Machine learning, data mining, and the world's largest manuscript traditions

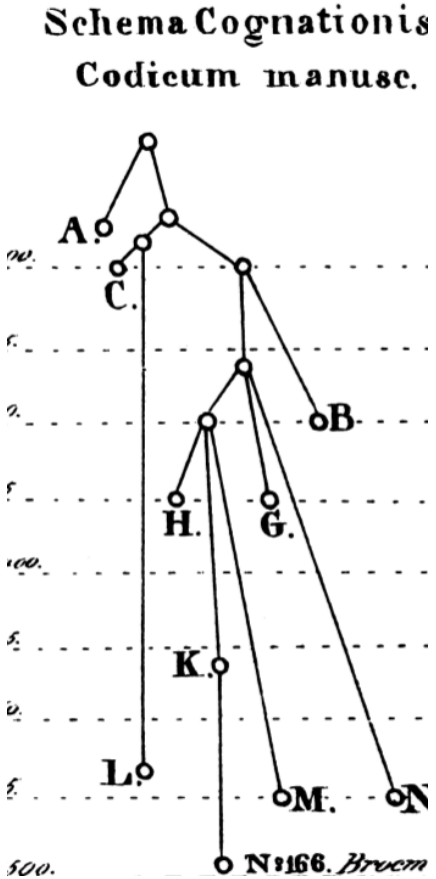
Pasi Hyytiäinen, Classical Texts in Digital Media II, Venice, 18.6.2025

In this presentation

- Stemmatology?
- Challenges
- Computer-assistance…
- The world’s largest manuscript traditions
  - Taking all data into account
    - Machine learning
    - Data mining

Stemmatology

- The classical texts were transmitted through the medium of manuscripts
  - There are different versions of a given literary work
- Stemmatology uses tree (or network) structures to depict the genealogical relationships between manuscripts
- The goal is to uncover the starting point of the manuscript tradition, the initial or the original text



Challenges

- Efficiency
  - Traditional Lachmannian stemmatology
    - Extremely time-consuming
      - Transcribing, collating, and dividing texts into variation units

A	the fox jumped over the hedge
B	-
C	the cat jumped over the fence
D	a man saw that the fox jumped over the hedge
E	a man saw that the fox jumped over the fence

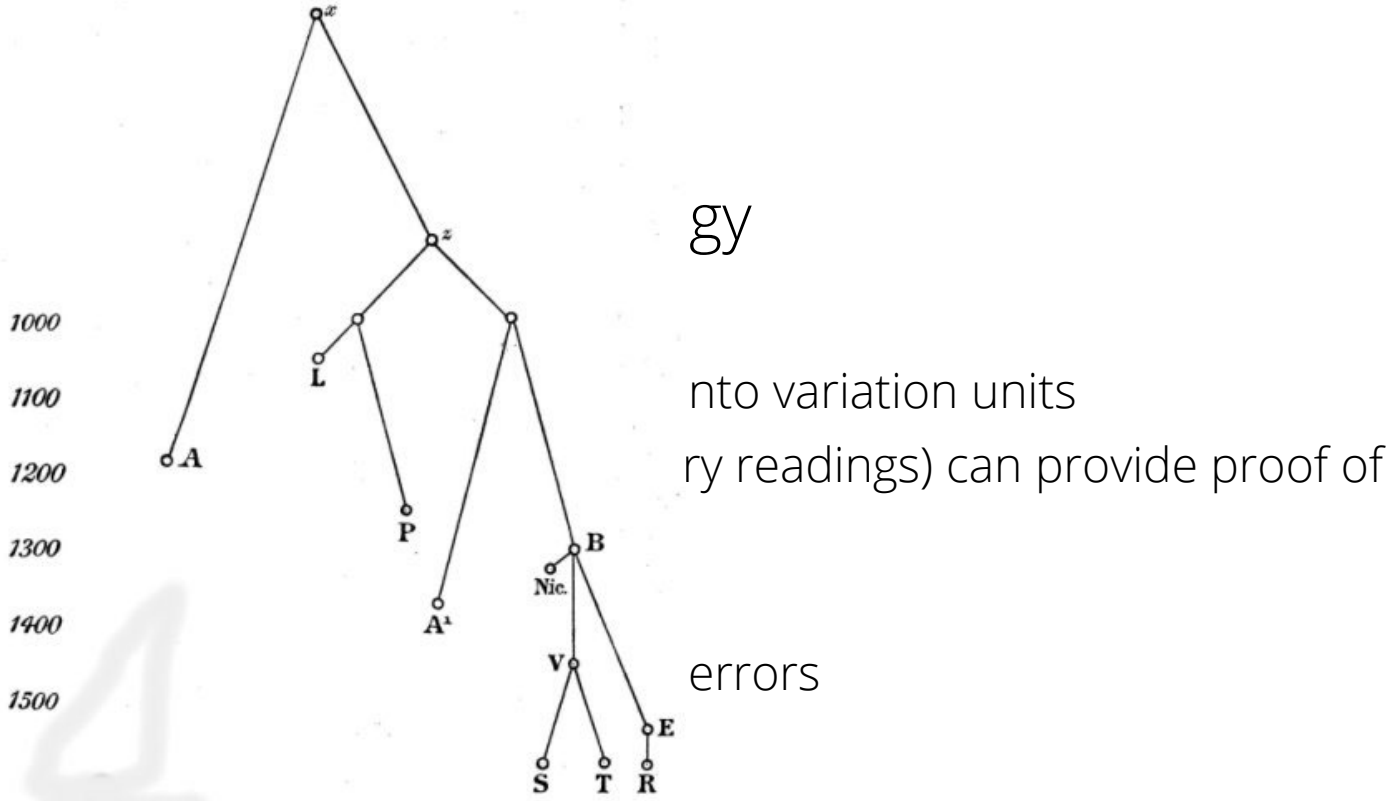
Solution 1		Solution 2		
–	the fox jumped over the hedge	–	the fox jumped	over the hedge
–	–	–	–	–
–	the cat jumped over the fence	–	the cat jumped	over the fence
a man saw that	the fox jumped over the hedge	a man saw that	the fox jumped	over the hedge
a man saw that	the fox jumped over the fence	a man saw that	the fox jumped	over the fence

Challenges

- Efficiency
  - Traditional Lachmannian stemmatology
    - Extremely time-consuming
      - Transcribing, collating, and dividing texts into variation units
      - Only shared errors (innovations, secondary readings) can provide proof of common ancestry
      - Distinguishing monogenetic variants
      - Constructing stemmata based on shared errors
        - » Stemma codicum

Challenges

- Efficiency
  - Traditional La gy
    - Extremely t
      - Transcrib nto variation units
      - Only sha ry readings) can provide proof of commor
      - Distingui
      - Construc errors
        - » Stem

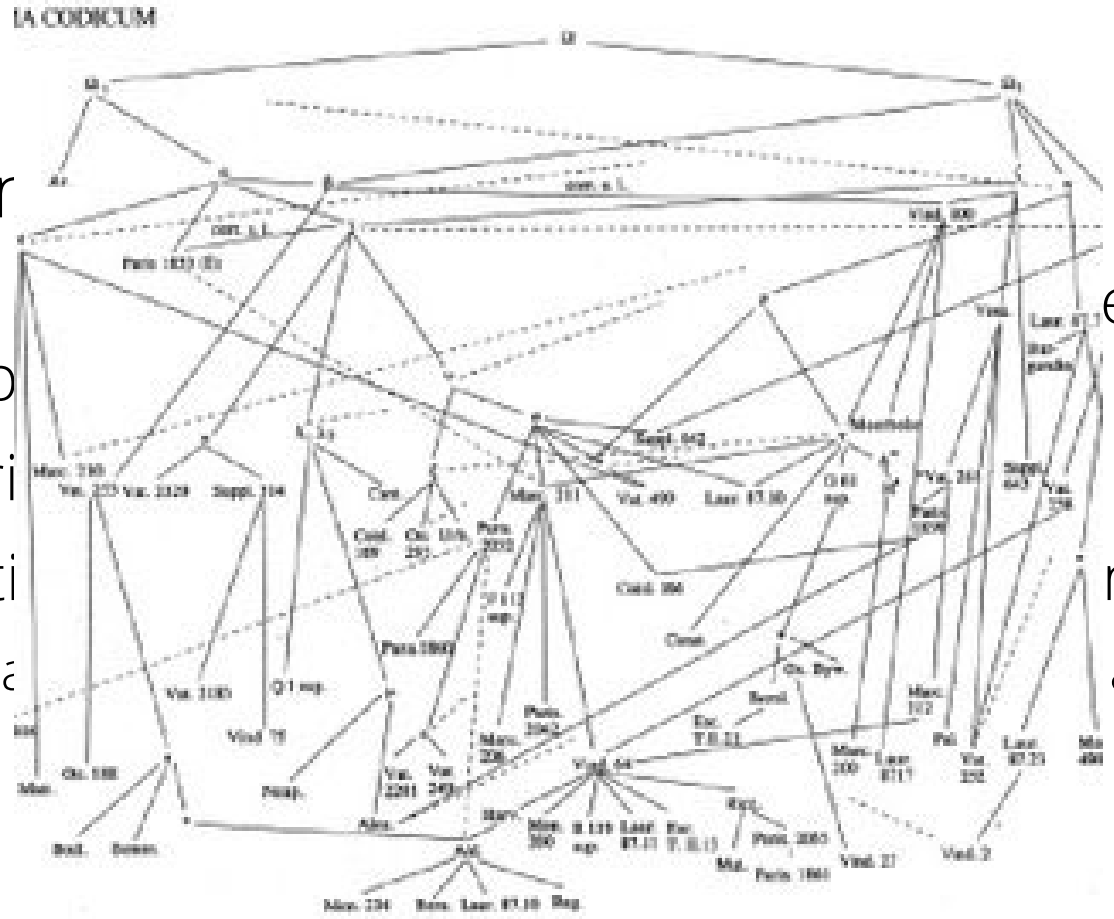


Challenges

- Contamination
  - “where contamination exists, the science of stemmatics in the strict sense breaks down.” (Paul Maas 1880-1964)
  - The scribes used more than one exemplar
  - A traditional stemma cannot depict a contaminated tradition
    - Contamination mixes the manuscript histories and ancestral lineages

Challer

- Contamir
  - “where emmatics in the strict sense b
  - The scri
  - A traditi nated tradition
    - Conta ancestral lineages

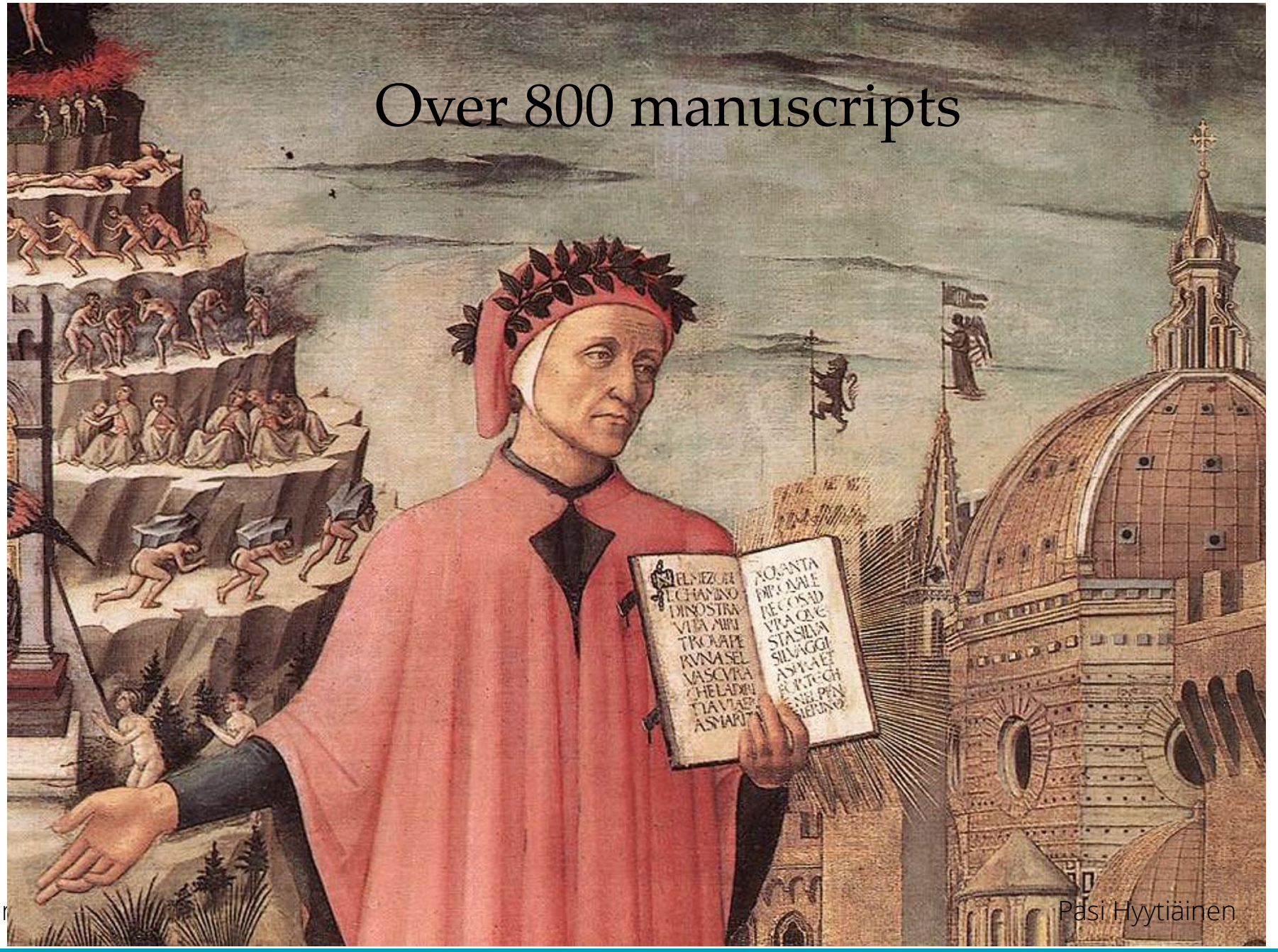


Computer-assistance…

- Digital methods have been used to study manuscript relations from the late 1960s (Griffith 1969)
- Resolving some of the challenges
  - Collations can be automated (Collate)
  - Trees can be automatically generated (phylogenetic methods)
    - Peter Robinson, Chris Howe etc.
- We need new methods in the face of large manuscript traditions

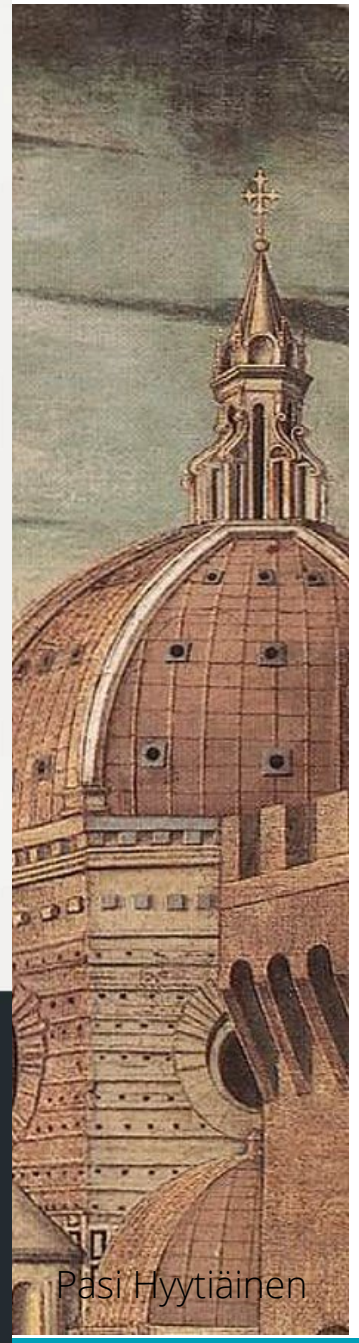
The world’s largest manuscript traditions

Over 800 manuscripts

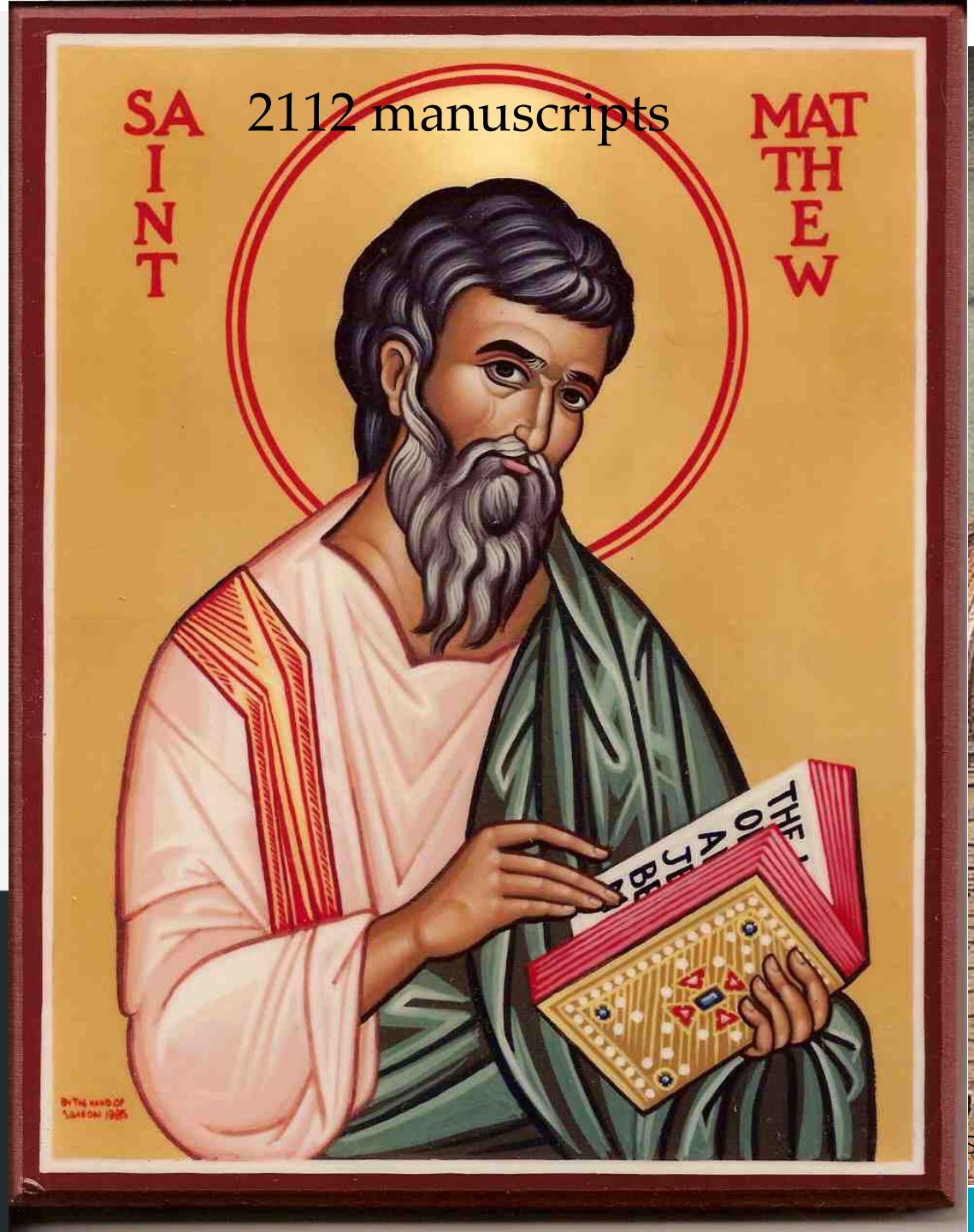


Over 2000 manuscripts

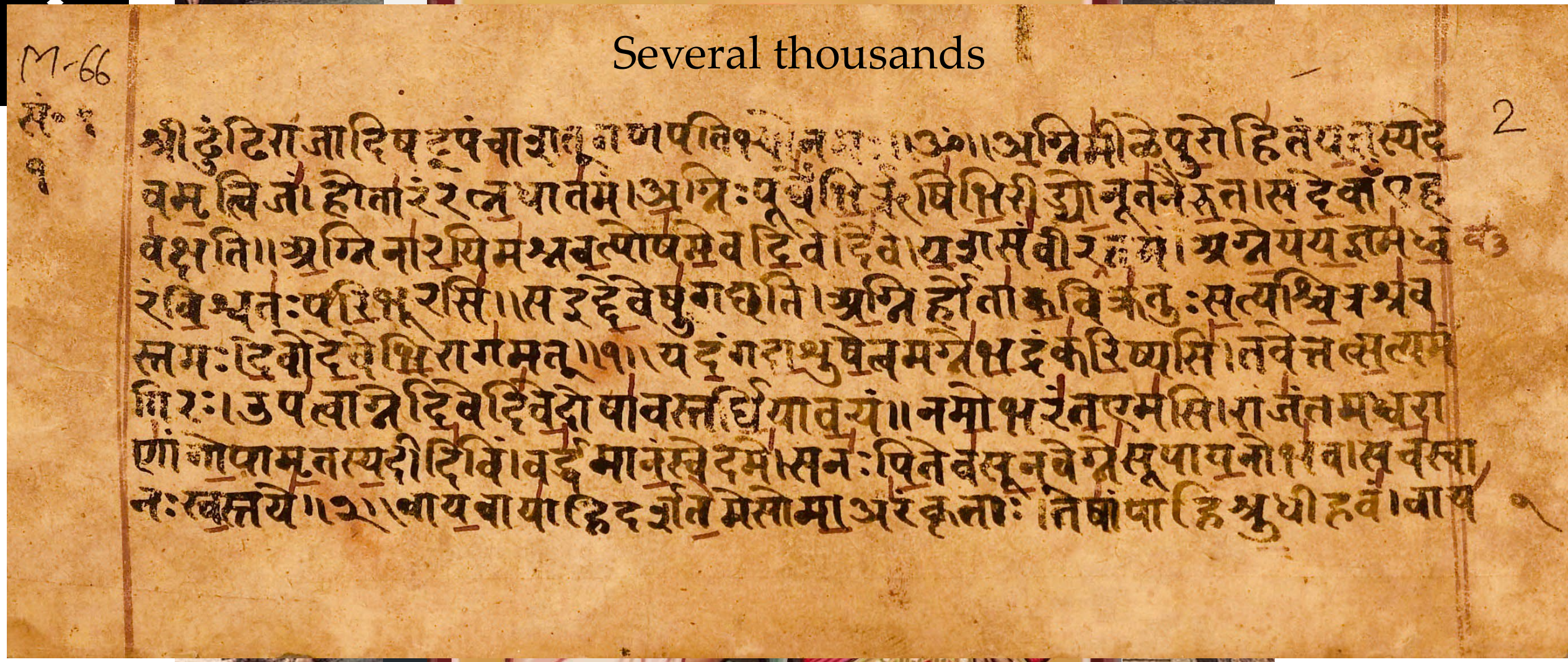
HOMER



2112 manuscripts



Several thousands



The world's largest manuscript traditions

- How do we handle these large traditions?
  - The ultimate goal is to take all data into account
- First step
  - Digital photography
    - All manuscripts need to be digitally photographed

The world's largest manuscript traditions

- Second step
  - Transcribing the texts into digital form
  - This step must be automated
    - Machine learning (and AI)
      - We generate a model for a given tradition (Latin, Greek, Church Slavonic…)
        - » We teach a program to differentiate text regions, lines, words, and letters from a photo
          - » Transkribus software

Valamon luostar... / XII.363 / p001.jpg   Ground Truth   Process with AI   0 unsaved changes   Save

Молитвы ѹ҆́тренніѧ.

Сла́ва тебѣ̀ бж҃е на́шъ, сла́ва тебѣ̀.
Цр҃ю̀ нбⷭ҇ный, ѹ҆тѣ́шителю, дш҃е и҆́стины,
и҆́же везⷣѣ сы́й, и҆ всѧ̑ и҆сполнѧ́ѧй, со_
кро́вище бл҃гъ, и҆ жи́зни пода́телю,
пріидѝ и҆ всели́сѧ въ ны̀, и҆ ѡ҆чи́сти ны̀
ѿ всѧ́кїѧ скве́рны, и҆ спасѝ бл҃же дꙋ_
шы̀ на́шѧ.
Ст҃ы́й бж҃е, ст҃ы́й крѣ́пкїй, ст҃ы́й безс_
ме́ртный, поми́лꙋй на́съ. Три́жды.
Сла́ва Ѻ҆ц҃ꙋ̀, и҆ Сн҃ꙋ, и҆ ст҃о́мꙋ Дх҃ꙋ, и҆
ны́нѣ и҆ при́снѡ, и҆ во вѣ́ки вѣкѡ́въ,
а҆ми́нь.
Прест҃а́ѧ Тр҃це поми́лꙋй на́съ: Гдⷭ҇и
ѡ҆чи́сти грѣхѝ на́шѧ: Вл҃ко простѝ
беззако́нїѧ на́шѧ; Ст҃ы́й посѣтѝ
и҆ и҆сцѣлѝ немѡ́щи на́шѧ, и҆́мене
твоегѡ̀ ра́ди.
Гдⷭ҇и поми́лꙋй, три́жды: Сла́ва: и҆ ны́нѣ:
Ѻ҆́ч҃е на́шъ, и҆́же є҆сѝ на нб҃сѣ́хъ, да
ст҃и́тсѧ

а҃

Молитвы ѹ҆́тренніѧ.

и҆́мѧ твоѐ: да прі́идетъ црⷭ҇твїе твоѐ:
да бꙋ́детъ во́лѧ твоѧ̀, ꙗ҆́кѡ на нб҃сѝ,
и҆ на землѝ. Хлѣ́бъ на́шъ насꙋ́щный
да́ждь на́мъ дне́сь. и҆ ѡ҆ста́ви на́мъ
до́лги на́шѧ, ꙗ҆́коже и҆ мы̀ ѡ҆ставлѧ́емъ
должникѡ́мъ на́шымъ. И҆ не введи́ на́съ
во и҆скꙋше́нїе, но и҆зба́ви на́съ ѿ лꙋка́вагѡ.
Та́же: Тро́ичны ст҃а́ѧ: Тропа́рь, гла́съ а҃:
Воста́вше ѿ снà припа́даемъ тѝ бл҃же,
и҆ а҆́гг҃льскꙋю пѣ́снь вопі́емъ тѝ си́льне:
ст҃ъ, ст҃ъ, ст҃ъ є҆сѝ бж҃е, богоро́дицею
поми́лꙋй на́съ. Сла́ва:
Ѿ о҆́дра и҆ снà воздвиглъ мѧ̀ є҆сѝ гдⷭ҇и,
ѹ҆́мъ мо́й просвѣтѝ и҆ се́рдце, и҆ ѹ҆стнѣ̀
моѝ ѿ ве́рзи, во є҆́же пѣ́ти тѧ̀ ст҃а́ѧ_
тро́ице: ст҃ъ, ст҃ъ, ст҃ъ є҆сѝ бж҃е, бцⷣею
поми́лꙋй на́съ. И҆ ны́нѣ.
Напра́снѡ сꙋ́дїѧ прі́идетъ, и҆ коегѡ_
ждо дѣѧ́нїѧ ѡ҆бнажа́тсѧ, но стра́хомъ

Region 1

Мол́итвы оу҆́треннїѧ.

Сла́ва тебѣ̀ бж҃е на́шъ, сла́ва тебѣ̀.

Цр҃ю̀ нбⷭ҇ный, ѹ҆тѣ́шителю, дш҃е и҆́стины,

и҆́же вездѣ̀ сы́й, и҆ всѧ̑ и҆сполнѧ́ѧй, со_

кро́вище бл҃гихъ, и҆ жи́зни пода́телю,

прі́иди̇̀ и҆ всели́сѧ въ ны̀, и҆ ѡ҆чи́сти ны̀

ѿ всѧ́кїѧ скве́рны, и҆ спасѝ бл҃же дꙋ_

шы̀ на́шѧ.

Ст҃ы́й бж҃е, ст҃ы́й крѣ́пкїй, ст҃ы́й безс_

ме́ртный, поми́лꙋй на́съ. Три́жды.

Сла́ва Ѻ҆ц҃ꙋ̀, и҆ Сн҃ꙋ, и҆ ст҃о́мꙋ Дх҃ꙋ и҆

ны́нѣ и҆ при́снѡ, и҆ во вѣ́ки вѣкѡ́въ,

а҆ми́нь.

Прест҃а́ѧ Тр҃це поми́лꙋй на́съ: Гдⷭ҇и

ѡ҆чи́сти грѣхѝ на́шѧ: Вл҃ко простѝ

беззакѡ́нїѧ на́шѧ; Ст҃ы́й посѣтѝ

и҆ и҆сцѣлѝ не́мощи на́шѧ, и҆́мене

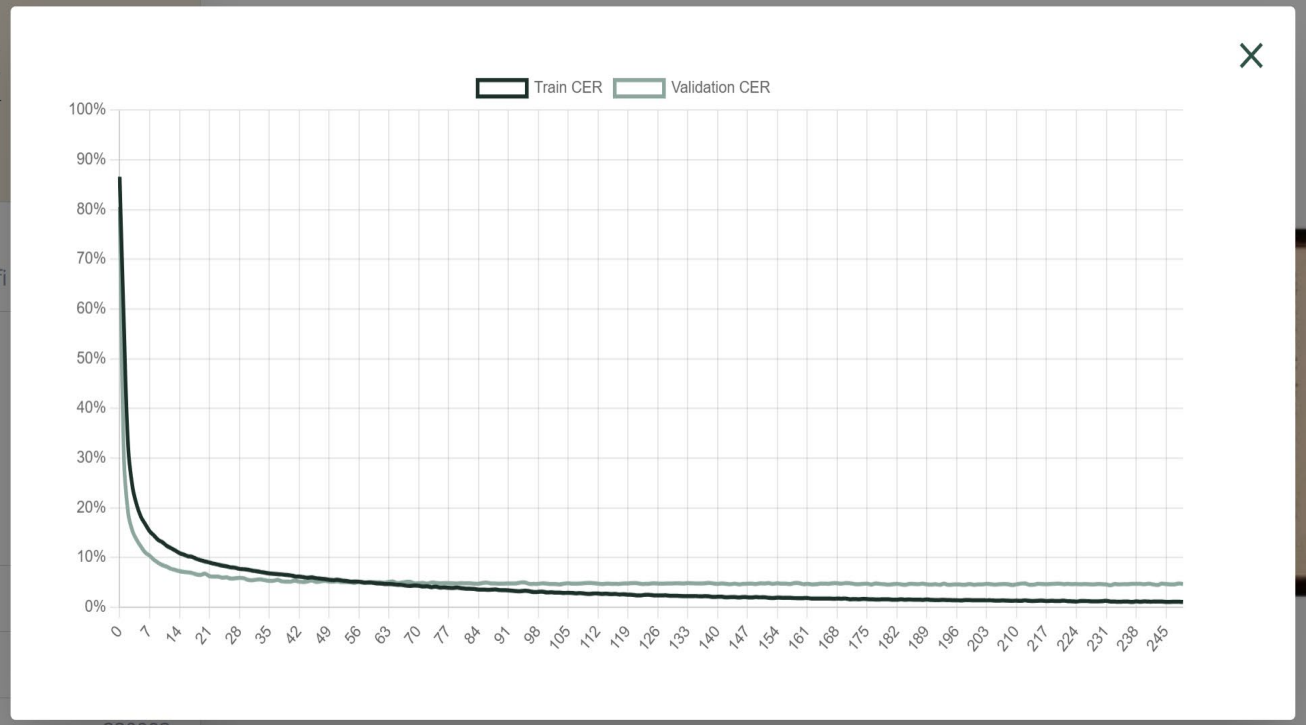
твоегѡ̀ ра́ди.

Гдⷭ҇и поми́лꙋй, три́жды: Сла́ва: и҆ ны́нѣ:

Valamo 0.7

Options

Model description



by pasi.hyytiainen@valamo.fi

Training Set Size

Training pages

Validation pages

Words

Lines

CER (Accuracy)

Centuries

Model ID 339993

See training chart

See training parameters

View all

AI Model: Valamo 0.7

Drop files here or browse files

The world's largest manuscript traditions

- Third step
  - Depicting the relations between the texts using distance values
  - Data mining techniques
    - No need to collate and divide texts into variation units
    - We calculate distances between pairs of texts
    - Introducing a very simple and fast method
      - K-shingling

K-Shingling

a man saw that the fox jumped over the fence

a man	man saw	saw that	that the	the fox	fox jumped	jumped over	over the	the fence

fox jumped	the fence	a man	over the	that the	saw that	the fox	jumped over	man saw

K-Shingling

a man saw that the fox jumped over the fence

a man saw that the fox jumped over the hedge

the fox	jumped over	saw that	that the	a man	fox jumped	man saw	over the	the fence

fox jumped	the hedge	a man	over the	that the	saw that	the fox	jumped over	man saw

K-Shingling

Set1	the fox	jumped over	saw that	that the	a man	fox jumped	man saw	over the	the fence
Set2	fox jumped	the hedge	a man	over the	that the	saw that	the fox	jumped over	man saw

Word bigram	Set 1	Set 2
fox jumped	1	1
man saw	1	1
jumped over	1	1
that the	1	1
saw that	1	1
the fox	1	1
a man	1	1
over the	1	1
the hedge	0	1
the fence	1	0

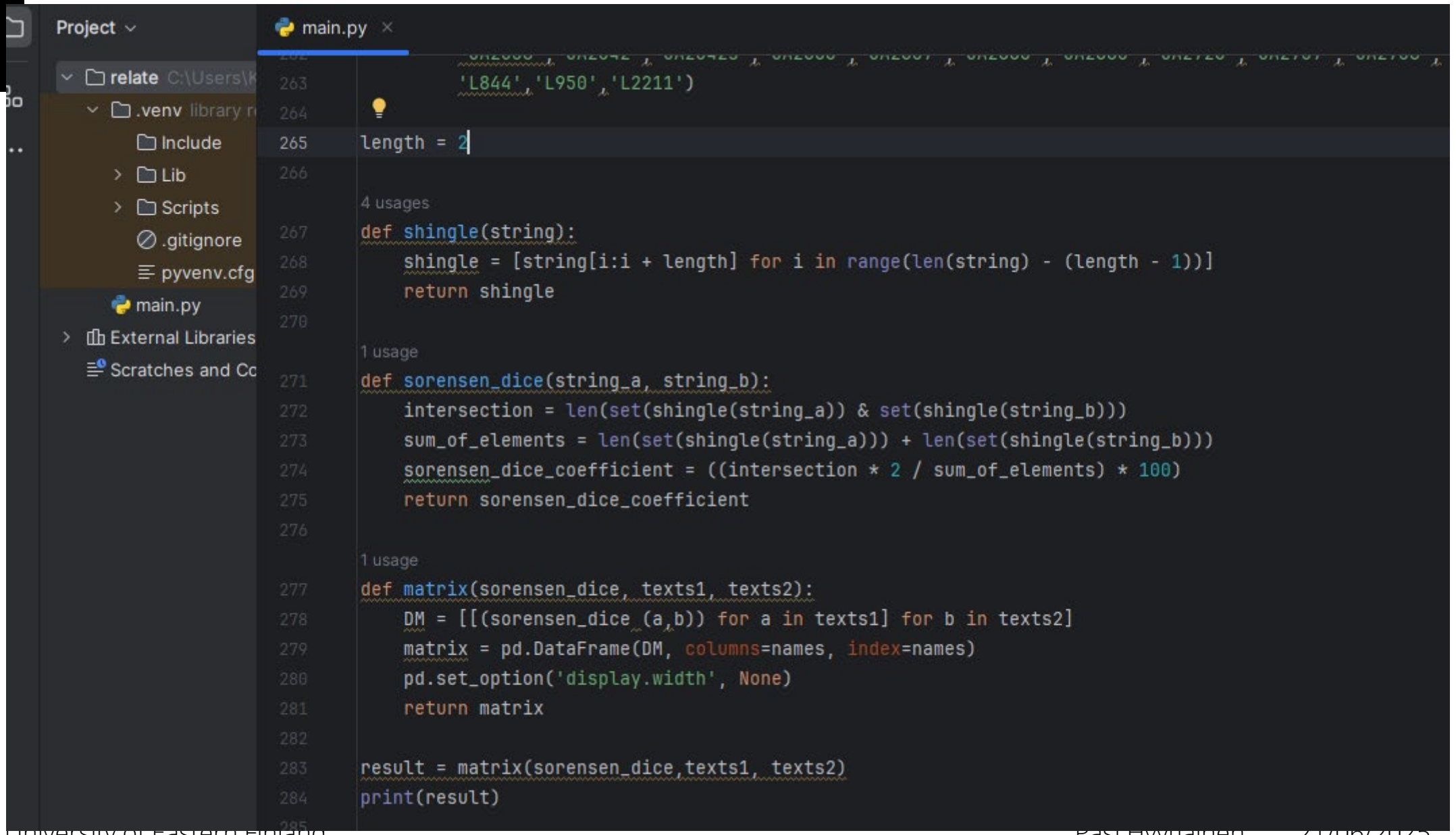
Intersection = 8

Union = 10

Sørensen-Dice Coefficient (SDC)

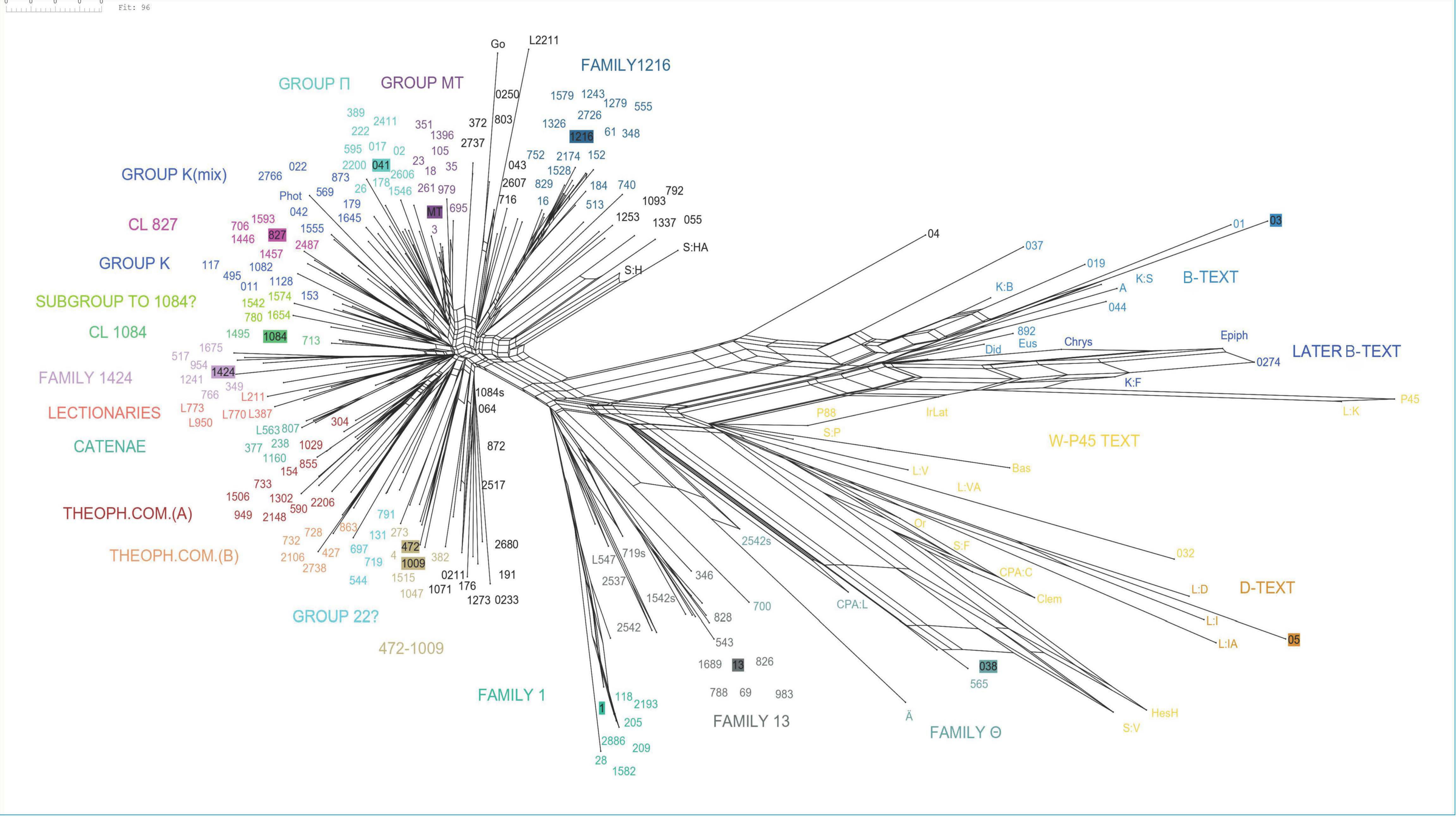
$$\frac{2\ x\ intersection}{sum\ of\ the\ number\ of\ elements\ in\ each\ set}$$

= 16 / 18 = 0.888 = 88 %


Project
main.py
relate C:\Users\K
.venv library r
Include
Lib
Scripts
.gitignore
pyvenv.cfg
main.py
External Libraries
Scratches and Cc
'L844','L950','L2211')
length = 2
4 usages
def shingle(string):
shingle = [string[i:i + length] for i in range(len(string) - (length - 1))]
return shingle
1 usage
def sorensen_dice(string_a, string_b):
intersection = len(set(shingle(string_a)) & set(shingle(string_b)))
sum_of_elements = len(set(shingle(string_a))) + len(set(shingle(string_b)))
sorensen_dice_coefficient = ((intersection * 2 / sum_of_elements) * 100)
return sorensen_dice_coefficient
1 usage
def matrix(sorensen_dice, texts1, texts2):
DM = [[(sorensen_dice (a,b)) for a in texts1] for b in texts2]
matrix = pd.DataFrame(DM, columns=names, index=names)
pd.set_option('display.width', None)
return matrix
result = matrix(sorensen_dice,texts1, texts2)
print(result)

The world's largest manuscript traditions

- Fourth step
  - Constructing trees or networks using the distance matrix
    - Tree inference algorithms
      - UPGMA
      - Neighbour-joining
    - Network algorithms
      - Neighbour-Net
      - Split decomposition


Fit: 96
GROUP Π
GROUP MT
FAMILY1216
GROUP K(mix)
CL 827
GROUP K
SUBGROUP TO 1084?
CL 1084
FAMILY 1424
LECTIONARIES
CATENAE
THEOPH.COM.(A)
THEOPH.COM.(B)
GROUP 22?
472-1009
FAMILY 1
FAMILY 13
FAMILY Θ
B-TEXT
LATER B-TEXT
W-P45 TEXT
D-TEXT
Go
L2211
0250
372
803
2737
043
2607
716
389
2411
222
595
017
02
2200
041
2606
26
178
1546
351
1396
105
23
18
35
261
979
MT
695
3
1579
1243
1279
555
2726
1326
1216
61
348
752
2174
152
1528
829
184
740
16
513
792
1093
1253
1337
055
S:HA
S:H
2766
022
873
Phot
569
179
042
1645
1555
706
1593
827
1446
2487
1457
117
1082
495
011
1128
153
1542
1574
780
1654
1495
1084
713
1675
517
954
1424
1241
349
766
L211
L773
L770
L387
L950
L563
807
304
377
238
1160
1029
154
855
733
1506
1302
2206
590
949
2148
791
728
863
131
273
732
427
697
472
2106
4
1009
382
2738
719
1515
544
0211
176
1047
1071
1273
0233
191
2680
1084s
064
872
2517
04
037
019
01
03
K:S
K:B
A
044
892
Eus
Chrys
Epiph
Did
0274
K:F
P45
L:K
P88
IrLat
S:P
L:V
Bas
L:VA
Or
S:F
CPA:C
Clem
032
L:D
L:I
L:IA
05
2542s
CPA:L
038
565
HesH
S:V
Ä
L547
719s
2537
346
1542s
700
2542
828
543
1689
13
826
788
69
983
1
118
2193
205
2886
209
28
1582

UNIVERSITY OF
EASTERN FINLAND

Thank you!

https://pasihyytiainen.com

A manuscript scholar,
theologian, digital humanist,
textual critic, and stemmatologist.