Wikipedia Clicks: Exploring Trends

Ljiljana Miljesic
VU University
Nr. 2573309
f.ricchiuti@student.vu.nl

Fabio Ricchiuti
VU University
Nr. 2573922
f.ricchiuti@student.vu.nl

1. INTRODUCTION

In the last century, the manner and the tools that the people decided to use to keep themselves informed changed frequently. Recently, the introduction of the new technologies influenced the society and its way to stay updated on what is happening in the world. The use of social media and websites that provide information on the current events in real time transformed how the population approach and react to the news. If a news generates a high level of interest than it is defined "viral". Who has an active role in the creation and the diffusion of viral content are the users of the social networks that share information trough their virtual channels.

At the same time, alternatively at this frantic world there exists other realities where the past and the present knowledge is stored in order to be freely accessed and used by everyone. One of the most important example is represented by Wikipedia[9]. Wikipedia is an online free encyclopedia, editable and accessible by everyone. The variety and size of the content and the opportunity to access it with the comfort of a smartphone or a computer, has given to Wikipedia a huge success, so much to make Wikipedia become integral part of the everyday life of everybody. Even though Wikipedia is an encyclopedia, it is often used to consult the updates of a topic of interest. Its nature that allows everyone to add or edit content instantaneously and, at the same time, to monitor the reliability of the information inserted through the inspection of the sources, makes Wikipedia, in fact, a sort of news websites.

The purpose of this project is to investigate the relationship between news websites and Wikipedia and how the users interact with these realities when a content become viral. Thus, an interactive visualization will be proposed. This interface gives an insight who among Wikipedia or news websites are the catalysts of the interest about trending topics. The interface allows selecting a period of time and, subsequently, receiving the trending pages of Wikipedia for that amount of time. For each page displayed, it is possible to visualize a chart that summarizes the views of that page along the whole year and all the articles that mentioned that argument.

This report is structured as follows: in the next section, an overview of the related literature of this work will be presented. In the section 3, the research questions that led the development of this work will be showed and described. In section 4 will be provided a description of the steps taken during the project. Eventually, in section 5 an early approach to the redirects resolution will be presented and in section 6 the findings and the research questions will be revisited and addressed.

2. RELATED WORK

One of the big first obstacles was to find a solution that would have made possible to identify the Wikipedia pages mentioned in an article. This task turned to be extremely hard given the dependency that a concept has from the context where it is located. In the natural speech, the same word can own multiple meanings and where a human, thanks to his human intuition, can easily identify the concept that underlies a word, a machine needs to carry out sophisticated analysis. One big step further in this direction was made by the creation of a tool called Dbpedia Spotlight[3][14]. DBbedpia Spotlight takes a text as input and extract all the probable mentions of Wikipedia pages contained inside. In order to obtain these results, Dbpedia Spotlight uses the data provided by Dbpedia [2]. DBpedia makes available for everyone all the content of Wikipedia in the form of structured linked data. This reorganization in a graph of labels allows the usage of queries and the creation of associations to different data from external sources. The underlying algorithm behind DBpedia Spotlight is divided in four main steps.:

- Spotting Stage: A lexicon was built from the linked data provided by DBpedia. The spotting stage uses the lexicon to link all the related surface forms to the each word in the text analyzed.
- Candidate Selection Stage: A first basic candidate selection is made in this phase. Depending on the surface form, a priority score is calculated. This score came from the relationship of the surface form according the disambiguation alternatives related to that surface form.
- Disambiguation Stage: This step represents the core of the algorithm. In the disambiguation stage the algorithm analyses the context around the word in order

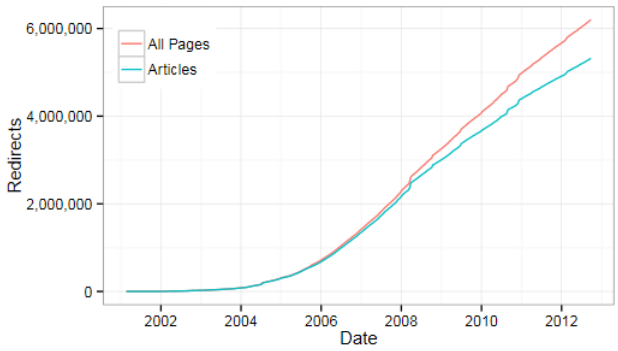

Figure 1: Growth of the number of redirects over the years

to narrow down further the possible candidates linked to that world

- Configuration Stage: This step applies the user defined filters. In the context of this project has been made use of the filter *confidence*. In section 4 will be provided more details about.

A more detailed explanation of the algorithm is provided in[14]

Dumps.wikimedia.org [12] hosts the periodic dumps of the content of Wikipedia and, in addition, provides several statistics. In particular, files with counters of the views of each page are provided. Unfortunately, the counters do not take in account that the pages in Wikipedia are sometimes subject to several redirects, therefore some operation was needed in order to solve the redirects. [13] gives a detailed overview of the importance of the redirects. Indeed, a substantial number of Wikipedia pages are actually redirects and, as showed in Figure 1 in the appendix, this number is dramatically increased in the last years.

Since we are dealing with numerous redirect pages, when analyzing the number of views their influence on the final result can became very relevant. Therefore, an important step in the processing of the data that contains number of views per Wikipedia page was to calculate the exact number of views concerning the page and all its possible redirects. Furthermore, [13] generated a file that contains This data was calculated by analyzing the text inside revisions of all articles. Using a research computing cluster at the Harvard MIT Data Center (HMDC), authors of the paper parsed the full text revision history of Wikipedia published in October, 2012, to create a list of all revisions that redirect. Even though this data set is slightly outdated for needs of our project, its exhaustiveness seemed more appealing than Wikipedia dump files.

3. RESEARCH QUESTIONS

The research question that led to carry out the analysis was: of *"Who is following who, are things first trending in Wiki and then in newspapers or the other way around?"* Therefore, the question leads to two other sub questions focused on the technological aspect:

- *"What kind of technology and techniques should be used to process the data?"*
- *"Does the solution provide enough scalability in order to expand the project in future?"*

4. PROJECT SETUP

4.1 Dataset

The aim of the project is to provide an interface that contains a view of the most accessed topics on Wikipedia filtered by time and an overview of the news that mention the topic. The range of time taken in consideration for the analysis is the year 2014. The datasets, taken from dumps.wikimedia.org[12], containing the page counts has a size of, approximately 820GB and it is divided in hourly generated files. Each row inside the page counts file contains the following information:

- The language of the page, e.g. en, nl, it;
- The wiki project in which the page belongs to, e.g. Wikibooks, Wikipedia, Wikiquote;
- The title of the page;
- The number of accesses at the page in the specific hour to which the file is referring. It is worth noticing that these access do not necessary match the exact views of the page. Indeed, the counter takes in account the amount of times the url of that page is requested to the website. However, even though the counter does not represent the exact views, it can be considered a good indicator of the popularity of that page.

The news were taken from the database of a popular dutch news website "de Volkskrant"[4]. For this project, only the news about 2014 were considered. The dataset consisted of a csv file of 190MB. Each row of the file contains:

- The date of the news
- the URL of the news
- The title of the news
- The content of the news

The third set of data that was required, for the purpose of gaining higher accuracy of the final result, was data about redirect pages on Wikipedia. Data used in this project was obtained from research that reflects on the importance of redirect pages when Wikipedia page views are explored, as presented in the referred paper[13]. The format of this data is:

- Page ID
- Redirect page title
- Target page title
- Date and time when the specific redirect became active
- Date and time when the redirect was changed or removed

4.2 Tools

The datasets were stored in a Hadoop cluster[5], kindly made availabe by Surfsara[8]. Hadoop is an open source framework that gives the tools to manage large data sets on several clusters. With Hadoop is possible to process high quantities of data in short times using the built-in file system *Hadoop Distributed File Systems* and the programming model *Hadoop MapReduce*. However, given its optimization, its compatibility with popular languages as Java and Scala and its easy of use, Apache Spark[1] was used to manage the elaboration of the data. Apache Spark is an open source engine built for large scale data processing.

As reported in section 2, DBpedia Spotlight was utilized to handle the identification of Wikipedia articles mentioned in the content of the news. DBpedia Spotlight is offered as a web service that can be queried through REST calls, and as a jar file deployable on an own server. For the sake of scalability and low response times the latter option is the optimal one. However, for simplicity and for the relative small size of the dataset concerning the news, it was eventually decided to make use of the web service version. In the requests of extraction of Wikipedia articles from a text, it is also possible to set a degree of confidence of the response. A higher degree of confidence gives greater guarantees to obtain an exact result but at the same time the chance is higher that some article that is eligible to be part of the response could be missing. In the context of our project a value of 0.6 has been decided since that in the tests carried out by [14], 0.6 resulted to be the value that provides the highest tradeoff between recall and precision.

The languages selected for data processing were Scala[7] and Java[6]. Java was chosen to process small amounts of data since that both member of the team had past experience with the language whereas Scala was chosen for its high expressivity and compactness and it was used in combination with Spark to elaborate the big-sized datasets.

4.3 Methodology

Figure 4 in the appendix illustrates graphically the process that was made on the dataset in order to provide the elaborated data needed to make possible the analysis of the trends. The shape with the cog indicates an operation of processing of the dataset and the technology used. The incoming arrows come from the files provided as input whereas the outgoing arrows are connected to the files generated in output by that process. The processing steps are the followings:

- Step 1: This step represents the first iteration of the process
  - Step 1a: The files that contain data about redirection between Wikipedia pages was originally in format where each row comprehends page ID, redirect page title, target page title, start date and end date as explained in more details in section Project Setup. Therefore, simple Java code was used to transform taking in consideration to exclude obsolete redirects. *Output of this step:* File that contains for each row a redirect page and its target page.
  - Step 1b: After careful observing of data about page views and researching about its format it becames clear that, in fact this is not only Wikipedia data but also Wikibooks, Wikimedia, Wikiquote and so on, and of course on every language suported by Wikipedia. Before any further processing it was necessary to remove all data that was not interesting for this project. Filters are applied to make sure that only dutch Wikipedia articles are taken into consideration. Also, rows that contain information about type of project page belongs to, language and size of the page are removed. *Output of this step:* Smaller version of Wikipedia page views counts cleaned of unnecessary information in the format page title and number of views.
  - Step 1c: Through the use of repeated REST calls to DBpedia Spotlight, a file containing all the Wikipedia Pages mentioned in each article given as input is generated. *Output of this step:* Json file that contains date of the news, URL of thew news, title of the news, wikipedia articles mentioned in the news
- Step 2: Files about Wikipedia page views in their original format contains number of views for pages per hour and the idea was to visualize changes in number of views per page for some period of time. If this visualization would be showed on the level of hours there would be big fluctuations since people generally use Wikipedia more during day hours than night hours. This was a reason to sum all the views in range of time of 24 hours and create visualization of views per day. Therefore, the second step is merging the 24 files that contains data about hourly views, creating a single file that contains an overall counter of views for every page that has been viewed that day. At the same time, while doing this processing, the resolving of redirects was applied using the file generated in Step 1a. *Output of this step:* The resulting set of files contained data in the form of page title and number of views for every page that has been viewed in one day, for every day of the year with resolved redirects. This data was much smaller than the original one, therefore it was used in the successive steps instead of the original data.
- Step 3: After obtained the files with number of views per day, applying a simple operation was produced files per week, month, quarter of a year and for the entire year. For example, summing all the 31 daily views file that belong to one page, it was calculated the number of views of January. After generating list of all pages that were viewed in January with the number of their views this list is sorted and top 100 lines were saved in separate files. The same approach was applied to all months, weeks, quarters and entire year. *Output of this step:* Wikipedia Page Counts of the views for all the weeks, months, quarters and year of 2014
- Step 4: For each page that shows up in files with top 100 pages that were generated in previous step (top 100 trending pages in every day, week, month, quarter and in the year) in final visualisation will be displayed chart that shows number of views for that page for each day in the year. Next step was to get a list of all pages that appear in these files. Thus, a scan was

made in the files that include the views of all pages for each day. If some page came under the top 100 viewed pages list, was included in the output. *Output of this step:* File that contains the views of all the pages in the lists of the top 100 viewed pages.

- Step 5: In the final step JSON files that included all final data were created. There is one JSON file for every time period for which we extracted top 100 viewed pages. *Output of this step:* Output files contain titles of top 100 pages and for each of them the list of views and list of news. List of views is an array of 365 numbers, one for each day of the year. This array is used to populate graph in visualisation. List of news contains 365 lists of news (their title and url), one list for each day.

5. EXPERIMENTS

After an initial review, the structure of the data inside the file concerning the Wikipedia pages views counts was considered fairly straightforward for purposes of this project therefore no additional preprocessing was needed. However, after first manual consultation of the files it became clear that there were certain information that needs to be excluded from the elaboration. For instance, some of the articles showed a not usual size that went from 0 to 20 kilobytes of size. After an investigation of the dataset, we came up to the conclusion that these articles, which were numerous, are actually pages that only redirect to other pages. In order to correctly address the mechanism of the redirects in our project, further researches were carried out among the literature. The research of Hill and Shaw [13] addressed the importance of the redirect. For further details about this paper, see section 2. To transform the original data about Wikipedia page views into more accurate data it was required to substitute all titles of redirect pages, in the data, with the titles of the final pages. Wikipedia dump files are, allegedly, providing this redirect table to the public. However, this data turned to be not complete. Every few weeks new dump files are generated and they consist of the only changes in redirects that happened during that period. If some page became a redirect page or stopped being a redirect page since the latest dump files were created, then information about that change is stored in the dump file created successively. Therefore, since these files do not contain information about redirects older than six months, by using Wikipedia dump data we would not have complete set of redirects. Moreover, mentioned redirects data is in the form of SQL script. First attempt to face all these obstacles was to run all available SQL scripts in database and create redirect data from generated tables. Eventually, this approach adjusted some of the redirect pages, but the result did not cover all the effective redirects. After discovering referenced paper [13], it was decided to use data created during the research described in the paper. Other data anomalies showed up only in later stages of the project. For example, special pages displaying meta-information about Wikipedia website, which are not interesting for the purposes of this project were removed.

6. CONCLUSIONS

This project provided an insight of the popular trends of interest comparing news and views of the Wikipedia articles related to the news.

The first research question wondered *"Who is following who, are things first trending in Wiki and then in newspapers or the other way around?"*. Of course there is not a single right answer for this question. As a matter of fact, who among Wikipedia and news website, in this case deVolkskrant.nl[4], leads the interest of a certain topic is strictly related to the context of the topic or news. For instance, in Figure 2 and in Figure 3 in the appendix two examples, taken from the interface are showed. They are respectively, the page of Winter Olimpics 2014 [11] (1st trending topic of February 2014) and the page of Nelson Mandela[10] (53rd trending topic of February 2014). In Figure 2 it is clearly visible how the views on Wikipedia precedes the news about the topic. A complete different situation is depicted in Figure 3 where the news and the views show a different correlation. In this case, the higher peak of visualizations on Wikipedia occurs in combination with single news in July.

Therefore, even though is not possible to summarize in a single statement the answer of the provided research question, the processing of the data and the visualization proposed in this report represents a little first step further to the direction of having a tool that allows to analyze and understand the complex bond that ties Wikipedia and websites of news. Unfortunately, the tool depicts only a limited overview. This is due to the fact that the dataset of the project contained only data about the dutch Wikipedia and the website deVolkskrant.nl of the year 2014. However, the dataset can be easily expanded. As regards the dataset of the deVolkskrant.nl, using the webservice provided by DBpedia Spotlight, the extraction of the wikipedia articles mentions in the news inside the database took around three hours to be completed. Keeping in consideration that with a local installation of DBPedia Spotlight the response times would be cut down, it is clear that the processing of the data of the news website does not represent a bottleneck and, therefore, more data can be added. On the other hand, it would be useful to consider expanding the visualization taking in consideration wikipedia views of other nationalities. This operation can be done broadening the filter in Step 1b that picks only the data of the select language and introducing a further step to aggregate the views of the same article of different nationalities. Therefore, concerning the research question *"Does the solution provide enough scalability in order to expand the project in future?"* the project can be considered scalable. Eventually, the question *"What kind of technology and techniques should be used to process the data?"* is answered in the section 4

7. REFERENCES

[1] Apache spark. http://spark.apache.org/. Accessed: 2016-03-23.

[2] Dbpedia. http://wiki.dbpedia.org/. Accessed: 2016-03-23.

[3] Dbpedia spotlight. https://github.com/dbpedia-spotlight/dbpedia-spotlight. Accessed: 2016-03-23.

[4] devolkskrant. http://www.volkskrant.nl/. Accessed: 2016-03-23.

[5] Hadoop. http://hadoop.apache.org/. Accessed: 2016-03-23.

[6] Java. https://www.oracle.com/java/index.html/. Accessed: 2016-03-23.

[7] Scala. http://www.scala-lang.org/. Accessed: 2016-03-23.

[8] Surfsara. https://www.surf.nl/en. Accessed: 2016-03-23.

[9] Wikipedia. https://en.wikipedia.org/wiki/Main$_{P}age.Accessed$: $2016-03-23$.

[10] The wikipedia article about nelson mandela. https://nl.wikipedia.org/wiki/Nelson$_{M}andela.Accessed$: $2016-03-29$.

[11] The wikipedia article about winter olimpics 2014. https://nl.wikipedia.org/wiki/Olympische$_{W}interspelen_{2}014.Accessed$: $2016-03-29$.

[12] Wikipedia dumps. https://dumps.wikimedia.org/. Accessed: 2016-03-23.

[13] B. M. Hill and A. Shaw. Consider the redirect: A missing dimension of wikipedia research. In *Proceedings of The International Symposium on Open Collaboration*, page 28. ACM, 2014.

[14] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

8. APPENDIX

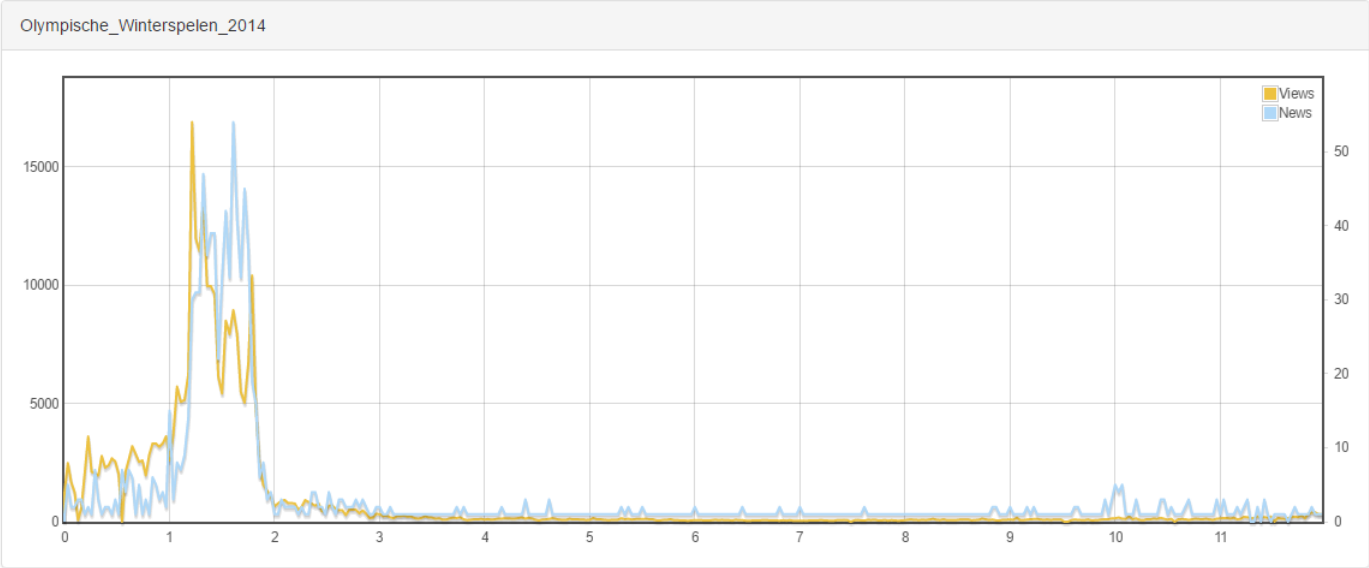

Figure 2: Comparison of the views on Wikipedia page of the Winter Olympics 2014 article and the news that mention it

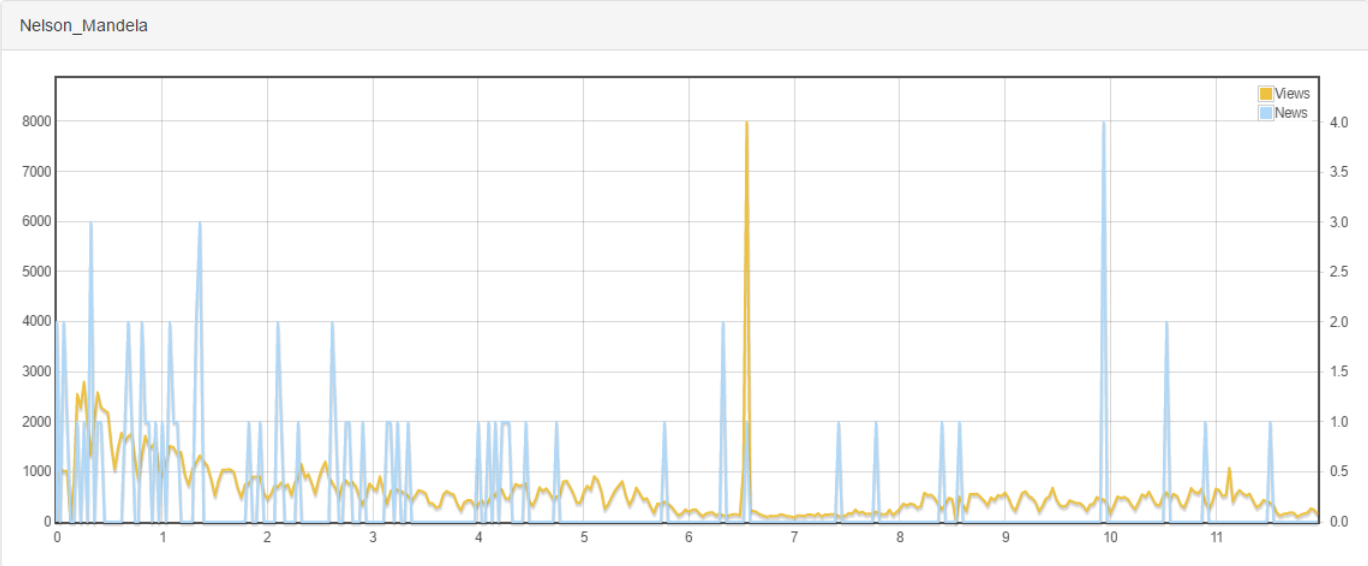

Figure 3: Comparison of the views on Wikipedia page of Nelson Mandela article and the news that mention it

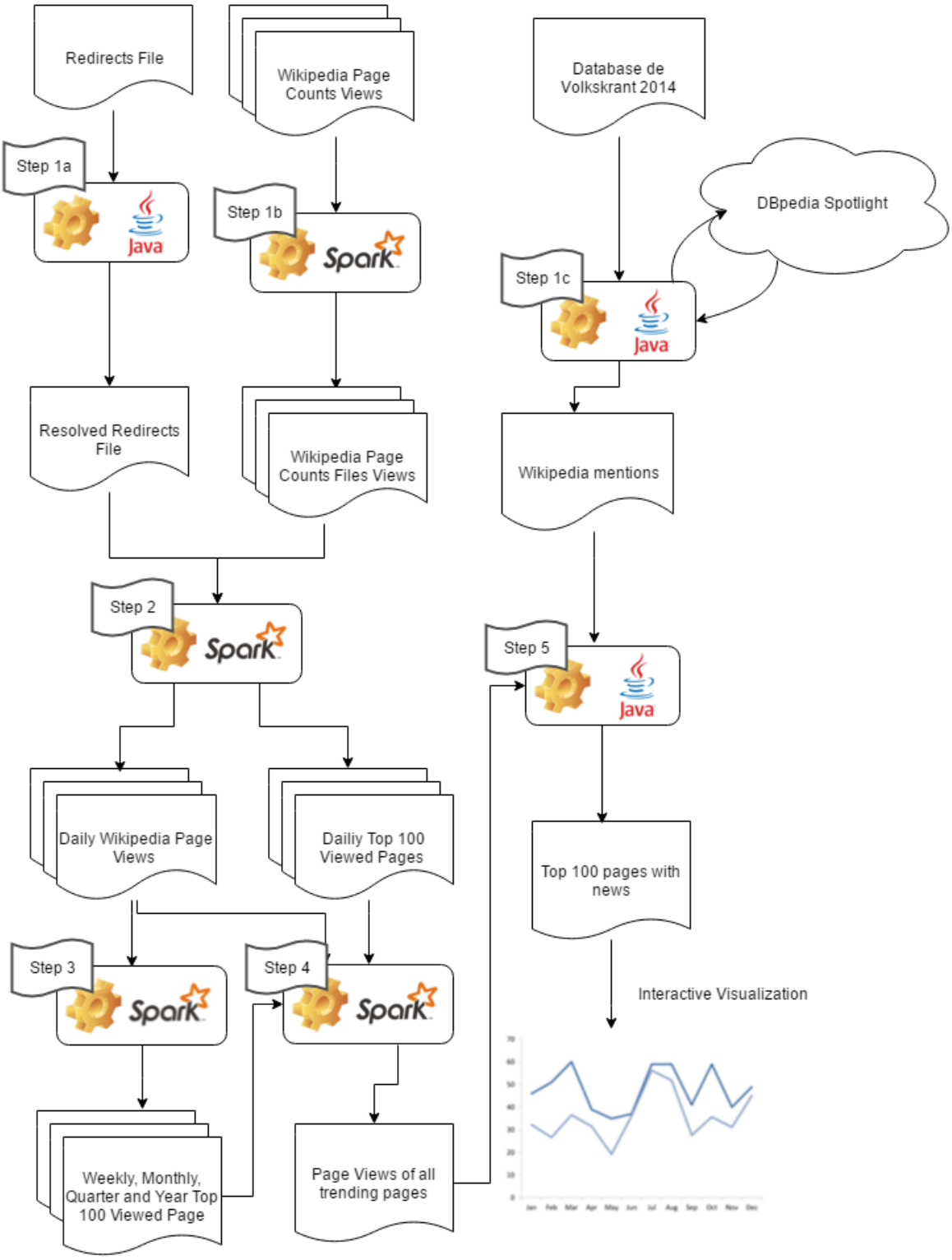

Figure 4: An abstract model representing the steps taken in the project