REWRITING LITERATURE HISTORY WITH BIG DATA

Mikko Koho
Aalto University

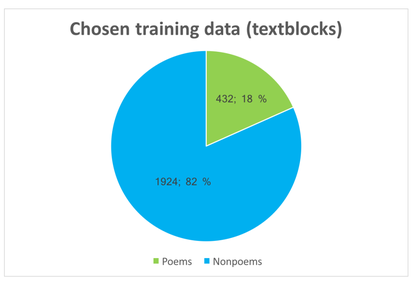
Elsi Hyttinen
Peter Ivanich
Elisabeth Oakes
Ilona Pikkanen
Leena Tulkki
University of Helsinki

Risto J. Turunen
University of Tampere

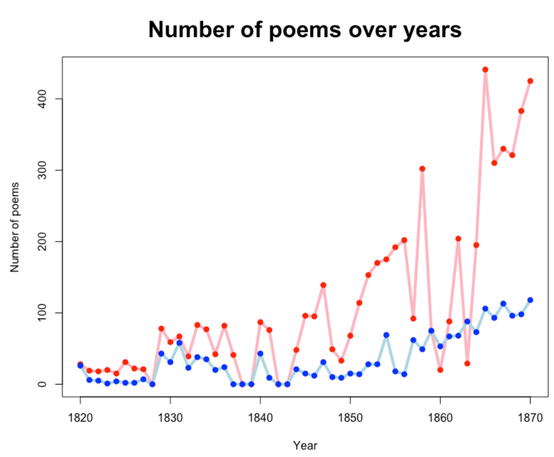
DIGITAL HUMANITIES HACKATHON
MAY 2017 HELSINKI

Data: National Library's digitalized newspaper archives - approx. 1,9M pages from years 1820-1910.

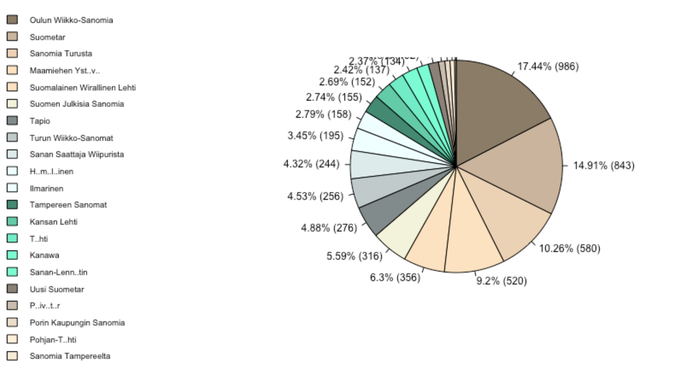
FINDING POEMS WITH MACHINE LEARNING

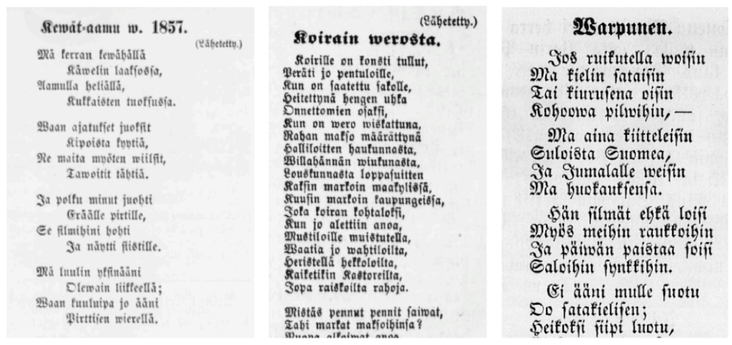


A Support Vector Machine classifier was used to classify textblocks into poem or non-poem classes. The word frequencies were used as features for the training classifier. The initial cross-validation error of 6 % was improved to 93 % by using an exhaustive search over the parameter values, and carefully choosing the training data.



The accuracy of the predictions on the early 19th century was over 50 %, however it decreased in the late 19th century. It would be interesting to take a closer look to investigate the reasons behind this observation.





Kewät-aamu w. 1857.
(Lähetetty.)

Mä kerran kewähällä
Käwelin laaksossa,
Aamulla heläällä,
Kukkaisten tuoksussa.

Waan ajatukset juoksit
Kirpeästa kyytiä,
Ne maita myöten wiilsit,
Tawoitit tähtiä.

Ja polku minut juohti
Eräälle pirtille,
Se silmihini hohti
Ja näytti tiistille.

Mä kuulin yksinääni
Olewain liikkeellä;
Waan kuuluipa jo ääni
Pirttisen wierellä.

(Lähetetty.)

Koirain werosta.

Koirille on konsti tullut,
Peräti jo pentuloille,
Kun on saatettu lakeille,
Heitettynä hengen uhka
Onnettomien olasta,
Kun on wero wiskattuna,
Rahan maksu määrättynä
Hallitoitten haukunnasta,
Willahännän wiukunasta,
Louskunnasta loppuisitten
Kaffin markoin maakyliässä,
Kuusin markoin kaupungeissa,
Joka koiran kohtalosta,
Kun jo alettiin anoa,
Mukiloille muistutella,
Waatia jo wahtiloilta,
Heristellä hekkoloilta,
Kaiketikin Kostoreilta,
Jopa raiskoilta rahoja.

Mistäs pennut pennit saiwat,
Tahi markat maksoihinsa?

Warpunen.

Jos ruikutella woisin
Mä kielin sataisin
Tai kiurusena oisin
Kohoowa pilwihin, —

Mä aina kiitteleisin
Suloista Suomea,
Ja Jumalalle weisin
Mä huokauksensa.

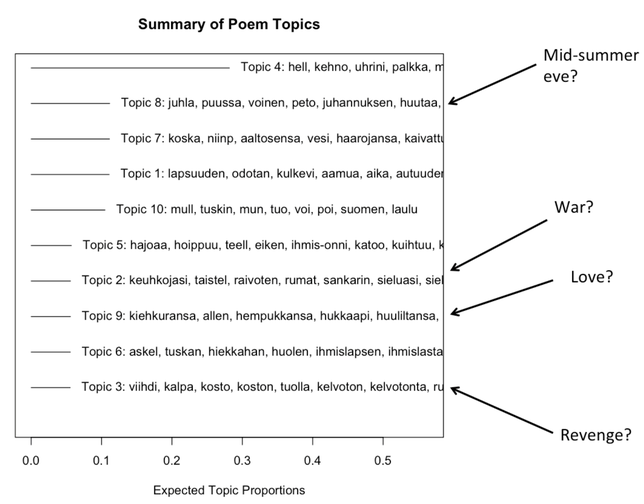
Hän silmät ehkä loisi
Myös meihin raukkoihin
Ja päiwän paistaa soisi
Saloihin synkkihin.

Ei ääni mulle suotu
Do satakielisen;
Heikoksi siipi luotu,

TOPIC MODELING

Structural topic modeling (stm) is a method to find topics in a set of documents. The user defines the number of topics (k) and the algorithm statistically analyses the frequencies of the words in the texts to find the groups of words which "go together".

With our training data of approximately 100 handpicked poems, we used the stm R package, which yielded promising results – the topics seemed like plausible poem themes.



We also performed the stm to lemmatised texts, which caused changes in the topics, but did not necessarily improve the model. There are several different types of stm algorithms, and it would be useful to compare their performance to determine which type would be best for analysing poems.

With the complete poem corpus and improved stm analysis, we could shed light to e.g. following questions:

- What kind of topics are there and what are their prevalences?
- Are there differences in poem topics between newspapers or geographical areas?
- How did the topics develop through decades?
- Are there sub-topics for common themes, such as nationalistic poems?

MORPHOLOGY OF POEMS

Could some morphological features highlight regional differences, changes in time, or allow us to differentiate types of poetry within the corpus? The features we wanted to look for were:

- Adjectives
- Verbs (the hypothesis being that poems containing large numbers of verbs would be more active in meaning)
- Interjections and exclamations such as "Oi!", which we know are typical for poetry

Steps of the analysis:

- Selecting the training data,
- Uploading the data to the server and running a morphological analysis tool (LAS) on the data (POS-tagging the data)
- Processing the JSON files in R
- Analysis of results

What we managed to do:

- POS-tagging on the corpus 1820-1870
- A small-scale test (manual)

Issues:

- Processing JSON files with parsed trees has proven difficult, and we have not been able to come up with a working script
- Problems within the data (OCR errors resulting in wrong tagging of words)

Aalto-yliopisto
UNIVERSITY OF TAMPERE
Turun yliopisto
University of Turku
KANSALLIS KIRJASTO
SKS Suomalaisen Kirjallisuuden Seura
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS