Contact: Måns Magnusson, STIMA. email: mans.magnusson@liu.se

3 Finding probabilistic journalistic news leads in large scale public data

3.1 Introduction

Many news rooms, and especially financially strained local news rooms, struggle to find ways to use data in the journalistic context. This field of journalism, often called data journalism, try to find or confirm stories using data and statistical methods. Examples range from handling the enormous corpus of the Panama papers leak to visualizing complicated data for readers.

The large upswing in using data in news rooms coincide with the fact that public data are becoming more and more easily accessible. In these large dataset there exist potential news leads, such as anomalies and new trends that are of journalistic interest to study further. There has been some few approaches to automatic surveillance of public data sets [5, 3] but the field is generally not very well researched.

Together with the trends in using data in journalism and the huge amounts of publicly available datasets there is also an increasing interest in automated journalism or robot journalism. To automate the journalistic work flow multiple parts in the process needs to be automated. The journalistic work flow can simplified be described as four steps [1], collect data, identify the news story, write the article and disseminate the article. All these parts are more or less difficult to automate. The writing parts needs a Natural language system and the identification of the news story often need statistical approaches.

3.2 Purpose

The purpose of this project is to study different Bayesian models for automatic data surveillance in the data journalistic context. The purpose is to study which types of bayesian unsupervised and semi-supervised models that can be used to identify potential news leads in univariate and/or hierarchical time series data. One of the questions to regard is if there are different statistical model considerations needed for the journalistic context compared to similar statistical fields such as statistical outbreak detection and statistical control theory.

3.3 Related work

This work is closely connected to statistical public health monitoring as well as statistical control theory.

References

[1] Dhrumil Mehta Al Johri, Eui-Hong (Sam) Han. Domain specific newsbots, live automated reporting systems involving natural language communication. 2016.

[2] Matthew James Denny and Arthur Spirling. Assessing the consequences of text preprocessing decisions. *Available at SSRN 2849145*, 2016.

[3] Måns Magnusson, Jens Finnäs, and Leonard Wallentin. Finding the news lead in the data haystack: Automated local data journalism using crime data. 2016.

[4] Alexandra Schofield and David Mimno. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4:287–300, 2016.

[5] Eirik Stavelin. Computational journalism. when journalism meets programming. 2014.