Tietokoneavusteinen tekijäntunnistus ja keskiajantutkimus: tuloksia ja haasteita

Dies Mediaevales, 17.11.2017
Reima Välimäki

Turku Centre for
Medieval and Early
Modern Studies
tucemems.utu.fi

Turun yliopisto
University of Turku

Profiling Premodern Authors (PROPREAU)

- Digitaalisten ihmistieteiden akatemiaohjelma, rahoituskausi 2016–2019
- Konsortion johtaja prof. Marjo Kaartinen, kulttuurihistoria
- Kulttuurihistoria: Teemu Immonen, Raija Vainio, Reima Välimäki, Anni Hella
- Tietojenkäsittelytiede: Sampo Pyysalo, Filip Ginter, Aleksi Vesanto
- Tekijäntunnistusta ja tekijöiden profilointia latinankieliseen kirjallisuuteen ja dokumentteihin antiikista (n. 100 eaa) myöhäiskeskiajalle (n. 1450)

Mikä tekijäntunnistus?

- Authorship attribution/verification/identification
- Authorship profiling
    - Author profiling tarkoittaa myös tekijän ominaisuuksien (esim. Ikä, sukupuoli, koulutustaso) tunnistamista teksteistä
- ”a form of style-based document authentication (Echtheitskritik),” (Stover & Kestemont 2016, 144)
- Tilastolliset menetelmät
- Koneoppiminen
- Tietokoneavusteinen tekijäntunnistus/profilointi laajasti käytössä modernien tekstien tutkimuksessa, myös rikosteknisiä sovelluksia
- PAN evaluation lab on digital text forensics http://pan.webis.de/clef17/pan17-web/index.html

Mikä tekijäntunnistus?

- *Basic attribution problem*
  - *Closed set of candidates, to determine who is the author*
  - *Substantial amount of texts extant from each candidate*
- *Verification problem "There is no closed candidate set but there is one suspect"*
- *Authorship verification* tarkoittaa myös kysymyksenasettelua, jossa kysytään, onko tekstin kirjoittanut joku tai ei kukaan esitetyistä tekijöistä.
- Profiling: *no suspected candidates. Are texts X & Y written by the same (anonymous) author?*
- Yhdistelmät mahdollisia

Stylometrisiä piirteitä tekstissä

- N-gram – jatkuva merkkijono (unigram, bigram, trigram etc.)
- Function words (funktiosana, kieliopillinen sana, vastakohta sisältösana)
  - Ei asiasisältöä, vaan keskusteluun liittyvä merkitys
  - Esimerkiksi pronominit ja partikkelit
  - Laskemalla funktiosanojen ja sisältösanojen suhdetta voidaan vertailla teksejä eri genreissä
  - Type-token ratio (lexical density)
- Lexomic analysis (niin sanottu “bag of words”, kaikki sanat)
  - Vaatii vähemmän esikäsittelyä
  - Tekstien aihepiiri vaikuttaa
- Sanapituus
- Lemma –taso vs. sanataso

Tekijäntunnistus antiikin ja keskiajan tutkimuksessa

Tilastollinen tekijäntunnistus ei uusi keksintö:

De imitatione Christi –tekstin tekijyydestä: Yule, G. Udny. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press, 1944.

Jo 1970-luvulla tietokoneavusteisia menetelmiä:

Marriott, Ian. “The Authorship of the Historia Augusta: Two Computer Studies.” *The Journal of Roman Studies* 69 (1979): 65–77. doi:10.2307/299060.

Myös muita tekijäntunnistuksen välineitä: käsikirjoitustraditio, käytetty sanasto, keskiaikaiset attribuutiot, harvinaisten lähteiden käyttö.

Vanhan tekstin tekijän tunnistaminen aina kokonaisuus, tietokoneavusteiset metodit yksi väline työkalupakissa

Tekijäntunnistus antiikin ja keskiajan tutkimuksessa

Tietokoneavusteiset tekijäntunnistukset tuore ilmiö antiikin ja keskiajan tekstien tutkimuksessa, keskustelu pääasiassa digitaalisten ihmistieteiden journaaleissa.

Kestemont, Mike. “Stylometry for Medieval Authorship Studies: An Application to Rhyme Words.” *Digital Philology: A Journal of Medieval Cultures* 1, no. 1 (2012): 42–72. doi:10.1353/dph.2012.0002.

Kestemont, M., S. Moens, and J. Deploige. “Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux.” *Digital Scholarship in the Humanities*, December 2, 2014. doi:10.1093/llc/fqt063.

Stover, Justin Anthony, Yaron Winter, Moshe Koppel, and Mike Kestemont. “Computational Authorship Verification Method Attributes a New Work to a Major 2nd Century African Author.” *Journal of the Association for Information Science and Technology* 67, no. 1 (January 2016): 239–42. doi:10.1002/asi.23460.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. ‘Authenticating the Writings of Julius Caesar’. *Expert Systems with Applications* 63 (30 March 2016): 86–96. doi:10.1016/j.eswa.2016.06.029.

Stover, Justin A., and Mike Kestemont. “The Authorship of the Historia Augusta: Two New Computational Studies.” *Bulletin of the Institute of Classical Studies* 59, no. 2 (December 1, 2016): 140–57. doi:10.1111/j.2041-5370.2016.12043.x.

PROPREAU: tuloksia

- Antiikin tekstien alahankkeesta artikkeli lähetetty Classical Philologyyn (1. kirjoittaja Raija Vainio): ”Reconsidering the authorship of Ciceronian corpus through computational authorship attribution”
- Asetelma *Rhetorica ad Herenniumin* tunnistus, testeissä toistuvasti Quintus Ciceron teos *Commentariolum petitionis* asettui M. Ciceron tekstien joukkoon - > uusi tunnistus

Pseudo-Augustine: a test case

Spring 2016: need for test data before the actual subprojects, i.e. easily available pseudo-text and background corpus

- > *Sermones ad fratres in eremo,* probably the most important pseudo-Augustinian work

Recognised unauthentic based on features that are NOT recognizable to the computer.

a) arguments used in the controversy between Aug. Eremites and canons after 1327
b) Late manuscript tradition
c) Single terms used in late medieval sense (*in spiritu libertatis*)

Composite collection, not single author:
-> possible to test the methodology with a difficult example

Sermones ad fratres in eremo

- The version in Patrologia Latina 40 is a late 15th cent. Compilation including 76 sermons.
- No critical edition, but Eric L. Saak (2012) has recognised a 14th-century core collection.

- **S1:** Core collection: of 25 sermons: sermons 1-22, 26 [Jordan 18], 43 [Jordan 21^2], 44 [Jordan 29] of the *PL* edition
- **S2**: PL edition of 76 sermons

Both based on the digital edition of PL at *Corpus Corporum* –database.

Corpus

- All major prose works of Augustine of Hippo (Patrologia Latina, CSEL, some recent editions of sermons: Dolbeau 1996, Weidmann 2015), total 113 works. Very short works of only few hundred words excluded, as well as those consisting mainly from quotations from other texts.

- Background corpus of late antiquity and medieval Latin prose, over 300 works, including Pseudo-Augustine

- Augustine: 4 344 775 words
- Others 9 667 819 words

Normalization

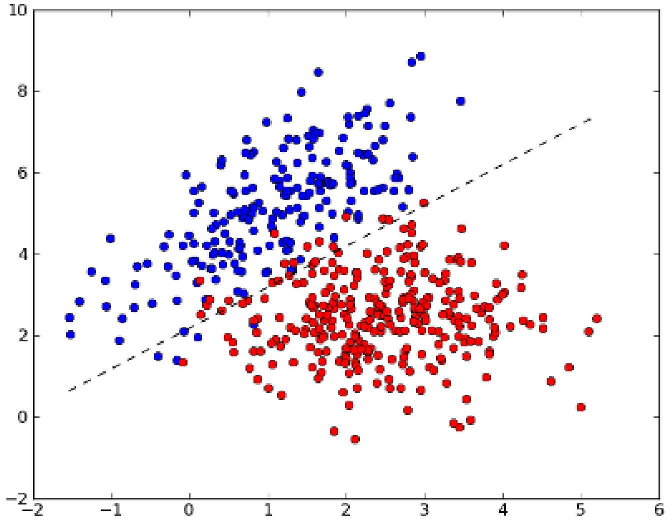
- In the preliminary test only minumum normalization of orthography
  - u/v variation normalized
- Mostly editions follow classical ortography, but some use medieval conventions
- In this case only editions, not transcripts from manuscripts
- Second test will be run where j/i variation and the most common diphtongs are normalized: ae->e, oe ->e
- Ti/ci & te/ce

How to proceed?

(Filip Ginter, Aleksi Vesanto, Reima Välimäki)

Machine learning

- Supervised machine learning
  - The classifier learns from annotated data
  - Classifier gets examples (books in our case) and the actual class (author) for them
  - Learns a model that can be used to classify the class for a completely new example



- Examples are represented as feature vectors
  - Features can be anything from the data that can be numerically represented
    - Word frequencies etc.

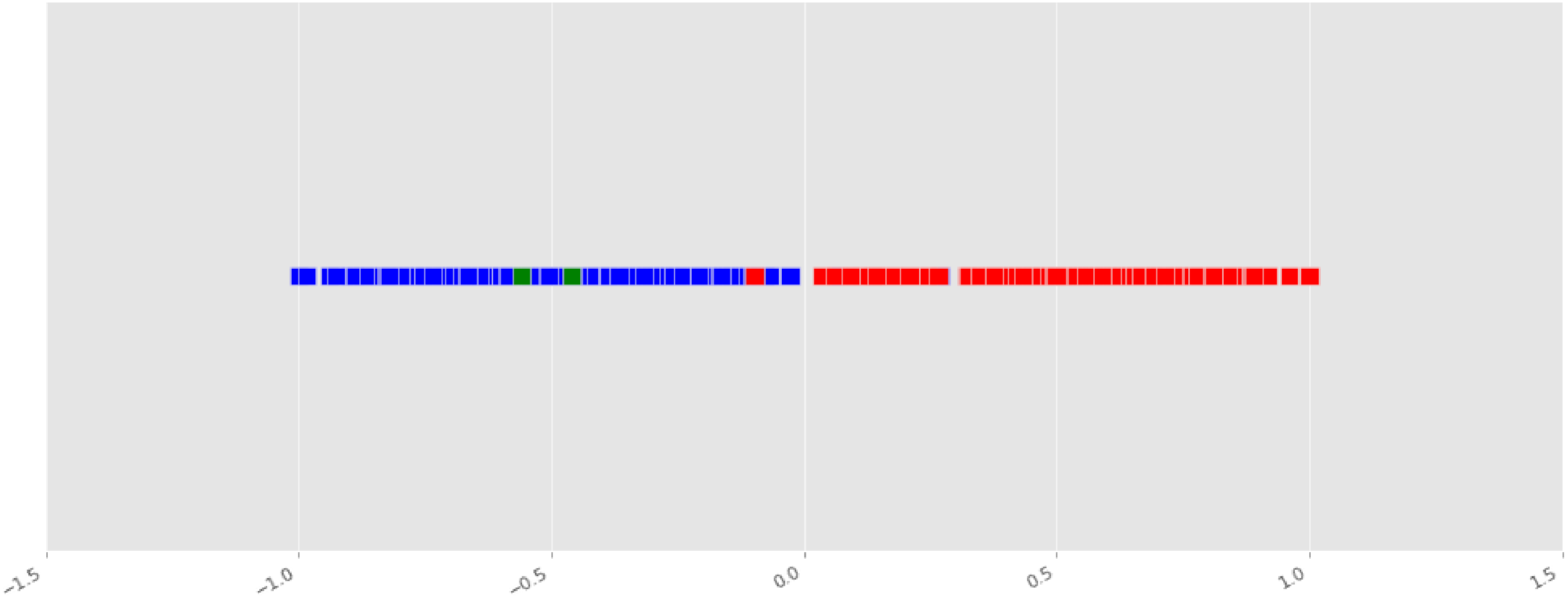
- A simple classifier will learn a weight for every feature
  - Useful features will get a high positive / negative weight

- Classifying new data:
  - Multiply every feature value from the test data with learned feature weights
  - Sum all these values together
  - If value is higher than a set threshold, give one class, if not, give another class

The first case

- A binary classification
    - Two classes
    - Positive class = Augustine
    - Negative class = Not Augustine (Any other author)
    - We are only trying to see whether a book was written by Augustine or not

Example of results (SVM, no mask,10.8.2017)



Turku Centre for
Medieval and Early
Modern Studies
tucemems.utu.fi

Turun yliopisto
University of Turku

The first approach – linear function

- SVM, Support Vector Machine
- A simple linear function
- Books are represented as vectors
    - Every element is the term frequency of a specific feature
- Features are word unigrams and bigrams
- This method learns a single weight for every feature
    - These weights tell how important a particular feature is
    - i.e. how strongly it represents Augustine or other authors
- Effects of content words controlled with masking. Only 1000 most common Latin words features as such, all others masked. The same classification was run to both masked and unmasked data.

The first approach – linear function

The strongest features of Augustine, unmasked data 10.8.2017

Commemoravi	34.22227297678724
qvod dixi	32.779259987671686
vtrvm	20.102265319292645
an forte	18.113550612904575
dicit apostolvs	17.853577974725066
neqve hoc	17.464898923001044
qvaliscvmqve	16.116328467928064
lavdas	14.626625504895799
etiamsi	13.968762434478839
non evm	13.753813130604286

The strongest features of other authors

nvnqvid	-44.96960122756715
cvnctis	-40.477173448990996
freqventer	-28.38878985239002
provt	-27.261090715803917
absqve	-21.861018106471548
idcirco	-21.591536693865734
psalmista	-17.892364804920437
dvm	-13.467083983618501
qvicqvid	-12.718505215442127
virtvtvm	-12.38967815553817

The first approach – linear function

The strongest features of
Augustine, masked data
10.8.2017

qvod dixi	13.602380873504515
an forte	12.397582306532254
dicit apostolvs	11.84989228957933
avt vero	9.77967596160497
vtrvm	8.846204804211647
nvmqvid enim	8.799727051240406
non evm	8.459725041685118
etiam ista	7.709849657591735
neqve hoc	7.441029311151708
vsqve xxxx	6.9859093437243125

The strongest features of
other authors

cvnctis	-25.834526520450325
absqve	-15.275308528983842
idem est	-12.304923961895735
pariter et	-11.939098370487928
idcirco	-8.67848556924062
virtvtvm	-6.839028947742164
in propria	-6.450075445900782
de ea	-5.975647080401483
ostendit dicens	-5.56228456156858
poterit xxxxxxxxx	-5.322886531579247

The second approach

- Convolutional neural network
- More complex method
- Hard to get weights for features
- Results are more accurate than with the SVM
- Features used here are character 5-grams

Results

- The classifier is difficult to fool: our first classification attempt used the attributions of the *Corpus Corporum* database. Several wrong attributions, which the classifier recognised (i.e. gave negative values to pseudo-Augustine texts)
- Not surprisingly, both S1 and S2 (*Sermones ad fratres in eremo*) get constantly negative in both masked and unmasked SVM and CNN values: between -0.45 and –0.98

Surprises

1) Evodius of Uzalis's *De fide contra Manichaeos* attributed strongly to Augustine in every test, gaining values between 0.27 (SVM no maks) and 0.99 (CNN no mask). SVM gives a more reliable attribution to Augustine with **masked** texts (0.41)

- The treatise includes long quotations from Manichean sources, used also by Augustine. These were extracted from the text before the test.

2) 13 sermons recently attributed to Augustine by C. Weidmann (CSEL 101, 2015) fall into the gray area in every test. SVM gives slightly negative values (-0.11, -0.10), CNN positive, but dubious (0.06 and 0.17)

Interpretation

1. the author(s)/compilers of *Sermones ad fratres in eremo commorantes* were not particularly good in imitating Augustine: most Pseudo-Augustinian texts get values much closer to the treshold

-> credibility of the text not in style but content?

2. Evodius of Uzalis's text bears very strong presence of Augustine.

3. Reconsidering the authorship of Weidmann's collection of 13 sermons

Proposition: integrated authorship attribution (qualitative + quantitative)

Weidmann’s attribution criteria included

1) language and style;
2) usage and form of bible quotations;
3) theological content and argumentation;
4) historical context;
5) manuscript circulation.

When four of the five criteria suggested Augustine’s authorship and no convincing contradictory evidence existed, the text was considered to be an authentic Augustinian sermon.

- **We add: 6) computational stylistics**. If strong attribution to Augustine, less qualitative evidence is needed. If dubious or negative result, all other evidence must be strong and without contradictions.

Sermones ad fratres in eremo – tekijä ?

Neljä kandidaattia 1300-luvulta

Anonymous Florentine,
Nicolas of Alessandria,
Henry of Friemar
Jordan of Quedlinburg

Alustavien tulosten perusteella Jordanus tekijä….mutta on myös yhden varhaisen *Sermones* -kokoelman kompilaattori

Haaste: keskiajan kirjallisen kulttuurin ominaispiirteet

Giles Constable:

> “ there are infinite shadings between correction, revision, imitation, and falsification and, in works of art, between repair, restoration, reproduction, and copying.”
>
> “Forgery and Plagiarism in the Middle Ages,” in Constable, Culture and Spirituality in Medieval Europe (Aldershot, 1996), 1–41, p. 3.

Miten binäärinen luokittelu pystyy vastaamaan tähän? Uuden filologian tekijyyskäsitys vs. kvantitatiivinen tutkimus

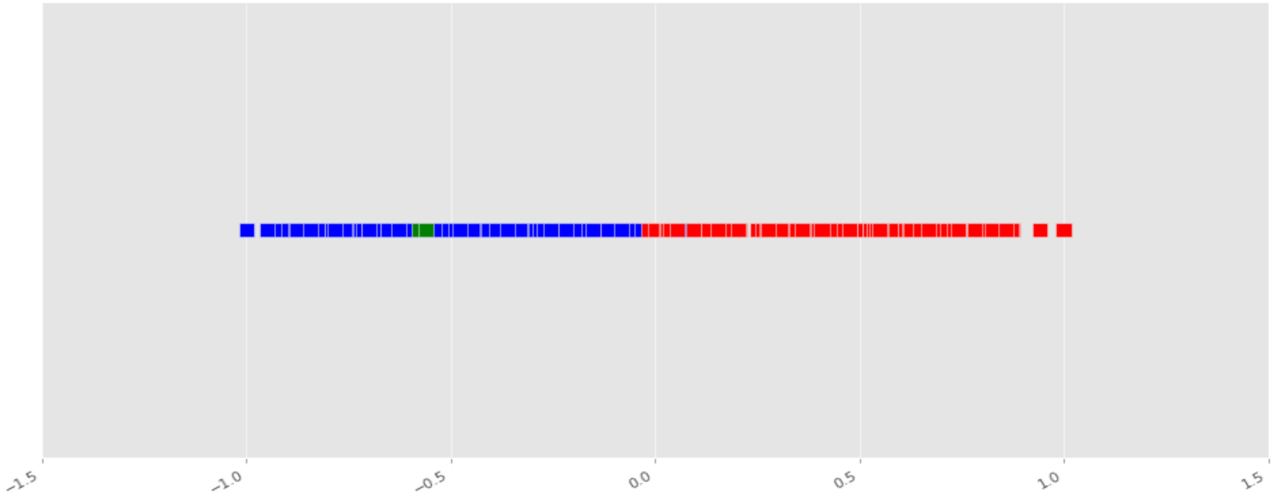
Jeroen De Gussem, ‘Bernard of Clairvaux and Nicholas of Montiéramey: Tracing the Secretarial Trail with Computational Stylistics’, *Speculum*, 92:S1 (2017), pp. S190–S225.

Osoittaa Nicholasin vaikutusta Bernardin korpukseen (kirjeet, saarnat), mutta tulkinnat alustavia

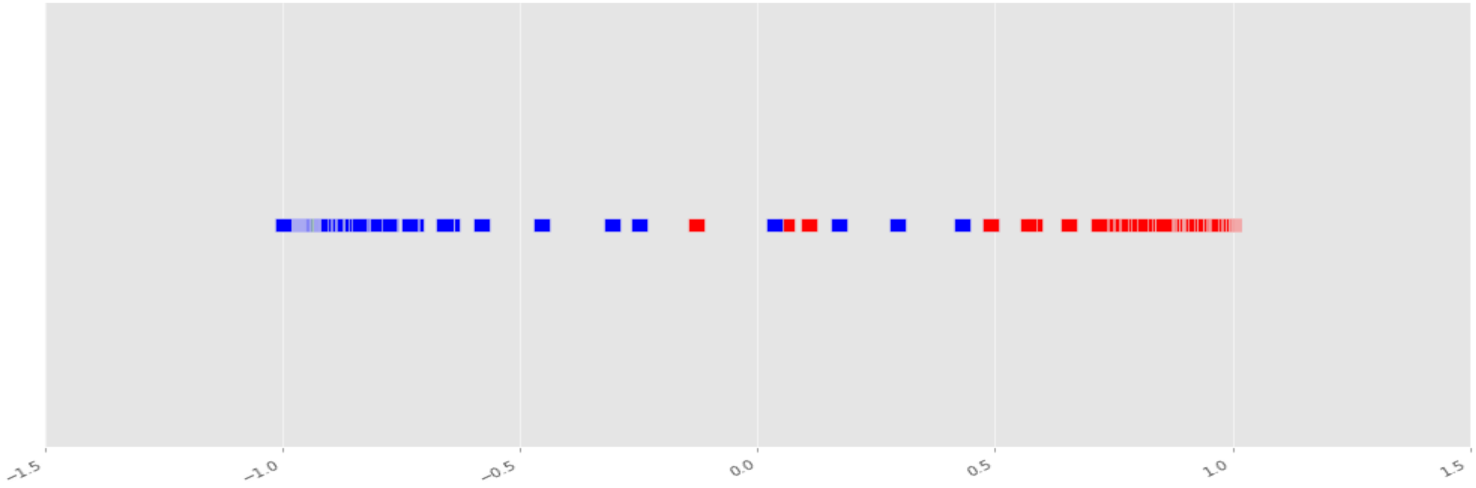
Tekstien eri redaktiot – kuinka pitkään tietokone tunnistaa tekijän tyylin samaksi?

Myöhäiskeskiajan ongelma

- Klassikkotekstit helposti saatavilla korpuksista, myös keskiaikaisia tekstejä n. vuoteen 1200 asti editoitu paljon.
- Myöhäiskeskiaikaisia tekstejä hyvin vähän valmiissa korpuksissa: käytettävä yhdistelmää valmiiksi koneluettavia tekstejä, skannattuja ja tarkistettuja vanhoja editioita ja transkriptioita käsikirjoituksista
- Häilyvä ortografia ja syntaksi
- Ratkaistava vielä, kuinka paljon tekstejä normalisoidaan
- Myös tekijyys-keskustelu painottunut 1000-1200 –luvulle – myöhäiskeskiajan ominaispiirteet kirjoittamisessa ja julkaisemisessa. Paperi!



SVM no mask



CNN no mask