„Automaatne info eraldamine eestikeelsest tekstist“ (projektijuhtidele)

08.10.2020 Metropol Spa Hotelli konverentsikeskuses

Üritust korraldab Riigi Infosüsteemi Amet Euroopa Liidu struktuuritoetuse toetusskeemist "Infoühiskonna teadlikkuse tõstmine", mida rahastab Euroopa Regionaalarengu Fond.

RIIGI INFOSÜSTEEMI AMET

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

TARK e-RIIK

Praktiline info

Koolitaja: Dage Särg (Tartu Ülikool / STACC OÜ)

Slaidid: https://tinyurl.com/nlp-tallinn

Notebook: https://tinyurl.com/nlp-notebook

EstNLTK: https://tinyurl.com/estnltk-repo

Kontakt: dage.sarg@ut.ee

Koolituse kava

8:30-9:00 Tervituskohv
09:00-10:00 Sissejuhatus, info eraldamise liigid ja rakendusvaldkonnad
Tekstitöötluse baassammud[NB]
10:00-10:15 Energiapaus
10:15-11:30 Eesti keele kasutusvalmis infoeralduse töövahendid[NB]
Töövahendite kohandamine ja uute loomine
Rühmatöö/individuaalne mõtlemisülesanne - keeletöötlusprobleemi lahenduse kavadamine
11:30-11:50 Toekas kohvipaus
11:50-13:00 ***Rühmatöö arutelu***
Tekstide sarnasus, klassifitseerimine ja klasterdamine
Sõnad ja vektorid[NB]
Nimeüksuste tuvastaja kohandamine[NB]
Kokkuvõte

Sissejuhatuseks

NLP - *Natural Language Processing* - loomuliku keele töötlus

Suurandmed

Andmekaeve

Masinõpe

Tutvumiseks

1) Kellele on kõik eelmise slaidi märksõnad igapäevaste tööülesannete osaks?
2) Kes on mingil moel puutunud kokku keele automaattöötluse valdkonnaga?
3) Kes on kasutanud mõnda programmeerimiskeelt?

Big Data - suurandmed

Andmestikud, mis on nii suured ja kompleksed, et traditsioonilistest andmetöötlusvahenditest ei piisa

Big Data - suurandmed

Eri tüüpi andmeid genereeritakse pea igal elusammul, nt:

- *Online*-suhtlus
- GPS-andmed nutitelefonidest
- Võrku ühendatud seadmete omavaheline suhtlus
- Kassasüsteemid, kaardimaksed
- Tööstusseadmed

Struktureeritud vs struktureerimata, sh keeleandmed

(Keele)andmed automaattöötlusel

Andmed peavad olema:

- kättesaadavad, sh juriidilisest vaatepunktist
- mahukad - kui mahukad?
- kasutatavad - puhtad ja struktureeritud (?)
- mõistetavad - mis andmed need on?
- hallatavad - peame suutma andmeid ja metaandmeid korduvalt ja arusaadavalt käsitleda

Keeletehnoloogia

- Keeletehnoloogia hõlmab arvutusmeetodeid, arvutiprogramme ning elektroonikaseadmeid, mis on loodud just inimkeele ja -kõne mõistmiseks, tekitamiseks ning teisendamiseks.

- Hans Uszkoreit

Keeletehnoloogia

- Keeletehnoloogia hõlmab arvutusmeetodeid, arvutiprogramme ning elektroonikaseadmeid, mis on loodud just inimkeele ja -kõne mõistmiseks, tekitamiseks ning teisendamiseks.

  - Hans Uszkoreit

- Kõnetöötlus

Keeletehnoloogia

- Keeletehnoloogia hõlmab arvutusmeetodeid, arvutiprogramme ning elektroonikaseadmeid, mis on loodud just inimkeele ja -kõne mõistmiseks, tekitamiseks ning teisendamiseks.

- Hans Uszkoreit

- Kõnetöötlus
- Tekstitöötlus

Keeletehnoloogia

- Keeletehnoloogia hõlmab arvutusmeetodeid, arvutiprogramme ning elektroonikaseadmeid, mis on loodud just inimkeele ja -kõne mõistmiseks, tekitamiseks ning teisendamiseks.

- Hans Uszkoreit

- Kõnetöötlus
- Tekstitöötlus

Keeletehnoloogia ja lähedased mõisted

Keeletehnoloogia (*language technology*) ~ loomuliku keele töötlus (*natural language processing*, NLP) ~ arvutilingvistika (*computational linguistics*)

Milleks?

Et säästa aega ja raha, teha elu mugavamaks ja toredamaks

Google Translate
vs inimesest tõlkija

Automaatne info eraldamine

- Avatud - eraldame kogu tekstis oleva info
- Piiratud - eraldame tekstist vajaliku /meid huvitava info

Avatud info eraldamine

- Teoorias ilus eesmärk
- Praktikas veel lahendamata probleem

Info eraldamist kasutavaid rakendusi

- Küsimus-vastus-süsteemid (*chatbot'*id)
- Masintõlge
- Automaatsed sisukokkuvõtted
- Meediamonitooring
- Rämpsposti filtreerijad
- Soovitussüsteemid
- ...

Info eraldamise alamülesandeid

- Kindla tehnilise struktuuriga elementide tuvastamine
- Kindla lingvistilise struktuuriga elementide tuvastamine
- Võtmesõnade ja teemade tuvastamine
- Nimeüksuste tuvastamine
- Kategooriate tuvastamine
- Sarnaste tekstide tuvastamine, tekstide liigitamine
- ...

Tehniline/ortograafiline struktuur

- Näiteks: telefoninumbrid, isikukoodid, veebiadressid, e-mailid …
- Üldiselt lahendatakse reeglipõhiselt - regulaaravaldiste abil

Arvandmete eraldamine vabast tekstist

Pulsi märkimine ühel isikul:

- Siinusrütm **59** lööki min
- Kiirenenud, Ekg-l siinus rütm , **160**/min
- vatsakeste tahhüarütmia fr -ga **150** lööki min
- EKG --> siinusrütm, fr.110xmin
- Ps 66/min
- Fr = 77 ' min
- Maksimaalne fr=196 , minimaalne fr=55
- V/v-l RR 133/105/90 , arütmia

Arvandmete eraldamine vabast tekstist

(?P<key>(((S|s)iinus)?r.tm(iline|ilised)?|[Ff]rekv?(ents)?|fr\.?|Fr|BPM|bpm|SR|SLS|FR|HR|(P|p)ulss(i)?|Ps)(\s*[xX]\s*)?)\s*-?:?(ca|u|[-,*x=])?\s*(?P<pulse>(([12][0-9]{2})|([3-9][0-9])))(\s*(?P<unit>(((l|x|X|lööki))\s*/?\s*(1\s*)?min(utis)?)|/min|x['´`]|bpm|BPM|x|X))?

Tehniline keerukus

PSA 2.19 ng/ml

PSA 2.46 (µg/L)

25.10.2010 PSA 1,71 ng/ml

PSA 03042012 - 0,83ng/ml perearsti poolt .

PSA 2010. 3ng/ml, PSA 2012.

PSA 1,53ng/ml . - Bx va

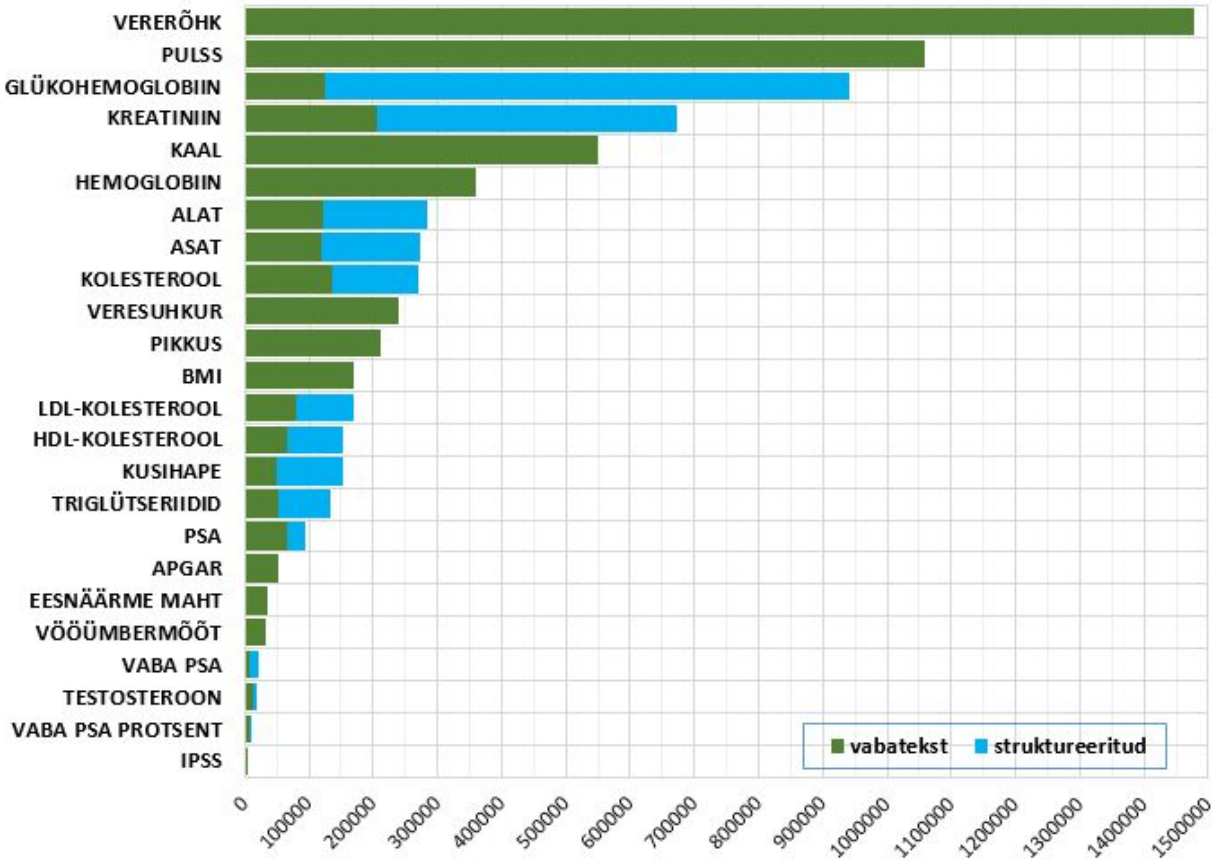
PSA 20105,99 ja 26.01.2012 uuesti .

PSA 2011 oli 0 , 4 nG7ml .

PSA 201222,25ng/ml

PSA 2 aastajooksuldünaamikata , eriuuring

Arvuliste väärtuste päritolu



Sisuline keerukus

Vererõhu **korrektne kirjeldus**?

Sisestusvead, ebatäpsused jne:

Vererõhk stabiliseerunud 150/210

Vererõhu väärtused ulatusid 180/190

Vererõhud 145-90/170-100 mmHg

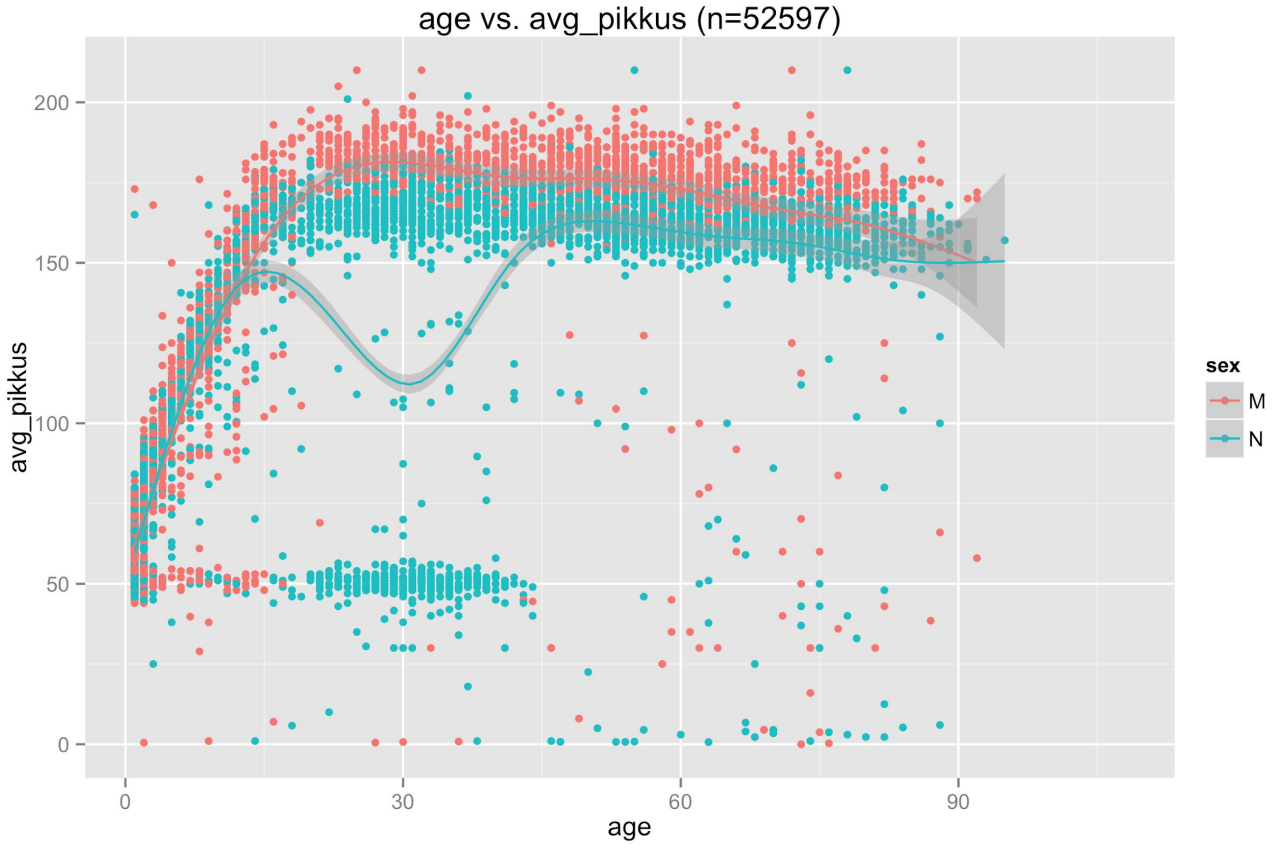
RR 120/70, kodus oma mõõtjaga 174/113 hommikul, seega pt aparaat valetab

⇨ Vajab sisendit spetsialistilt

1. 60% - väärtus ühene, arsti mõõdetud
2. 10% - väärtus patsiendi mõõdetud
3. 18% - väärtus on antud vahemikus või mõõdetud kolmanda osapoole poolt
4. 12% - valepositiivne tulemus

1. ja 2. sobiks kasutamiseks otsusetoele

Probleemid andmetes



Lingvistiline struktuur

- Nimisõnafraaside tuvastamine

  Suurepärane etendus, aga lühike vaheaeg ja kallid hinnad kohvikus rikkusid pisut tuju.

Võtmesõnade ja teemade tuvastamine

- Info otsimise lihtsustamine
- Tekstide liigitamine
- Baastasemel - lemmatiseerimine

Nimeüksuste tuvastamine

- Named Entity Recognition (NER)
- Kes? Kus? => Sageli olulisim info tekstis

Ajaväljendite tuvastamine

- Millal? Kui kaua?

Tekstide anonümiseerimine

- Inimese otsest identifitseerimist võimaldava info (ees- ja perekonna nimed, isikukoodid, telefoninumbrid jms) eemaldamine tekstist

Tekstide anonümiseerimine

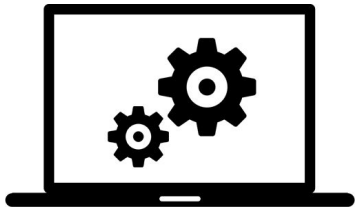
Näiteks teadustööks Digiloo andmetel:

- Andmete valdajal ja uuringute läbiviijal lasub kohustus tagada patsientide ja ka meditsiinipersonali anonüümsus
- Palju sensitiivset infot
- Sageli vabatekstides (isa nimi, isikukood, tel nr…)
- Ei saa teadlastele sellisel kujul kätte anda

Nimeüksuste tuvastamine - anonümiseerimine

Sisendtekst

Patsient **John Doe** Vanus 44 a. IK – **77771478888** võeti statsionaarsele ravile. Asjaolude täpsustamiseks helistada dr. **Hämarikule** tel: **7177765**, kell 10.00-13.00.



Anonümiseerija

95% isikuinfost eemaldatud

Anonümiseeritud tekst

Patsient **XXX** Vanus 44 a. IK – **XXX** võeti statsionaarsele ravile. Asjaolude täpsustamiseks helistada dr. **XXX** tel: **XXX**, kell 10.00-13.00.

Tekstiotsingud

Nt google otsingud, Eesti keele seletav sõnaraamat, ...

Sarnaste tekstide tuvastamine

- plagiaadituvastus;
- rämpspostituvastus,
- kliendimeilide liigitamine ja suunamine õigele inimesele
- uudiste rubriikide tuvastus

Meelestatuse analüüs

- Mida kliendid/kasutajad/tudengid/kolleegid/… minust/firmast/tootest/konkurendist arvavad?
- Kuidas mind/firmat/.. kajastatakse?

Info struktureerimine/tekstikaeve

Tekstiliste andmete “teisendamine” struktureeritud kujule

Tekstikaeve näide

- *Patsiendi eluanamnees: suitsetaja. ei joo. varasemalt diagnoositud hüpertoonia, diabeet. kerge ülekaal.*

Tekstikaeve näide

- *Patsiendi eluanamnees: suitsetaja. ei joo. varasemalt diagnoositud hüpertoonia, diabeet. kerge ülekaal.*

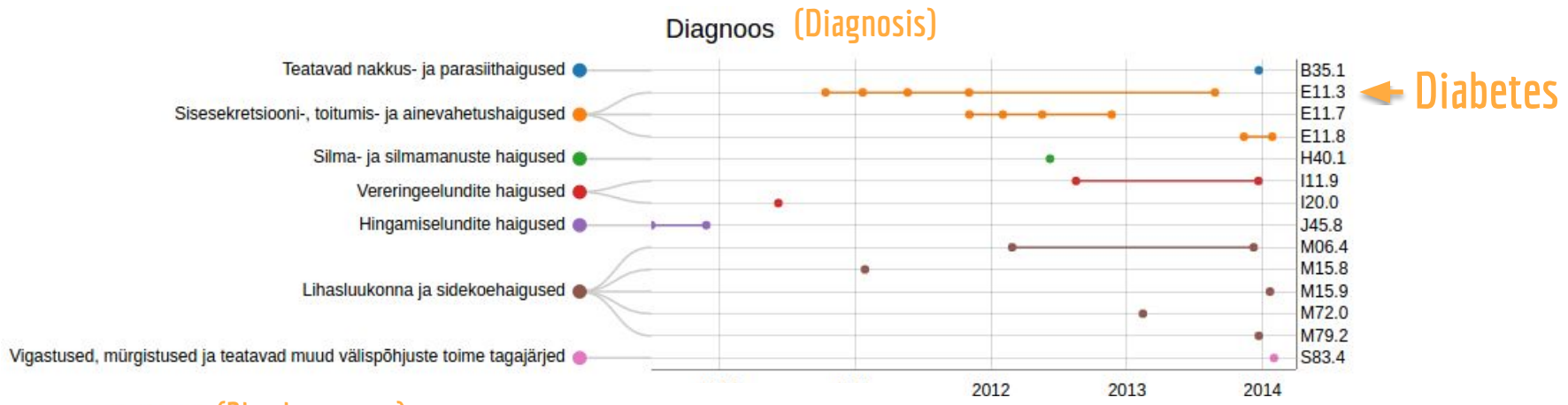
Patsient	Suitsetamine	Alkohol	Varasemad diagnoosid	Ülekaal	...
1	TRUE	FALSE	I10; E11	TRUE	...

Tekstikaeve näide

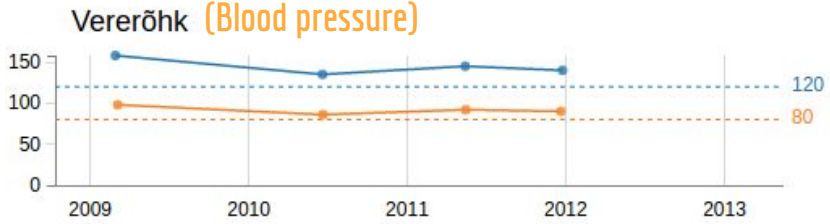
- *Patsiendi eluanamnees: suitsetaja. ei joo. varasemalt diagnoositud hüpertoonia, diabeet. kerge ülekaal.*

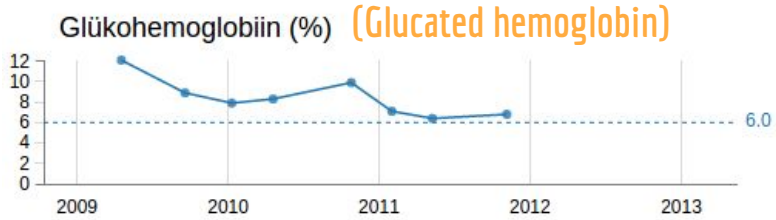
Patsient	Suitsetamine	Alkohol	Varasemad diagnoosid	Ülekaal	...
1	TRUE	FALSE	I10; E11	TRUE	...
2	FALSE	TRUE	NULL	TRUE	...
3	TRUE	TRUE	F32; F33	TRUE	...
N	...	...	...	...	...

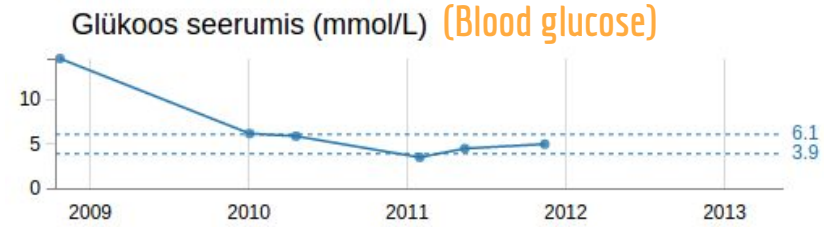
Patsiendi ülevaade

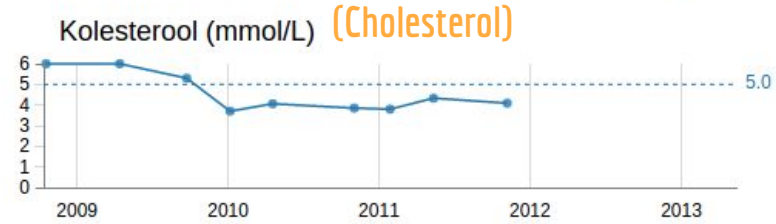


Diabetes









Tekstilised andmed

Suunatud PA-lt , probleemiks kõhuvalu, mis ca aasta vältel olid 2 x nädalas, aga aprilli algul äge haigus seedehäiretega.

Sellest ajast kaebab iga päev, rohkem hommikul ärgates või enne und. liveldust, oksendust sel ajal ei ole.

Iste tavaliselt regulaarne, eile-täna iste vedel.
Anamneesis ka peavalud.

Saadetud gastroskoopiasse.

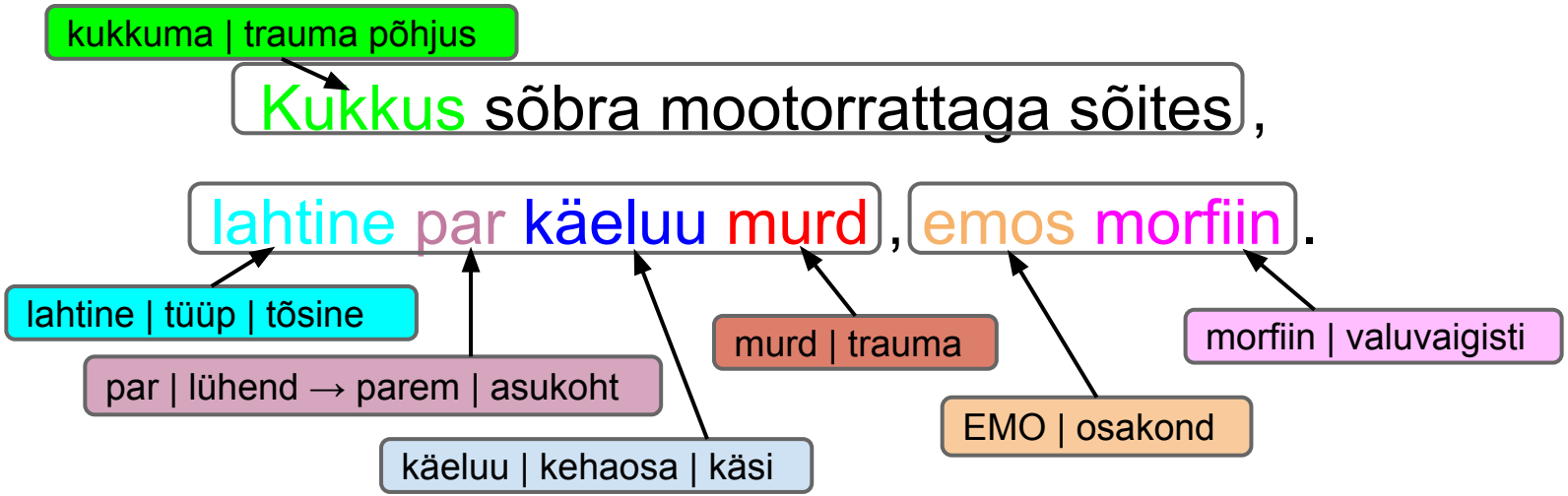
Kaebused:

- Kõhuvalu
- Seedehäire
- Iiveldus (neg)
- Oksendamine (neg)
- Iste vedel
- Peavalu

Uuringud:

- gastroskoopia

Tekstilised andmed

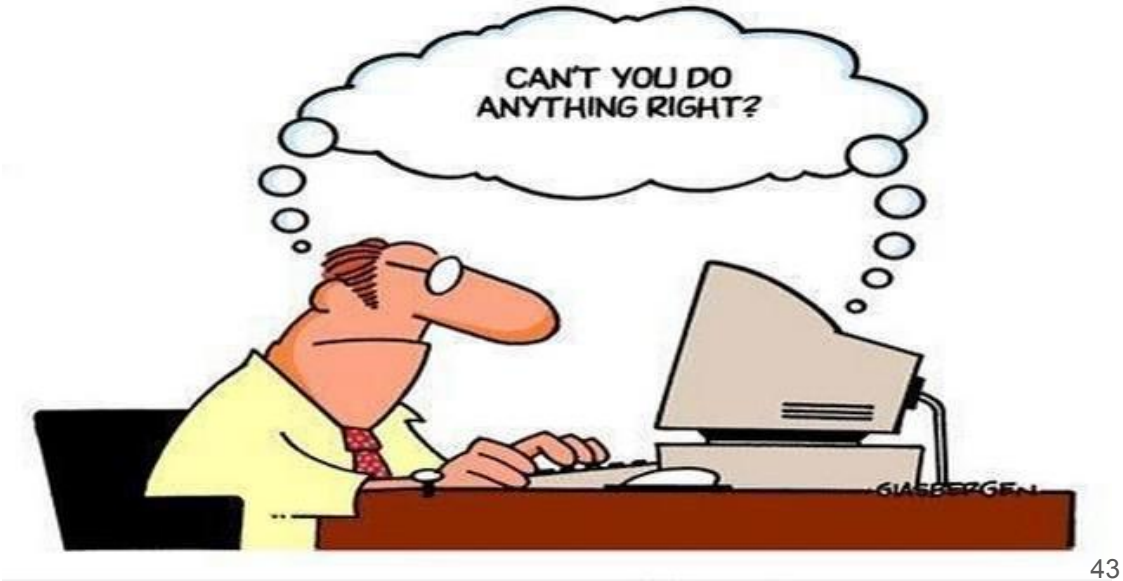


Kuidas?

Inimesed oskavad keeli, mõistavad tekste

Arvutid?

Mõistavad teisi keeli



Eesti keeletehnoloogia ~1995-2014

C++, Java, Perl, Bash, Awk, Vislcg3, Python, etc…

‘*available on demand*’

EstNLTK

- Kogumik Pythoni teeke eestikeelsete tekstide töötluseks
- Olemasolevate tööriistade omavaheline liidestamine, kättesaadavaks muutmine + uute loomine
- https://github.com/estnltk/estnltk

EstNLTK

- 2014-2016 EKT57 EstNLTK: Pythoni teegid eestikeelsete vabatekstide lihtsamaks töötlemiseks (TÜ, Sven Laur)

  Siim Orasmaa, Timo Petmanson, Aleksandr Tkatšenko

- 2017: EKT110 EstNLTK teegi täiendamine ja selle rakendamine praktikas (TÜ, Sven Laur)

  Siim Orasmaa, Aleksandr Tkatšenko, Uku Raudvere, Dage Särg

EstNLTK

- 2018-2021: EKTB14 "EstNLTK teegi ja sellega seotud veebiteenuste arendamine (TÜ, Sven Laur)

  Siim Orasmaa, Paul Tammo, Dage Särg, Rasmus Maide, Birgit Sõrmus, Claudia Kittask

Miks Python?

- Loetav
- Kompaktne
- Lihtne (?)
- Sisseehitatud andmestruktuurid
- Palju standardteeke
- *De facto* keel andmeteaduses (koos R-iga)
- Üldotstarbeline keel
- Õpetatakse ülikoolides
- Juba olemasolevad tekstitöötluse teegid (NLTK...)

Kuidas?

Aeg järgi proovida :)

Google colab notebook: https://tinyurl.com/nlp-notebook