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?

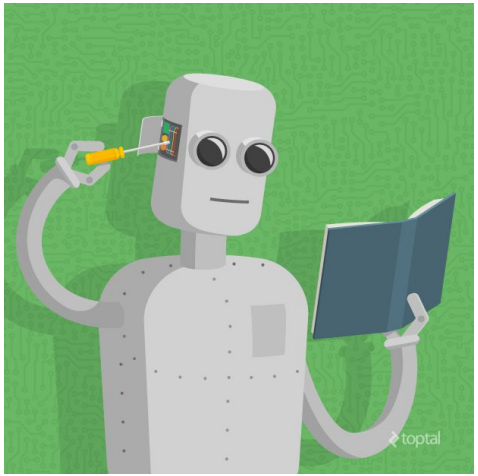
Reeglipõhised vs andmepõhised meetodid

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

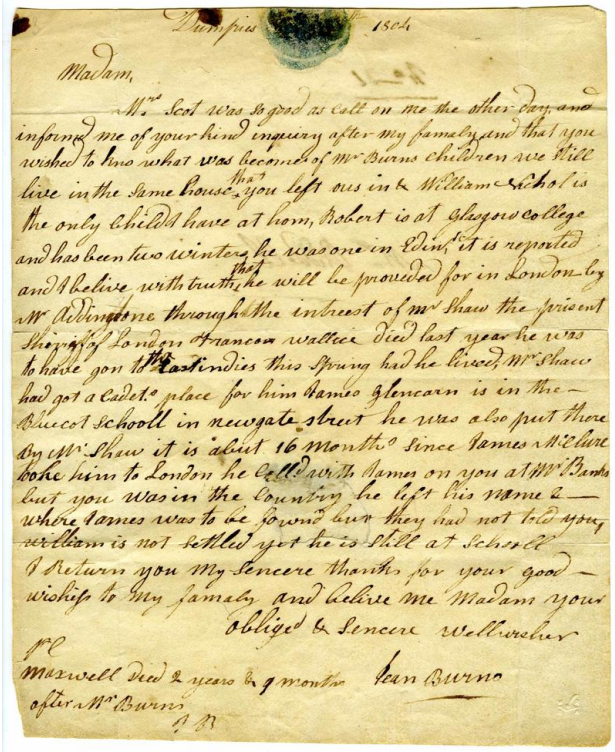
- Statistilised meetodid tuginevad vektorite sarnasusele ega võta arvesse tunnuseid, mis vektorites ei kajastu
  - Võime vastavalt oma soovile ka tunnuseid lisada

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

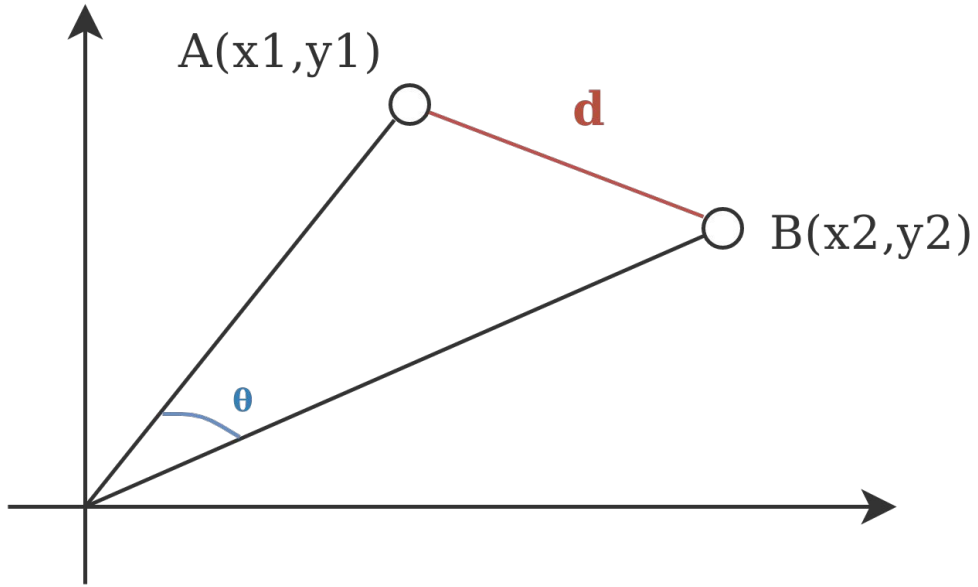
Keskmine sõna pikkus

Mainitud nimed

Teksti meelestatus

Dokumentide võrdlemine

- Eukleidese kaugus
- Koosinuskaugus



Dokumentide automaatne liigitamine

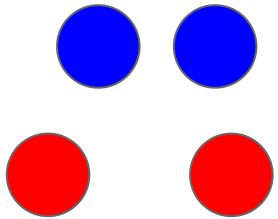
- Klassifitseerimine
  - Eeldefineeritud liigid, nt žanride, autorite vms kaupa, vajab (käsitsi märgendatud) treeningandmeid
- Klasterdamine
  - Tüüpiliselt pole teada võimalike gruppide arv või kuidas neid nimetada

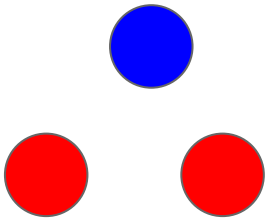
Masinõppe nõudmised andmetele

- 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.

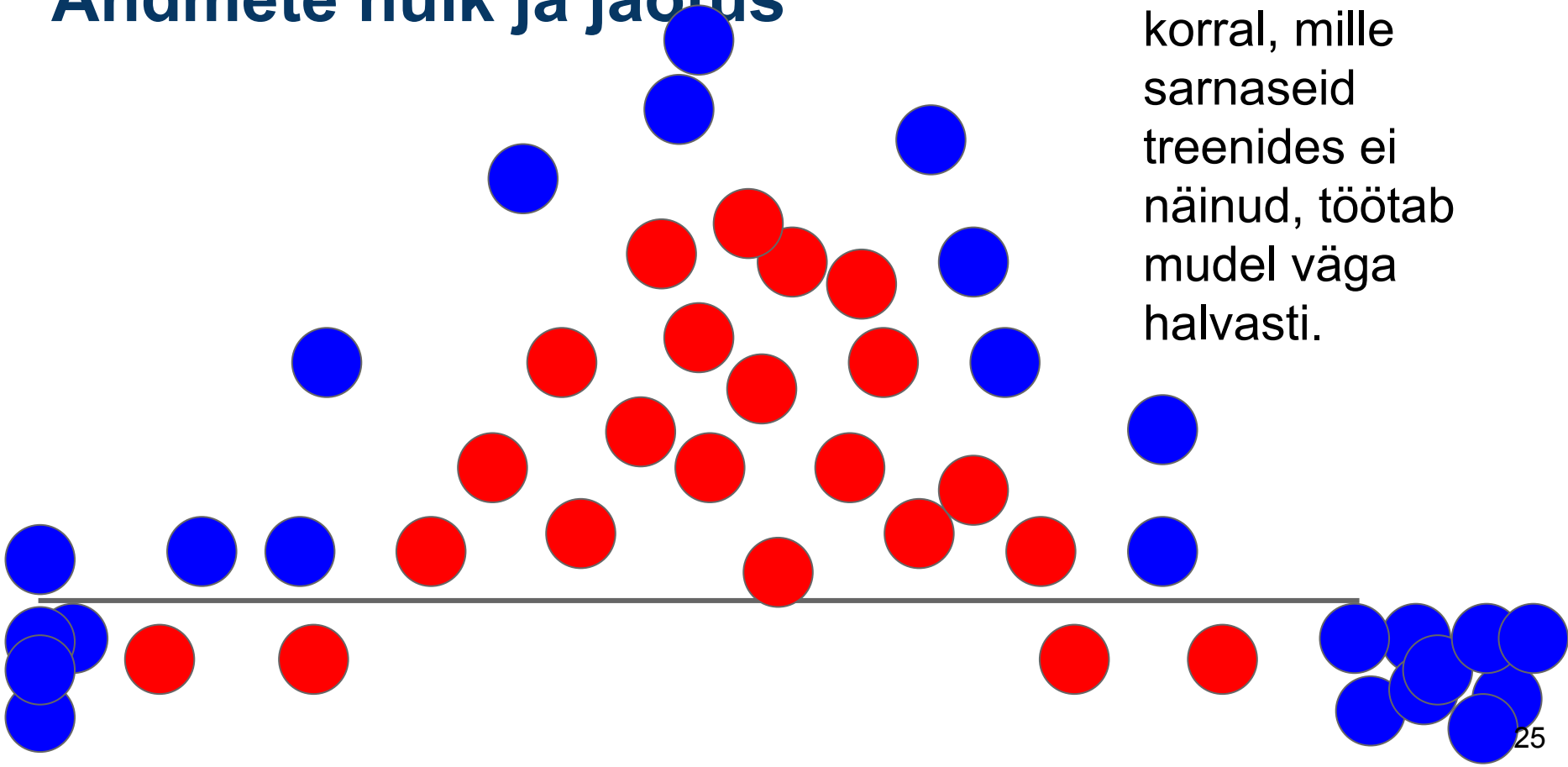




Andmete hulk ja jaotus

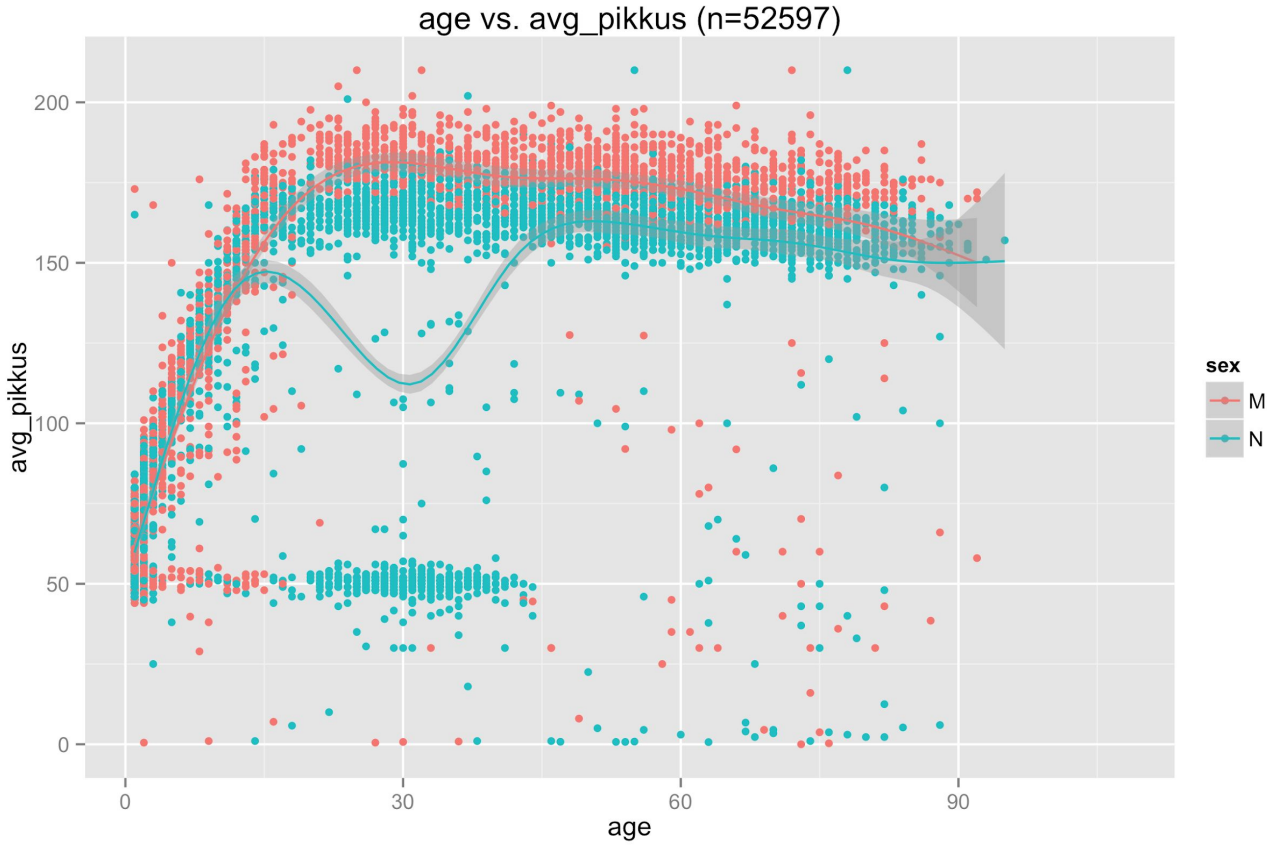
Väikese treeningandmestiku korral võime leida lihtsa kirjeldava mudeli.

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

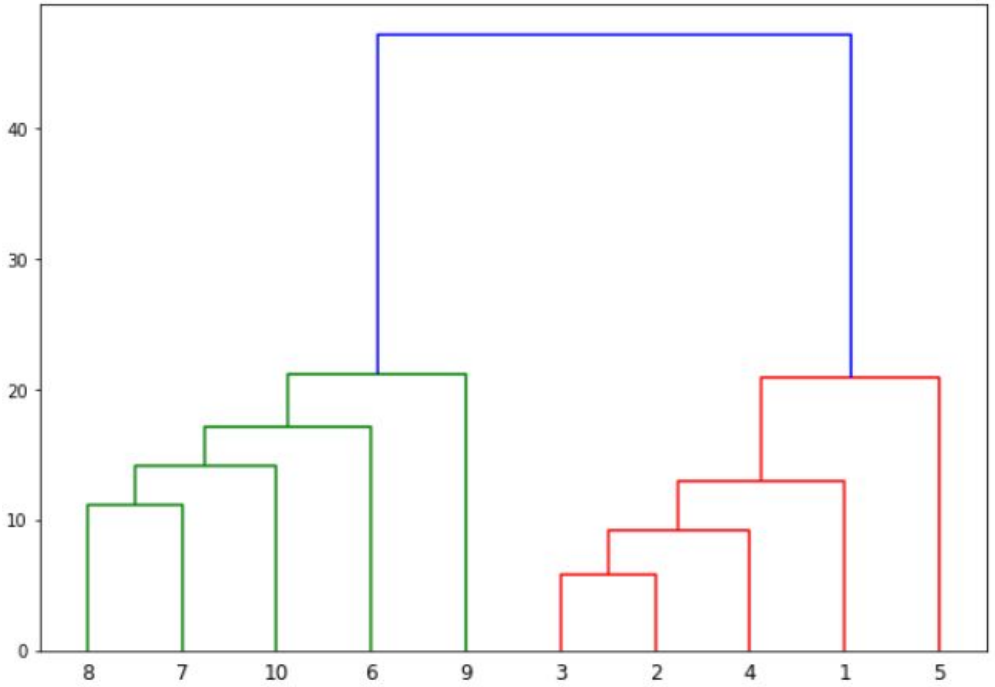
- võimaldab leida sarnaseid dokumente/sõnu
- ideaalne vahend andmestikuga tutvumiseks
- harva süsteemides ilma inimese sekkumiseta

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



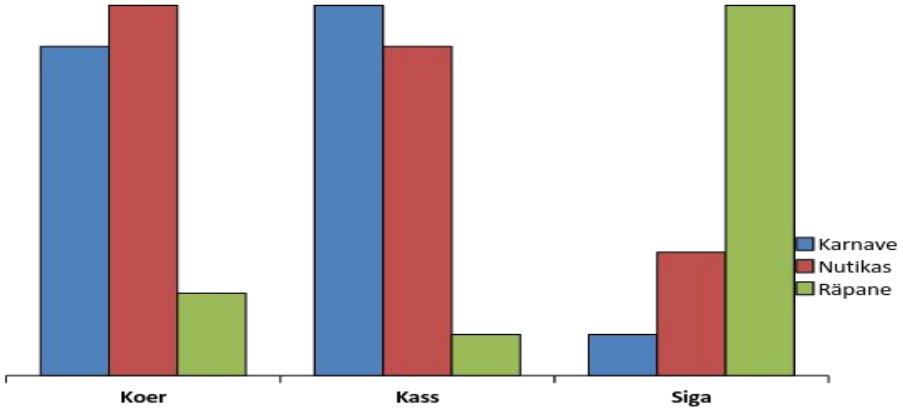
Teemade modelleerimine

- 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