* <voyant-tools.org> OS-työkalu, yleisesti käytetty digitaalisen humanismin projekteissa
* selitä mitä korpus tarkoittaa
* <https://hfroehli.ch/2014/05/11/intro-bibliography-corpus-linguistics/>
* [https://agricolaverkko.fi/tietosanomat/milta-ursula-von-der-leyenin-puhe-nayttaa-etaalta-luettuna/](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fagricolaverkko.fi%2Ftietosanomat%2Fmilta-ursula-von-der-leyenin-puhe-nayttaa-etaalta-luettuna%2F&data=04%7C01%7C%7C7d9dde0327c14931313e08d979b0dedd%7C87879f2e73044bf2baf263e7f83f3c34%7C0%7C0%7C637674622830177324%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=PTyX69katJMHd7dkNiNCPtqeitHWGUKhLcdvqtI8XoU%3D&reserved=0)
* <https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html>

# Demo 1

Vertaillaan kahta 1800-luvun lopun henkikirjoitusasetusta. Nämä löytyvät esimerkiksi Kotuksen [sivuilta](https://kaino.kotus.fi/korpus/1800/meta/asetus/asetus_coll_rdf.xml).

1. Kopioi vuoden 1878 asetuksen teksti Voyantiin.
2. Kopioi myös vuoden 1894 asetusten teksti (**Documents** > **Modify** > **Add**)
3. Lisää poistettavien sanojen lista (<https://github.com/stopwords-iso/stopwords-fi/blob/master/stopwords-fi.txt>).
4. Lisää *Summary*-välilehti. Mitä *Distinctive words* -kohta kertoo tekstien eroista?

# Tutkimusdatan organisointi

Data tiedostossa: mieti, onko koneen luettavissa -> silloin yleensä myös saavutettavaa ja helpommin muunnettavissa formaatista toiseen

avoimet ja ”agnostiset” (platform agnostic) formaatit suositeltavia

esimerkki: utf-8 [[video utf-8](https://www.youtube.com/watch?v=MijmeoH9LT4)] ja mojibake (korjattavissa pythonilla: [ftfy](https://ftfy.readthedocs.io/en/latest/index.html))

jos data pelkkää numeroita tai tekstiä: suositeltavaa käyttää tekstipohjaista tiedostomuotoa (esim. md, txt, csv, tsv), ei binääristä tiedostomuotoa

Siisti (tidy) data:

1. Jokainen sarake on muuttuja
2. Jokainen rivi on havainto
3. Jokaisessa solussa on yksi arvo

Kansiorakenne on mietittävä huolellisesti

Kansioiden ja tiedostojen looginen ja informatiivinen organisointi, käytä hierarkiaa hyväksesi

/artikkelit/2021/Tohmajarven\_historiaa/data/raaka/1991-06-02\_Punkalaidun.csv

tarkemmin eroteltu tai käsitelty data nimetään samalla tyylillä, erottava tieto lisätään loppuun

/artikkelit/2021/Tohmajarven\_historiaa/data/raaka/1991-06-02\_Punkalaidun\_raty.csv

[ISO-standardin](https://fi.wikipedia.org/wiki/ISO_8601) 8601 mukainen pvm: vvvv-kk-pp (jos joku haluaa välttämättä käyttää tätä standardia muinaisiin asioihin, niin vuosi 1 eKr. on 0000, vuosi 2 eKr. -0001 jne.)

pvm alussa, vuosi-kk-pp -> tietokone laittaa tiedostot automaattisesti kronologiseen järjestykseen

|  |  |
| --- | --- |
| camel case | suurtenMuinaistenAika |
| snake case | suurten\_muinaisten\_aika |
| kebab case | suurten-muinaisten-aika |

# Powershell

UNIX-filosofia: yksi ohjelma tekee yhden asian hyvin

komentorivi on raivaussaha, tekstieditori on oksasakset tai jotain

komentoriviputki

navigointi: ~, ., ..

putkioperaattori |

uudelleenohjausoperaattori, > ja >>

cp ./tiedosto/polku/teksti1.txt ./toinen/tiedosto/polku/teksti2.txt

paetaan ctrl-c

iterointi: for loop

rekursio: hyödyllistä mutta vaarallista

|  |  |  |
| --- | --- | --- |
| Pitkä nimi | Lyhenne | UNIX-vastine (karkeasti) |
| Get-Help | help | man |
| Set-Location | sl | cd/chdir |
| Get-Location | gl | pwd |
| Get-Content | gc | cat |
| Get-ChildItem | gci | dir/ls |
| Rename-Item | rni | mv |
| Copy-Item | copy, cpi | cp |
| New-Item | ni | mkdir |
| Write-Output | write | echo |
| Select-String | sls | grep |
| For-Each | foreach | for … in |
| Measure-Object | measure | grep/wc |
| Select-Object | select |  |

# Demo 2

pwd

echo ”Saksa on paska maa” > raty\_1.txt

cat raty\_1.txt

cat raty\_1.txt > raty\_2.txt

cat raty\_2.txt

echo “Caterpillar” > raty\_2.txt

echo “Ooppera?” > raty\_3.txt

cat raty\_3.txt

echo ”Hyi saatana!” >> raty\_3.txt

cat raty\_3.txt

cat raty\_1.txt, raty\_2.txt, raty\_3.txt

cat raty\_\*.txt > raty\_kaikki.txt

(systemaattinen nimeäminen kannattaa!)

# Demo 3

cat raty\_1.txt | measure

cat raty\_1.txt | measure -l -w -c -ig

cat raty\_1.txt, raty\_2.txt | measure -l -w -c -ig

cat raty\_\*.txt | measure -l -w -c -ig

ls | foreach {cat $\_ | measure -l -w -c -ig}

# Demo 4

Lataa riksdagtryckistä säätyvaltiopäivien talonpoikaissäädyn valtiopäiväaineisto ja pura se kansion *data* sisään.

sls -Path .\data\bondeståndet\\*.txt -Pattern “finland” | echo

sls -Path .\data\bondeståndet\\*1720\*.txt -Pattern “finland” > finland\_bonde\_1720-1727.txt

cat finland\_bonde\_1720-1727.txt

sls -Path .\data\bondeståndet\\*.txt -Pattern "landtdag" | select Filename, LineNumber, Line, Path > lantdag.csv

Import-Csv lantdag.csv

# Demo 5

Lisää tilisi ympäristömuuttujiin MALLET\_HOME ja aseta sen arvoksi mallet-kansion sijainti

cd ~\mallet-2.0.8

bin\mallet import-dir --input sample-data\web\en --output tutorial.mallet --keep-sequence --remove-stopwords

bin\mallet train-topics --input tutorial.mallet --num-topics 20 --optimize-interval 20 --output-state topic-state.gz --output-topic-keys tutorial\_keys.txt --output-doc-topics tutorial\_composition.txt

# Demo 6

cd ~\mallet-2.0.8

bin\mallet import-dir --input runot --output kalevala --keep-sequence --remove-stopwords --stoplist-file stoplists\fi.txt

bin\mallet train-topics --input kalevala --num-topics 20 --optimize-interval 20 --output-state runot\_state.gz --output-topic-keys runot\_keys.txt --output-doc-topics runot\_composition.txt

# Harjoitus 1

Etsi [Gutenberg-projektin](https://gutenberg.org/) sivulta tekstimuodossa oleva kirja tai pamfletti. Analysoi sitä [Voyant](https://voyant-tools.org/)illa. Voit käyttää myös lyhyempää tekstinpätkää, jos kokonainen teksti on liian raskas. Katso apua [ohjeista](https://voyant-tools.org/docs/#!/guide/start).

# Harjoitus 2

Etsi talonpoikaissäädyn valtiopäiväaineistosta mainintoja Karjalasta.

# Harjoitus 3

Analysoi itse valitsemaasi tekstiaineistoa Malletilla.