NWA Startimpuls VWData

P2: Accuracy

Report P2.1:

CaptureBias Datasets

1 July 2019

Markus de Jong, Xander Wilcke, Panagiotis Mavridis, Alessandro Bozzon, Alec Badenoch, Antoaneta Dimitrova, Honorata Mazepus, Jesse de Vos, Johan Oomen, Lora Aroyo, Tobias Kuhn

Commit 2 Data

connecting business and science

Introduction

The primary dataset used in this WP consists of online news videos and their metadata, gathered from YouTube: the popular online video platform. In this rapport, we describe this dataset in more detail, as well as the various subsets and derivatives thereof, and the task we use them for.

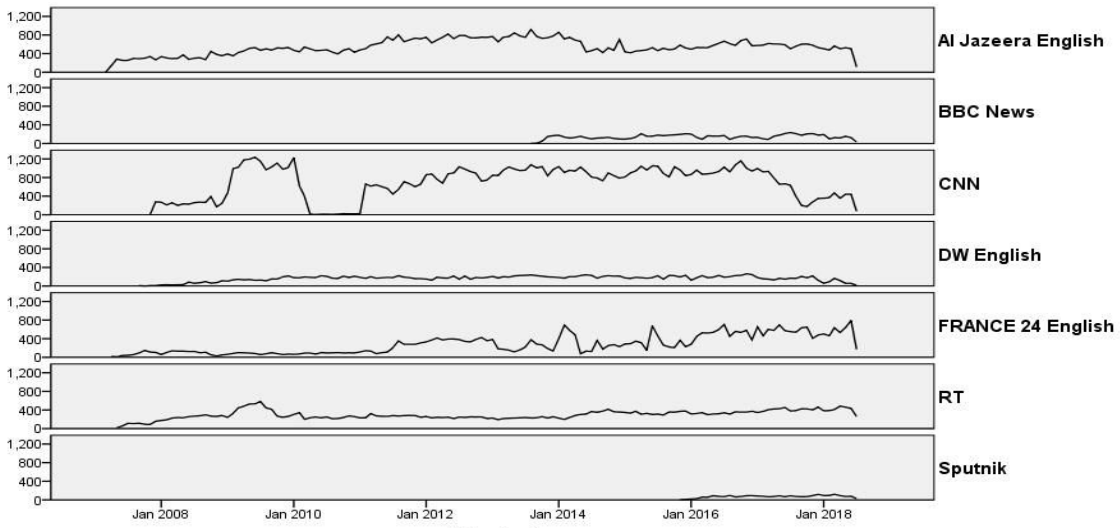

Figure 1: Video publishing timeline per channel

YouTube 280k Dataset

At the time of writing, the entire dataset comprises all videos that were published by 7 selected video channels (Table 1) from their first publication (ca. 2007) up till 2018 (Fig. 1). The selected channels were chosen such that they were likely to hold slightly to largely different views of the same events, and would therefore be suitable to analyse bias with.

Table 1: Selected channels and the country they operate from.

Channel	*Country*
RT (previously Russia Today)	Russia
BBC News	United Kingdom
CNN	United States of America
Sputnik	Russia

Al Jazeera English	Israel
DW News / Deutsche Welle English	Germany
France 24 English	France

The combined dataset from the 7 selected video channels holds close to 275.000 videos. We will refer to this as the **YT280k** dataset.

This number of 275.000 is split unevenly over the channels, with *CNN* (ca. 89k) and *Al Jazeera* (ca. 74k) contributing the most, and with *Sputnik* (ca. 2.3k) and *BBC News* (ca. 8.5k) contributing the least. A similar difference can be seen in their support of captions: only *DW* and *Al Jazeera* seem to take captions seriously, with a coverage of 96% in both cases. However, these captions have almost all been generated automatically by YouTube, which may lowers their quality. In contrast, only *CNN* seems to actively supply their videos with manual captions.

Table 2: YT280K data set video count and caption coverage.

channel	Total videos	Manual captions	Auto-generated captions	Caption coverage (%)
RT (Russia Today)	39952	75	30321	76%
Sputnik	2308	0	2	0%
CNN	89448	24560	24554	55%
France24	37679	252	487	2%
Al Jazeera	74004	267	70567	96%
DW	21338	79	20425	96%
BBC News	8554	1	6	0%
Totals	273283	25234	146362	

Video Metadata

The metadata belonging to these videos has been retrieved along with the video streams themselves. For each video, the metadata lists technical information of the video stream (e.g. duration and dimension), as well as contextual information about its publication (e.g. uploader and like count). A complete overview of the information covered in the metadata is listed in Table 3. The metadata themselves can be accessed on our google drive (link).

Table 3: The YT280k dataset metadata attributes.

Attribute	Description	Sample data
display_id	Youtube video identifier (also used to construct video URL)	*IVVILTD3B10*

title	Given title	*Yemen southerners push for independence*
fulltitle	Given full title	*Yemen southerners push for independence*
description	Given video description	*Thousands of Yemeni people, who are pushing for the southern part of the country to break away from the North, have camped out in the port city of Aden's Independence Square. Al Jazeera's Hashem Ahelbarra reports from Aden. Subscribe to our channel http://bit.ly/AJSubscribe Follow us on Twitter https://twitter.com/AJEnglish Find us on Facebook https://www.facebook.com/aljazeera Check our website: http://www.aljazeera.com/*
upload_date	Date of upload	20141130
duration	Duration (Seconds)	93
uploader	Channel that uploaded this video	Al Jazeera English
thumbnail	URL to video thumbnail	https://i.ytimg.com/vi/IVVlLTD3B10/maxresdefault.jpg
tags	Video tags	*3916124021001+Ali Salim al-Beidh+aden+package+yemen+Al Jazeera English+News+al Jazeera+jazeera+youtube+unblocked+independence+Hashem Ahelbarra+hassan baoum+aljazeera+south yemen*
categories	Video category	*News & Politics*
average_rating	Video rating (scale 0-5)	4.199999809
view_count	View count	1251
like_count	Number of likes	12
dislike_count	Number of dislikes	3
width	Video width	1280
height	Video height	720
ext	Video file format	mp4
format	Textual format description	136 - 1280x720 (DASH video)+140 - audio only (DASH audio)

acodec	Audio codec	mp4a.40.2
vcodec	Video codec	avc1.4d400c
automatic_captions	Does this video have automated captions?	0
subtitles	Does this video have manual captions?	0

Dataset Subcollections

The entire **YT280k** dataset is too vast and too diverse to allow for focussed experiments. Therefore, we have selected two proper subsets of this dataset onto which we run our experiments: the **YT15k** dataset, and the **YT2014** dataset. In case of the latter, we have additionally created two related datasets: 1) a collection of audio transcriptions (**YT2014T**), and 2) an enriched metadata knowledge graph (**YT2014KG**) to analyse the videos' *context*, as opposed to purely the *content* as provided by the **YT15k** and **YT2014** datasets. A brief description of these datasets will be given next.

YT15k Dataset

The **YT15k** subset is a collection of roughly 15 thousand randomly selected videos from the **YT280k** dataset (Table 4). This subset is specifically created for the task of text analysis, or Natural Language Processing, and therefore contains only videos that feature captions (Table 2). To remove caption quality from the list of potential variables, this subset only considers videos for which the captions have been automatically generated.

Table 4: Video distribution in the YT15k subset

channel	Total videos
RT (Russia Today)	2918
CNN	2570
France24	52
Al Jazeera	7225
DW	2108
Total	14871

YT2014 Dataset

The **YT2014** subset holds all online news videos from the **YT280k** dataset that were published in 2014. This time period was specifically chosen due to the occurence of two important events: the MH17 plane tragedy, and the beginning of the Crimea Crisis. We have chosen these two events as case studies given their frequent coverage and the controversies surrounding both events.

In total, this subsets contains roughly 28 thousand videos, the majority of which was published by *CNN* (Table 5). Missing from this subset is the Russian outlet *Sputnik*, which only started publishing online videos from 2016 onwards (Fig.1).

Table 5: Video distribution in the YT2014 subset

channel	Total videos
RT (Russia Today)	3235
CNN	10635
France24	3629
Al Jazeera	6262
DW	2261
BBC News	1504
Total	27522

YT2014T Dataset

As the main coverage of the Annexation of Crimea is contained between January and April of 2014, we have created a complete transcript set for the videos that were published in these four months. As YouTube only provides a subset of these transcripts, we have supplemented this incomplete set by running dedicated speech-to-text software on the audio feeds from the missing cases. The **YT2014T** dataset contains these audio transcriptions (Table 6).

Table 6: Video distribution in the YT2014T subset

channel	Total videos
RT (Russia Today)	150
CNN	1826
France24	2010
Al Jazeera	42
DW	7
BBC News	563
Total	4598

YT2014KG Dataset

The **YT2014KG** dataset is a knowledge graph, build upon the RDF data model, which contains all metadata from the videos in the **YT2014** subset. As such, this dataset, of which the construction is still in process, represents the videos' context, rather than their content. Said contexts holds various information on the video streams themselves, such as definition (SD or HD), dimension (2D or 3D), and projection (portrait or landscape), as well as related information, such as the channel that published them, the categories they belong to, and the countries they are restricted in (Fig. 2). This contextual information gives us an idea how the videos might be perceived by their viewers.

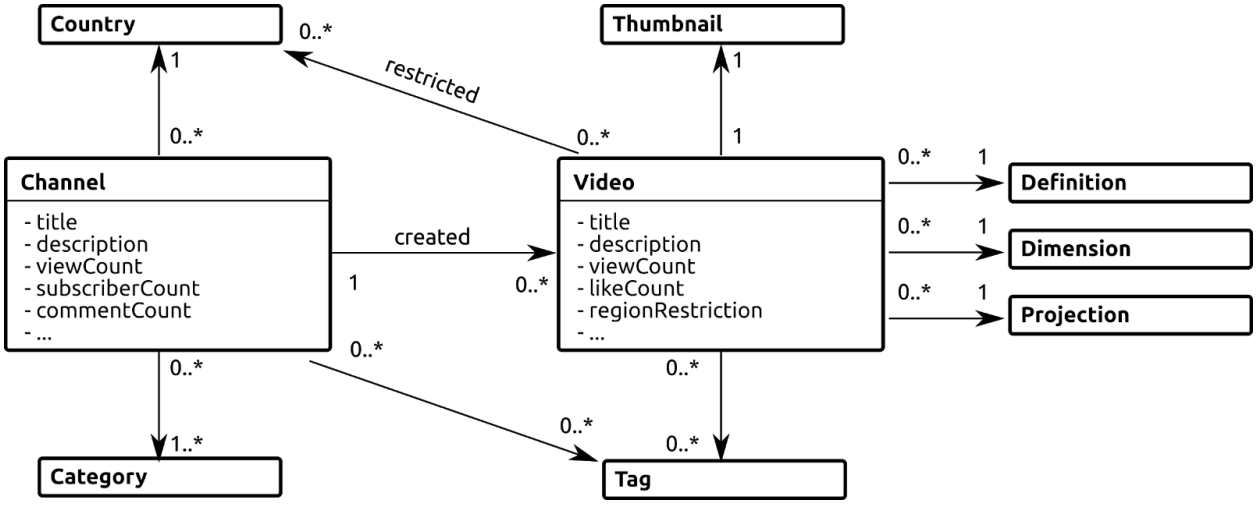

Figure 2: A simplified model of the pure metadata graph: channels hold multiple videos, and both channels and videos have shared tags, categories, and (restriction on) countries which link otherwise distant data points together by context similarity.

Being tied to the **YT2014** dataset, the knowledge graph contains an equal number of videos---27532 to be exact---which are represented as entities (vertices) in the graph (Table 7). Each of these video entities has roughly 15 attributes, and connects to at least eight other entities (channel, category, thumbnail, etc.). In total, the entire metadata graph contains roughly 55 thousand entities with ca. 248 thousand edges between them, as well as about 413 thousand attributes with as many edges. The schema used to define all these elements is also under development, and is published under the GPL3.0 license on our GitHub repository.

Table 7: Overview of graph elements in the YT2014KG dataset. Final figures may vary slightly as mutations to the dataset can still occur.

Graph element	Count
Entities (vertices)	55364
- Channels	6
- Videos	27532
- Countries	250
- Categories	32
- Thumbnails	27538
- Video stream info	6
Literals (attributes)	413070
Properties (edges)	660876
- Object-type	247806
- Data-type	413070

By using the RDF data model, we can easily extend our dataset on demand by linking it to external RDF datasets, such as WikiData and Geonames (Fig. 3). Doing so enables us to add additional context to the graph, for example by incorporating that the Crimea peninsula was once a part of the USSR, as was Russia.

The different datasets variants will be used in future experiments, which enables us to study the effects of varying levels of context on the performance of our models to capture bias.

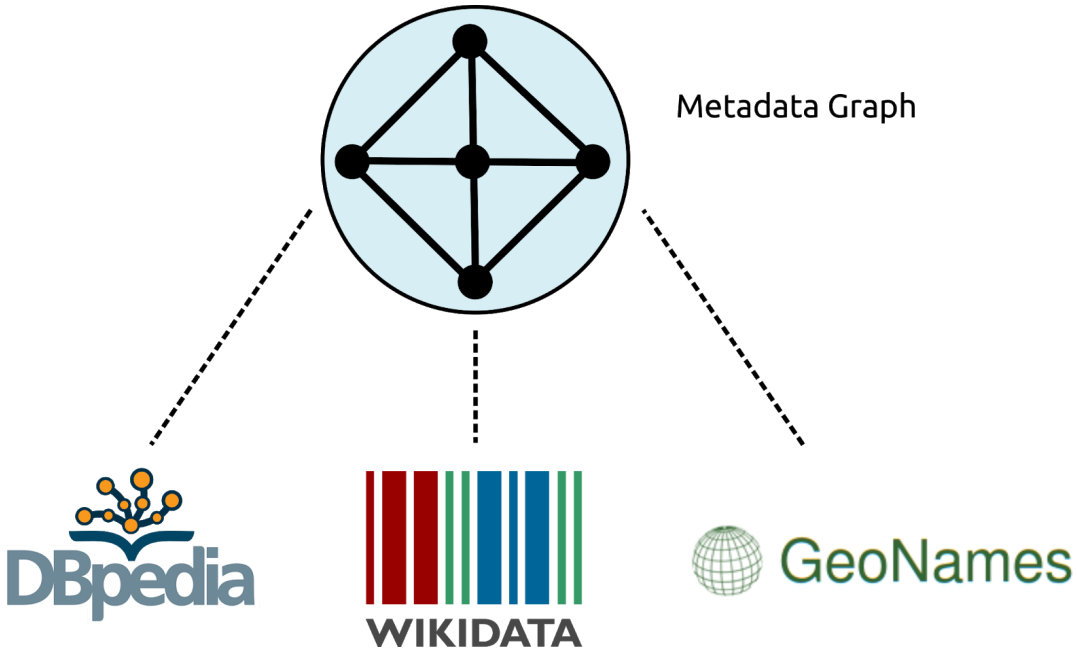

Figure 3: The pure metadata graph (Fig. 2) can be extended on demand by linking shared data points to external RDF sources (e.g. tags to DBpedia, categories to WikiData, and countries to Geonames).