Software Newsroom

Company partners

Sanoma News
2General
Lingsoft
STT-Lehtikuva

Research partners

University of Helsinki
Aalto University

Juhani Huovelin, Department of Physics, University of Helsinki

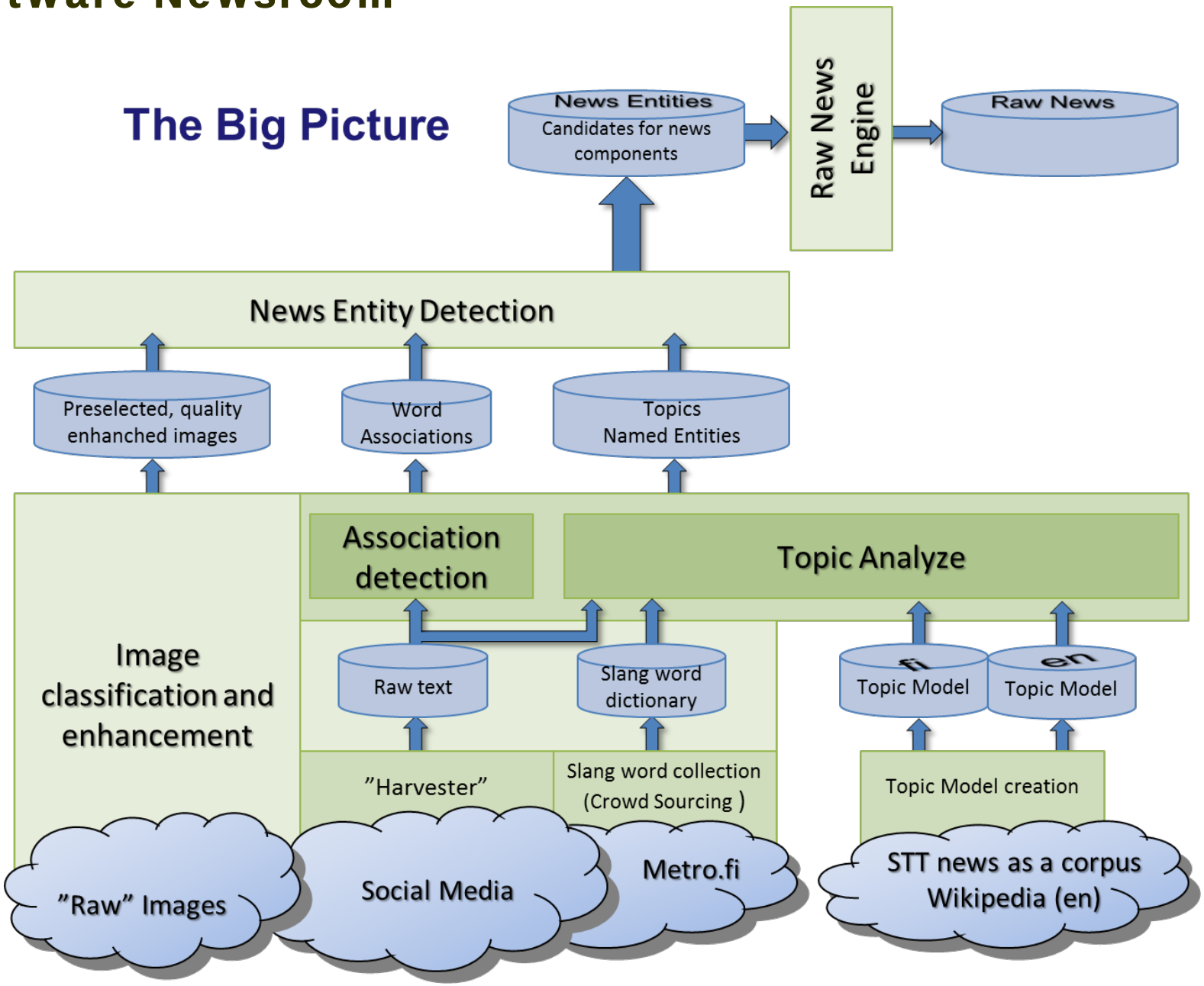
QUESTIONS AND TOPICS

- Is a computer algorithm able to find news topics from open data in the internet ?
- Searching news from social media
- How much of the editors' work (news authoring) can be done automatically by computer ?
- Commercial applications of results ?

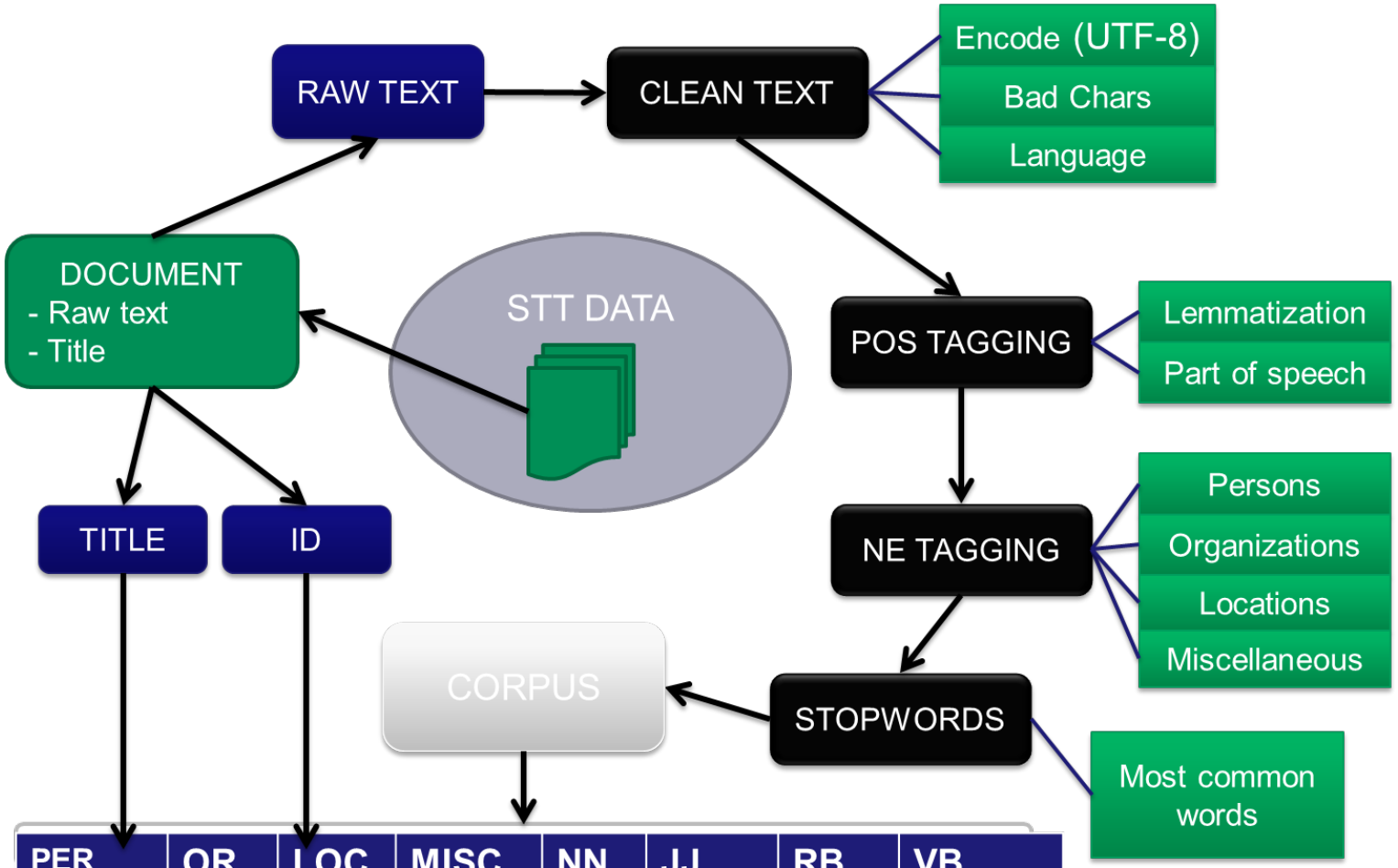
Vision

SEARCH ENGINE
Data mining
Algorithms
Machine learning
NEWS SEARCH
SOURCES
Social media
News agencies
Traditional media
Hobby
Voluntary co-operators
SEARCH RESULTS
SOFTWARE NEWSROOM
News material collection and preprocessing
Textual and image material as news sources
Topic and association detection and tracking
"Semi-automatic" news production
NEWS COMPUTING

Software Newsroom



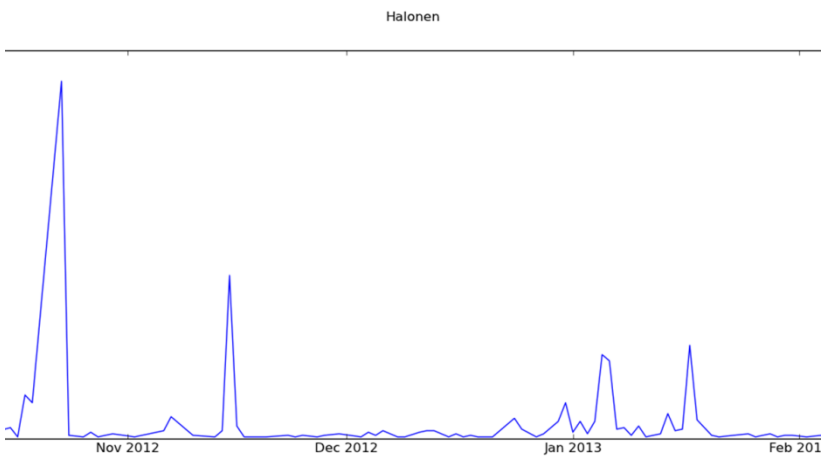
Topic model creation

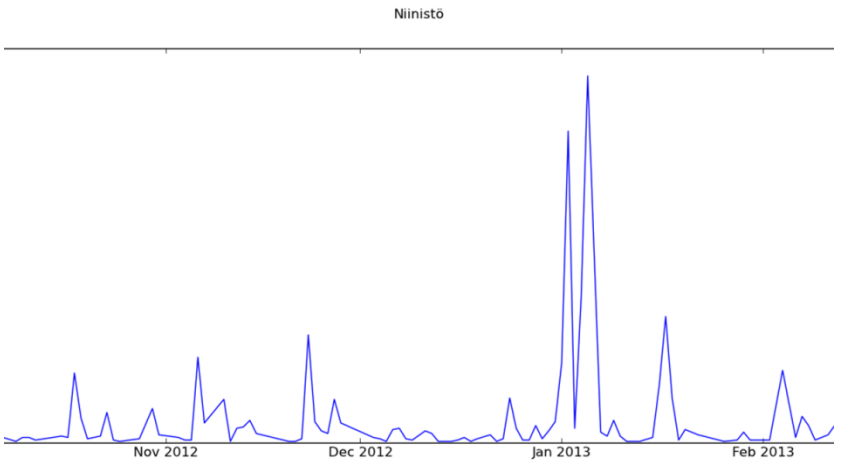


PER	ORG	LOC	MISC	NN	JJ	RB	VB
Napolitano	Hallitus	Italia	presidentti	valinta	radikaali	lisää	Valinta

News search

1. Projection of individual articles to a topic model to find the associated (most probable) topic and named entities = what / who is written about in each of them
2. If a topic or same named entity occurs frequently in a short time window, or the topic is otherwise interesting or anomalous (e.g. “accident”), the associated material is labeled as a “news entity candidate”
3. The candidate topic short term occurrence frequency is compared to longer term average in the chosen history, i.e. “normal” occurrence frequency
4. Significant deviation from “normal” or trend → potential news





Software Newsroom

Association discovery from social media, another method for finding news candidates

Hannu Toivonen et al. (UH, Helsinki Institute of Information Technology)

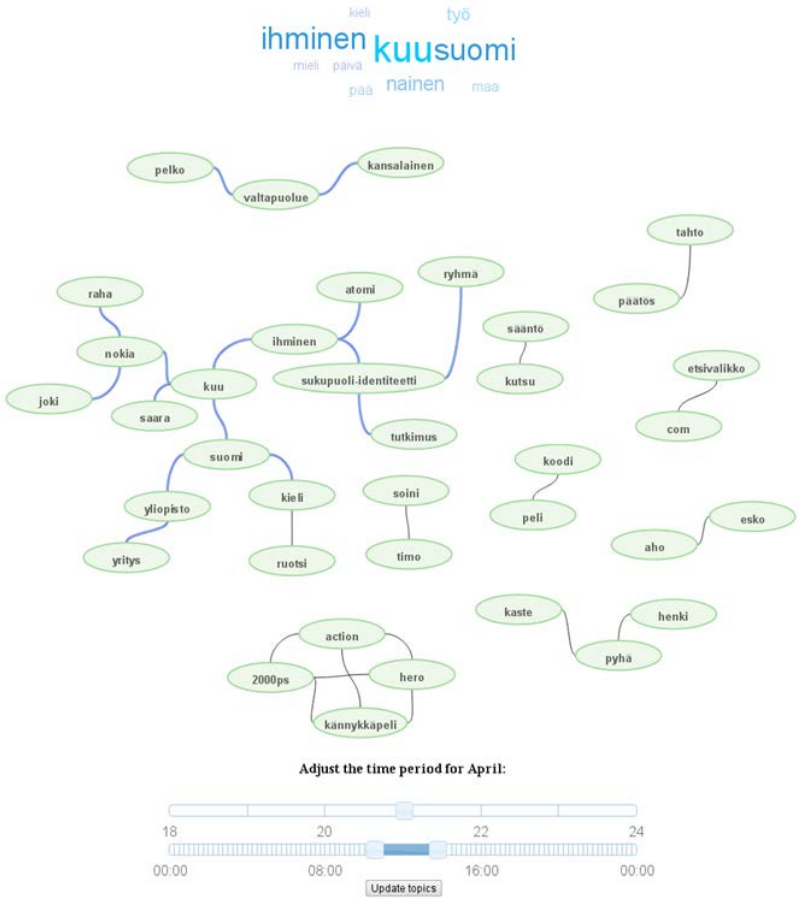
Goal: giving a quick overview of ”what's hot” in social media in a given time frame

Approach: Extract word pairs from social media posts written in the time frame

Problem: Lot of word pairs are ”common sense” and thus non-informative

Solution: Highlight connections between concepts which are specific to the time frame by contrasting the current connections to the ones in the history

Methods: Statistical analysis, e.g. “Log-Likelihood Ratio” (LLR, G squared statistics) and further developments.



Ongoing and future research

Improvement of topic models

- Improved Finnish corpus and named entity database
- Understanding of Finnish social media language (“slang word” dictionary)
- Iterative / hierarchical topic modelling and analysis
  - e.g. sports → football, ice hockey, volleyball, golf, ...
- English and Swedish corpora and named entities

Combining of topic mining and association search

- Improvement of the efficiency and quality of news search

How much the editing of news can be automated ?

- Investigation of methods for automated authoring of news articles from unstructured data (i.e. raw news entities)

Software Newsroom

Business ideas and commercialization prospects

- 2General has been developing and iterating on various approaches for commercialization of the Software Newsroom results

- Integration of the Software Newsroom technology to a complete in-house developed publishing solution that would add a new category of articles based on data gathered from social media

- Standalone social media analysis software that would allow mining the social media data collected and analyzed with the Software Newsroom technology

- Hyperlocal news publishing platform that would allow a local media publisher to operate a profitable online business with one journalist only

- Mobile application that allows freelance journalists to manage their multi platform publishing workflow from start to end

- Several spinoff projects utilizing the media insight gathered in the project, especially in the field of newsroom operations automation

Software Newsroom

Contacts (research)

Juhani Huovelin, juhani.huovelin@helsinki.fi

Henri Kivinen, henri.kivinen@aalto.fi

Krister Lindén, krister.linden@helsinki.fi

Hannu Toivonen, hannu.toivonen@cs.helsinki.fi