# 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?

# 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