Detecting (viral) news in digitized historical newspaper archives

Slides authored: Aleksi Vesanto, TurkuNLP, University of Turku
Slides presented: Filip Ginter, TurkuNLP, University of Turku
https://turkunlp.org

This presentation

Explain data / task

How to find text reuse

Show some example results

Find viral news

If enough time, show a search interface where one can search the results

Starting point

- Digitized Finnish newspapers and journals from 1770-1920
    - About 5 million pages
    - Availabe for public access at Kansalliskirjasto.fi
    - OCR-read text

The task

- Finding viral reused passages
  - First find all kinds of reused passages from these
- News, advertisements, poems, timetables, etc.
- Can be used to see how news spread
  - Starting location → ending location
  - How fast a particular piece of news spread
  - How often it appeared
  - How often it was reprinted in the same publication
  - Lots of different applications
- Also possible to see long term reuse
  - Over 100 years between reuse
  - Impossible to do manually

The task

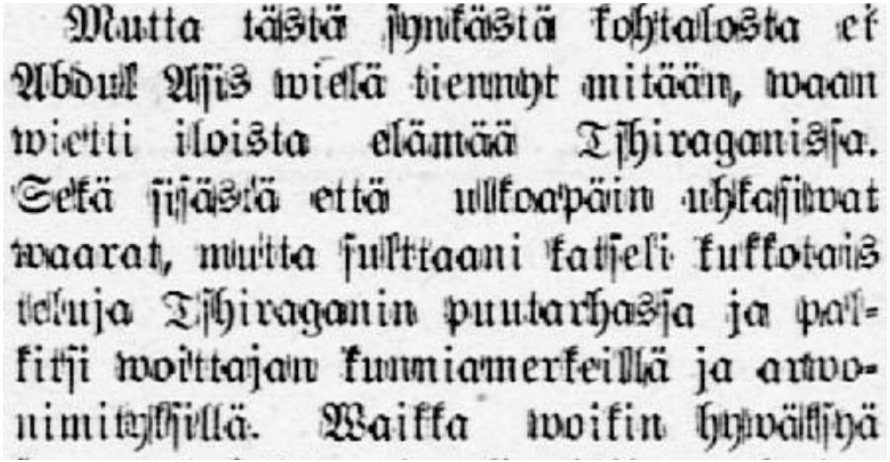
- Should be simple to find them, right?
- Several text document similarity measures exist
  - ngram overlap etc.
- Well, no
- The OCR quality is at times pretty abysmal
- Mostly due to the fraktur font

Example pair - Finnish

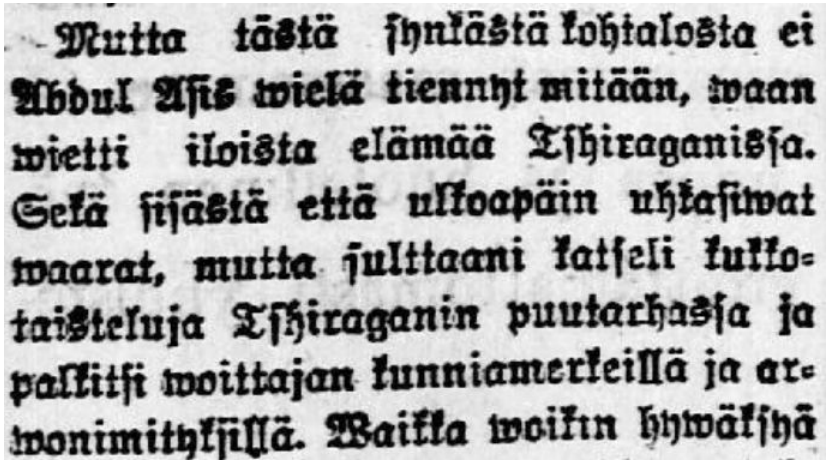
Multa t\ä@tä fyNlkÄsiii kchtalostu ,ct , Abouil Asi,3 wic!lä ticiun't>t ,mitää>«, » vaalii luiftti iloista M,m<iä Tshiragauissa, ©elä fi:föf3>i'öi että uiUfatfpäim -uhkaisiloui i Hviarat, miinto fu^tiaani 'fatifefi- fuffotai» lÄuja THi roinin, puutarhassa ja, ipici'ilitsi hwi'tt<iiöii fmmiamcrk^iUi ja anoo» »imilyMla,

Mutta tästä synkästä kohtalosta ei Abbul Asib »ielä tiennyt mitään, vaan »ietti iloista elämää TshiraganiSsa. Sekä sis»Stä «ttä ulkoapäin uhkasivat «aarat. mutta sulttaani katseli lukkotaisteluja Tfhiiaaanin puutarhassa ja palkitsi voittajan lunnicnnerleillä ja ar° vonimityksillä.

Example pair - Finnish



Mutta tästä ſynkästä kohtalosta ei Abdul Aſis wielä tiennyt mitään, waan wietti iloista elämää Tſhiraganiſſa. Sekä ſiſäsklä että ulkoapäin uhkaſiwat waarat, mutta ſulttaani katſeli kukkotaiſteluja Tſhiraganin puutarhaſſa ja palkitſi woittajan kunniamerkeillä ja arwonimityſhillä. Waikka woikin hywälſyä



Mutta tästä ſynkästä kohtalosta ei Abdul Aſis wielä tiennyt mitään, waan wietti iloista elämää Tſhiraganiſſa. Sekä ſiſästä että ulkoapäin uhkaſiwat waarat, mutta ſulttaani katſeli kukkotaisteluja Tſhiraganin puutarhaſſa ja palkitſi woittajan kunniamerkeillä ja arwonimityksillä. Waikka woikin hywäkſyä

How to go about it then?

- BLAST
  - Program designed for comparing and aligning biomedical sequences, like proteins
  - Finds overlapping sequences in a large sequence database (used on whole genomes)
    - Overlap means a sub region in two sequences that are similar
    - Whole sequence does not have to be similar
  - 25% overlap in protein sequences significant
  - Does exactly what we want
- Problem: the data is not protein sequences...

BLAST

- We need to encode our data to proteins
- 23 distinct amino acids to work with
- Find the 23 most used characters from the data
- Form character → amino acid mapping

BLAST

- Using the mapping, encode the data into proteins, discarding characters that don’t have a match in the mapping

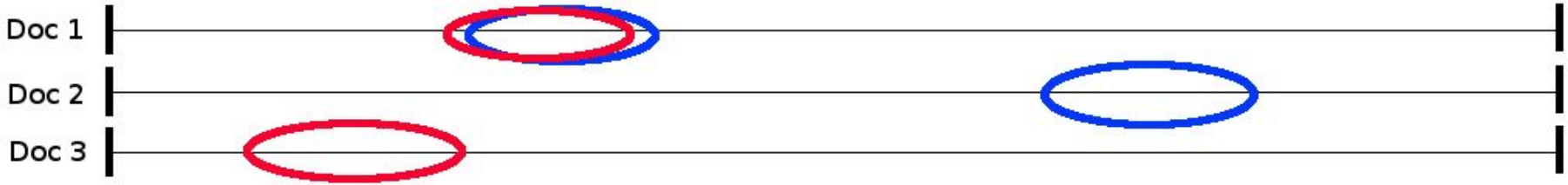
- "This is an example sentence” → “DSCHCHBEGBNQFGHGEDGEG”
- We then feed our proteins to BLAST

BLAST

- BLAST outputs a pairwise alignment for all sequences
  - Meaning for all pairs we know the regions that are similar, i.e. that contain text reuse
- Uses heuristic methods to decide which parts of the sequences should be aligned
- Example:
  - This is ---- an example sent----.
  - This is not an example sentence.

BLAST

- We know which parts of the sequences are the hits
- Using the sequence offsets, we cluster all hits that overlap enough to be part of the same cluster



BLAST - cluster

- Example of a cluster

Filename: fk10435_1913-08-02_2.txt
Style: Newspaper
Date: 1913-08-02T00:00:00Z
Year: 1913
Location: Savonlinna
Language: fin
Cluster ID: 10634174
Title: Savolainen
URL: http://digi.kansalliskirjasto.fi/sanomalehti/binding/1290172?page=2
Text: a ja muita fa« manlaista tarpeitaan marten ja faistmat yksin ne kustantaa. jo näiden suurien onnettomuuksien estämiseksi täytyy jokaisen suomalaisen puolueen äänimaltaisen miehen ja naisen tänään mennä äänestämään. meillä ei olemaraa menettää ainoatakaan ääntä. ia yhdes» tä äänestä saattaa maalin tulos ja eduskunnan kokoonpano suuresti riippua. isoaro

Filename: 1458-2740_1913-08-02_2.txt
Style: Newspaper
Date: 1913-08-02T00:00:00Z
Year: 1913
Location: Pori
Language: fin
Cluster ID: 10634174
Title: Satakunta
URL: http://digi.kansalliskirjasto.fi/sanomalehti/binding/1287306?page=2
Text: a ia muita 'samanlaisia tarpeitaan marten ja saisiwat yksin ne kustantaa. jo näiden suurien onnettomuulksien estämiseksi täy- wy sokaisen suomalaisen puolueen ääniwaltaisen miehen ia naisen tänään mennä äänestämään. meillä ei die maraa imettää ainoatakaan ääntä. ia yhdestä äänestä laattaa waalin tulos sa eduskunnan kokoonpano juu» resti riippua. ku

Filename: fk10116_1913-07-31_1.txt
Style: Newspaper
Date: 1913-07-31T00:00:00Z
Year: 1913
Location: Helsinki
Language: fin
Cluster ID: 10634174
Title: Iltalehti
URL: http://digi.kansalliskirjasto.fi/sanomalehti/binding/1201344?page=1
Text: a ja muita samanlaisia tarpeitaan varten ja saisivat yksin ne kustantaa, jo näiden suurien onnetlomuujksiien fjstiärniseksii täytyy jokaisen äänivaltaisen miehen ja naisen huomenna tai yiihucmenna mennä äänestämään. meiiiä ei ole varaa menettää ainoatakaan ääntä ja yhdestä äänestä saattaa vaalin tulos ia eduskunnan kokoonpano suuresti riippu

BLAST

- BLAST can find even very short matches
  - Like 5 characters long hits

- We limited matches to minimum 300 characters

- Shorter than that tends to be boilerplate
  - Would increase the amount of results several times

Results

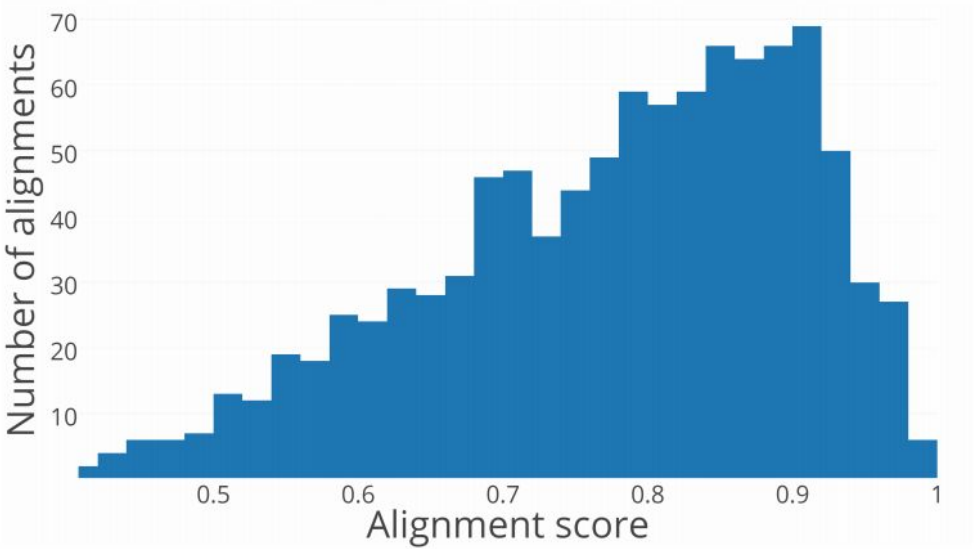
- Found 73,922,354 hits, consisting of 13,797,868 clusters

- Took nearly 500,000 CPU core hours
  - Running on a single laptop would take 14 years

- Thankfully CSC provided access to supercomputers
  - Took only 2 weeks

Results

- How similar the matches were:

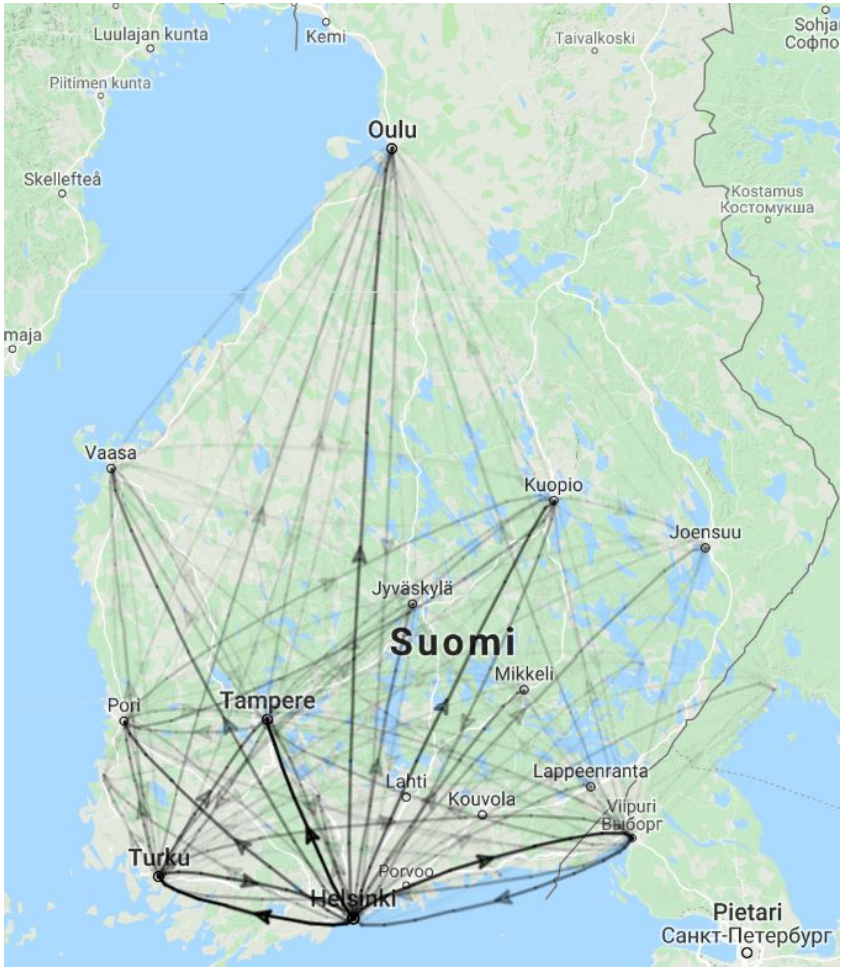


Example

News of a bank robbery

Place	Date	Title
Helsinki	1906-11-07	Uusmaalainen
Helsinki	1906-11-07	Helsingin Sanomat
Turku	1906-11-08	Uusi Aura
Helsinki	1906-11-08	Elämä
Tampere	1906-11-08	Tampereen Sanomat
Turku	1906-11-08	Sosialisti
Helsinki	1906-11-08	Uusi Suometar
Jyväskylä	1906-11-09	Suomalainen
Oulu	1906-11-09	Kaleva
Kuopio	1906-11-09	Pohjois-Savo
Tampere	1906-11-09	Kansan Lehti
Viipuri	1906-11-09	Karjala
Sortavala	1906-11-10	Laatokka
Heinola	1906-11-10	Heinolan Sanomat
Savonlinna	1906-11-10	Keski-Savo
Joensuu	1906-11-10	Karjalatar
Lahti	1906-11-11	Lahden Lehti
Kemi	1906-11-12	Pohjois-Suomi
Kristiina	1906-11-12	Etelä-Pohjanmaa
Lahti	1906-11-13	Lahti

Map


Luulajan kunta
Kemi
Taivalkoski
Piitimen kunta
Oulu
Skellefteå
Kostamus
Костомукша
Vaasa
Kuopio
Joensuu
Jyväskylä
Suomi
Mikkeli
Pori
Tampere
Lahti
Lappeenranta
Kouvola
Viipuri
Выборг
Turku
Porvoo
Helsinki
Pietari
Санкт-Петербург

Viral news

- Massive database of results
- How to find the viral ones?
- Largest cluster = most viral?
- Maybe highest span?

Viral news - highest count

- Count: 1013
- An add for Apothecary
- Clearly not “viral”
- Spread to:
  - 10 locations
  - 19 titles
  - 9713 days

Apotekaren

Al. Mertns sockrade **Carlsbader Piller.**

Ett beprövadt medel mot

Förstoppning

och deraf uppstående lidande, såsom svindel, hufvudvärk, andnöd, beklämning, tarmtröghet, slembildning, aptitlöshet, lefverlidande, hemorroider m. fl. m. fl. Deras värkan är mycket angenäm och mild, hvarför de gerna tagas af fruntimmer och t. o. m af barn samt föredragas framför de skarpt värkande salterna, bittervatten och dyl.

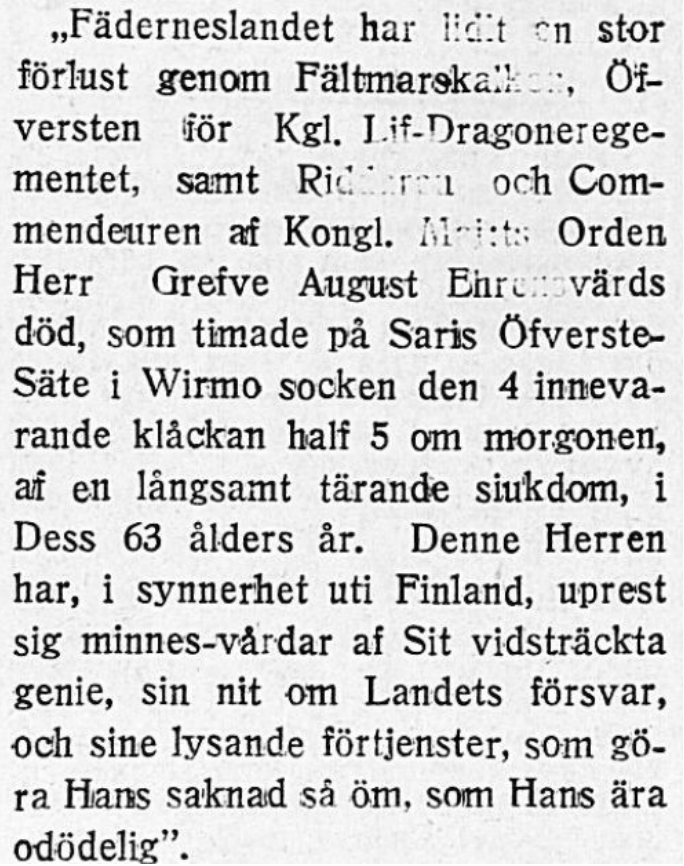
På hvarje ask är det här aftryckta skyddsmärket inpregladt och hvarje bruksanvisning är försedd med namnteckningen. Pris per ask 1 mark. Står att erhållas hos

442

alla landets Apotekare.

Viral news - highest span

- Death announcement
  - Probably
- Again, not “viral”
  - Printed in 1772, 1872 and 1918
  - 1 location
  - 2 titles
- Example of long term reuse



„Fäderneslandet har lidit en stor förlust genom Fältmarskalken, Öfversten för Kgl. Lif-Dragoneregementet, samt Riddaren och Commendeuren af Kongl. Maj:ts Orden Herr Grefve August Ehrensvärds död, som timade på Saris Öfverste-Säte i Wirmo socken den 4 innevarande klåckan half 5 om morgonen, af en långsamt tärande siukdom, i Dess 63 ålders år. Denne Herren har, i synnerhet uti Finland, uprest sig minnes-vårdar af Sit vidsträckta genie, sin nit om Landets försvar, och sine lysande förtjenster, som göra Hans saknad så öm, som Hans ära odödelig”.

Viral news - viral score

- Calculate a virality score for each cluster
  - Take in account the number of unique locations, unique titles and the time time it took for the reuse to spread

$$score = \frac{uniquelocations}{totaluniquelocations} * \frac{uniquetitles}{totaluniquetitles} * \frac{1}{numberofelapseddays}$$

- Scaled the number to be between 0 and 100
  - 0 least viral
  - 100 most viral

Viral news - viral score

- One problem:
  - Even a single “late” reprint will completely destroy the score, even if it would otherwise be viral

- Omit outliers
  - If a reprint is clearly not part of the main “cluster” of reprints, ignore it
  - Keeps the viral score realistic

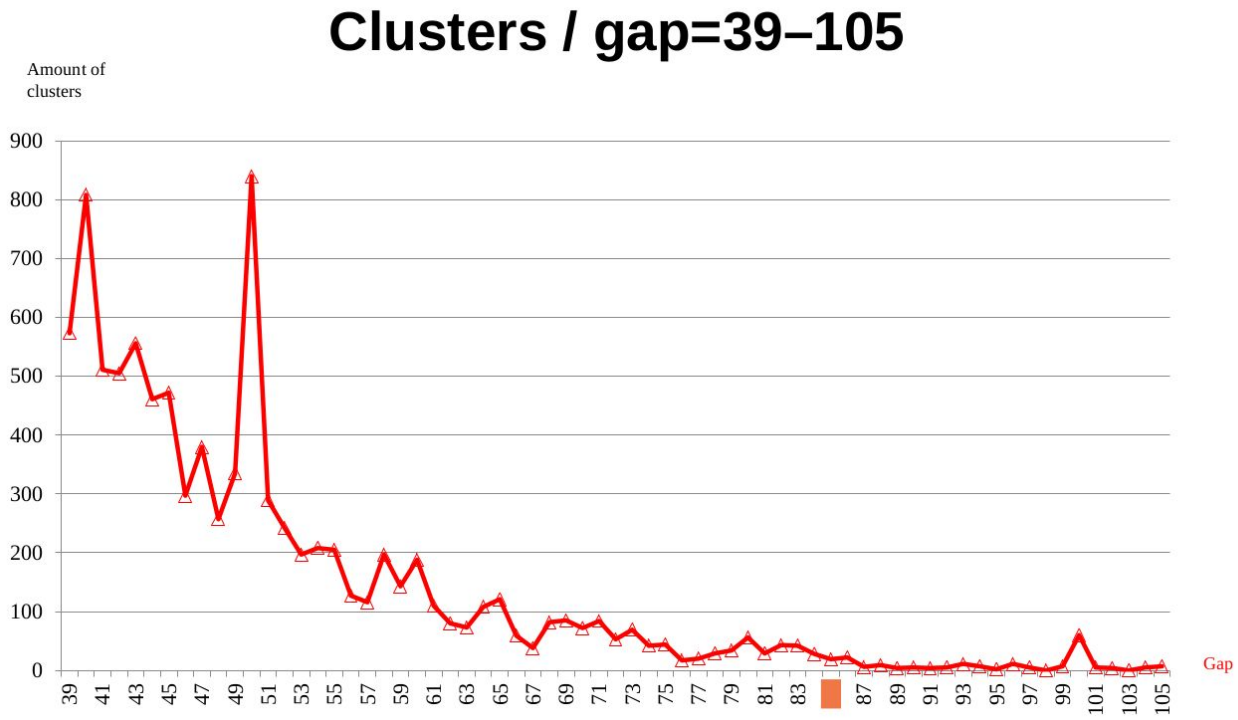
Viral news - Least viral

- Viral score 0
- Printed once in 1802 and then again in 1902
- Another long term example
  - Naturally, long term reuse is the opposite of viral

Den hithörande notis, som vi åsyftat i Å. T, och som tyckes innebära en antydan till vederbörande om den dräkt, de vid det stundande konungabesöket böra bära, är af följande lydelse: „Beskrifning på den af Kongl. Maj:t i nåder fastställda paraduniformen". Lång rock af mörkblått kläde med krage, reverer och uppslag af lika färg, foder af mörkblått schalong, hvit väst, blå byxor, gula knappar med länets vapen, svart trekantig hatt med guldtrens och svart ros samt en något större knapp, (de, som varit officerare, bära gul ros), svarta kragstöflar, hvartill kunna brukas sporrar, och värja med svart balja. Hvita byxor brukas till skor eller strumpor. Strumporna böra vara hvita och af silke vid underdåniga uppvaktningar hos Deras Kongl. Maj:ter. Denna dräkt nyttjas af 1) Ridderskapet och Adeln. 2) Ämbets- och tjänstemän med Kongl. Maj:ts höga fullmakter, och med fullmakter, som medföra lika värdighet, samt af landtmäntmästare. 3) Possessionater. 4) Kronofogdar. Kronobetjäning utom fogdarne nyttja blå underkläder och hvita i stället för gula knappar samt hvit eller siltvertrens på hatten." *)

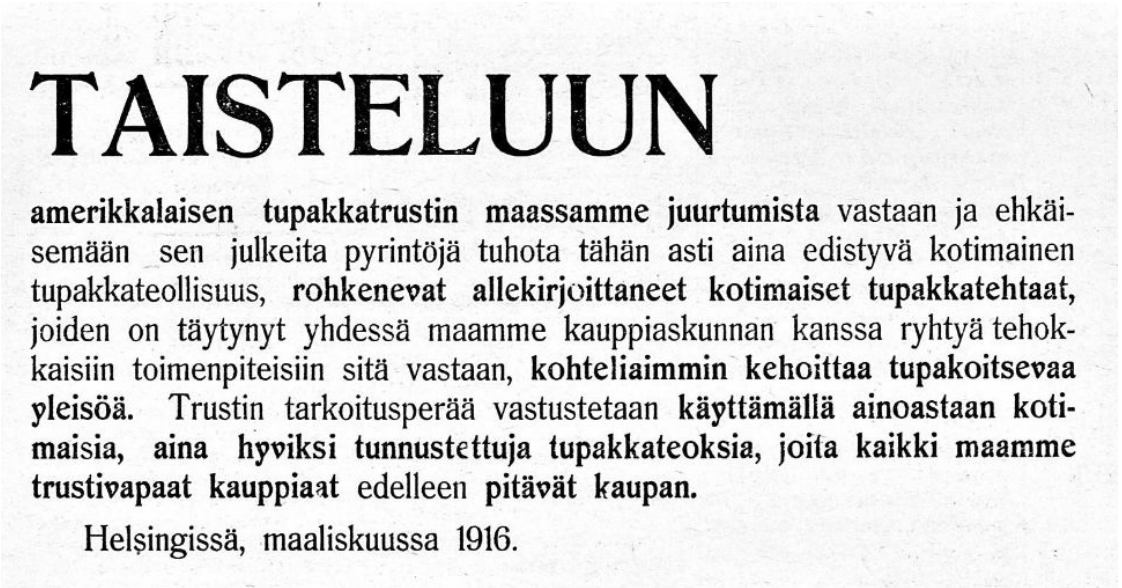
Viral news - Non viral

- Gap where no reprints occurred
- Clear spikes in 50, 100



Viral news - Most viral

- Viral score 100
- Appeared in:
  - 26 unique locations
  - 45 unique titles
  - in 1 day
- Actually a paid Ad



TAISTELUUN

amerikkalaisen tupakkatrustin maassamme juurtumista vastaan ja ehkäisemään sen julkeita pyrintöjä tuhota tähän asti aina edistyvä kotimainen tupakkateollisuus, **rohkenevat allekirjoittaneet kotimaiset tupakkatehtaat,** joiden on täytynyt yhdessä maamme kauppiaskunnan kanssa ryhtyä tehokkaisiin toimenpiteisiin sitä vastaan, **kohteliaimmin kehoittaa tupakoitsevaa yleisöä.** Trustin tarkoitusperää vastustetaan **käyttämällä ainoastaan kotimaisia, aina hyviksi tunnustettuja tupakkateoksia, joita kaikki maamme trustivapaat kauppiaat** edelleen **pitävät kaupan.**

Helsingissä, maaliskuussa 1916.

- Idea works
  - Would require a cluster classification first
  - Ignore ads

Examine clusters

- http://comhis.fi/clusters
-

Conclusion

- A robust method to detect text reuse even through heavy OCR noise
- Viral score to rank the results and find the viral ones
- Search interface to search the found clusters

Thanks for listening!