„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

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

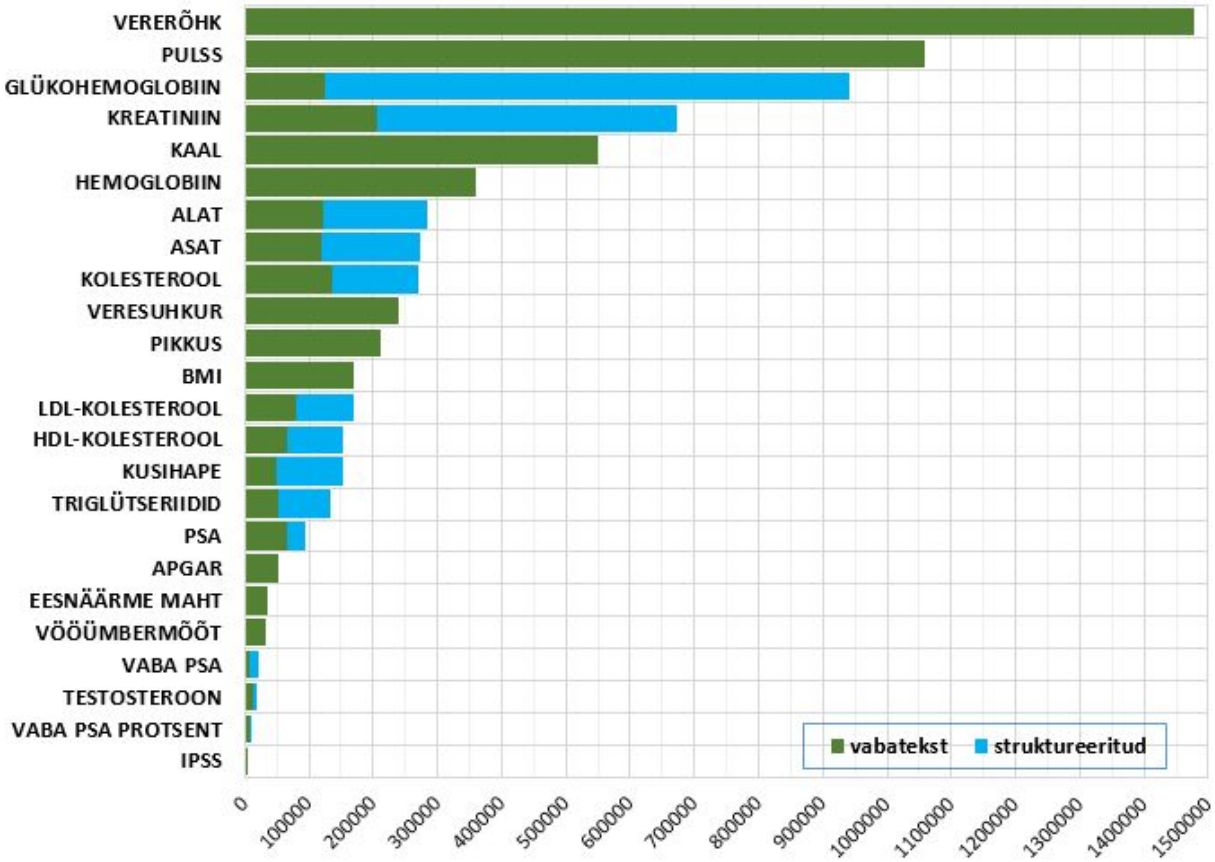
PSA 20105,99 ja 26.01.2012 uuesti .

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

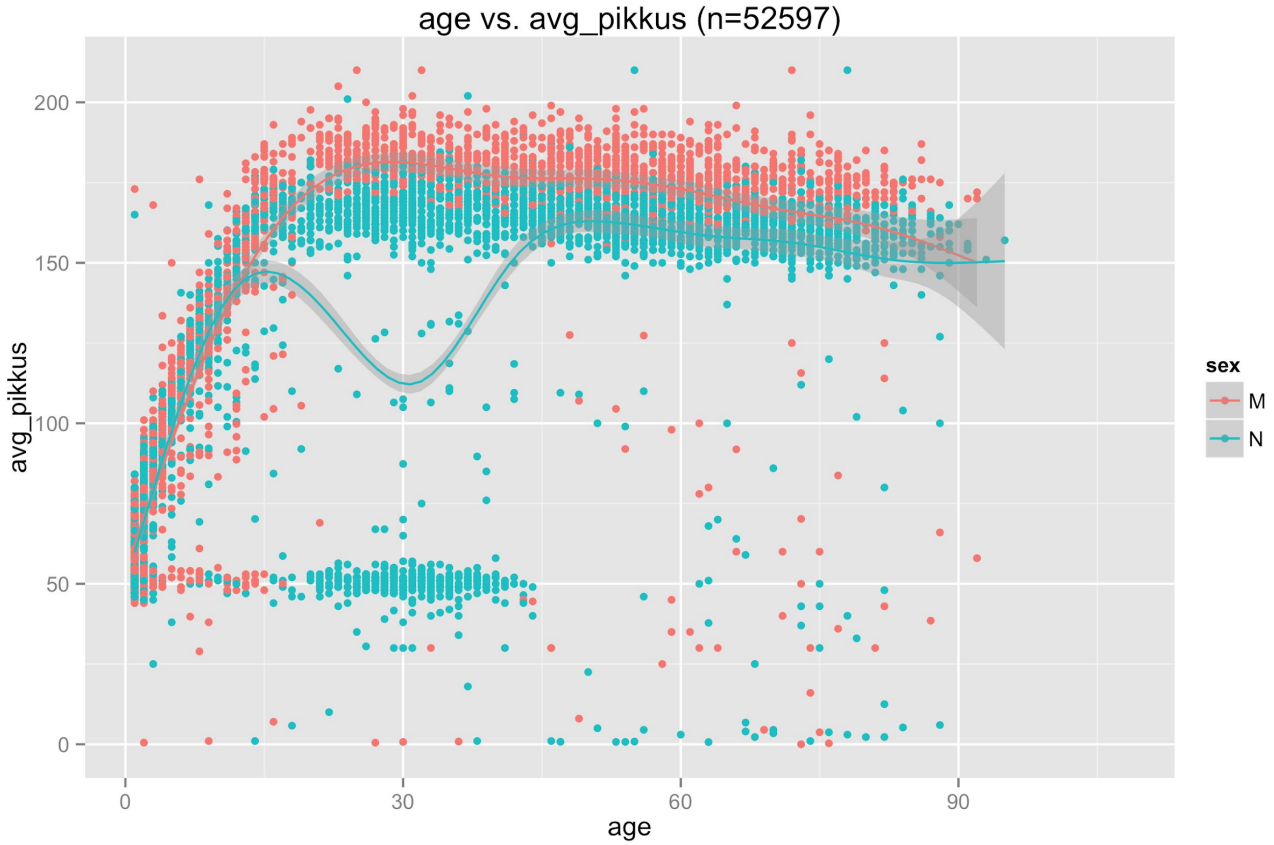
RR 120/70, kodus oma mõõtjaga 174/113 hommikul, seega pt aparaat valetab

⇨ 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

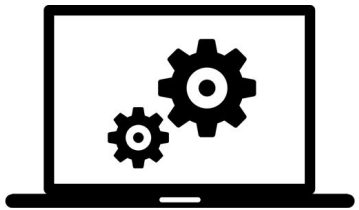
Näiteks teadustööks Digiloo andmetel:

- 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.*

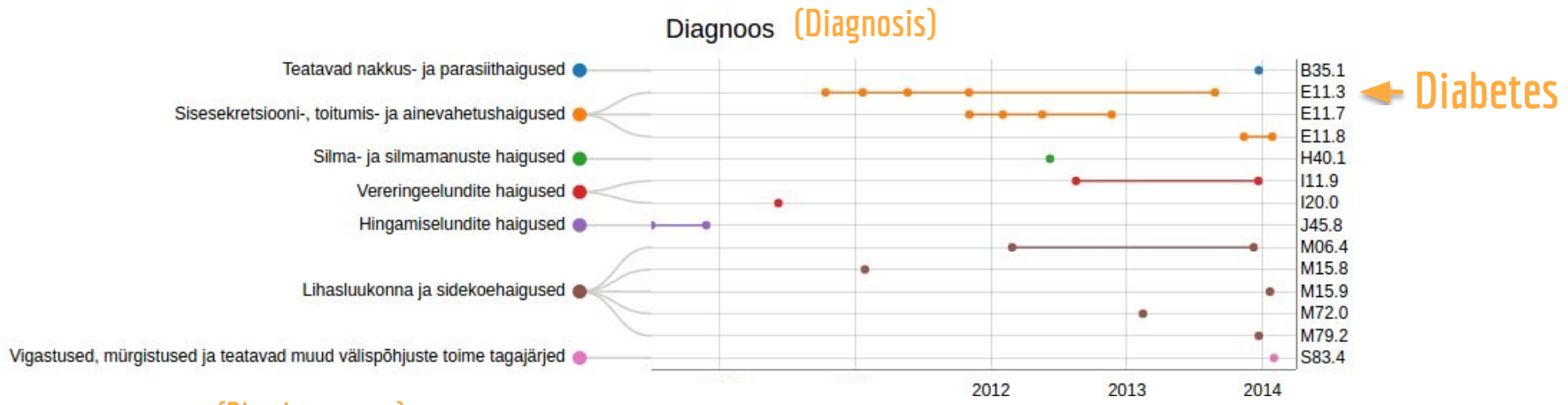
Patsient	Suitsetamine	Alkohol	Varasemad diagnoosid	Ülekaal	...
1	TRUE	FALSE	I10; E11	TRUE	...

Tekstikaeve näide

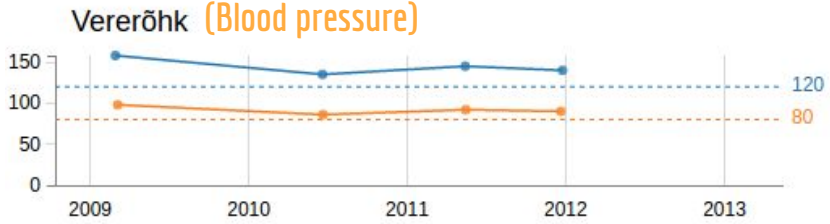
- *Patsiendi eluanamnees: suitsetaja. ei joo. varasemalt diagnoositud hüpertoonia, diabeet. kerge ülekaal.*

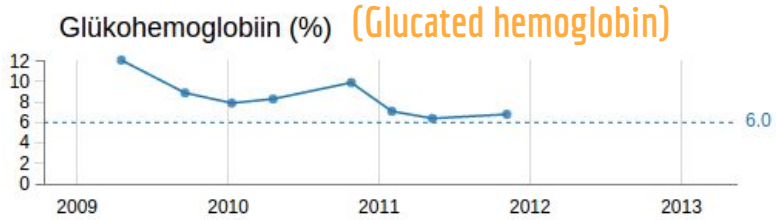
Patsient	Suitsetamine	Alkohol	Varasemad diagnoosid	Ülekaal	...
1	TRUE	FALSE	I10; E11	TRUE	...
2	FALSE	TRUE	NULL	TRUE	...
3	TRUE	TRUE	F32; F33	TRUE	...
N	...	...	...	...	...

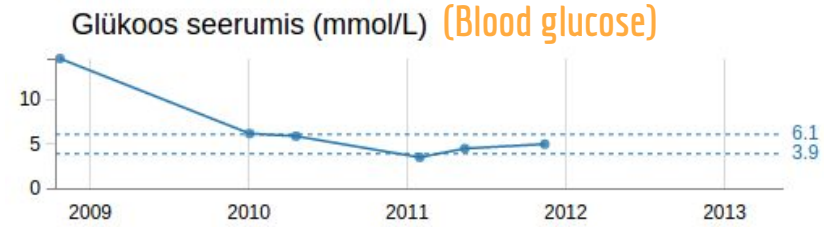
Patsiendi ülevaade

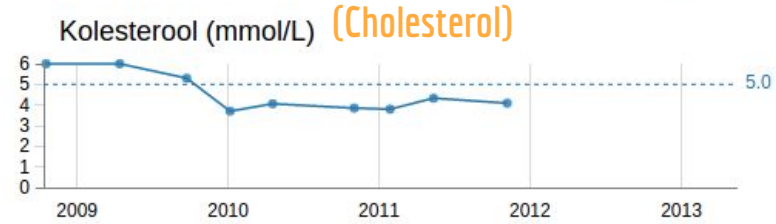


Diabetes









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. Iiveldust, oksendust sel ajal ei ole.

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

Saadetud gastroskoopiasse.

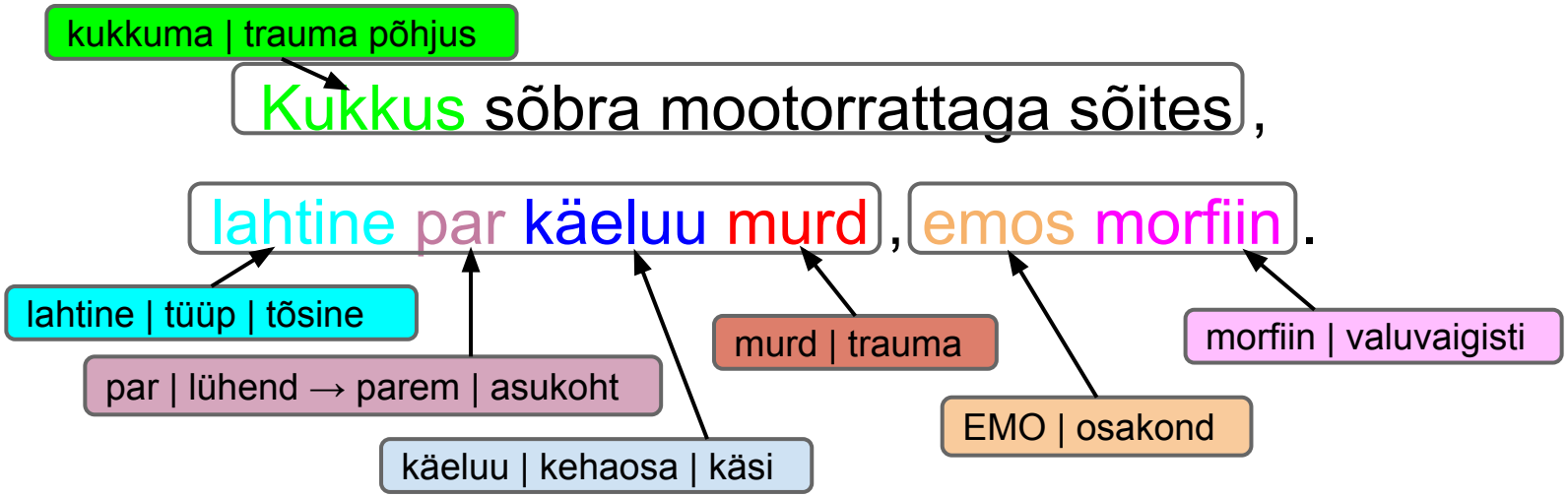
Kaebused:

- 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: TODO

Vahekokkuvõte ja mõtlemishetk

Praeguseks oleme rääkinud, kuidas...

...kätte saada laused,sõnad, lemmad, morfoloogiline info, süntaktiline info, teha sagedusloendeid

...kätte saada asukohad, isikud, organisatsioonid, aadressid, ajaväljendid, verbiahelad, nimisõnafraasid

...luua grammatikapõhiseid märgendajaid (nt nimisõnafraasid)

...kasutada enda huvides ära Wordneti jm leksikaalseid ressursse

Mõtlemisülesanne

Keeletöötlusülesande lahenduse kavandamine

Olukord: meil on hulk eestikeelsete uudiste tekste, kuhu oleme standardse nimeüksuste tuvastaja abil peale märkinud isikud, asukohad ja organisatsioonid. Tahame lisaks juurde märkida, kas tekstides esinevad persoonid on eestlased või välismaalased.

Kuidas võiks seda lahendada?

Lahendusvariandid

- ?

Lahendusvariante

- Kasutada ära Eesti statistikaameti nimeloendit
- Kõrvutada nimesid mõne muu riigi uudistes esinevate nimede sagedustega
- Leida kohanimeandmebaasi abil, kas tekstides esinevad asukohad on Eestis või mitte
- Leida üle kogu korpuse isikute ja asukohtade koosesinemise sagedused
- Kui isik esineb Wikipedias, kasutada ära sealset infot
- Koostada loend meid huvitavatest isikutest ja jälgida ainult neid
- Kasutada masinõppemeetodeid (räägime kohe)

Lahendusvariante

- Treenida keelemudel nimede “keeletuvastuse” jaoks
- Märgendada käsitsi hulk tekste ja loota, et saame konteksti põhjal masinale õpetada
- Leida kontekstide põhjal sarnaste sõnavektoritega sõnad nt word2veci vm kasutades

Keeleandmed ja masinõpe

Kontekst?

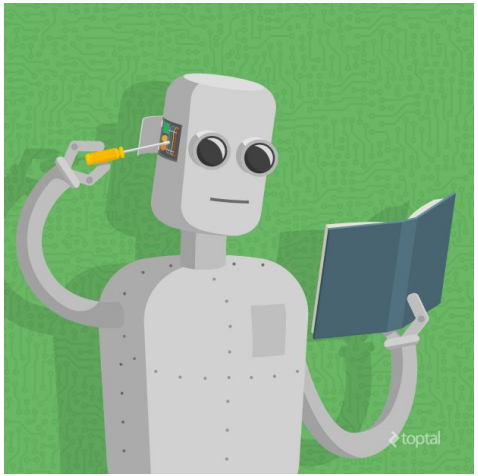
Reeglipõhised vs andmepõhised meetodid

Ei ole otseselt eesti keele spetsiifilised, aga rakendatavad eesti keelele sobiliku eeltöötluse korral

Kuidas?

Arvutid võivad ‘õppida’ keeli

Masinõpe



Inimesed võivad ‘tõlkida’ oma teadmuse programmeerimiskeelde

Reeglipõhised süsteemid

Andmed vs reeglid

I ___ to school. [go|goes]

Every day I go to school.

I go to school at 8am.

I go to school to learn new things.

...

```
go = ['I', 'you', 'we', 'they']
goes = ['He', 'She', 'It']
if word in go:
    gap = 'go'
elif word in goes:
    gap = 'goes'
else:
    gap = '____'
```

Reeglipõhised vs statistilised meetodid

Reeglipõhine: morfoloogiline analüüs

Kala|*de*|*le* - kala + mitmus + alaleütlev

Andmepõhine: morfoloogiline ühestamine

*Kohtusime metsarajal peesitava **maoga**.* -> madu/magu?

Masinõpe

Masinõpe - tehisintellekti valdkond, mille eesmärgiks on “õpetada” arvuteid võtma olemasolevate andmete põhjal vastu otsuseid, ilma et neid peaks selleks otseselt programmeerima. Masinõppes kasutatakse matemaatikat ja statistikat, et avastada mustreid treeningandmetest ja rakendada neid uutele, **sobival kujul** olevatele andmetele.

Masinõppe liigid

- Juhendatud õppimine (*supervised learning*)
  - Andmestikus on kirjas oodatav väljund
    - Õpime ennustama: nt rämpsposti filtreerimine, tekstiliikide tuvastus, keeletuvastus, autorituvastus, meelsusanalüüs
- Juhendamata õppimine (*unsupervised learning*)
  - Andmestikus pole kirjas oodatavat väljundit
    - Otsime andmestest struktuuri

Dokumentide automaatne liigitamine

- Automaatsed meetodid (masinõpe) kasutavad vektoreid (arvude järjendid) ja maatrikseid (arvude tabelid) => tekst vaja “teisendada” vastavale kujule

Dokumentide vektoresitus

dok1: ‘Koerad ja kassid ei salli üksteist.’
dok2: ‘Vanaema kudus eile salli.’
dok3: ‘Koerad ja kanad ei saa ühes puuris elatud.’

	koerad	ja	kassid	ei	...
dok1	1	1	1	1	
dok2	0	0	0	0	
dok3	1	1	0	1	
...					

TF-IDF skoor

- Vektoriteks ei pea olema sõnade esinemissagedused
- Laialdaselt kasutusel on TF-IDF skoor, mis näitab, kuivõrd iseloomulik on sõna mingile tekstile
  - TF - term frequency - sõna esinemissagedus tekstis
  - IDF - inverse document frequency - kajastab, kui paljudes korpuse dokumentides see sõna esineb
  - TF-IDF = TF*IDF

Dokumentide vektoresitus

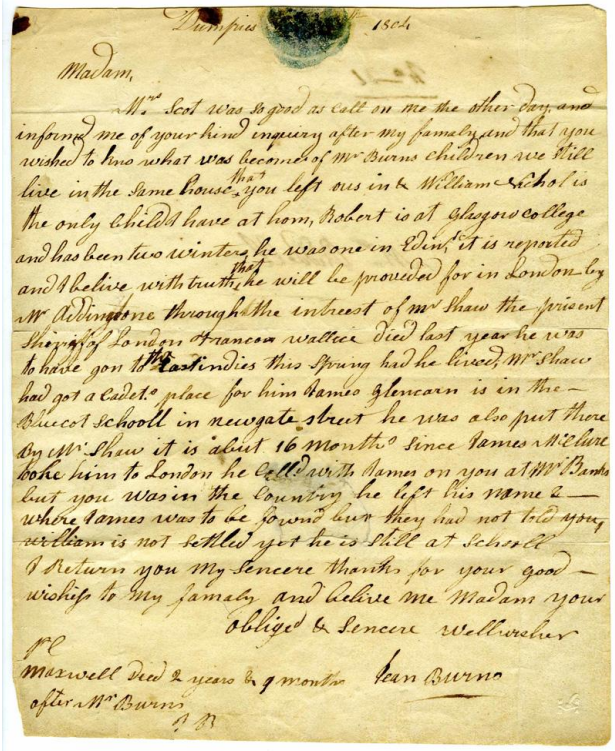
- Statistilised meetodid tuginevad vektorite sarnasusele ega võta arvesse tunnuseid, mis vektorites ei kajastu
  - Võime vastavalt oma soovile ka tunnuseid lisada

Dokumentide vektoresitus

dok1: ‘Koerad ja kassid ei salli üksteist.’
dok2: ‘Vanaema kudus eile salli.’
dok3: ‘Koerad ja kanad ei saa ühes puuris elatud.’

	koerad	ja	kassid	ei	[Liitsõnade arv]
dok1	1	1	1	1	1
dok2	0	0	0	0	1
dok3	1	1	0	1	0
...					

Feature engineering



Lemmik kirjavahemärk

Kirjutamise aeg

“M”-ga algavate lausete suhteline arv

Sõnavara keerukus

autor

ortograafia

Kirja värv

Kasutatud sõnad

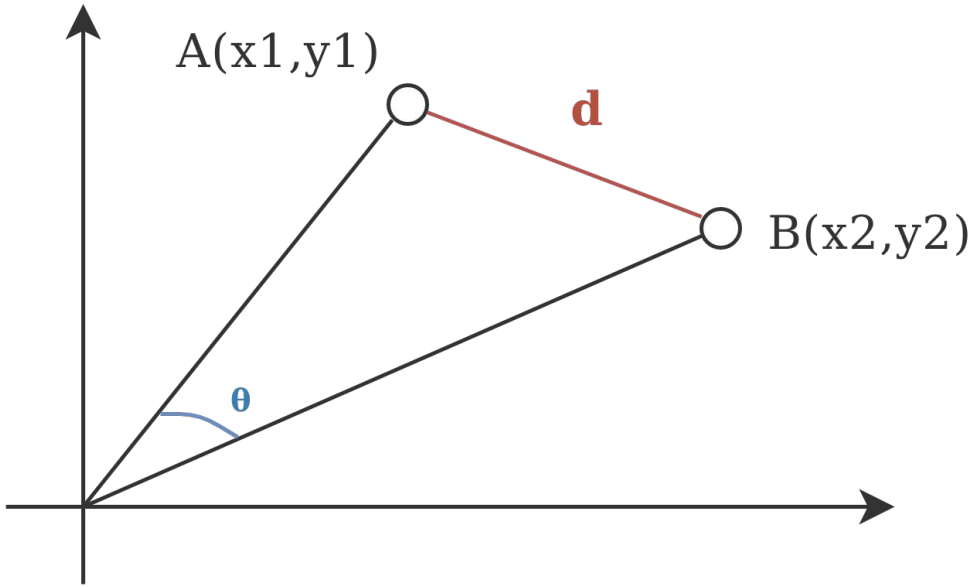
Keskmine sõna pikkus

Mainitud nimed

Teksti meelestatus

Dokumentide võrdlemine

- Eukleidese kaugus
- Koosinuskaugus



Dokumentide automaatne liigitamine

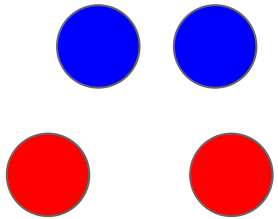
- Klassifitseerimine
  - Eeldefineeritud liigid, nt žanride, autorite vms kaupa, vajab (käsitsi märgendatud) treeningandmeid
- Klasterdamine
  - Tüüpiliselt pole teada võimalike gruppide arv või kuidas neid nimetada

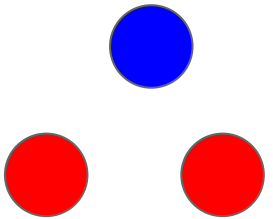
Masinõppe nõudmised andmetele

- Andmete hulk
  - Mida rohkem, seda parem. Mida keerulisem mudel, seda enam andmeid vajab.
- Andmete jaotus
  - Treeningandmed kirjeldama reaalsust
- Andmete kvaliteet
  - Vigaste andmetega saame vigase mudeli
- Tunnuste relevantsus sõltuva muutuja suhtes
  - Teose žanrit ei saa ennustada ainult autori abil

Andmete hulk ja jaotus

Väikese treeningandmestiku korral võime leida lihtsa kirjeldava mudeli.



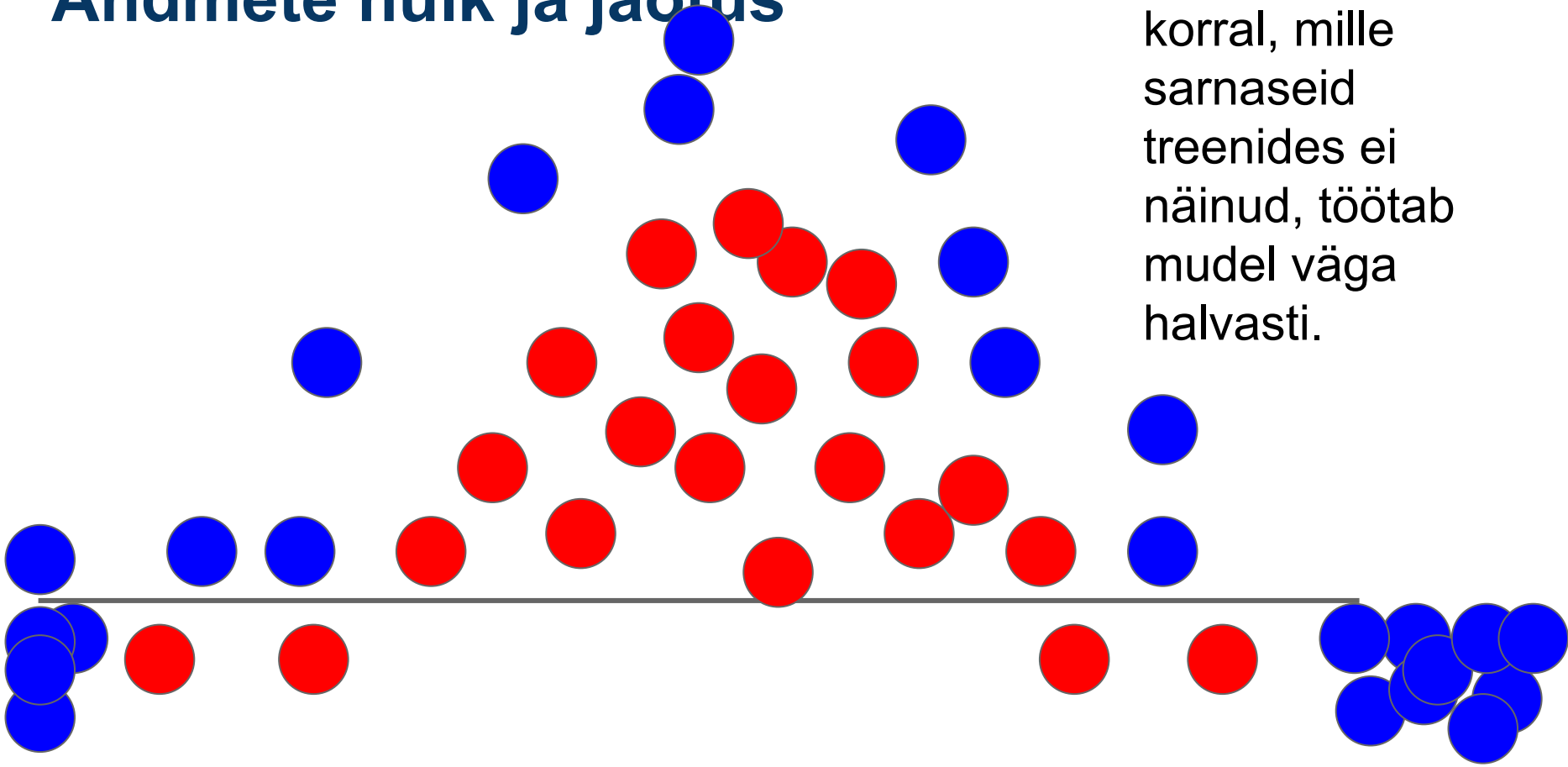


Andmete hulk ja jaotus

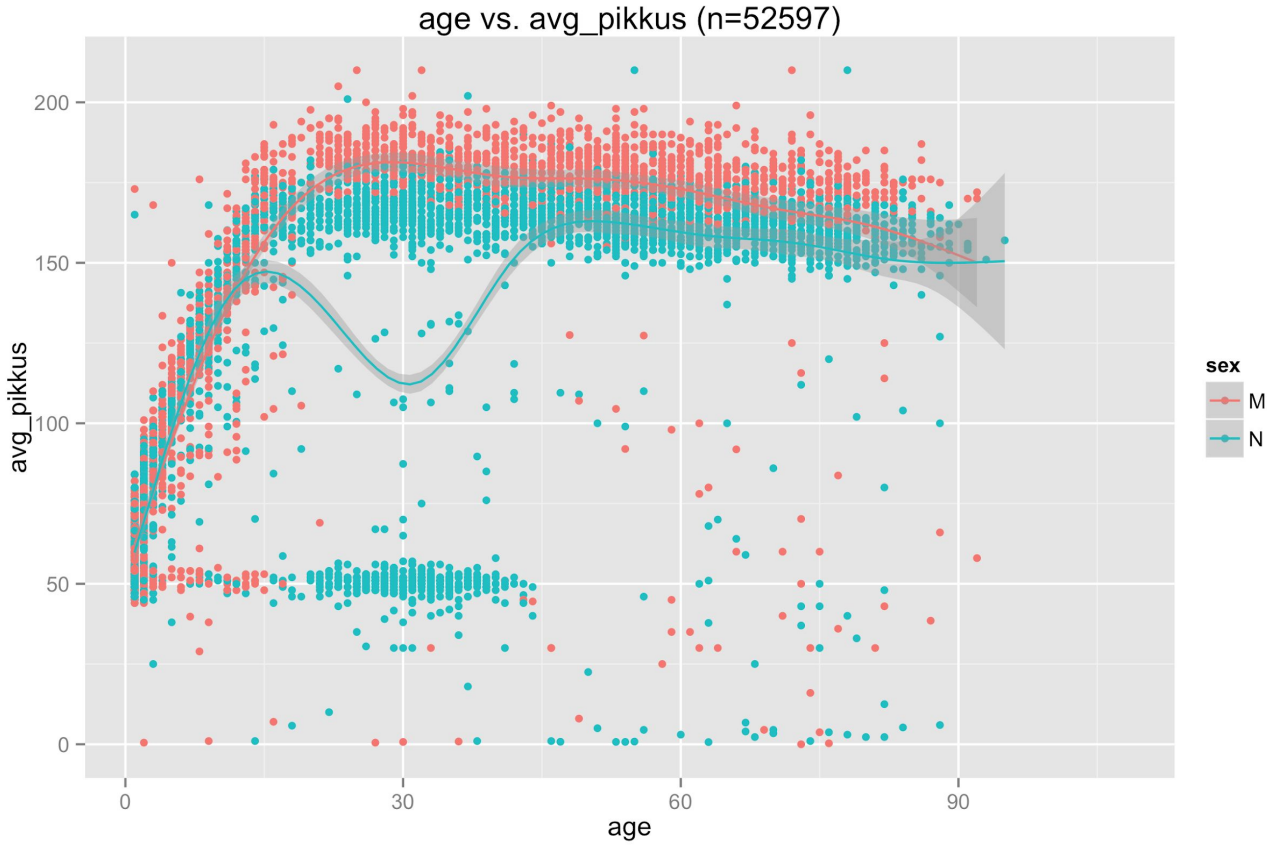
Väikese treeningandmestiku korral võime leida lihtsa kirjeldava mudeli.

Andmete hulk ja jaotus

Uute andmete korral, mille sarnaseid treenides ei näinud, töötab mudel väga halvasti.



Andmete kvaliteet



Klasterdamine (*Clustering*)

- automaatne objektide grupeerimine
- samas klastris lõpetavad objektid on üksteisele lähemal kui erinevates klastrites lõpetavad objektid
- hõlmab palju algoritme

Klasterdamine

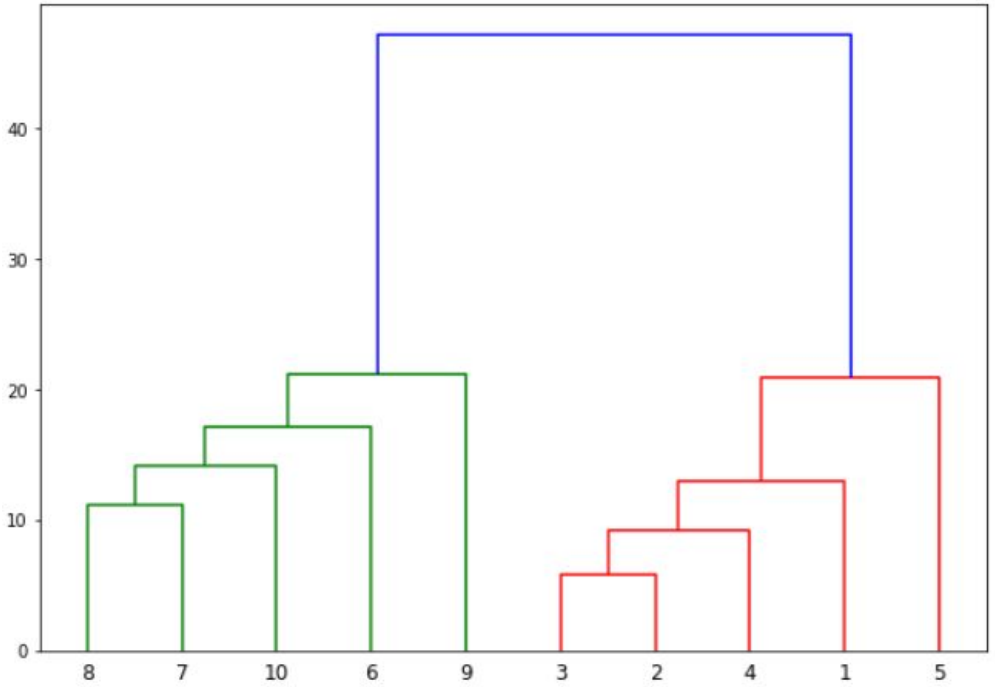
- võimaldab leida sarnaseid dokumente/sõnu
- ideaalne vahend andmestikuga tutvumiseks
- harva süsteemides ilma inimese sekkumiseta

Hierarhiline klasterdamine

- määrab ükshaaval lähimad klastrid/punktid ühte vanemkastrisse ning moodustab klastrite puu
- sageli keelekontekstis kasutusel
- võimaldab klasterdada erinevate omadustega andmeid, sõltuvalt parameetritest
- populaarne:
  - aglomeratiivne

Dokumentide klasterdamine

- Eesmärgiks moodustada korpuses olevatest dokumentidest grupid
- Hierarhiline klasterdamine annab ülevaate sarnasustest, ilma et peaksime teadma gruppide arvu - moodustatakse hierarhia



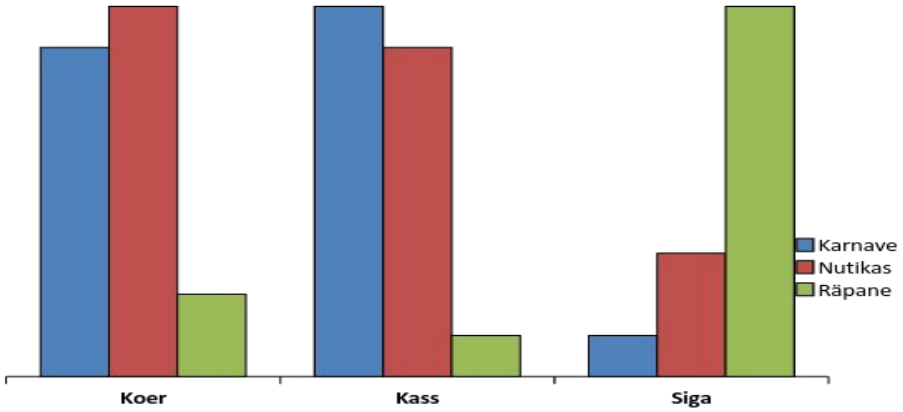
Teemade modelleerimine

- Eesmärk anda ülevaade dokumentidest
- Eeldus: erinevatest teemadest kirjutades kasutatakse mingil määral erinevaid sõnu
- Korpusest leitakse esmalt seal esinevad teemad ehk sõnade tõenäosuslik jaotumine. Seejärel võimalik leida igale dokumendile, milline teema seda kõige paremini kirjeldab
- Vaja teemade arv ette määrata

```
[(0, '0.010*"tulema" + 0.009*"meeldima" + 0.008*"hea" + 0.008*"töö" + 0.007*"palju" + 0.007*"või
ma" + 0.007*"aeg" + 0.007*"välja" + 0.006*"elu" + 0.006*"suutma" + 0.006*"ütlema" + 0.006*"teadm
a" + 0.005*"tundma" + 0.005*"uskuma" + 0.005*"otsus" + 0.005*"tahtma" + 0.004*"oluline" + 0.004
*"tunne" + 0.004*"plaan" + 0.004*"uus" + 0.004*"keegi" + 0.004*"jääma" + 0.004*"nägema" + 0.004
*"andma" + 0.004*"hästi" + 0.004*"hakkama" + 0.004*"tihti" + 0.004*"rohkem" + 0.004*"ette" + 0.0
04*"minema"'), (1, '0.010*"meeldima" + 0.010*"töö" + 0.009*"elu" + 0.008*"võima" + 0.008*"hea" +
0.008*"tulema" + 0.007*"teadma" + 0.007*"otsus" + 0.006*"palju" + 0.006*"püüdma" + 0.006*"välja"
```

Distributiivne semantika

- „You shall know a word by the company it keeps“ (John R. Firth 1957)
- Distributiivhüpotees: sarnaste distributsioonidega lingvistilised üksused omavad semantilist sarnasust



	karvane	nutikas	räpane
koer	98	99	10
kass	102	96	5
siga	5	22	81

Sõna kui objekt eukleidilises ruumis

	karvane	räpane
koer	98	10
kass	102	5
siga	5	81

- Sõnade sarnasusest võib mõelda kui vastavate objektide vahelisest kaugusest
- St kass ja koer on omavahel sarnasemad kui kass (või koer) ja siga, kuna asuvad ruumis üksteisele lähemal

Distributiivne semantika - proovime ise järgi