Wiktor WERNER
Adam Mickiewicz University in Poznań

CAN IDEOLOGICAL IDENTITY BE MEASURED? RESEARCHES OF MENTAL PHENOMENA IN THE PERSPECTIVE COMPUTATIONAL SOCIAL SCIENCE, DATA SCIENCE AND GOOGLE TRENDS DATA

Abstract: *The article describes research on ideological identity and historical awareness based on the methods of computational social science and data science. The object of the analysis is data from search engines.*

Key words: *identity, historical awareness, data science, Google trends.*

Computational Social Science (shortened sometimes as CSS, although this term can be decrypted also as Cascading Style Sheet and therefore is ambiguous) is relatively a new branch of science, it's main features are:

- interdisciplinary character;
- strong orientation toward specific tools of research – data mining software;
- strong orientation toward particular sources digitalized and digital-born data[1] (Claudio Cioffi-Revilla 2014: 6-7).

Interdisciplinary character of Computational Social Science is rather self-explanatory but term «data-mining» needs explanation. This word describes, also, very recent method of discovering patterns in large data sets with help of algorithms which can operate both with or without constant human supervision (machine-learning algorithms).

The character of sources used in Computational Social Science researches is obvious but one should distinguish between digitalized sources which existed before their digitalization in the oral, written, printed or any non-digital form and born-digital – which emerged in cyberspace. This differentiation can be very important because of category of 'genuineness' which is applied to this two forms of information's existence differently.

The source which exists in «real» space in the form of written or printed

[1] C. Cioffi-Revilla, *Introduction to Computational Social Science. Principles and Applications,* London 2014, pp. 6–7.

document is genuine as singular phenomenon created in particular time and space. This singular being can be later «copied» or «falsified» but copies and falsifications are historically dependent to the one «genuine» source. Digital-born artifacts aren't tied to any particular space nor even the particular time because can exist in many parallel entities though «copies» are not physically distinguishable from «an original». Only the context of their presentation in cyberspace can be «genuine», «secondary» or «falsified».

The research of digital information can be conducted in many ways and there are many tools to obtain interesting outcome.

From the perspective of social sciences there is only one but very important problem – if effects of the research of digital artifacts/informations are relevant to social and mental phenomena which construct a society and relations between individuals within a society. Cyberspace is a dimension where those phenomena show themselves (and is a product of social activity as well). A social scientist examines cyberspace not as such but as a source to obtain knowledge about social interactions and mental states which are outside it (because humans are not digital-born beings yet) although social relations and states of consciousness can appear in cyberspace.

How do individuals build and express their identities in cyberspace? There are many ways of course. One of them is by their activities in social networks (Facebook, Twitter). This field of social research is, at this point, considered the most promising not so much to explain human mentality as to study and predict behavior[2].

The other way (on it we would like to examine in this article) is connected with obtaining knowledge about world by using search engines, especially the most popular one – Google.

The behaviorist pattern of using Internet searcher for acquiring knowledge (any knowledge) is important for cyberspace social research because of the possibility of measuring searching activities creating by the service trends.google.com. The output is relative that means this service shows us how often given term is searched in time or in comparison with other terms. That powerful tool is widely used to find out what kind of product is/was the most desired by consumers in given area or time-period[3] but it can be also used in many other purposes like predicting outbreaks of diseases[4] although it cannot be treated as «oracle» as shows us the example of it's failure in predicting the flue outbreak in 2013[5]. One of them, which is connected with our topic of researching identity and mental phenomena in cyberspace, is purpose of predicting results of elections, especially elections where electors chose between ethical and social values as well as between political options. If google search index can show as in advance the results of elections, we can assume that there is a correlation between searching activity of people

[2] M. Kosinski, S. Matz, S. Gosling, V. Popov, D. Stillwell, "Facebook as a Research Tool for the Social Sciences", *The American psychologist*, 2015, vol 70, pp. 543–556,

[3] H. Choi, H. Varian, Y. Carrière-Swallow, F. Labbé, "Predicting the Present with Google Trends", *The Economic Record*, 2012, vol 88: Special Issue, (June), pp. 2–9.

4 H. A. Carneiro, E. Mylonakis, "Google Trends: A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks", *CID*, 2009, vol 49 (15 November): SURFING THE WEB, pp. 1556–1564.

5 D. Lazer, R. Kennedy, G. King, A. Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, 2014, vol 343 (14 March), pp. 1203–1205.

in cyberspace and their decisions as voters. Voting decisions are, on the other hand, important manifestation of social and political identity especially in case of elections where ethical and cultural values are involved. Example of such election was the presidential election in United States of America in 2016 where two main candidates: Hillary Clinton (Republican) and Donald Trump (Democrats) represent not only their parties but different sets of cultural values as well. The well known fact is that a Trump victory was considered unlikely by majority of media forecasts but analysis of trends.google data can show that it could be foreseen.

Illustration 1. Search indexes of Donald Trump and Hillary Clinton for year 2016 (year of Election). Data from trends.google.com, visualization by W.W.:

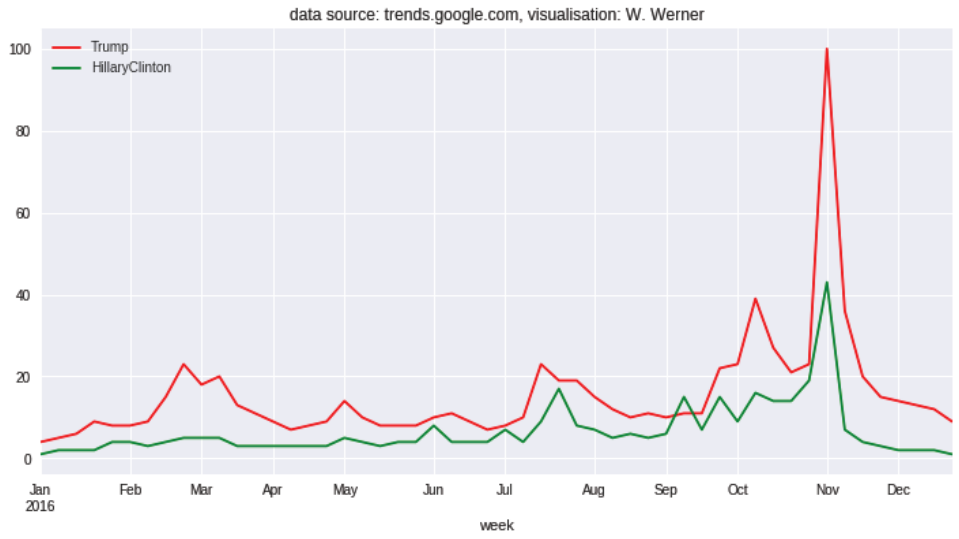

Illustration 1.

One could interpret this data also in a way that Donald Trump is more colorful character and therefore more people searched information about him, so we checked if similar correlation took place in previous presidential elections when sharp distinction between values was also involved but it didn't have so much personal character as in 2016.

Also in this example the result of election (although not that surprising) was visible from the perspective of Computational Social Science analyzing search indexes from the most popular search engine.

Illustration 2. Search indexes of Barack Obama and John McCain for year 2008 (year of Election). Data from trends.google.com, visualization by W.W.:

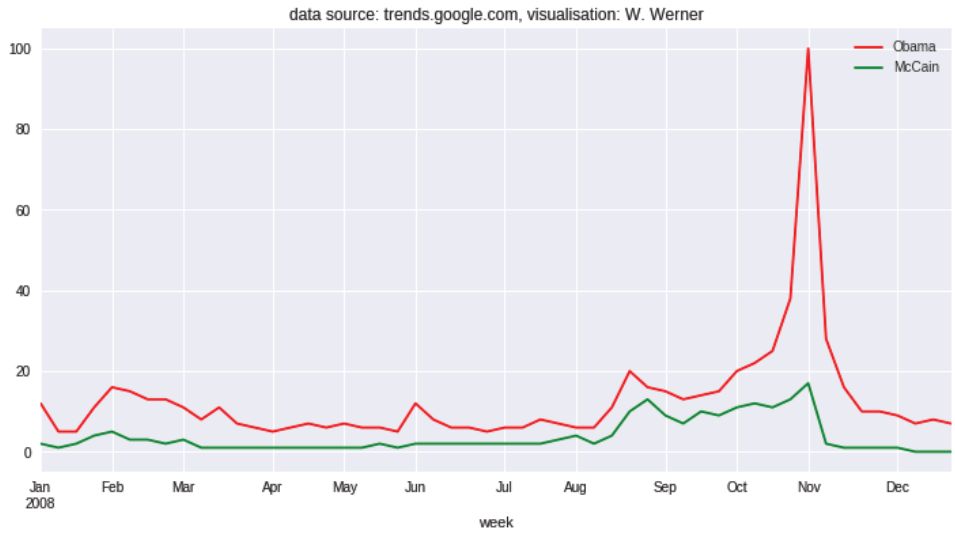

Illustration 2.

Similar correlation took place in year 2012 (illustration 3), when searching measurements from trends.google also predicted election's results:

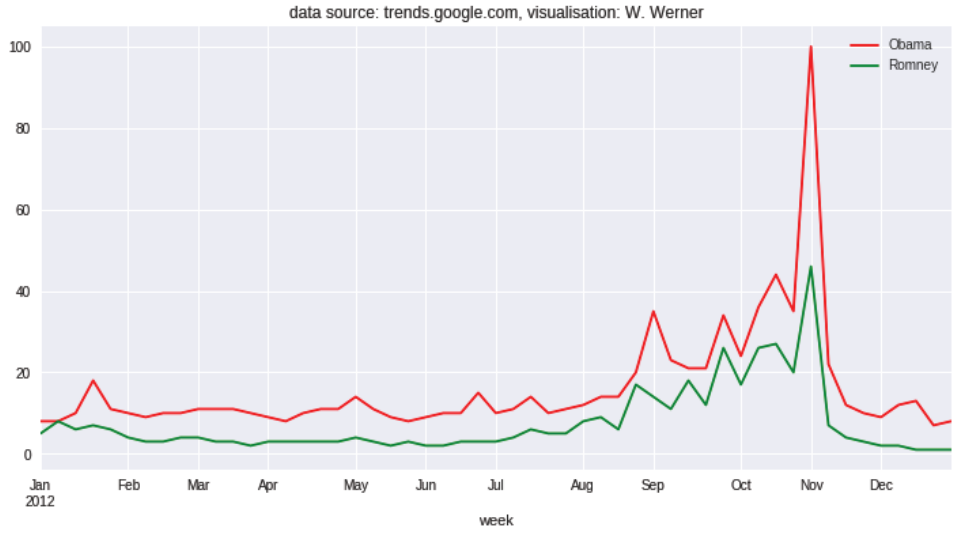

Illustration 3.

We cannot take this correlation as a general law of course. Sometimes, as in the example of Czech first direct presidential election of 2013, the concern of internauts was equally distributed between two political competitors: Miloš Zeman

and Karel Schwarzenberg because of the specific context of elections – it's new way of conducting (illustration 4):

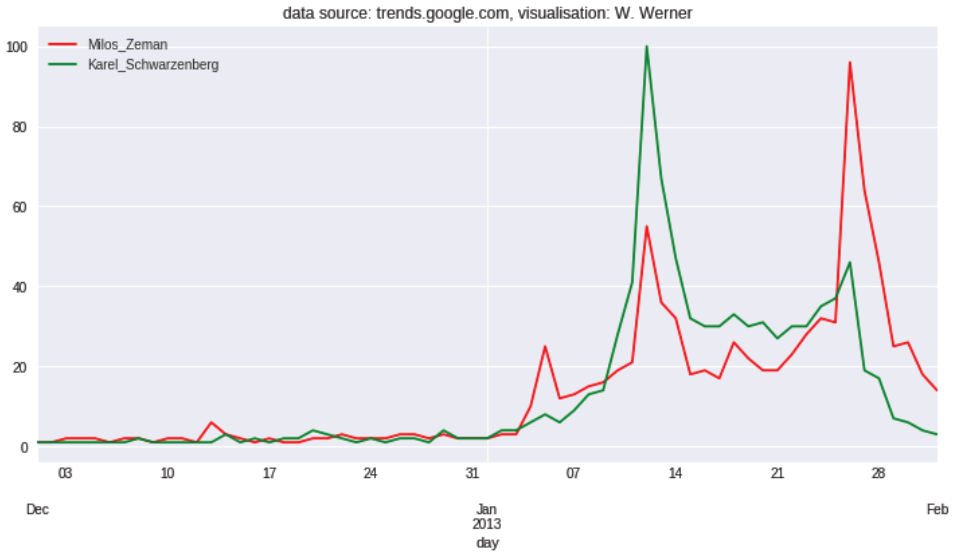

Illustration 4.

but we can see, that the election was won by the candidate who took over the lead at the finish of the campaign (Zeman), we can also check that the mean of the value of the search index is slightly bigger in his case then for Schwarzenberg who eventually lost (illustration 5):

Illustration 5.

On the other hand, in the case of the latest presidential election in France

(2017), we cannot say that the winner was the most weird and therefore attracting attention candidate but he was the most searched in the google search engine (illustration 6):

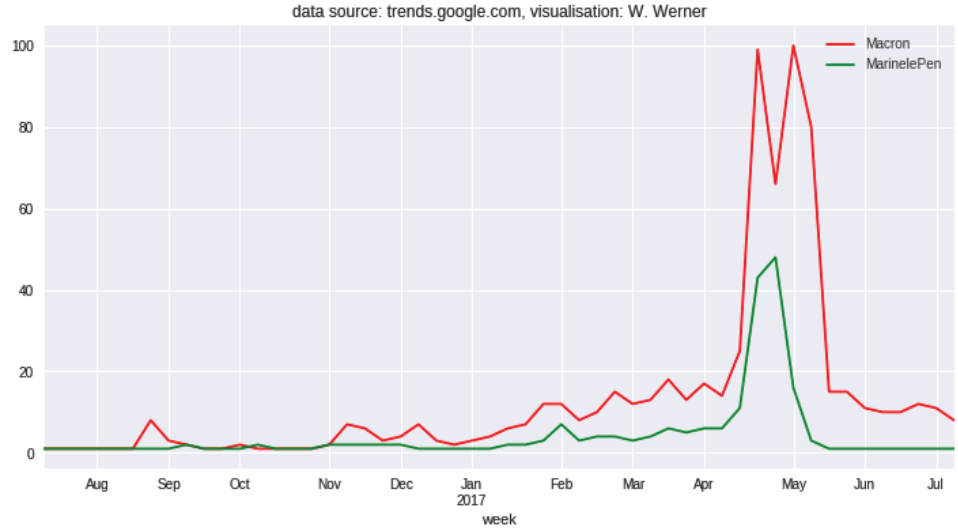

Illustration 6.

As a final example let's take the «Brexit» referendum of year 2016 which shock European community. Like in case of Trump vs. H.Clinton election most commentators argue for staying UK in European Union's result of this voting. It could be wishful-thinking or part of the campaign but could be also underestimating of the meaning of social-attention's trend hidden in the searching activities through Internet.

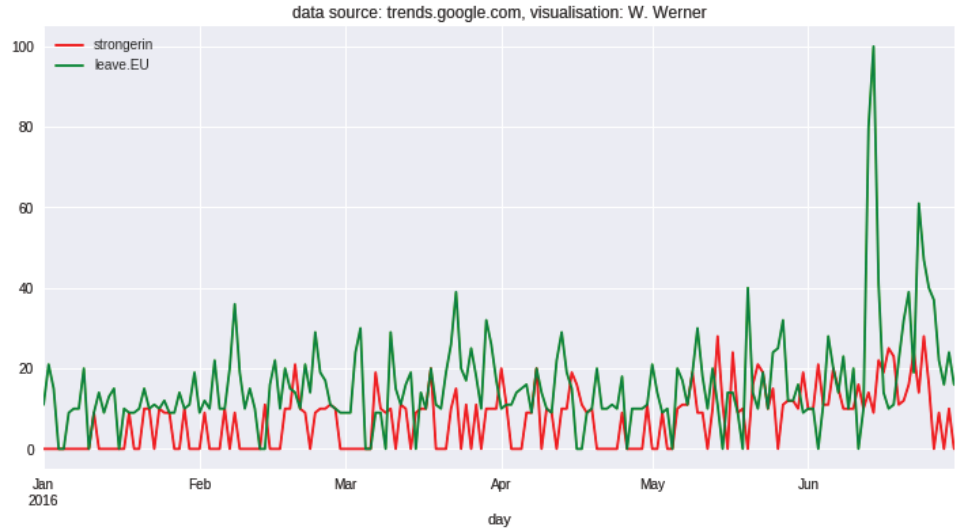

Illustration 7.

We can see that search-trends (illustration 7) of «leaving» slogan was gradually gaining advantage over «Stay» slogan («strongerin») and the mean values of search activities in the cyberspace are also compatible with the «real -life» political decision of voters (illustration 8) although not proportionally compatible: the gap between «leave» interests and «stay» interests is bigger then real-life results of the referendum: (means of *strongerin* searches: 7.329670 and leave.EU searches: 15.703297 which is relation 1:2 vs real proportion of votes: 51.89% for leave and 48.11% for remain).

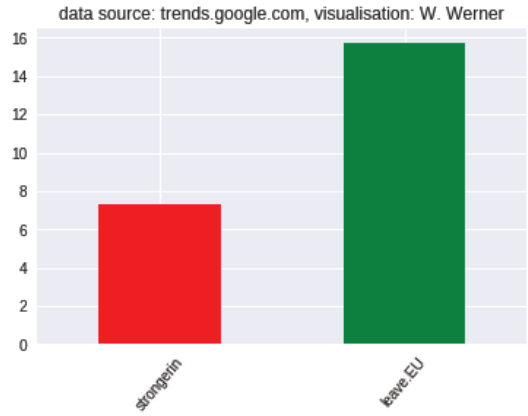

Illustration 8.

What is the meaning of those examples of correlations between Internet's searching trends and real-life political and ideological decisions?

We can say that in contemporary world obtaining knowledge from Internet search engines is, for a civilized and prosperous society, a form of activity as ordinary as using bathroom tap or electric socket for obtaining water and electricity or going shopping to the supermarket. We are using various goods which our civilization delivers to satisfy our needs. One of these goods is information which we need for making decisions good for us and coherent to values we share. It is natural for contemporary voters to seek information about candidates in Internet, it is also probable that the one who is more convenient and more appropriate gains more concern and should be an object of the more intensive searching activities through Internet.

It is quite normal for people who grow up along with world wide web (being in their early teens in nighties of XX century and younger) to seek information by search engines as their grand parents looked for it in encyclopedias and dictionaries[6]. The search engines look like they were more neutral than television, newspapers or most books which are explicitly tied to somebody's particular point of view. In the case of search engines most people expect that information will flow from them as pure as water from the source. Because of this belief search engines are so important in shaping political and ideological identities of members

[6] *Society of the Query Reader: Reflections on Web Search*, ed. R. König, M. Rasch, Amsterdam 2014, pp. 40–46.

of contemporary societies.

For researches in humanistic and social studies data acquired from Internet (especially search engines and social media) are potentially very rich source of information about mentality of contemporary people, too rich to be ignored. The problem lies in finding relevant methods for analyzing those sources and theoretical model which can help in linking Internet traffic's data with consciousness of contemporary people.

Let's check one of the problems of common historical consciousness which is it's concentration on internal affairs of particular national and political communities. Idea that historical knowledge is used for creation of ethnic and national identities was widely discussed in theory of history[7] as well as in intellectual history and history of social identity[8]. Using a mass data from internet search engines (in this example it will be google.search statistics data from trends.google service) we can find out if it is true that common people are interested in their national history more than in world history.

In our research we will check relation between trends in searching for «world history» and particular national history in USA and chosen countries in Europe and Asia.

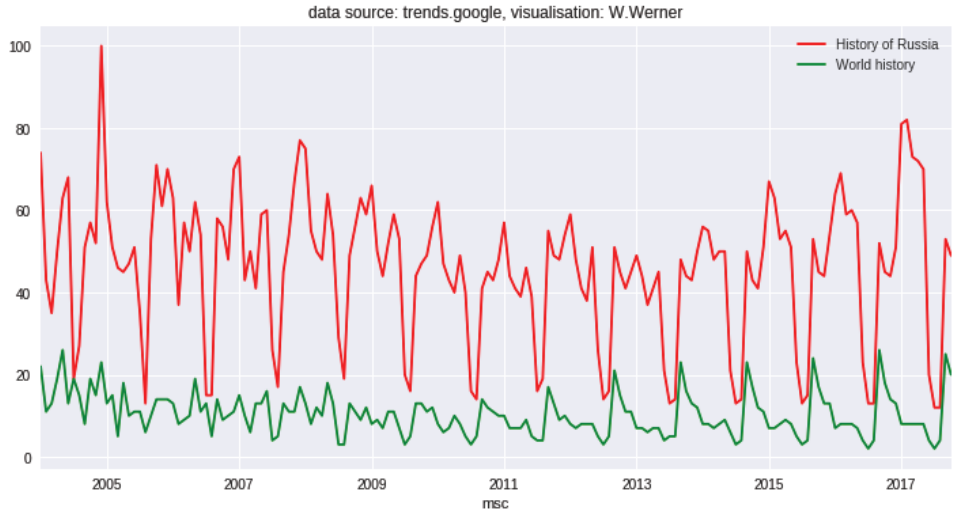

Illustration 9.

As we can see on the illustration 9 trend of searching history of Russia dominates the other trend. Mean of values for all data points shows relation 46:10 in favor of nation-history. We can also observe that this trend is practically stable from 2004 year until now:

[7] W. Wrzosek, "Historiography as a vehicle for the nationalist idea", [in:] *Nationalisms across the globe. An overview of nationalisms in the state – endowed and stateless nations,* vol. 1: Europe, ed. W. J. Burszta, T. Kamusella, S. Wojciechowski, Bydgoszcz 2005, pp. 43–48.

[8] B. Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, New York 2006, pp. 155–163.

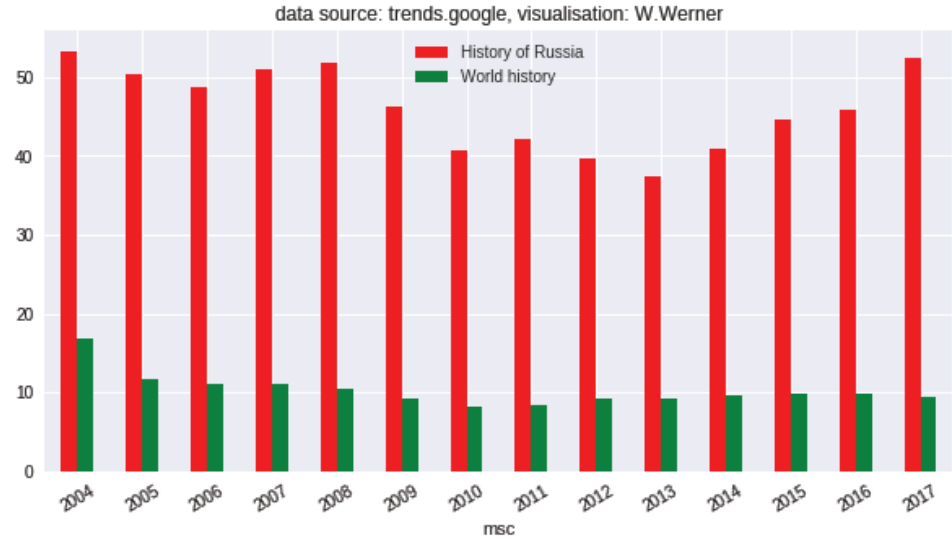

Illustration 10.

Comparing this data to data coming from other country we can see not only different trend but also cultural difference between communities generating those results:

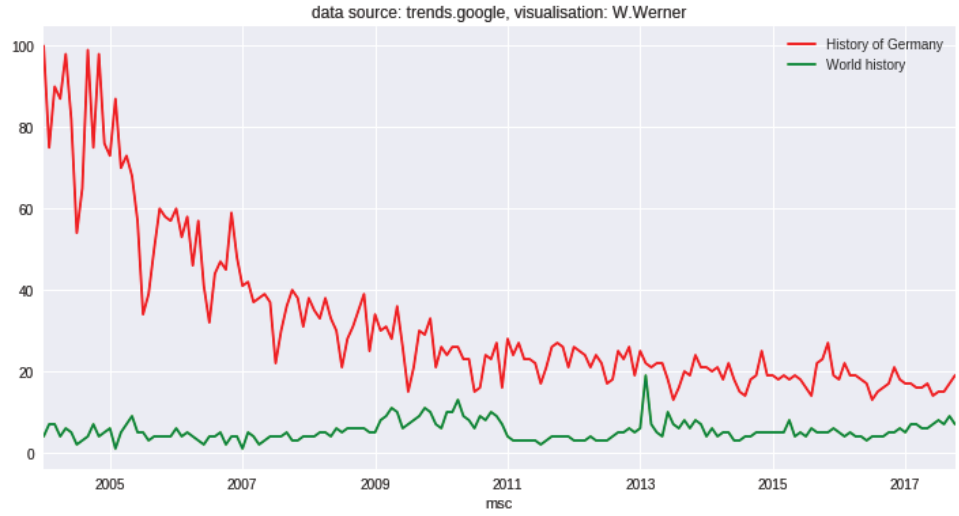

Illustration 11.

As it is visible in Germany trend of searching for history of Germany in gradually falling down from 2004 while the trend of interest in world history is slightly going up – what we can see in means counted for each year separately:

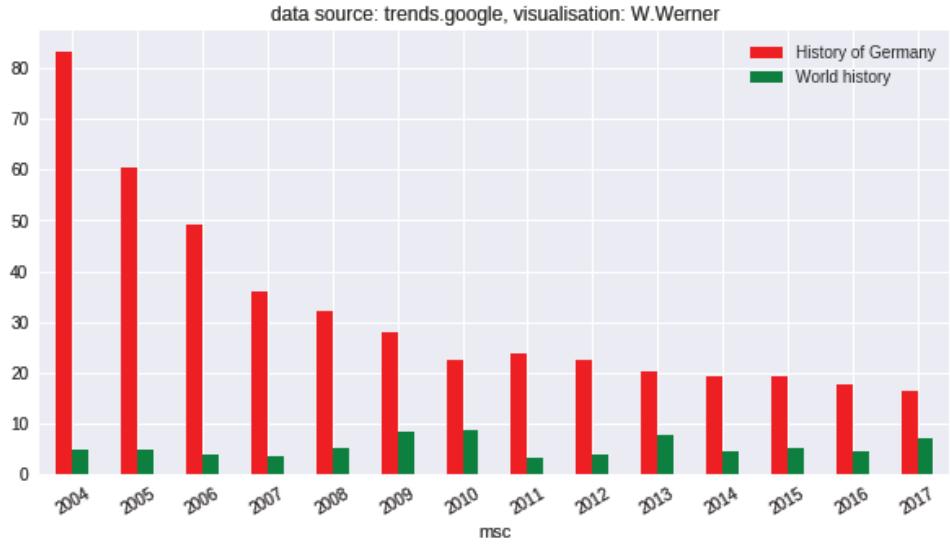

Illustration 12.

Similar tendency we can watch in results from France:

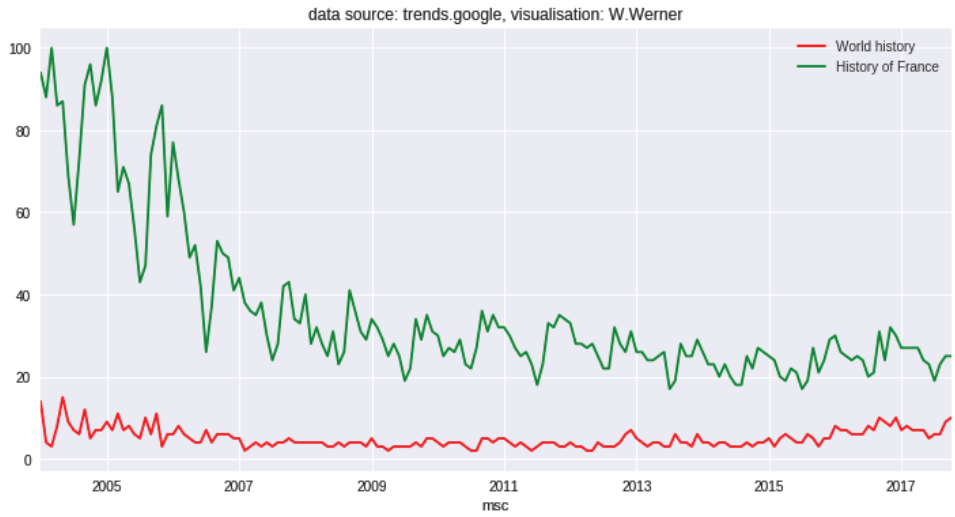

Illustration 13.

and in Italy, where tendency to reduction of disproportion between interest in national history and world history is observed;

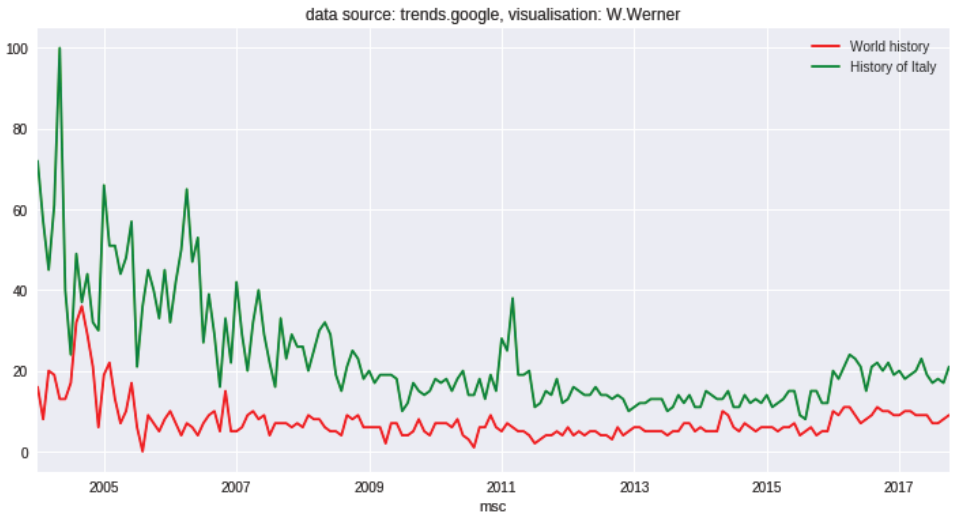

Illustration 14.

but from the other side of Atlantic Ocean we can observe tendency very similar not to the west-european trends but to the Russian tendency where domination of involvement in national history is absolute :

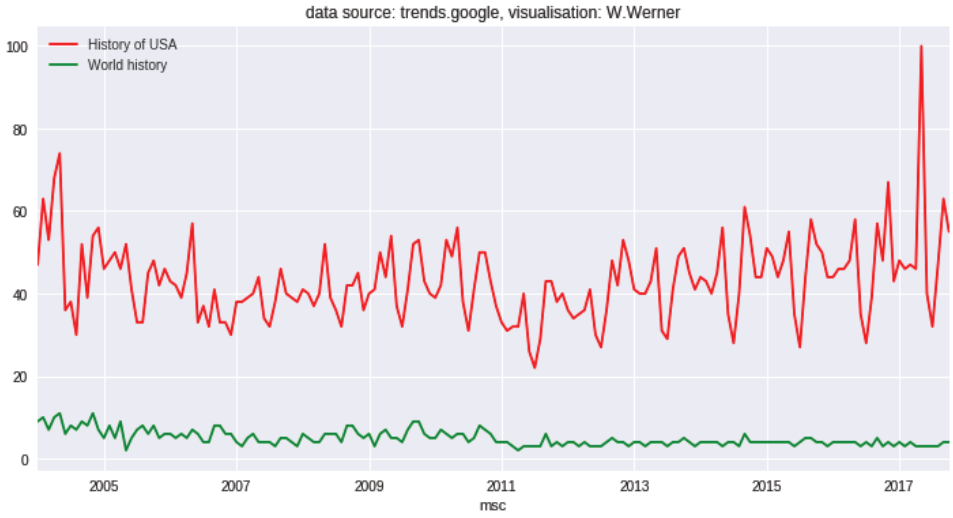

Illustration 15.

As for China we can observe tendency to approaching those two trends but only until year 2014 andthe change after this year:

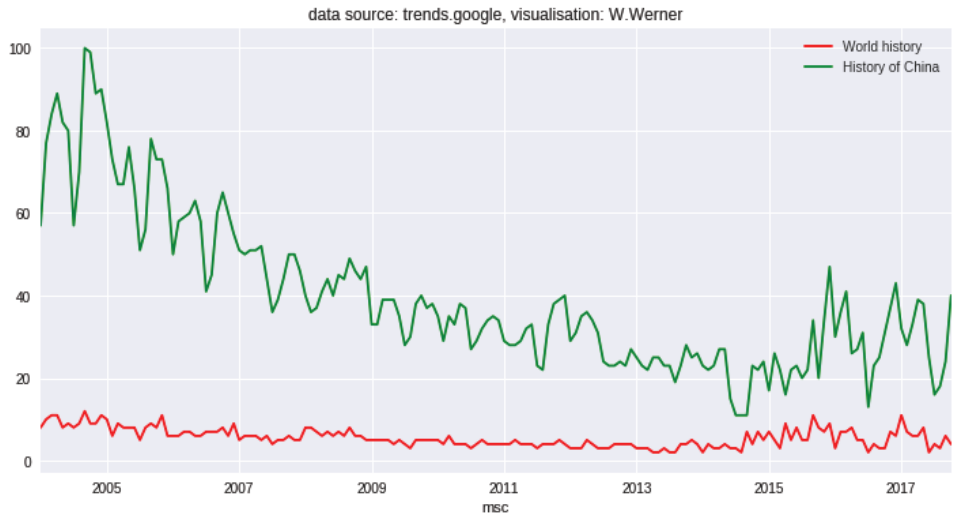

Illustration 16.

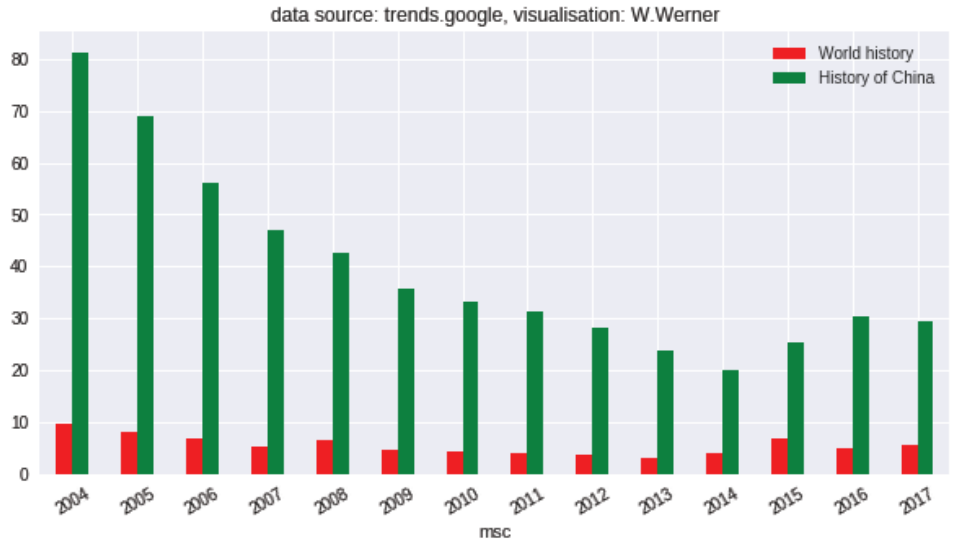

Illustration 17.

Are those data good enough to make transcultural comparison? It depends on what kind of meaning we are able to tie them. Searching activities of vast populations cannot be considered as meaningless, there are serious causes for existing those trends. Can we observe trend for de-ethnocentrisation in western-european countries and strict tendencies to concentrate on national perspective in global-power countries: USA and Russia? If our interpretation is correct there is also a shift of paradigm in Chinese historical consciousness from tendency to «european-like» internalization of historical knowledge to return to the road of «imperial» ethnocentrism.

The last example come from Japan and shows us european-like tendency to shrinking discrepancy between national-history interest and world-history interests:

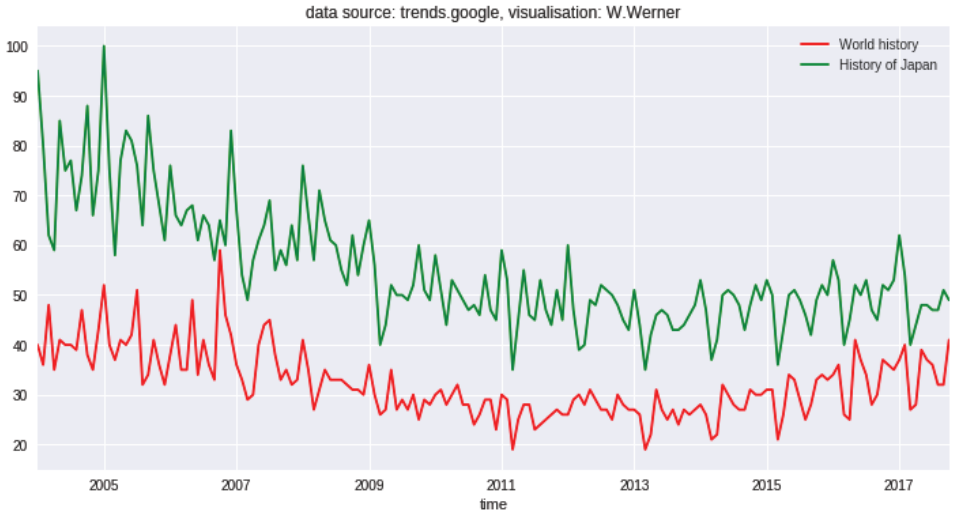

Illustration 18.

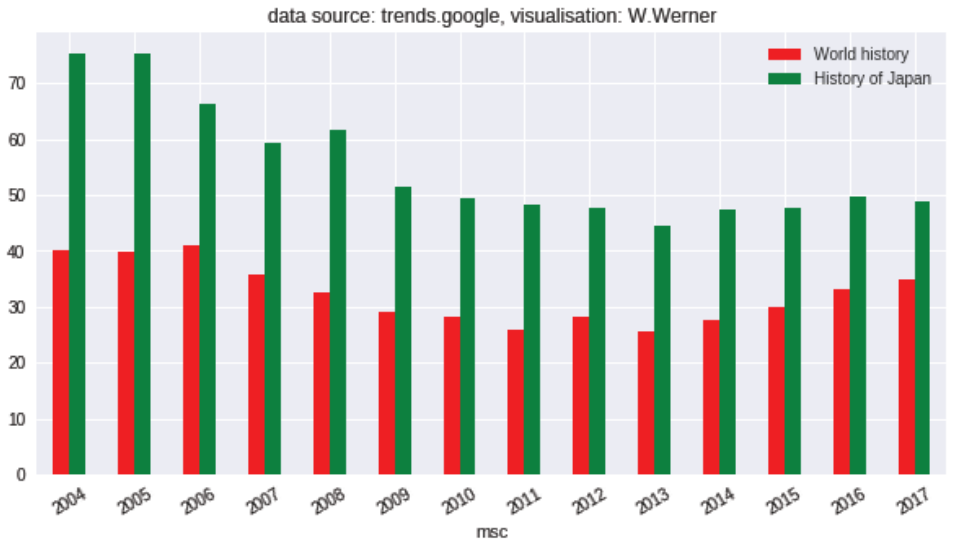

Illustration 19.

In those visualizations of data the one thing more attracts attention. It is the similarity of lines showing dynamics of trends on the illustration 18. It provokes to asking additional question about association between those two trends showed in each dataset (national-history interest and world-history interest). If those trends are

uncorrelated (low correlation coefficient) this can means that trends are created by various and unrelated agents e.g. interest in national history is caused by education and the state's historical politics and interest in world history is a spontaneous people's activity. We are going to count Pearson correlation coefficient for each of analyzed data-set using python programming language's data-science modules: pandas, numpy and seaborn[9]. Pearson correlation coefficient is a measure of the linear correlation between two variables X and Y. It has a value between +1 and −1, where 1 is total positive linear correlation (they have the same values), 0 is no linear correlation (they are unrelated), and −1 is total negative linear correlation (variables are inverse)[10].

For dataset from Japan, visual similarity between lines on the graph is confirmed by the result of calculation: Pearson correlation coefficient is 0.73 which is relatively high value in scale from -1 to 1. Results for remaining datasets are as follows:

China = 0.71
Japan = 0.73
USA = 0.28
Russia = 0.49
Germany = -0.14
Italy = 0.52
France = 0.48

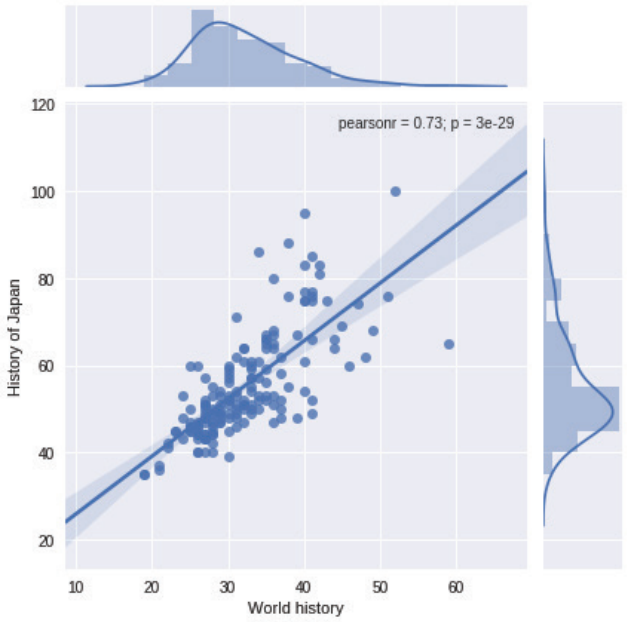

Illustration 20.

[9] H. K. Mehta, *Mastering Python Scientific Computing*, Birmingham 2015, p. 180.
[10] J. L. Rodgers, W. A. Nicewander, "Thirteen Ways to Look at the Correlation Coefficient", *The American Statistician*, 1988, vol. 42, no. 1. pp. 59–66.

Those results indicate a slightly different view on the problem of historical identity in analyzed regions. In countries with the highest correlation between interest in national-history and interest in world-history (China, Japan) the seeking knowledge about both topics can be correlated as two elements of the same process – gaining knowledge about historical reality. Therefore when tendency for looking for knowledge about national history is going up – the seeking knowledge about world's history is going up altogether as we can see on graphs comparing those tendencies in Japan (ill. 20) and China (ill. 21):

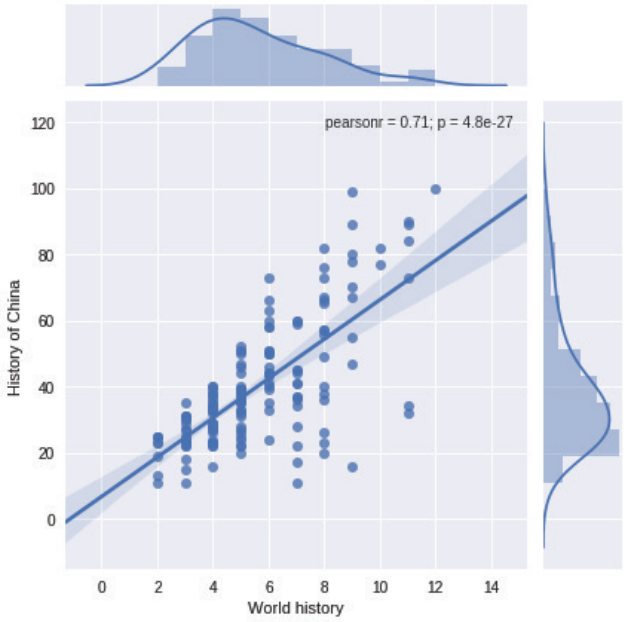

Illustration 21.

On the other hand for countries with low correlation coefficient between those trends we can assume that looking for knowledge about national history and seeking information about world's history are elements of different processes and therefore their dynamics are unrelated in time. For visualizing differences in structure of analyzed two trends we grouped data-points by months and counted mean for each month separately. As we can see on following graphs dynamics of searching information about world's history and national history are quite different also from perspective of monthly distributions of interests:

In Germany (the lowest Pearson's coefficient) interest in national history has it's monthly dynamics probably connected with educational schedule (the lowest rate is in holidays months). Quite similar situation is present in USA where the lowest interest rate in national history is during vacations but highest level takes place in may – month of American «Memorial Day». On the other hand interest in world history in USA is monthly little varied – so we can assume it is connected with spontaneous men's curiosity about different subjects.

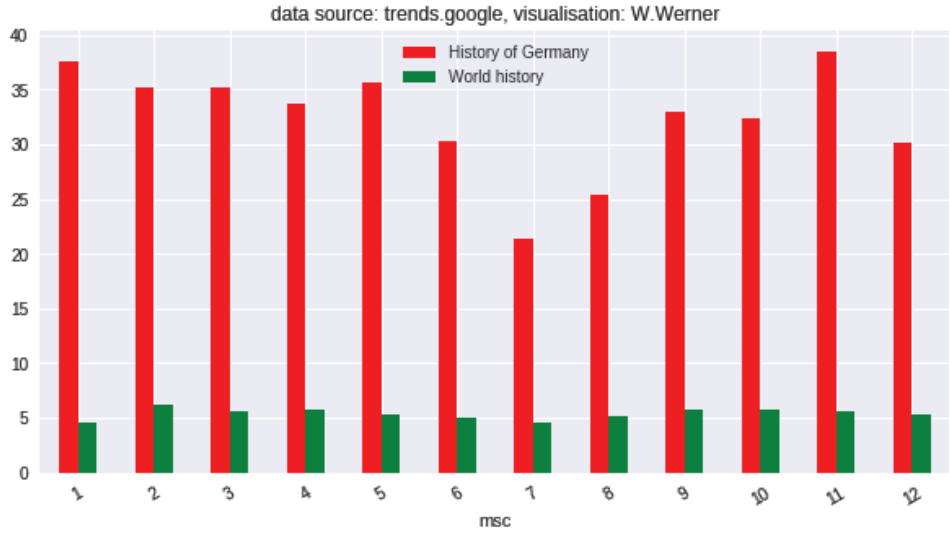

Illustration 22.

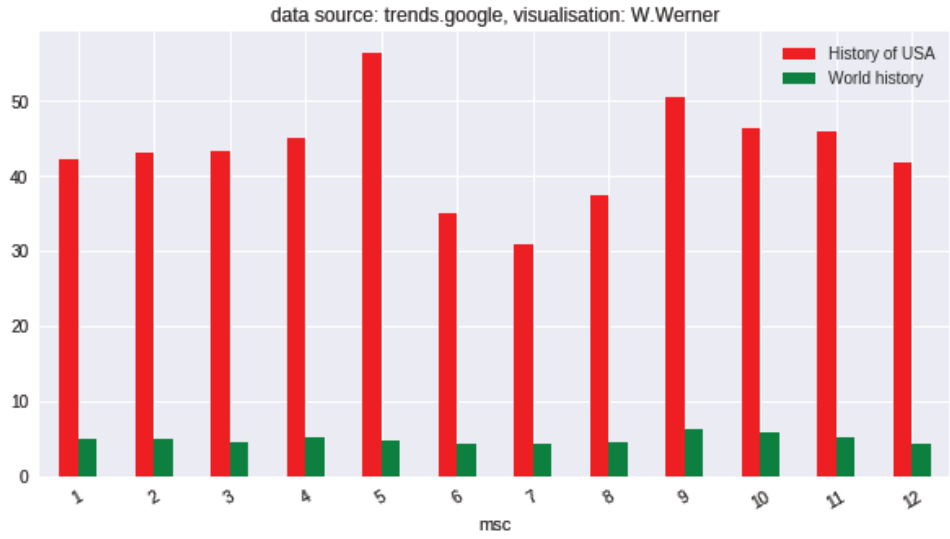

Illustration 23.

Similar configurations is in Russia, so we can assume that distinction between searching for knowledge about national history and world's history has similar base.

We can check the results of our quantitative research based on statistical data by means of parallel qualitative research, i.e. about the analysis of keywords entered into the search engines associated with the search term.

Illustration 24.

Illustration 25.

In this case, we mean associations with the word “history”. Quantitative data show the (chronological) correlation of this concept with the practice of school education and the ethnocentric nature of interest in history.

Let us take a closer look at the most frequent questions accompanying key words history entered in English (at US territory), German, Russian lan-

guage with the use of a web search-engine. For visualizing the frequency of the data obtained (the list of the most frequently entered key words), I have prepared a 'word cloud'. Visualizations have been prepared with the use of a script written in python language with the modulus 'word_cloud' authored by Andreas Mueller11. The size of the words in the visualization represents the frequency of its occurrence in the data:

Illustration 26.

Can we suppose that analyzed keywords, together with quantitative data illustrating trends, can be the basis for transcultural comparison of historical consciousness? Of course, they provide a lot of material for interpretation. We can see that in the American (ill. 24) case the degree of ethnocentrism is smaller than we expected. Key words point to linking interest in history with media events - it is more popular to search for information about specific films or TV series (Vikings) than to search for information about events in American history that is not the subject of a media event. A strong correlation of interest in history with schooling practice is evident in the German case (ill. 25). Here a lot of words appear regarding the curriculum in schools («Gymnasium», «Realschule», «Abitur», «Klasse»). In the Russian case (ill. 26), apart from the links with the school practice, there is a strong correlation between the interest in history and the search for information specifically about the history of Russia («история», «государства», «российского»).

[11] A. Mueller, 2018, https://github.com/amueller.

References

1. Anderson B., *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, New York, Verso, 2006.
2. Carneiro H. A., MylonakisE., Google Trends "A Web-Based Tool for Real-Time Surveillance of Disease Outbreaks", *CID*, 2009, vol. 49 (15 November): SURFING THE WEB.
3. Choi H., Varian H., Carrière-Swallow Y., Labb F., "Predicting the Present with Google Trends", *The Economic Record*, 2012, vol. 88: Special Issue, (June).
4. Carrière-Swallow Y, Labb F., "Nowcasting with Google Trends in an Emerging Market", *Journal of Forecasting*, (2011).
5. Cioffi-Revilla C., *Introduction to Computational Social Science. Principles and Applications*, London 2014.
6. *Society of the Query Reader: Reflections on Web Search*, ed. R. König, M. Rasch, Amsterdam 2014.
7. Kosinski M., Matz S., Gosling S., Popov V., Stillwell D., "Facebook as a Research Tool for the Social Sciences", *The American psychologist*, 2015, vol. 70.
8. Lazer D., Kennedy R., King G., Vespignani A., *"The Parable of Google Flu: Traps in Big Data Analysis"*, 2014, vol. 343 (14 March).
9. Mehta H. K., *Mastering Python Scientific Computing*, Birmingham 2015.
10. Mueller Andreas, (2018), https://github.com/amueller.
11. Rodgers J. L., Nicewander W. A, "Thirteen Ways to Look at the Correlation Coefficient", *The American Statistician*, vol. 42, no. 1. (Feb., 1988).
12. Wrzosek W., "Historiography as a vehicle for the nationalist idea", [in:], *Nationalisms across the globe. An overview ofnationalisms in the state – endowed and stateless nations*, vol. 1: *Europe*, ed. W. J. Burszta, T. Kamusella, S. Wojciechowski, Bydgoszcz 2005.