



# Tekstinlouhinnan mahdollisuudet Digin historiallisessa sanomalehtiaineistossa

Kimmo Kettunen

Dimiko (Digra-projekti)

# Tekstinlouhinta

- Tekstinlouhinnassa pyritään saamaan tekstimassoista automaattisesti esiin niiden sisältämää informaatiota
- Apuna käytetään yleensä tilastollisia ohjelmia, jotka etsivät tekstimassoista toistuvia hahmoja tai malleja
- Esimerkkejä: tekstin eristäminen ja luokittelu, tekstien klusterointi, käsitteiden/nimien eristäminen, dokumenttien tiivistelmien tuottaminen jne.
- Viime kädessä on kyse aineiston jalostamisesta ja helpommasta pääsystä kiinni aineistoon

# Digin sanomalehdet

- Vuosien 1771-1910 sanomalehtiaineistossa on noin 1.95 M sivua aineistoa (= useita kymmeniä miljoonia artikkeleita)
- Aineistoa voi hakea tällä hetkellä hakusanoilla ja tuloksena saadaan digitoitu lehden sivu, jolla hakusanan osumat on korostettu
- Ensimmäinen järkevä tekstinlouhinnan askel on alkaa tunnistaa artikkeleita digitoiduilta sivuilta automaattisesti → artikkelien eristäminen/erottaminen → indeksointi (haettavuus ja käytettävyys paranevat)
- Työtä on tehty alustavasti Digra-projektissa (Srikrishna Raamadhurai)



# Artikkelien eristäminen

Helsinki.fi x Flamma x Mail :: Search Results (Inb... x Bolzano\_2015 - Dropbox x Sanomalehdet - Digitoidu... x 0355-3787\_1982-05-18\_56... x 0355-3787\_1983-02-10\_17... x

file:///C:/Users/kikettun/AppData/Local/Temp/Temp1\_Maaseudun\_Tulevaisuus%20(1).zip/0355-3787\_1983-02-10\_17.html

## Tärkkelysperunalle yhtenäiset ehdot

Kaikki tehtaat käyvät ensi kesänä tärkkelysperunakaupaa yhtenäisin ehdoin. Tehäiden on noudatettava maa- ja metsätalousministeriön laadunvalvontapäätöstä sekä yhtenäistä viljelysopimusta. Maa- ja metsätalousministeriö, MTK ja tärkkelysteollisuus ovat yhdessä valmistelleet tärkkelysperunalle tarkkelys edellyttämän laatu hinnoittelupäätöksen, jota tehtaiden on pakko noudattaa. Samassa yhteydessä on sovittu yhtenäisestä viljelyso- pimuksesta.

Ylitarkastaja Ilkka Ruuska maa- ja metsätalouden ministeriöstä pitää yhtenäistä käytäntöä eteenpäin menona. "Kaikki viljelijät ovat ensi kesänä samassa asemassa toimitettavien satojen osalta", hän sanoo.

Ruskan mukaan tärkkelysperunakauppa on ollut tähän saakka hajanaista. Tehtaat ovat voineet itse päättää esimerkiksi tilitystavasta ja Ylitarkastajan maksamisesta. Nyt tärkkelysperuna tulee samaan asemaan kuin esimerkiksi öljykasvit.

## Lakot vaikuttavat lannoitetoimituksiin

Lakot jatkuivat eilen Kemi-rannan Siilinjärven ja Kokkolan tehtailta. Alkuvuoden tuotantoa jarruttaneet lakot ovat häirinneet lannoitetoimituksia ja pitkäaikaan jatkuessaan ne vaikuttavat keväällä lannoitetoimituksiin kotimaassa. Vaikutukset ulottuvat myös vientitoimituksiin. Tällä hetkellä työt ovat käynnissä Uuden kaupungin, Oulun ja Harjan vallan tehtailta, joten lannoitetoimitukset näiltä osin pyörivät, todetaan Kemirasta.

## VUOSIKOKOUKSET

Etelä-Savo: Puumala

Maataloustuottajain yhdistysten

Hartola

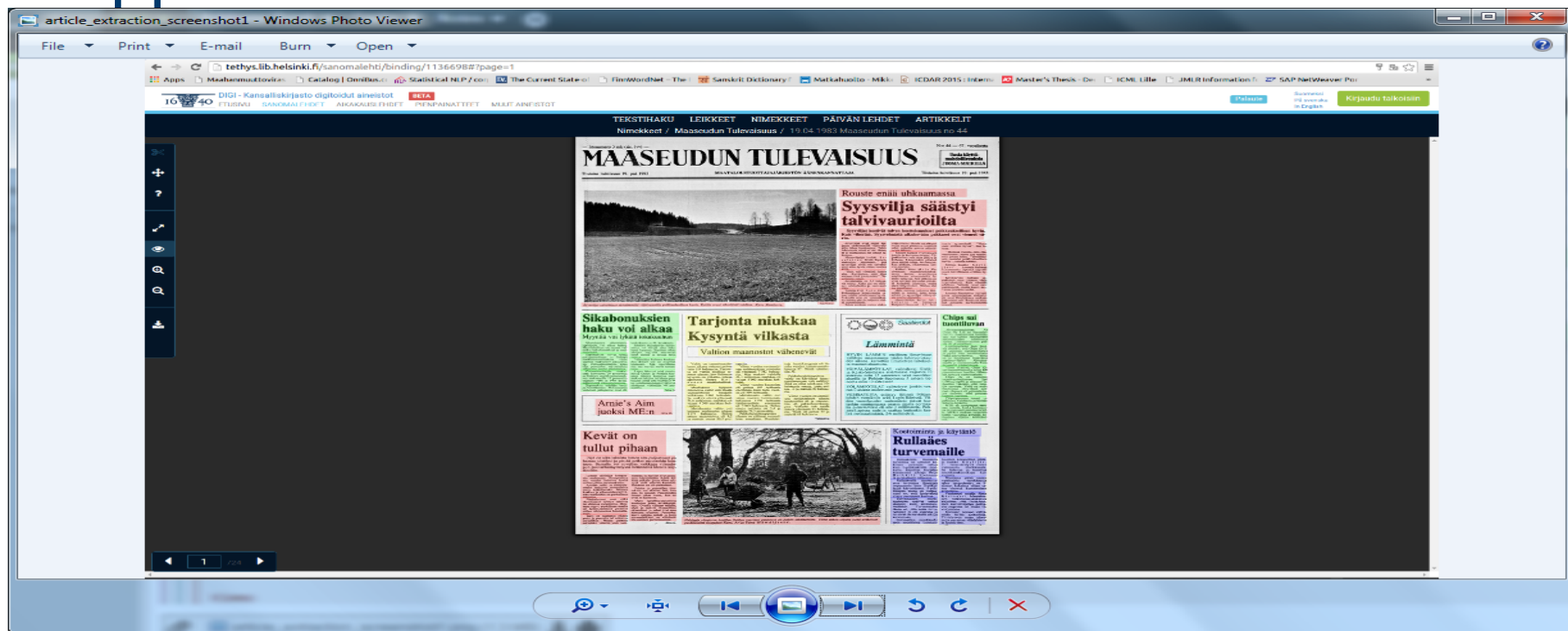
Joroinen

Kainuu: Kuhmo Puolanka Kajaani Paltamo Hyrynsalmi Keski-Suomi Sumiainen Konnevesi Pylkämäki Kivijärvi

Maaseudun\_Tulevaisuus...zip

Show all downloads...

# Artikkelien eristäminen – mahdollinen lopputulos



## Artikkelien eristäminen

- Kehitetty menetelmä perustuu koneoppimiseen: ohjelma oppii malliaineistoista artikkelien ominaisuuksia ja osaa sen jälkeen erotella artikkeliteita lehdestä
- Kukin lehti tarvitsee oman opetusaineistonsa, koska lehtien layout vaihtelee (myös lehden sisällä eri vuosikymmeninä/vuosina on muutoksia)
- Artikkelien eristäminen sisältää myös artikkelin osien luokittelun (otsikko, kuvateksti, kirjoittaja, ilmoitus, teksti jne.)
- Eristämisen jälkeen artikkelit voi indeksoida ja näyttää Digin käyttöliittymässä

# Artikkelien eristäminen

- Toimivan artikkelien eristämisen etuja, esimerkkejä:
  - Haut voisi käyttöliittymässä kohdistaa tarkemmin (esimerkiksi ilmoitukset)
  - Käyttäjä saa artikkelit helpommin käyttöön (ilman leikkaamista&liimausta)
  - Sisällön indeksointi hakukoneisiin toimisi paremmin
  - Jne.



## Nimien tunnistaminen

- Yksi sovellus tekstinlouhinnalle on nimien tunnistus tekstistä/ artikkeleista (NER, named entity recognition)
- Henkilöt, paikat, valtiot, yritykset, ajankohdat, määrät jne.
- Nimien käyttö hakusanana on tyypillistä Digin käyttäjille: kolmen vuoden käyttäjälökin 1000 yleisimmän hakutermin joukosta 80 % on nimiä:

Etunimet	30 %
Sukunimet	30 %
Paikannimet	20 %

## HFST-SweNER – A New NER Resource for Swedish

Dimitrios Kokkinakis<sup>§</sup>, Jyrki Niemi<sup>±</sup>, Sam Hardwick<sup>±</sup>,  
Kristen Lindén<sup>±</sup>, Lars Borin<sup>§</sup>

<sup>§</sup>Språkbanken, Department of Swedish  
University of Gothenburg, Sweden  
E-mail: first.last@svenska.gu.se

<sup>±</sup>Department of Modern Languages  
University of Helsinki, Finland  
E-mail: first.last@helsinki.fi

# HFST-SweNer (LREC 2014)

1. **Person** (PRS): people names (forenames, surnames), animal/pet names, mythological etc.;
2. **Location** (LOC): functional, geographical, geo-political, astronomical, street names;
3. **Organization** (ORG): political, athletic, media, military, transportation, education etc.;
4. **Artifact** (OBJ): food/wine products, prizes, means of communication (vehicles), etc.;
5. **Work&Art** (WRK): printed material, names of films, novels and newspapers, sculptures, etc.;
6. **Event** (EVN): religious, athletic, scientific, cultural, races, championships, battles, etc.;
7. **Measure/Numerical** (MSR): volume, age, index, dosage, web-related, speed etc.;
8. **Temporal**<sup>2</sup> (TME).

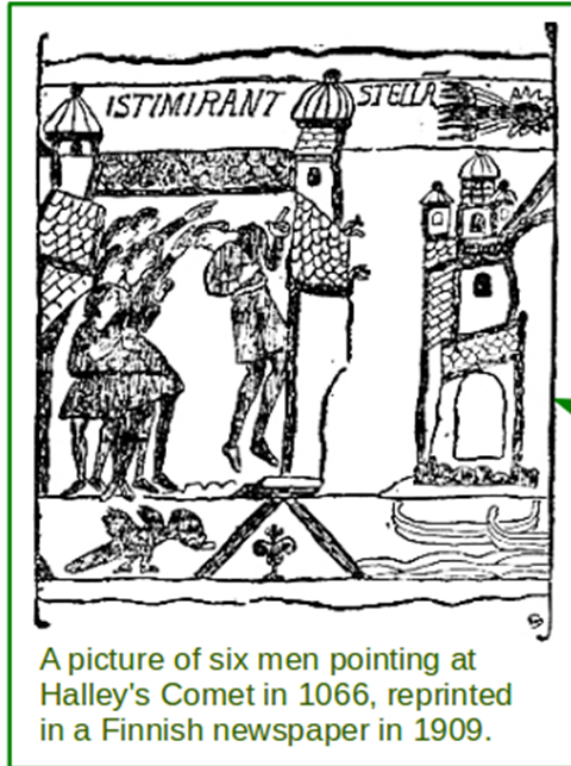
## HFST-SweNer (LREC 2014)

(a) <s id=jc04-041> Han kom till <ENAMEX TYPE="LOC" SBT="PPL">Stockholm</ENAMEX> <TIMEX TYPE="TME" SBT="DAT">1885</TIMEX> , fick en organisatorisk bas i <ENAMEX TYPE="ORG" SBT="PLT">Socialdemokratiska klubben</ENAMEX> ( senare <ENAMEX TYPE="ORG" SBT="PLT">Socialdemokratiska förbundet</ENAMEX> ) , kunde starta <ENAMEX TYPE="PRS" SBT="HUM">Social-Demokraten</ENAMEX> och fick medarbetare som <ENAMEX TYPE="PRS" SBT="HUM">Axel Danielsson</ENAMEX> , <ENAMEX TYPE="PRS" SBT="HUM">Fredrik Sterky</ENAMEX> och , <TIMEX TYPE="TME" SBT="DAT">året därpå</TIMEX> , <ENAMEX TYPE="PRS" SBT="HUM">Hjalmar Branting</ENAMEX> . <s>

# Miksi käyttää NERiä?

- Nimien tunnistus on oleellinen osa informaation eristämistä
- Informaation suodatus (käyttäjälle näytetään kaikki artikkelit, joissa esiintyy Sibelius)
- Informaation linkitys (esimerkiksi kaikki Sibeliuista käsittelevät artikkelit linkitetään toisiinsa kun nimet tunnistettu)

Uutiskartta  
- konsepti ,  
jossa  
samaa  
teemaan  
liittyvät  
ovat  
lähekkäin.



## AURORA BOREALIS

19.02.1887  
Kaiku 14:2

13.07.1910  
Työmies 157:5

14.02.1908  
Tornion Uutiset 12:3

19.10.1907  
Ilkka 120:4

05.01.1897  
Mikkelin Sanomat 1:3

21.11.190  
Otava 133:3

28.09.1909  
Uusi Aura 223B:1

## ZEPPELIN

21.10.1909  
Karjalan Sanomat 118:3

04.07.1900  
Päivälehti 153:3

02.11.1881  
Finlands Allmänna Tidning 253:3

## HALLEY

## Yhteenveto

- Digitaalinen sanomalehtikokoelma tarjoaa rikkaan aineiston tekstinlouhintaan, esitetyt kaksi esimerkkiä kertovat työstä jota on vasta aloitettu tai ollaan aloittamassa myöhemmin
- Kaiken Digissä tehtävän tekstinlouhinnan päämäärä on
  - jalostaa aineistoa
  - helpottaa ja monipuolistaa aineiston käyttöä
  - tuottaa parempia aineistoja jotka mahdollistavat paremman tutkimuksen