Our Findings

From Objects to Data

From Objects to Data
dr. M.H.A Koolen
Culturele informatiewetenschap
Universiteit van Amsterdam

19 october 2014

Conall Sleutel
Jason Lam
Sjoerd Bergmans
Stefan Wisselink
Warner Nanninga

Table Of Contents

Preface	5
1. Research question	6
2. Approach	7
2.1 Initial approach	7
2.2 final approach	8
3. Results	10

PREFACE

This is the final report of Warner Nanninga, Stefan Wesselink, Conall Sleutel, Jason Lam en Sjoerd Bergmans. This report offers a short recap of what our research group has achieved during the *'Objects to Data'* course. In order to truly get a grip on what has been done, we would like to encourage anyone to visit our website:

http://objects-to-data-group.github.io/page/

The website contains a bit of background information about the relation between our research group and the 'Objects to Data' course in general, it also allows you to download our source-code in which a readmefirst file (READMEFIRST.md) can be found that explains how to install and run all of our final scripts.

The website shows some of our findings/results, though it is not as complete as this report.

This report contains information about our main research question, which will be addressed in the first paragraph. It also offers some insights into both our initial and final approaches. This can be read in the second paragraph. The third paragraph describes our current results and conclusions and the forth and final paragraph will detail some of our reflections.

1. RESEARCH QUESTION

Our research group started pursuing an answer to the question:

> *"What could be the most likely explanation(s) for the fact that the humanities stopped being featured within front-page articles from the 1990's and onwards?"*

This research question was based on a preliminary finding of Sjoerd that showed a deep drop of the amount of humanities articles that reached the front-page during the 1980's, while they seemed to completely disappear from the front-page after 1990. This drop was even more remarkable given the fact that before this moment the total number of published articles about humanities seemed to increase.

Combined with the knowledge that this preliminary research question would manage to score the highest grade, compared to the individual questions of the other group members, we decided to adopt this particular question for the group's research.

2. APPROACH

2.1 Initial approach

We decided that in order to explain what could be the cause of the decline of humanities front-page articles, we should first allow ourselves some time to gain a more detailed image of the evolution of front-page articles over time. To do so, we decided to start working on scripts that could answer some of the sub questions that where already addressed in Sjoerds preliminary proposal. For example, one of the questions is:

> *"How does the ratio between front-page and other articles (non front-page articles) evolve over different periods of time? (e.g. per month / year / day of the month, etc.) And how do they compare percentage wise? (e.g. what percentage of the total amount of humanities articles within the individual time periods reached the front-page?)"*
>
> *"How did the ratio between regular front-page articles and regular articles evolved over the same time-periods? And how does this evolution compare to that of the "humanities" articles?"*

These findings where meant to give us some preliminary proof or disproof that the decline of humanities front-page articles was or wasn't as dramatic as it seemed in concrete numbers, or that it was or wasn't as dramatic when compared to the evolution of regular front-page articles.

After answering these questions, our goal had been to develop some form of topic modeling in order to enable us to either prove or disprove a content-based cause to the drop in humanities front-page articles. With that, we would have covered both the 'outside' and the 'inside' of the problem.

However, our initial approached failed when the apparent drop in humanities front-page articles was labelled invalid by our teacher who pointed out that the data-field containing the front-page indication was simply changed after 1981.

2.2 final approach

Our initial approach landed us with 5 different scripts:

Script 1

Script 2 (old)

Script 2_2

Script 2_2 MAC OSX

Script 2_3

Script 2_3 MAC OSX

After this we gained a new insight and concluded that we needed to find a new research question. But, because we did not want to give up on all of our scripts, we decided to adapt them so that they would get their data from the correct data-field after 1981.

One might say that after this adaption, we sort of lost direction for a while. Now that our preliminary research question was invalid we started developing even more script to provide us with an even more detailed insight into the evolution of the humanities front-page articles, since we hoped that this would somewhere offer unexpected results that would then somehow justify the formation of a new research question based upon them. And so, more script where developed:

Script 4

Script 5

Script 6 (this is a very raw and unfinished script, that was meant to form a basis for a future script that would enable us to translate a date-stamp into a day of the week (e.g. Monday, Tuesday, etc.)).

In hindsight, the development of those scripts might now in some ways be labelled as a waste of time, since they did not deliver the results we hoped for, but also the initial idea behind them now seems pretty vague as well. Since, if script 4 tells us (which it does) that during the 1960's and 1970's December was the

month in which most humanities articles reached the front-page, then what does that actually tell us? And, why should we even want to look for a reason behind that?

It was clear that we had lost our direction. Therefore, the group contacted one of the visiting teachers, Jan-Heijn, asking him for help. We told Jan-Heijn that our efforts to look into the development of humanities front-page articles from many different angles had not delivered a glitch in the data that we considered interesting enough to justify further research. Jan-Heijn then pointed out that if there were no strange glitches in the data at all, then that was a very interesting find in itself, since in the past there had often been discussions about the 'decay of humanities'. Jan-Heijn thus advised us to go and search for the same results within other, humanities-related, keywords, like science. And see how they would match-up (or not).

The final two weeks of the project we spent a lot of time with trial-and-error attempts to retrieve the required data from the API, in such a way that it would allow us to compare the development between humanities (front-page) articles and science (front-page) articles. For this goal, the final scripts where written:

Script 7

Script 8

Script 9

3. RESULTS

For a complete overview of all our results, including the old and unused scripts and description files, please visit our group-page on Github: https://github.com/objects-to-data-group.

Unfortunately, in the end, we did not get around to a meaningful interpretation of our results. We feel that our results are still much to raw to allow such an interpretation. For example, the articles that are in the datasets that where created via the API sometimes hardly have anything to do with science or humanities at all. For instance there is this article about a professor who was suspended after a misplaced joke, although the professor was part of the humanities-faculty it is hard to conclude that this article has anything to do with "the humanities" in general.[1]

Also, the dataset contains a lot of articles that where not published in the newspaper at all, the "print_page" field of those articles is set to "null". Those articles negatively affect the validity of the ratio between the total number of articles and front-page articles. An affect that we are unable to prevent, since the API does not allow us to search the "print_page" field, and our downloaded dataset of 'science' does not cover the pre-1980 period. Therefore we are unable to effectively filter the "print_page:null" articles out. Yet another conflict arises when some of the other workgroups, during their final presentations, mentioned that they had found some edition of the New York Times that had over a 1000(!) pages. For an article to reach the front page in such a newspaper would hold a lot more weight, then an article within a 12 page paper.

As a final example, the physical size as well as the total number of front-page articles of the newspaper have to be considered, since changes in any of them would have affected the chance for any article to reach the front-page.

We had hoped to gather all of this data during our research, but in the end that turned out to be a bit utopian. Therefore, considerations like these have led us to

[1] www.nytimes.com/2012/09/13/nyregion/merchant-marine-academy-professor-wont-be-fired-for-colorado-shooting-joke.html?_r=1&gwh=D87321CA82A85B6E79C3B19D261F0114&gwt=pay

the decision to steer away from trying to establish a meaningful interpretation. Instead we have documented some hypothetical steps that we feel that will be needed to be taken to continue our research.

We offer them as a guideline for future researchers that might be tackling the same research-question:

1. Think about how to develop a script that would help you to define whether query-results from the New York Times API are actually related, topic-wise, to the query.
2. Figure out a way to effectively filter out all the articles that where not in the actual (paper edition) newspaper.
3. Establish a script that will offer more detailed insight into the development of the ratio between total number of articles & co-relating total number of front-page articles in general over time.
4. Perform some research towards the (changes in the) physical size of the news-paper.
5. Perform more extensive research into the history of the articles and/or books that have been written in different contexts about 'the decay of humanities'. And see if you can pinpoint a specific period in which most of these where published. Then check if your data transforms in a significant way during that period as well.
6. Investigate different ways of topic modeling to see if they can be effective in establishing a better understanding not only of the number of humanities front-page articles, but also in a much more detailed way about what those front-page articles where about. Try to then compare those results to the reg-ular humanities articles.
7. Etc.