
# Tekstinlouhinnan mahdollisuudet Digin historiallisessa sanomalehtiaineistossa

Kimmo Kettunen

Dimiko (Digra-projekti)

1640 KANSALLIS KIRJASTO

# 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

16 40
KANSALLIS KIRJASTO

# Digin sanomalehdet

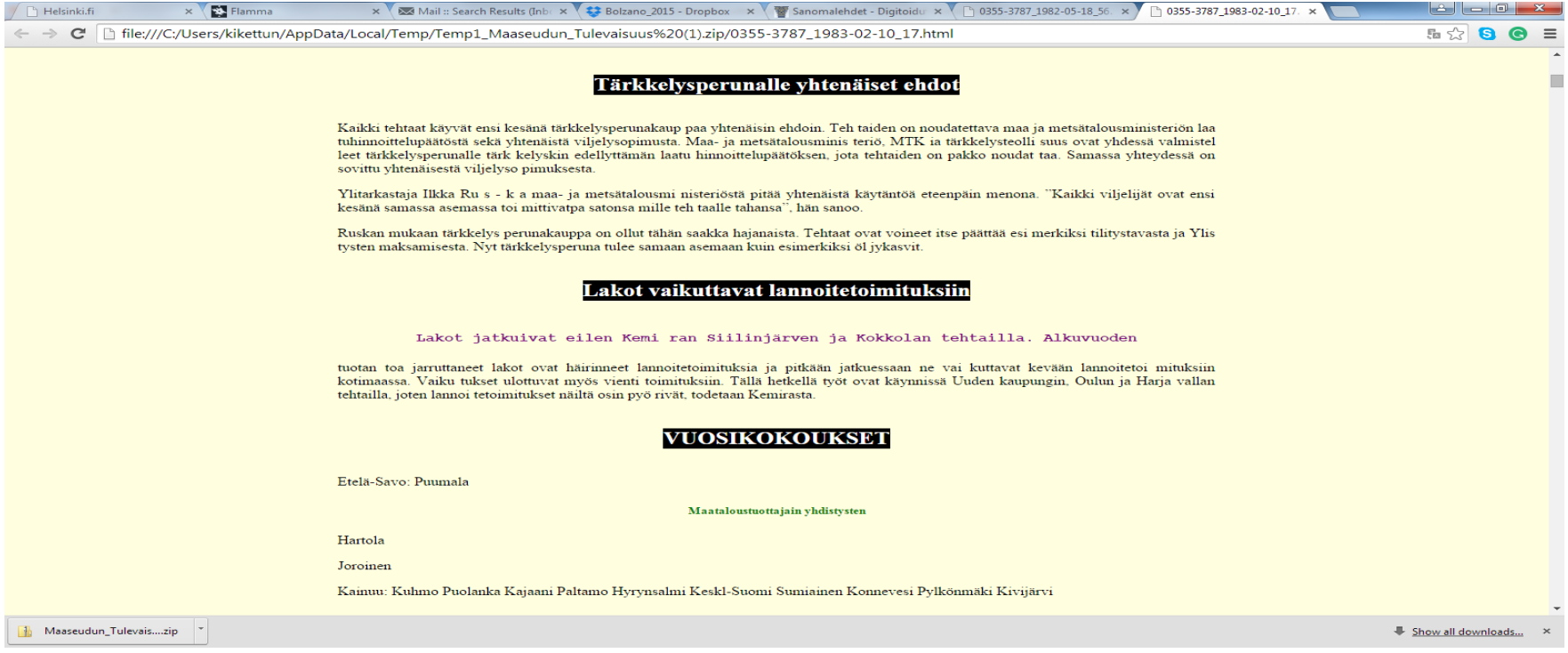
- Vuosien 1771-1910 sanomalehtiaineistossa on noin 1.95 M sivua aineistoa (= useita kymmeniä miljoonia artikkeleita)
- Aineistoa voi hakea tällä hetkellä hakusanoilla ja tuloksena saadaan digitoitu lehden sivu, jolla hakusanan osumat on korostettu
- Ensimmäinen järkevä tekstinlouhinnan askel on alkaa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

# Nykytilanne – 1.95 miljoonaa sivukuvamöykkyä


DIGITOINTI- JA KONSERVOINTIKESKUS

KANSALLIS KIRJASTO

# Artikkelien eristäminen



# Artikkelien eristäminen – mahdollinen lopputulos

1640 KANSALLIS KIRJASTO

# Artikkelien eristäminen

- Kehitetty menetelmä perustuu koneoppimiseen: ohjelma oppii malliaineistoista artikkelien ominaisuuksia ja osaa sen jälkeen erotella artikkeleita 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

1640 KANSALLIS KIRJASTO

# 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 käyttäjälokin 1000 yleisimmän hakutermin joukosta 80 % on nimiä:

| | |
|---|---|
| Etunimet | 30 % |
| Sukunimet | 30 % |
| Paikannimet | 20 % |

KANSALLIS KIRJASTO

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

**Dimitrios Kokkinakis[§], Jyrki Niemi[±], Sam Hardwick[±], Krister Lindén[±], Lars Borin[§]**

[§]Språkbanken, Department of Swedish
University of Gothenburg, Sweden
E-mail: first.last@svenska.gu.se

[±]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**[2] (TME).

16 40
KANSALLIS KIRJASTO

# HFST-SweNer (LREC 2014)

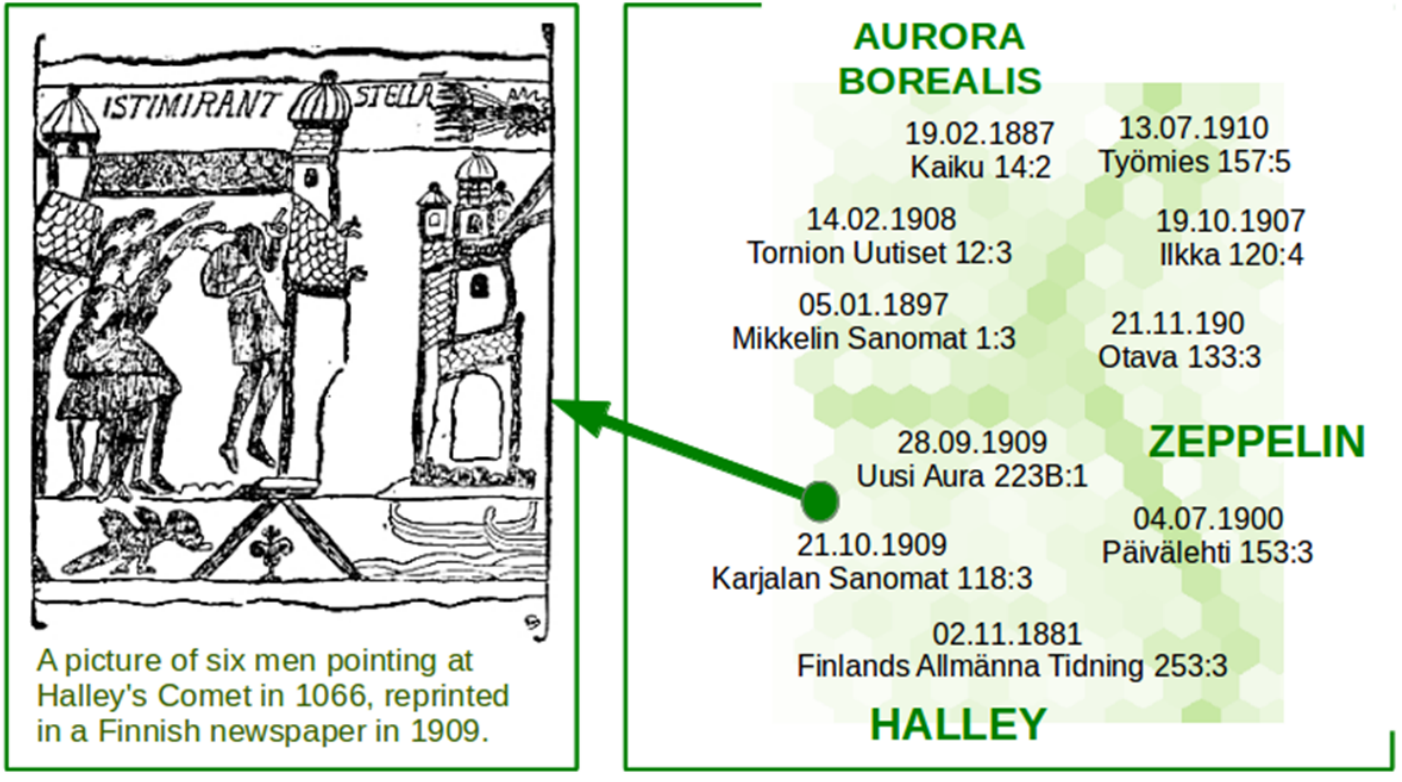
(a) <s id=jc04-041> Han kom till <ENAMEX TYPE="LOC" SBT="PPL">Stockholm</ENAMEX> <TIMEX TYPE="TME" SBT="DAT">1885</TIMEX> , fick en organisatorisk bas i <ENAMEX TYPE="ORG" SBT="PLT">Socialdemokratiska klubben</ENAMEX> ( senare <ENAMEX TYPE="ORG" SBT="PLT">Socialdemokratiska förbundet</ENAMEX> ) , kunde 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>

KANSALLIS KIRJASTO

# Miksi käyttää NERiä?

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

Uutiskartta - konsepti , jossa samaan teemaan liittyvät ovat lähekkäin.



1640 KANSALLIS KIRJASTO

# 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