Data Processing Pipelines in Data Journalism: Preparing Insights for Editorial Use

Text Mining and Data Engineering in the Field

Abstract

We present two trends in data engineering applications that pose novel challenges. On the one hand, Data Journalism is an emergent idea of applying regular data analytics techniques to data of public interest, with the intent to generate newsworthy insights. On the other hand, publishers are increasingly relying on data acquisition techniques combined with Natural Language Processing (NLP) to aid editors.

Data Acquisition and Preparation

There are a wide variety of potential sources that can be considered for scraping, differing both in the usual response time or publication frequency and the editorial quality of the content. Some formats, such as audio recordings, impose unique challenges when integrated into a textual publishing workflow. Other formats, such as Tweets or public databases, require precise filtering to be useful or even just usable.

Analysis

The nature of the analysis methodology obviously depends on the type of source material at hand.

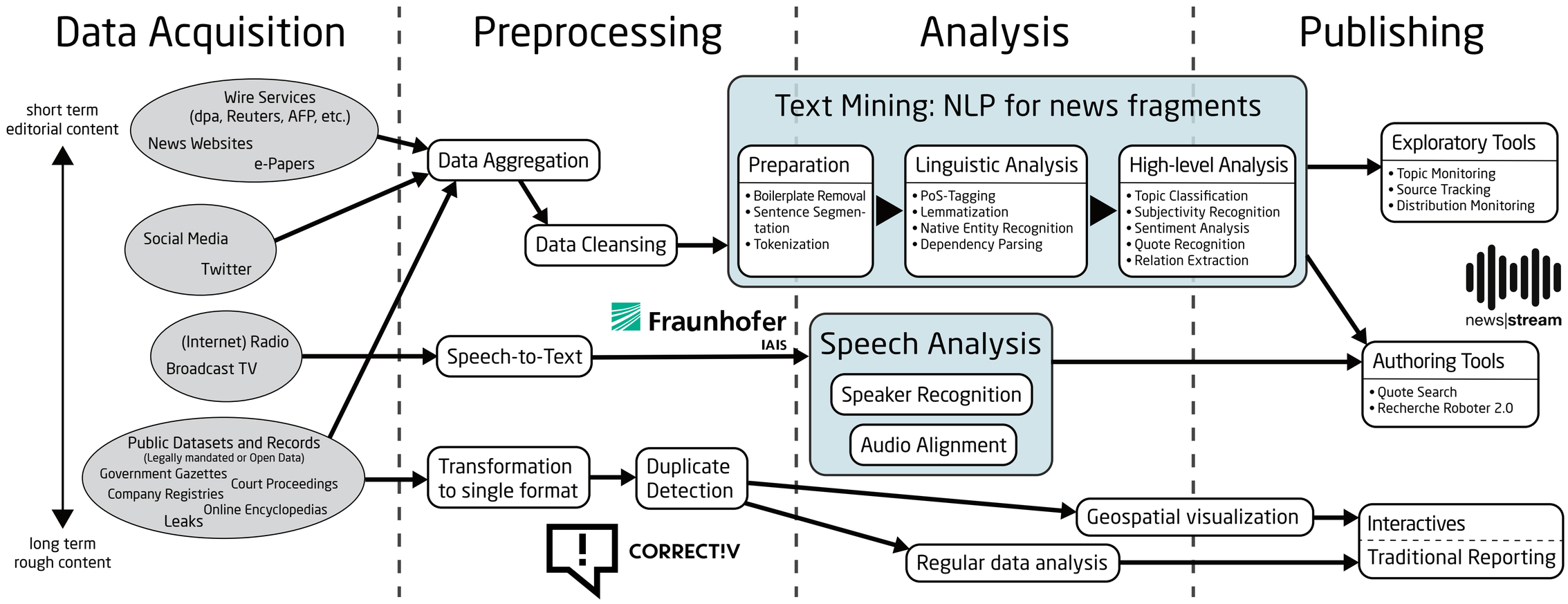
In the case of newsstream, textual records are put through a Text Mining pipeline that generates insights:

- General Information, such as topic, language, and guessed industry/sector
- Subjectivity and Sentiment of the author
- Mentioned named entities, their roles and quotes

Publishing

How does a newspaper take advantage of newsstream's analytics? By using tools provided by newsstream that aid the editor:

- Aiding the exploratory process by informing about the current trending topics and press coverage
- Helping the editor by providing background information and quotes when needed



Tobias Markus (Bachelor)
E-Mail: tobias.markus@student.hpi.de

References

Wehrmeyer, Stefan: *Data Engineering in the Newsroom*, 11/2017
Adolphs, Peter: *Eating News from the Web*, 01/2018

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam