Information Retrival and Text Mining Proposal

Sam Sweere (i6231098), Alexander Reisach (i6197692)

April 15, 2020

1 Introduction

This project will be a part of the course KEN4153 Text Mining and Information Retrieval. As such, we will use the methods explained in said course to mine an extensive data set so to be able to automatically answer content related questions about actors, locations, events, and trends. Given the recent global developments, we will use a large publicly available text corpus on Covid-19 news coverage as the basis for our analysis.

2 Data Set

Our data set is a pre-aggregated selection of 500,000 news articles on the Covid-19 global pandemic[1]. It covers articles published between November 1st 2019 and April 6th 2020. The data set comprises publications from 400 prominent news sources with an Alexa [2] rank below 5,000. It contains English language publications only. Moreover, the data set has been pre-processed and enriched by entities recognised, topical category tags, a sentiment measure and summaries. Source information and date of publishing are included as well. The data set contains the following keys:

{*'author', 'body', 'categories', 'characters_count', 'entities', 'hashtags', 'id', 'keywords', 'language', 'links', 'media', 'paragraphs_count', 'published_at', 'sentences_count', 'sentiment', 'social_shares_count', 'source', 'summary', 'title', 'words_count'*}

Usage of the data set id open for researchers, scientist and any other form of non-commercial analysis.

[1] https://aylien.com/coronavirus-news-dataset/
[2] https://www.alexa.com/topsites/category/Top/News

3 Overview over Approach

In this section we describe what information we want to extract out of our data set. A lot of the appropriate technical approaches are still to be discussed in the lectures. It is therefore hard to specifically determine what techniques could work well for our data set. For every part of this project we will determine the best technique when we have discussed it in the lecture or when we need it to reach our deadlines described in section 4.

3.1 Pre-Processing

Before doing text analysis we first have to do some pre-processing of the data. The dataset already contains separate keys for headlines, body, authors, etc. This saves us a lot of time, since we do not have to text mine these articles ourselves.

However the text still consist out of plain text which we will have to structure to be able to run our data enrichment procedures. Depending on what we need for each of the types of data enrichment we plan to do, our pre-processing of the news articles will consists of normalization, case folding, grouping of similar words/synonym detection, tokenization/word indexing, etc.

3.2 Data Enrichment

Our dataset already contains some extra information which was not part of the source material such as a sentiment measure. We will not use this for our analysis except to possibly compare our results. There is also a lot of information yet to be extracted. In this section we will describe what information we want to extract form the data set.

3.2.1 Topics & Entities

First off, we will examine what the Covid-19 global pandemic is in itself. We will aim to detect how the development from a local cluster of respiratory diseases to a global pandemic is reported by leading global news outlets by asking the question "What is happening" and "How much is happening" throughout the Covid-19 health crisis.

Furthermore, we will extrapolate a time line of key locations and that receive the most global coverage. For example, we expect to see a rise in coverage of affected locations starting with Wuhan, then Italy and now new York.

Finally, we will also collect the topics that most affect people globally throughout the development of the epidemic. This will likely include topics such as nation-wide lock-downs, the number of infections, the health care system, and economic repercussions of the crisis.

3.2.2 Trends and Events

For all of these topics we will analyse how they develop over time to gain insights into trends and the events that may be shaping them. In terms of events, we expect to find for example the detection of the first cases outside China, inaction of curfews and state of emergency declarations by major countries.

3.2.3 Sentiment Analysis

We will track the sentiment of the articles over time. We expect that it will start more neutral, become more negative and might become more positive towards the end. This could be correlated to specific trends and events to see what possibly changed the sentiment.

3.3 Visual Presentation of Results

If the scope of the project allows it, we would like to present our results in a way that is visually appealing and intuitive for a wider audience. This could include a time-animated world map of locations most associated with the Covid-19 health crisis, or a topic stream visualization of global trends and events. Furthermore, it could be interesting to combine our results with data on the spread of the epidemic measured in reported number of infections.

4 Planning

Milestones / Dates	24.04	06.05 Intermediate Report	15.05	20.05	27.05 Final Report
Code Framework	X				
Pre-Processing	X				
Topics and Entities		X			
Trends and Events			X		
Sentiment Analysis			X		
Visual Presentation				X	
Final Report				X	X

Table 1: Gantt chart for our project planning. All dates for the year 2020. Milestones in italic will be added if the timing and scope of the project allows it.

Table 1 shows our the milestones/deadlines. Among the deadlines, the ones not further specified are self-imposed. We already have our data set and are able to read it in from a data stream. Thus we will start by developing an access framework to conveniently extract specific articles. Alongside, we plan to implement the first stage of basic text pre-processing. We set the deadline for this part to the 24.04. Thereafter we will start working on extracting topics and entities. We would like to complete this segment before the intermediary update on 06.05. Next, we will start to extract trends and events from the change in topics and entities over time. If time allows it, we will also use this period to conduct an analysis of sentiment. We set the deadline for this part on 15.05. We hope that this will allow us to work on a more elaborate visual presentation of our results until the 20.05. so as to make them more accessible to a wider audience. This leaves us one week to finalize the report and prepare the presentation until the 27.05.