistab, siis algavad analüüside tabelid erinevate kombinatsioonidega sõnadest "Analüüside Tellimus", "Materjal", "Vastused" ja "Märkus". Seejuures pole ükski komponentidest kohustuslik. Lõpu regulaaravaldise puhul on kaks võimalust:

- kui järgneval real on kuupäev, tähendab see seda, et eelnev analüüsi tabel on läbi ja uus algab
- kui tegu on teksti lõpuga ja see lõppeb kas märgiga ")" või tähega.

Näide 9 on tüüpiline tüüp 2 alla kuuluva analüüsi teksti näide. Seejuures on tekstid tavaliselt pikad ja '...' tähistab sarnase struktuuriga ridade jätkumist, mille välja toomine siinkohal pole oluline.

```
ANALÜÜSIDE TELLIMUS nr: 400942

MATERJAL:
IP0011032640 21.12.2015 16:35 (võetud: 21.12.2015 00:00)
IP0011032634 21.12.2015 16:35 (võetud: 21.12.2015 00:00)
...

VASTUSED:
Glükohemoglobiin 5.7 (<11.0 %)
Erütrotsütaarsete antikehade sõeluuring kahe erütrotsüüdiga Antikehi ei
     leidu (Antikehi ei leidu )


Triglütseriidid paastuseerumis 5.2 (<6.0 mmol/L)
...
Hemogramm viieosalise leukogrammiga
WBC 9.43 (3,5 .. 10,8 E9/L)
NEUT# 7.52 (2,9 .. 9,3 E9/L)
...
AB0 veregrupp
AB0 veregrupp 0
AB0-veregrupi ja Rh(D) kinnitav määramine
...
```

Näide 9: Analüüsi tekst tüüp 2

4.2 Vale positiivsed

Valepositiivsed tekste iseloomustab see, et need vastavad regulaaravaldise tingimustele, aga ei sisalda analüüside kohta informatsiooni (näide 10, regulaaravaldis leiab ülesse punaseks märgitud tekstid). Selliseid juhte on raske välja jätta, sest regulaaravaldise ümber kirjutamine tõstaks valenegatiivsete arvu (välja jääb oluline informatsioon). Hinnanguliselt on valepositiivseid 500 ringis.

```
HAIGLAS  TEOSTATUD ANONYM JA  LABORI  ANALÜÜSIDE  VASTUSED:
VASTUS:
  CT-HRCT  ANONYM NATIIVIS RINDKEREST JA MEDIASTIINUMIST
Saatedgn: TBC. Dünaamikas  CT 2011.a.
Pleuraõõntes ega perikardi õõnes vedelikku ei esine. ...
ANONYM (D01795)

09.12.2014 21:00
```

Näide 10: Analüüsi tekst tüüp 2

4.3 Vale negatiivsed

Eraldamata jäävad tekstid, mis on kõik ühel real (näide 11). Hinnanguliselt on selliseid tekste umbes 800 ja tulevikus võiks regulaaravaldise ümber kirjutada, et selised analüüsid kaduma ei läheks.

```
<arsti/patsiendi nimi>ANALÜÜSIDE TELLIMUS nr: 1440456MATERJAL:
    V100199510045 13.04.2013 08:51 (võetud: 13.04.2013  00:00)VASTUSED:
    Hemogramm WBC 8.23 (4,1 .. 9.2 E9/L ) RBC 3.61 (4,1 .. 6,9 E12/L )
    HGB 142 (121 .. 169 g/L ) HCT 44 (32 .. 47 % ) MCV 79.0 (80 .. 93
    fL ) MCH 28.0 (27 .. 39 pg ) MCHC 343 (321 .. 368 g/L ) PLT 184
    (170 .. 385 E9/L ) RDW-CV 13.9 (10,3 .. 15 % ) 13.04.2013 09:08
```

Näide 11: Analüüsi tekst tüüp 2

5 Tagger teksti tüübile 3

5.1 Kirjeldus

Tüüpi 3 tekste sai eraldatud tabelitest 4616, mis on vaid 1% kõikidest eraldatud tekstidest. Selle tekstitüübi iseloomulikumaiks omaduseks on püstkriipsud. *Taggeri* regulaaravaldised näevad välja järgnevad

```

start_regex = r"(\|(\s*[Uu]uring\s*|<ANONYM.*>)\|(" + date_regex + "\|)
    +Mõõt(ühik)?\|)"  # 1. pealkirjaga rida
    + "|(^\s*\|.*)|"  # 2. pealkirjata rida
    + "\n\s*\n\|.*\|",  # 3. pealkirjata rida

end_regex = "((?!.*\n\|).+)|(\|\s*$)"
```

Tabelid algavad kahte erinevat moodi:

- pealkirja reaga (näide 12 ja 13)
- ilma pealkirjata reaga (näide 14).

Alguse regulaaravaldis üritab kõigepealt leida pealkirjaga rida. Täpsemalt rida tekstiga "|Uuring|*Kuupäev*|Mõõtühik|", kusjuures kuupäevi võib olla rohkem kui üks. Samuti võib olla uuringu nimi anonümiseeritud ehk pealkiri on "|ANONYM ... |Kuupäev|Mõõtühik|".

```
|Uuring|01.07|12.07|Mõõtühik|
|Adrenokortikotroopne hormoon p...|19.3|  |pg/ml|
|Insuliinisarnane kasvufaktor 1...|  |140.1|ng/mL|
|Kasvuhormoon seerumis|  |0.639|mU/L|
|Kilpnääret stimuleeriv hormoon...|0.79|  |mIU/L|
```

Näide 12: Analüüsi tekst tüüp 3

```
|<ANONYM id=3 morph=_S_ sg n/>|08.11.2013|Mõõtühik|
|Glükohemoglobiin|3.3|%|
|Glükohemoglobiin veres|39|mmol/mol|
|Glükoos seerumis|6.4|mmol/L|
```

Näide 13: Analüüsi tekst tüüp 3

Kui tegu on ilma pealkirjata tabeli algusega, siis

1. otsitakse kas teksti esimene rida algab püstkriipsuga või
2. rida sisaldab kahte püstkriipsu ja neile eelnevad tühjad read.

Näiteks

```
...

|Aeroobide uuring||Normaalne mikrofloor...||
|C-reaktiivne valk seerumis/pla...|9.4|1.8|mg/L|
```

Näide 14: Analüüsi tekst tüüp 3

Tabeli lõpu tuvastamisel on samuti kaks võimalust:

- Esimesel juhul kontrollitakse, kas järgnev rida sisaldab püstkriipsu, kui ei sisalda, siis on tabel läbi
- Teisel juhul kontrollitakse, kas tegu on faili lõpu reaga, mis sisaldab püstkriipsu. See on vajalik, kuna pole järgmist rida, mida kontrollida (nagu teeb esimene juht).

5.2 Vale positiivsed

Hinnanguliselt on 4620 eraldatud tekstist 83 ehk 1.8% valepositiivsed. Valepositiivsed koosnevad tavaliselt täpselt ühest püstkriipsust ja sellele järgnevast infost, mis tavaliselt pole aga soovitud analüüsi tekst (näide 15).

```
|
|                        Ananeesis varasem müokardi infarkt  2017 koos
    PTKA + 2BMS  C(6)-->(1) ja LAD (6)-->( 02.09.2012.a.)2013.a.
    Angiograafia leid: Ühe soone koronaartõbi (mitte vasak  peatüvi). C
    (Cx) mid 80% Discrete lesioon.
```

Näide 15: Analüüsi tekst tüüp 3

Samas regulaaravaldises ei taha nõuda, et esimesel real oleks vähemalt kaks püstkriipsu, sest mõningatel huvipakkuvatel analüüsi tabelitel on samuti vaid üks püstkriips. Seega kui on vaja valida, kas tekitada rohkem valepositiivseid või valenegatiivseid, siis eelistada valepostiivseid (võimalikult palju infot eraldada).

5.3 Vale negatiivsed

Teadaolevalt valenegatiivseid on umbes 300-400 ringis. Eraldamata jäävad tekstid, kus tabelil pole pealkirja veergu ja eelneb tekst. Sellised kohad tuleks tulevikus ära parandada. Näiteks

```
UURINGUD JA PROTSEDUURID
|C-reaktiivne valk|3.1|mg/L|
```

Näide 16: Analüüsi tekst tüüp 3

6 Tagger teksti tüübile 4

6.1 Kirjeldus

Tüüpi 4 tekstid on üpris struktureerimatud, sõnad on tihtipeale keskelt suvalise koha pealt tühiku(te)ga eraldatud. Tuleb mainida, et tüüpi 4 tekste leidub andmetabelitest kõige vähem (275 tükki). Seega pole mõistlik panustada nende tuvastamise parandamiseks liigselt aega.

Taggeri regulaaravaldised näevad välja järgnevad

```

start_regex = "((Proovi\s*v\s*õtmise\s*kuupäev|Tellimise\s*kuupäev)\s*
    Teostamise kuupä\s*ev\s*U\s*u\s*r\s*ing\s*Lühend\s*Tulemus\s*Ühik\s
    *Referents\s*HK\s*ko\s*o\s*d\s*)|"
    + "(Näita\s*Refe\s*rentsväärt\s*usi\s*ja\s*ühikuid\s*Analüüsid\s*)|
    "
    + "(Analüüsid\s*Kliiniline vere ja glükoh\s*emoglobiini analüüs\s*
    Rerentsväärtus\s*"
    + date_regex
    + ")|"
    + "(Analüüs\s*Parameeter\s*Referentsväärtus\s*ja\s*üh\s*ik)|"
```

```
    + "(Analüüsid\s*Nimetus\s*Referentsväärtus\s*Tulemused\s*Ühi\s*k)"

end_regex = "\n(.*)\n(.*)\n(?="
    + "\s*(Väljastatu\s*d\s*dok\s*umendid|"
    + "[Tt]öövõimetusleht|"
    + "[Rr]avi\s*a\s*rst\s*/ko\s*ostaja|"
    + "[Tt]eosta\s*ja|"
    + "Kokku\s*võte patsiendi ra\s*vist)" + ")"
```

Tekstid algavad järgnevate pealkirjadega

- Proovi võtmise kuupäev Teostamise kuupäev Uuring Lühend Tulemus Ühik Referents HK kood
- Tellimise kuupäev Teostamise kuupäev Uuring Lühend Tulemus Ühik Referents HK kood
- Näita Referentsväärtusi ja ühikuid Analüüsid
- Analüüsid Kliiniline vere ja glükohemoglobiini analüüs Rerentsväärtus *kuupäev*
- Analüüs Parameeter Referentväätus ja ühik
- Analüüsid Nimetus Referentsväärtus Tulemused Ühik

Tabeli lõpu tuvastamiseks kontrollitakse, kas järgnev rida algab fraasiga

- Väljastatud dokumendid
- Töövõimetusleht
- Raviarst/koostaja
- Teostja
- Kokkuvõte patsiendi ravist.

Näited 17, 18 demonstreerivad kahte eraldatud teksti. Näite 17 puhul eraldatakse table ridadelt 1-17, näite 18 puhul ridadelt 2-8. See tähendab, et sõnad "Väljastatud dokumendid" ja "Teostaja" märgivad teksti lõppu aga need jäetakse välja eraldatud tekstist.

```
Analüüsid



Kliiniline vere ja glükoh
emoglobiini analüüs

Rerentsväärtus

04.04.2019 09:35:45
```

```
11
12
13 Materjal
14 B-CBC-5Diff  tehtud
15 WBC 4 -
16 10*10*9/l 9.67
17 ...
18 Väljastatu
19 d dokumendid
20 ...
```

Näide 17: Analüüsi tekst tüüp 4

```
1
2 Proovi võtmise kuupäev Teostamise kuupäev U
3 uring Lühend Tulemus Ühik Referents HK kood
4
5 07.10.2018 19:30 07.10.2018 19:38 D-dimeerid
6 P-D-Di 0,24 mg/L <=0,38 66306
7
8 ...
9
10 Teosta
11 ja: 13.09.2019 10:38
12 ...
```

Näide 18: Analüüsi tekst tüüp 4

6.2 Vale positiivsed

Eraldatud 275 analüüsitabeli tekstist polnud mitte ükski valepositiivne. Küll aga ei sisalda 11 tabeli teksti kasulikku informatsiooni. Need kõik näevad välja nagu näites 19.

```
1 Näita Referentsväärtusi ja ühikuid
2         Analüüsid
3         ?
4         ?
5         Analüüs
6         Parameeter
```

Näide 19: Analüüsi tekst tüüp 4

6.3 Vale negatiivsed

Valenegatiivseid leidub kindlasti veel tekstiväljades, sest algust/lõppu märkivad sõnad võivad olla suvalise koha pealt tühikuga pooleks löödud. Seega on väga tõenäoline, et kõik võimalused pole kajastatud kirjutatud regulaaravaldises ja absoluutselt kõik tüüp 4-le vastavad tabelid ei saa eraldatud.

Tulevikus võiks lõpu regulaaravaldist täiendada nii, et see kontrolliks kas lõpu reale järgnev rida ei sisalda numbrit. Kui järgnev rida ei sisalda numbrid on analüüside tabel läbi, sest iga tabeli rida peab sisaldama mõõtmiste tulemusi.

7 Tagger teksti tüübile 5

7.1 Kirjeldus

Tüüpi 5 analüüsi tekstide tabelid on saransed tüübile 4, kuid sisaldavad vähemalt ühte kuupäeva pealkirjas. Selliseid tabeleid on samuti andmetes väga vähe - eraldatud sai 1645 analüüsi tabelit (0,4% kõikidest eraldatud tabelitest).

Taggeri regulaaravaldised näevad välja järgnevad

```
1
2  start_regex = r"((Analüüsid\s*)?Analü\s*üs\s*(\\s*Tellitud\s*)?Ühik\s*
       Refer\s*ents\s*("  # 1.
3      + date_regex  # 1.
4      + "\s*)+)|"  # 1.
5      + "((Biokeemia analüüs|"  #2.
6      +  "Kliiniline vere\s*(ja glükohemo\s*globiini\s*)?analüüs|"  #2.
7      +  "Hemostasiogramm|"  #2.
8      +  "Immuunmeetoditel\s*uuringud|"  #2.
9      +  "Uriini analüüs)"  #2.
10     +  "\s*Rer\s*entsväärtus\s*("  #2.
11     + date_regex  #2.
12     + ".*?\s*)+)",  #2.
13
14 end_regex = "\n\s?\n\s?\n|"
15         # mainly for start 1.
16         + "\n(.*)(?="
17         + "Raviarst|"  #1.
18         + "Re[ _ ]iim ja ravialased soovitused|"  #1.
19         + "[Rr]avimid|"
20         + "Operatisoon|"
21         + "Kokkuvõte patsiendi ravist|"  #1.
22         + "Retsept|"  #1.
23         # mainly for start 2.
24         + "Biokeemia analüüs|"  #2.
25         + "Kliiniline vere ja glükohemo\s*globiini analüüs|"  #2.
26         + "Hemostasiogramm|"  #2.
27         + "Uriini analüüs|"  #2.
28         + "Immuunmeetoditel\s*uuringud)",  #2.
```

Peamiselt kuulub siia alla kahe erineva struktuuriga tabeleid. Kuna kumbagi tüüpi pole liialt palju, siis olen need kokku pannud ühe tüübi alla.

Esimene tüüp algab alati pealkirjaga "Analüüs Tellitud Ühik Referents" ning sellele järgneb vähemalt üks kuupäev. Tabel lõppeb, kui tekstis on kolm tühja rida järjest või järgmisel real on üks järgnevatest sõnadest:

- "Raviarst"
- "Režiim ja ravialased soovitused"
- "Ravimid"
- "Operatsioonid"

- "Kokkuvõte patsiendi ravist"
- "Retsept".

Näide 20 demonstreerib eelenvalt kirjeldatud teksti. Seejuures regulaaravaldis tagastab read 1-5. Rida 6. küll märgitakse ära tabeli lõpuna, aga kuna tegu on ebavajaliku infgoga, siis seda ei tagastata.

```
Analüüsid
Analüüs Ühik Referents 31.11.2018
S,P-UA Kusihape seerumis umol/L 220-430 400
...
S,P-CRP C-reaktiivne valk mg/L <=7 3

Raviarst
...
```

Näide 20: Analüüsi tekst tüüp 5

Teisel juhul algab tekst mõne analüüsi nimega, millele järgneb sõna "Rerentsväärtus" ja seejärel tuleb vähemalt üks kuupäev. Tekst lõppeb, kui järgneval real on uue analüüsi nimi.

Näide sellisest tekstist, esimene tabel on ridadel 1-6 ja järgmine tabel algab realt 7.

```
Kliiniline vere ja glükohemoglobiini analüüs Rerentsväärtus 08.06.2013
    02:30:00 09.06.2013 06:56:01
WBC 6 - 10*10*9/l 5.10
RBC M3,5-7,5; N:4,0-6,5 x 10*12/l 5.79
HGB N:130-175; M:130-190 g/l 141
...
S-NT-proBNP (B-tüüpi natriureetilise propeptiidi N-fragment <300pg/mL-
    kroonilise südamepuudulikkuse välistuspiir; >2000pg/mL- diagnost.
    otsustuspiir 915
Hemostasiogramm Rerentsväärtus 01.09.2011 04:30:00 02.09.2011 07:01:41
...
```

Näide 21: Analüüsi tekst tüüp 5

7.2 Vale positiivsed

Ühtegi vale positiivset analüüsi tabelit ei eraldatud. Siiski mõned eraldatud tekstidest ei sisalda mõõtmiste informatsiooni ja seega ei oma erilist väärtust (näide 22).

```
Immuunmeetoditel uuringud

Rerentsväärtus

31.02.2017 21:47:00
```

Näide 22: Analüüsi tekst tüüp 5

7.3 Vale negatiivsed

Hinnanguliselt on valenegatiivsete ehk eraldamata jäänud tabelite arv umbes 180. Peamine põhjus mitte tuvastamisel on lõpu regulaaravaldis. Täpsemalt seisneb probleem kolme reavahetust kirjeldavas regulaaravaldises ("$\backslash n \backslash s? \backslash n \backslash s? \backslash n$"). Nimelt on mõnikord tabeli lõpus hoopis üks või kaks reavahetust, selle kirjutamine regulaaravaldisse tooks aga valepositiivsete hüppelise kasvu.

Teine sarnane probleem on HTML enkodeerimisest tulenev *tag* "NBSP", mis tähistab tühikud. Andmebaasis paistab see sõnena, pythonis aga konverteeritakse kujule "$\backslash t \backslash t \backslash t \backslash t \backslash xa0$". Selle arvesse võtmine võiks valenegatiivsete arvu vähendada.

8 Tagger number 6

8.1 Kirjeldus

Tüüpi 6 analüüsi tekstides leiduvaid tabeleid on kõige rohkem ehk umbes 222 000 (49% kõikidest eraldatud tabelitest). Põhjus peitub selles, et regulaaravaldised on kõige "lõdvemad" ehk märgivad ära väga palju osi tekstidest. Seda probleemi on püütud reguleeritud regulaaravaldiste prioriteetidega, mida käsitletakse peatükis "Prioriteedid".

```
analysis_substitutions = {
    "AN_NAMES":
    r"([Hh]emogramm|[Uu]riini\s*ribaanalüüs|[Uu]riini\s*analüüs\s*
    testribaga|[Uu]riini\s*ribatest|"
    + r"[Vv]ereäige\s*mikroskoopiline\s*uuring|[Kk]liiniline\s*veri|[Vv
    ]ere\s*biokeemia|[Uu]riini\s*sademe\s*mikroskoopia|"
    + r"[Hh]ematoloogilised\s*ja\s*uriini\s*uuringud|[Kk]lii\s*nilise\s
    *keemia\s*u\s*uringud|"
    + r"[Hh]ematoloogilised\s*uuringud|[Ii]mmuunmeetoditel\s*põhinevad\
    s*uuringud|Hematoloogia\s*labori\s*analüüsid|"
    +  r"[Bb]iokeemilised\s*uuringud|MP\s*Hormoonid,\s*kasvajamarkerid\
    s*jm\.\s*immuunuuringud\s*|"
       r"[Vv]ereäige mikroskoopia)\s*",
    "AN": r"(analüüsid)\s*",
}
analysis_name_regex = r"({AN_NAMES})\s*".format(**
    analysis_substitutions)
analysis_regex = r"({AN}|{AN_NAMES})\s*".format(**
    analysis_substitutions)

start_regex = analysis_regex  + date_regex + "|"  # 1.
    + date_regex + analysis_regex + "|"  # 2.
    + analysis_name_regex + ".*",  # 3.

# 1. matches a line that is followed by two empty lines
# 2. end of the file (ending parenthesis or word)
# 3. if next line start with word "LEID"
# 4. if next line starts with optional date and another analysis name
end_regex = r"(.\n\n\n)|"  # 1.
    + r"((\)|\w*)\s*$)|"  # 2.
```

```
    + r"(\n(.*)(?=\s*LEID))|"  # 3.
    + r"\n(.*)(?=\s*" + date_regex  # 4.
    + r"?\s*{AN_NAMES})".format(**analysis_substitutions),
```

Tüüp 6 põhineb erinevatel analüüsinimedel, mis on defineeritud ridadel 3-8. Tabeli alguseid on kolme tüüpi:

- esimene eraldatava teksti rida algab analüüsiga ja sellele järgneb kuupäev
- esimene eraldatava teksti rida algab kuupäevaga ja sellele järgneb analüüs
- esimene eraldatava teksti rida algab analüüsi nimega.

Tabeli lõputingimusi on defineeritud 4 (tegelikkuses veelgi rohkem):

- reale järgneb 3 reavahetust
- teksti lõpp
- järgmine rida algab sõnaga "LEID"
- järgmine rida algab valikulisel kuupäevaga ja analüüsi nimega.

Näites 23 on alguse tingimuseks kuupäev + analüüsi nimi ning lõputingimuseks on kolm reavahetust (eraldatakse read 1-5).

```
meetod alates 01.05.2011
Hemogramm viieosalise leukogrammiga
WBC 8.23 (3,6 .. 9,4 E9/L)
...
PLT 281.0 (160 .. 460 E9/L)


EKG kaasa antud
```

Näide 23: Analüüsi tekst tüüp 6

Näites 24 on alguse tingimuseks sõne "ANALÜÜSID", millele järgneb kuupäev. Lõpuntingimuseks on uus analüüsi nimi (eraldatakse read 1-4).

```
ANALÜÜSID 13.01.15
WBC 9.03 (2,9 .. 9,2 E9/L)
...
Erütrotsüüdid 10 / L  ! (NEG )
Uriini sademe mikroskoopia
...
```

Näide 24: Analüüsi tekst tüüp 6

8.2 Vale positiivsed

Kuna eraldatud tekste on üles 200 000 on raske täpset valede tekstide arvu hinnata. Siiski saab kindlalt öelda, et vähemalt 1100 teksi ei sisalda informatsiooni patsiendi analüüside kohta. Sellised eraldatud tekstid näevad välja nagu näites 25. Tegelikkuses on algne tekst nagu näites 26 (punasega on märgitud *taggeri* poolt märgitud algus ja lõpp). Selliste tekstide jaoks, kus iga rida algab kuupäevaga tuleks defineerida eraldi tüüp.

```
Hematoloogilised uuringud:
```

Näide 25: Analüüsi tekst tüüp 6

```
*11.09.2017 <ANONYM id="0" type="per" morph="_Y_ ?;_H_ sg n"/>: Väsimus
    ,halb enesetunne,peas halb tunne,esinenud vererõhu tõusu maksim.
    145/88mmHg.Anamneesis pankreatiit.
*23.12.2017: *Hematoloogilised uuringud:
 Hemogramm
11.09.2017 10:24: Hemoglobiin 139 g/L
```

Näide 26: Analüüsi tekst tüüp 6

8.3 Vale negatiivsed

Tüübi 6 regulaaravaldised on väga paindlikud ja probleemiks on pigem liiga paljude tekstide eraldamine (suur valepostiivsete arv). Teadaolevaid eraldamata jäänud tekste pole või polnud otsing nende leidmiseks piisavalt põhjalik.

9 Tagger teksti tüübile 7

9.1 Kirjeldus

Tüüp 7-le vastavaid tabeleid sai eraldatud 66 615, mis moodustab 15% kõikidest eraldatud tabelitest. Olulisem tüüp 7 *taggeri* juures on see, et see tegeleb n.ö. ülejääk tabelite korjamisega. Täpsemalt, tabelite otsinguga, mida teised *taggerid* tuvastada ei ole suutnud. Seega on selle puhul väga raske teha üldistusi alguse ja lõppu regulaaravaldise kohta, sest need varieeruvad seinast seina. Sellest tingituna on enamik eraldatud tabelitest poolikud. Siiski on üldistus tehtud, et võimalikult vähe tabeleid eraldamata jääks.

Taggeri regulaaravaldised näevad välja järgnevad (***date_subsitutions* on defineeritud peatükis "Kuupäevad").

```
date_regex_with_time = r"({DAY}\.\s*{MONTH}(\.\s*{YEAR}\s*)?\s*({HOUR
    }:{MIN})?)".format(**date_substitutions)

start_regex =  "(" + date_regex_with_time +")?\s*(S,P-Na|WBC|S,P-Glü
    koos)\s*\d"

```

```
end_regex = r"(PLT|S,P-CRP|S,P-Kreatiniin|Protrombiini\s*%|S,P-CA|fS,fP
    -Triglütseriidid|RDW-CV).*?\n"
```

Regulaaravaldistest on näha, et need otsivad väga konkreetseid nimesid tabeli alguse ja lõpuna. Tabeli alguseks loetakse sõnu:

- S,P-Na
- WBC
- S,P-Glükoos.

Alguse regulaaravaldise lõpus võib täheldada "\d", mis teeb kindlaks, et neid nimesid ei mainitud niisama teksti sees, vaid neile järgneb ka numbriline mõõtmistulemus. Võimalusel võtab *date_regex_with_time* kaasa ka lisainfona analüüsi kuupäeva, kui see on olemas. Tabeli lõpuks loetakse järgnevaid sõnu:

- PLT
- S,P-CRP
- S,P-Kreatiniin
- Protrombiini*%
- S,P-CA
- fS,fP-Triglütseriidid
- RDW-CV.

Lõpu regulaaravaldises võtab ".*?" osa rea lõpuni kogu info. Järgmises näites (number 27) on hästi demonstreeritud ".*?" eesmärk. Regulaaravaldis märgib lõputingimuseks "RDW-CV", ".*?" aga võtab kogu info kuni rea lõpuni ja seega eraldatakse kogu info kuni PLT 351 (172 .. 385 E9/L)-ni. Punasega on märgitud eraldatav info.

```
19.07.2012 10:19
WBC 17.3 (2,3 .. 9,2 E9/L ) RBC 2,9 (4,1 .. 6,7 E12/L ) HGB 98 (119 .. 164 g/L ) HCT 38
    (31 .. 49 % ) MCV 76.0 (78 .. 91 fL ) MCH 21.5 (26 .. 38 pg ) MCHC 361 (323 .. 365
    g/L ) RDW-CV 19.1 (12,3 .. 16 % ) PLT 351 (172 .. 385 E9/L )
```

Näide 27: Analüüsi tekst tüüp 7

Samamoodi toimib regulaaravaldis ka juhtudel, kui iga analüüt on eraldi real. Näites 28 eraldatakse kõik read (1-9).

```
WBC 8.04.9 - 7.3 E9/L Venoosne veri - 31.08.2017 10:12:39
% kp ja väärtused asendatud
RBC 5.13 3.9 - 5.2 E12/L
Hb 147 121 - 168g/L
Hct 45 32 - 54 %
```

```
MCV 92 84 - 107 fL
MCH 32.1 24 - 39 pg
MCHC 339 327 - 369 g/L
RDW-CV 15.7 13.1 - 16 %
```

Näide 28: Analüüsi tekst tüüp 7

9.2 Poolik tulemus

Probleemseks osututub aga järgnev tekst (näide 29). Siin eraldatakse info alates "10.03.2014 WBC"-st kuni "RDW-CV 14.2 (12,3 .. 16 %) " ehk eraldamine pole täielik. Eraldatav tabel on märgitud punasega, tegelik tabeli lõpp rohelisega.

```
10.03.2014
WBC 5.12 (2,9 .. 7,4 E9/L )
RBC 6.75 (4,5 .. 7,9 E12/L )
HGB 154 (139.. 157 g/L )
HCT 39 (38 .. 52 % )
94.2 (78 .. 103 fL )
MCH 29.4 (26 .. 38 pg )
MCHC 351 (323.. 365 g/L )
RDW-CV 14.2 (12,3 .. 16 % )
PLT 352 (172 .. 385 E9/L )
erütroblastid
NRBC% 0.0 (/100WBC )
NRBC# 0.00 (E9/L )
```

Näide 29: Analüüsi tekst tüüp 7

Samas ei saa lõputingimusest "RDW-CV" eemaldada, sest sellisel juhul jääksid näite 30 sarnased tabelid eraldamata.

```
21.07.2015
WBC 15.8 (5,6 .. 10,1 E9/L )
RBC 5.74 (3.9 .. 8.1 E12/L )
HGB 160 (159.. 190 g/L )
HCT 47 (46 .. 53 % )
RDW-CV 214 (130 .. 370 E9/L )

<muu tekst>
```

Näide 30: Analüüsi tekst tüüp 7

Kokkuvõtteks võib öelda, et tüüp 7 puhul on enamik eraldatavatest tabelitest poolikud. Seda probleem on raske parandada, sest analüütide järjekord pole fikseeritud ja seega ei ole võimalik väga üldisi regulaaravaldisi kirjutada.

10 Probleemsed tekstid

Pärast tekstide eraldamist ilmnes tekstitüüp, mida algses tekstide läbivaatusel ei täheldatud, aga millele võiks kirjutada eraldi *taggeri*. Nimelt praegused taggerid

eksivad totaalselt tekstide eraldamisega, mis sisaldavad palju (või lausa igal real) kuupäevi.

Näiteks näites 31 eraldatakse hetksesisuga 4 tabelit:

- tüüp 2 tabel (punane), read 1-2
- tüüp 6 tabel (sinine), read 3-6
- tüüp 7 tabel (lilla), read 9-12,

kuigi oleks optimaalne eraldada üks tabel ridadel 1-15.

```
*??.02.2013: LISAN: AS HAIGLAS TEOSTATUD JA LABORI ANALÜÜSIDE VASTUSED:
19.02.2013 erütrotsüütide settekiirus B-ESR 13 mm/h -
19.02.2013 vereäige mikroskoopia B-Smear
Keppt. 5 <7
...
basof. 7 -
08.02.2012 hemogramm B-CBC-3Diff
B-RBC 5,03 10 12 /l 4-8,1
B-WBC 5,1 10 9 /l 3,9-11,2
B-HCT 39,8 % 34-45
...
B-PLT 189 10 9 /l 140-460
B-MPV 6,9 fl 8-10
...
19.02.2013 glükoos seerumis/plasmas gly 6,7 mmol/l 5-8
 LISAN: AS HAIGLAS TEOSTATUD JA LABORI ANALÜÜSIDE VASTUSED:
```

Näide 31: Probleemne tekst

11 Segmentide tagger

Segmentide taggeri eesmärk on panna kokku vastavad alguse ja lõpu märgendid, et kätte saada kogu nende vahele jääv tekst. *Tagger* otsib kõigepealt ülesse "START" märgendi. Kui sellele järgneb "END" märgend, siis märgendatakse segmendiks "START" ja "END" vahel olev tekst, kaasaarvatud algus ja lõpp (näiteid on toodud kaustas *code*).

Oluline on märkida, et *tagger* võtab arvesse ainult järjestikkuseid "START" ja "END" märgendeid. Muus järjekorras olevad märgendeid arvesse ei võeta.

12 Prioriteedid

Prioriteedid mängivad olulist rolli *taggerite* omavahelisel koostoimimisel. Tihti juhtub, et mitu *taggerit* märgendavad sama teksti osa. Prioriteedid teevad kindlaks, et jäetakse alles märgend, mille *taggeri* prioriteet on madalam. Vastasel juhul eraldatakse ühte infot mitu korda ja tekivad duplikaadid. Üritame leida kõigepealt väga struktureeritud tabeleid (väiksem prioriteedi number) ja alles siis vähem struktureerituid (suurem prioriteedi number).

Näiteks tüübid 6 ja 7 märgendavad mõnikord poolikut/vale informatsiooni ja seeläbi eraldavad mitte sobivat infot. Samas tüüp 2 on väga kindla struktuuriga ja seega on mõistlik eelistatada alati eraldada tüübile 2 *taggerile* vastavat teksti. Järgnevas tabelis on kõigi tüüpide prioriteedid välja toodud.

Tüüp	"START" prioriteet	"END" prioriteet
type1	1	1
type2	0	0
type3	1	1
type4	1	1
type5	1	1
type6	2	3
type7	4	5

Eelnevat prioriteetide kirjeldust saab illustreerida näitega 32. Selles teksti leidub nii tüüp 4 alguse (punane), tüüp 7 alguse (sinine) ning tüüp 4 ja 7 lõpu (lilla) märgendeid. Kuna tüüp 4 prioriteet on madalam, siis eelistatakse selle poolt märgendatud teksti. Seega eraldatakse read 1-19 (tüüp 7 puhul oleks eraldatud ainult read 15-19).

```
Analüüsid:



Kliiniline vere ja glükoh
emoglobiini analüüs

Rerentsväärtus

04.04.2019  09:35:45


Materjal
B-CBC-5Diff  tehtud
WBC 4 -
10*10*9/l 9.67
MCHC 372 - 395 g/l 357
RDW-CV 9 - 10
g/l 13,0

Väljastatu
d dokumendid
...
```

Näide 32: Analüüsi tekst mitmene tüüp

13 Tulemused

Pärast 37 miljoni tekti läbikäimist, sai nendest eraldatud umbes pool miljonit teksti (458 274). Tekstid jaotustest tüüpide järgi on näha järgnevast tabelist.

Sellest ilmneb, et tüüpidele 4 ja 5 pole vaja väga palju tähelepanu pöörata, sest nende osakaal andmestikes on väga väike. Samal ajal tüüp 6 moodustab pea poole kõikidest tekstidest. Siiski on seal väga palju vale positiivseid, mille eemaldamisele tuleks tähelepanu pöörata.

Tüüp	Sagedus	osakaal (%)
type6	222 693	48.6
type2	129 249	28.2
type7	66 615	14.5
type1	33 181	7.2
type3	4 616	1.0
type5	1 645	0.4
type4	275	0.1

14 Kokkuvõte

Tekstidest mustrite otsimine ja nende põhjal regulaaravaldiste kirjutamine on väga ajamahukas töö. Mida rohkem tekste läbi vaadata, seda paremaid ja täpsemaid tulemusi on võimalik saada. 37 miljonit teksti on aga väga suur hulk, nii et selles projektis sai tehtud esmane töö. Regulaaravalidstes on palju täiendamise võimalusi, potentsiaalsed parandamise kohad on ka selles projektis välja toodud. Siiski on pool miljonit eraldatud teksti väga hea tulemus, sest varasemalt pole neid andmeid olnud võimalik sellisel kujul üldse kasutada.