International Symposium on Visual Analytics Science and Technology (2010), pp. 1–5
The definitive version is available at diglib.eg.org

DYNEVI - DYnamic News Entity VIsualization

F. Wanner[1], M. Schaefer[1], F. Leitner-Fischer [1], F. Zintgraf[1], M. Atkinson[2] and D. A. Keim[1]

[1]University of Konstanz, Germany
[2]EC Joint Research Centre Ispra, Italy

Abstract

Dynamic news entity visualization shows an implementation of visualizing news entity data to give an overview as well as to display emerging and vanishing news topics. We present a robust and dynamic visualization system with case studies that show its benefits and high functionality.

Categories and Subject Descriptors (according to ACM CCS): H.4.m [Information Systems]: Information Systems Applications—Miscellaneous

1. Introduction

News data with associated entities are flooding us daily, hourly and even every few minutes or seconds from news agencies all over the world. To get an overview on the one hand and detect the most important topics on the other, is impossible just by reading, therefore visual tools are essential.

The goal of our work is to create a news illustration by using the word clouds technique. The advantages of our system are the intelligent use of size, color mapping, shapes and ordering to show the entities and meta information including a search possibility to achieve a high functionality. The main task is to display an overview of all entities mentioned in the news and to show the emerging and vanishing ones at a single glance. We also use a sentiment value of the entities which judges the tonality of the underlying news story. We have implemented a very robust system with a dynamic update in a customizable user defined interval in order to get a good picture of what is happening over time. The findings and case studies of this work will show some real examples taken in May and June 2009.

2. Related Work

Our work uses EMM (Europe Media Monitor) news data. The current web interface of the "EMM News Brief" can be found on http://emm.newsbrief.eu. The EMM NewsBrief provides a summary of automatically classified breaking and live news stories from all around the world showing the 10 top stories of the past 4 hour or 24 hour window that is recalculated every 10 minutes.

EMM also provides the "EMM News Explorer" (http://emm.newsexplorer.eu). It automatically generates daily news summaries, allowing users to see the major news stories (news clusters) in various languages for any specific day and to compare how the same events have been reported in the media written in different languages. It also provides a list of most mentioned names and further automatically derived information (eg. variant name spellings, titles and phrases, list of the most recent articles and list of related persons and organisations (entities). We are using entities, extracted from english, french and german news.

Tag clouds have become a more and more popular technique for visualizing text on the internet. Their simple technique: word frequencies are mapped on font size [VW08]. The more frequent the word, the bigger the font size. In recent years users were allowed to generate clouds with their own content [VWF09]. In this context they refer to a number of examples like "single-purpose tools" such as The Tag Cloud Generator [TCG], TagCrowd [SKK*08] or wordle [WDL] as well as "more general visualization sharing sites such as Many Eyes [VWvH*07]". The latter was predominatly created for interaction and data analysis, whereas particularly wordle may support non-experts to visualize and arrange "personally meaningful information" [VWF09]. Pousman et al. [PSM07] use the term "casual infovis" for that kind of usage of visualizations. Another related work is "Ten by ten" [TBT] , which tries to capture news at the moment in

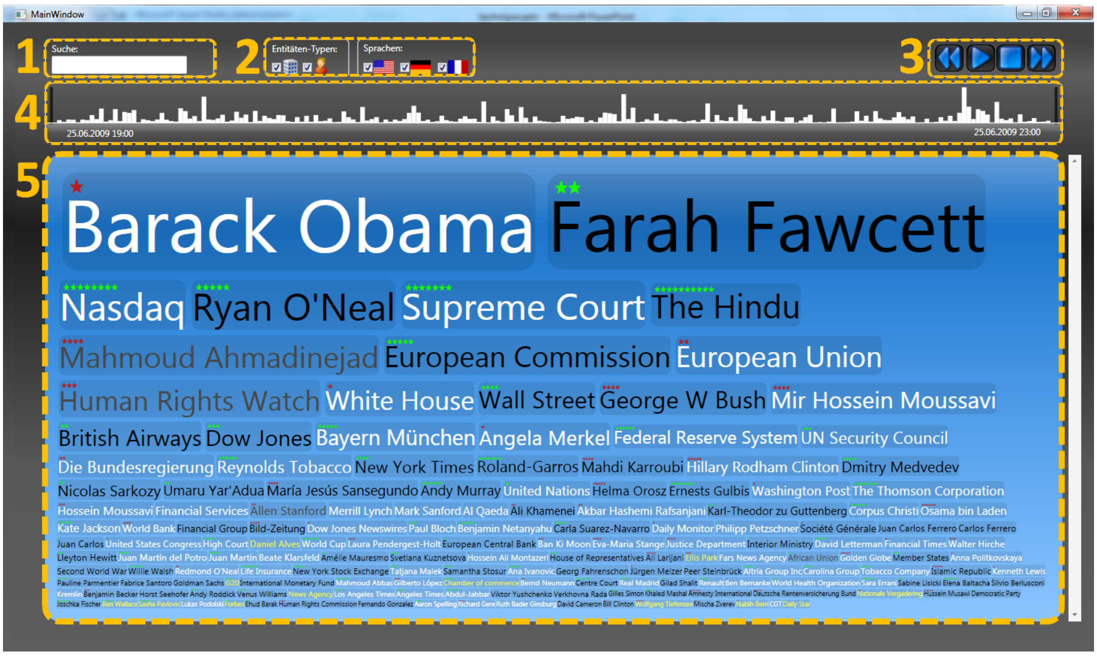


Figure 1: *Visualization of the June 25th 19:00h to June 25th 23:00h. A raw overview of the look of the entire application's components. The main visualization is located in the middle (5). It can be changed by either interacting within the timeline visualization (4) beyond, or by using the media buttons on the upper right side (3). On the upper left side the user can search for entities he is interested in (1), whereas on the right the data can be filtered by entity type (object/person) and by news language (english/german/french)(2).*

words and images. The 100 images are arranged in a 10 x 10 matrix. On the right side of the matrix is a list of the top 100 words in descending order. The correspondending image is highlighted when you shift through the words. In 2005 Ong et al. showed their "newsmap" [OCSZ05], which visualizes the "landscape" of the Google News news aggregator in a treemap visualization.

Vuillemot et al. [VCPK09] mention, that tag clouds are critized in the internet: "time captions are missing and it is not possible to determine what is encoded with what, such as if colors or word orders have meanings". Our approach is related to their work, where they colored words according to their part of speech. Our goal was to modify a well-known "casual" visualization that way, to use it predominantly as an analysis visualization. Therefore, we take the representation of a tag cloud (frequency = font size) and color the terms according to their frequency over time. The frequency at each point in time also determines the ordering of the entities. [VWF09] also mention that other scientists tried to improve conventional tag cloud arrangements. They refer to Seifert et al. [SKK*08] who developed an algorithm to create more compressed word clouds and to Gambette and Veronis [GV09] who presented a visualization called "Tree Cloud". In our application a Likert-style star rating shows the sentiment score per entity. Furthermore we allow the user to search and filter the entities. A time-slider to explore movements of entities over time is also included.

3. Data Source

The Internet gives us the possibility to access a large amount of online news articles originating from thousands of different sources, like reuters, heise or google news. The preprocessed and visualized data in this work is Europe Media Monitor (EMM) data. This is a news aggregator, which collects news articles from over 2,500 sources in 42 languages. These sources include media portals, government websites and commercial news agencies. EMM processes 80,000 - 100,000 articles per day, enriching them with various metadata, such as entities, categories, geo-tags of news source and source country, URL of the article, source, publication date and time, date and language, on which we perform our analysis. The entities are extracted people or objects that occur in the news text. All articles are classified into different categories based on combinations of trigger words. Association of a news article to certain categories gives additional semantic information that can be analyzed further. The sentiment score is an automatically calculated value, which is based on using word weighted adjective lists. It describes the general sentiment of the news text.

In our analysis, we have worked with 5 weeks of data, which was collected in May and June 2009. In total, 243765 articles were collected, with 43125 entity records in English, German and French. In total, our collection is created from 710 news sources.

4. Visualization Technique

Figure 1 gives a raw overview of the look of the entire application's components. The main visualization is located in the middle (5). It can be changed by either interacting within the timeline visualization (4) beyond, or by using the media buttons on the upper right side (3). On the upper left side the user can search for entities he is interested in (1), whereas on the right the data can be filtered by entity type (object/person) and by news language (english/german/french) (2). We display the entities located in the news items within a specified time span. The visualization technique we use is a word cloud where one entity is represented as word (Fig. 1 - 5). The font size of the word depends on how often the entity appears within the period. Since the word size is also affected by the length of the word, the entities are additionally displayed in descending order determined by their frequencies. The main visualization is also zoomable so small objects can be observed easier.

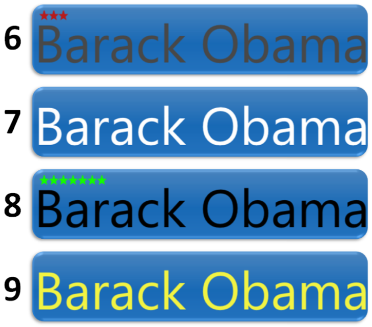


Figure 2: *Examples: Entities with no position change are displayed with a gray color (6). Emerging entities are marked white (7), vanishing entities are marked black (8). If a new entity appears the first time it will be marked yellow (9). The amount of stars expresses the intensity and the red color shows the negative sentiment orientation of the news where the entity was mentioned in (6). (8) illustrates an extremly positive sentiment orientation. Neutral sentiments (zero-values) do not own a star rating (7)+(9).*

Additionally every entity obtains the average of its sentiment values. It is represented through a star shape that leads to a star rating. The amount of stars expresses the intensity and the color expresses the positive (green) or negative (red) orientation[†]. Neutral sentiments (zero-values) do not own a star rating. These three states are displayed in Fig. 2. The main aspect of the application is not to display a static picture of one period but the alteration between a period change. Therefore the user can interact with the timeline (Fig. 1 - 4) by using the mouse. It displays the amount of news per minute in a simple bar diagram. The user is able to move and resize the time span and even to pick a certain date. In addition, the period changes can be automized like consumer electronics: the alteration can be replayed, paused, stopped, fast-forwarded and rewinded. These alterations can be recognized by the color-changes, the position changes and the size-changes of the entities. The different change types can be looked-up in Fig. 2. Climbing entities are marked white (7), falling entities are marked black (8). Entities with no position change are displayed with a gray color (6). If a new entity appears the first time it will be marked yellow (9). To recognize every change technically, the entities are stored within a hashtable. After a change the entity's new position is compared to its old position. The inter-entity relationships can be identified by the user searching for entity names. The entities which do not share any connection with the news items delivered by the search term become transparent but are still visible to the user. Additionally he can also select certain entities to have a better track on its position changes.

5. Application and Comparison

In the following two subsections, we present two case studies which give an overview of the capabilities and features of the visualization. On the one hand we show, how relations between entities in news can be visualized and on the other hand, we discuss how the overall frequency of an entity in the news and sentiment can be visualized. Additionally, we compare our visualization with the Wordle[‡] visualization. The input data for Wordle was created with our application.

5.1. Who is mentioned with whom

DYNEVI offers the ability to show the relations between different Entities. In Fig. 3 we see the news of the evening of the 25th June 2009. The top mentioned person is the actress Fara Fawcett, who passed away after a long illness. The second top mentioned person is the president of the united states of America Barack Obama. To see who of the other persons is mentioned together with Barack Obama, we type "Barack Obama" in the search filter. All persons who are not mentioned with Obama in one article are now displayed semitransparent, consequently all persons that are mentioned with Obama in one article move to the foreground but remain still in context (see Fig. 4). The visualization shows that the

[†] The green/red coloring is not optimal as some people are not able to distinguish these two colors. But for the bigger part of people green implicates positive and red negative values. Therefore the green/red coloring is the best choice in this case.

[‡] www.wordle.net

Iranian president Mahmoud Ahmadinejad and the former US president George W. Bush have negative sentiment values as indicated by the red stars in the upper-left corner of their names. Dmitry Medvedev, the president of the Russian federation has a positive sentiment value indicated by the green stars.

Figure 5 shows the Wordle visualization of the same data as in Figure 3. We used the color scheme "Wordly" and the layout mode horizontal. Note that even though the entities are colored, the color is random and there is no relation to either sentiment or frequency. The size of the names, represents the frequency of the names, but the random position makes it hard to compare the size, for instance it is difficult to say whether the frequency of Mir Hossein Moussavi is higher as the frequency of George W Bush. It is much easier and faster to make this comparison with our application. Because of the generality of the Wordle visualization it is not possible to achieve a similar result as in Fig. 4.

5.2. The Death of Michael Jackson

On June 25th at 23:26 MESZ Michael Jackson was declared death in the UCLA Medical Center in Los Angeles. Within minutes, Michael Jackson is dominating the news as can be seen by the big size on the first position in Fig. 6.The Los Angeles Times who first published the Story about Michael Jacksons death is the most quoted news paper. There are two reasons for the relatively small amount of entities in this visualization: The first is that only news originated from Europe are considered for the visualization and during the night there is a relatively small amount of news items. The second reason is that only the upper 90 percent of the mentioned entities are visualized. In the time period of interest, Michael Jackson dominates this 90 percent. In other words, the news of Michael Jackson's death supersedes all other news stories.

The data of Figure 6 is visualized with Wordle in Figure 7, even with a very small number of names it is hard to make a comparison of the frequency.

6. Conclusion and Further Work

In this paper, we presented a dynamic visual system that was developed by the Data Analysis and Visualization Group at the University of Konstanz, Germany and the European Commission's Joint Research Centre in Ispra, Italy. The case studies showed the advantages of the system. The comparison with the aesthetically oriented Wordle visualization [VWF09] illustrated, that a specialized news visualization approach is functionally superior. Our future work focuses on an dynamic integration of the system in a news monitor system for large high definition displays.

References

[GV09] Gambette P., Véronis J.: Visualising a Text with a Tree Cloud. In *IFCS'09: International Federation of Classification Societies Conference* (2009), Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin / Heidelberg, p. 8. 2

[OCSZ05] Ong T.-H., Chen H., Sung W.-k., Zhu B.: Newsmap: a knowledge map for online news. *Decis. Support Syst. 39*, 4 (2005), 583–597. 2

[PSM07] Pousman Z., Stasko J., Mateas M.: Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics 13* (2007), 1145–1152. 1

[SKK*08] Seifert C., Kump B., Kienreich W., Granitzer G., Granitzer M.: On the beauty and usability of tag clouds. In *IV '08: Proceedings of the 2008 12th International Conference Information Visualisation* (Washington, DC, USA, 2008), IEEE Computer Society, pp. 17–25. 1, 2

[TBT] 10x10 ('ten by ten'), http://www.tenbyten.org/, accessed april 23rd, 2010. 1

[TCG] Tag cloud generator, http://www.tagcloud-generator.com/, accessed april 23rd, 2010. 1

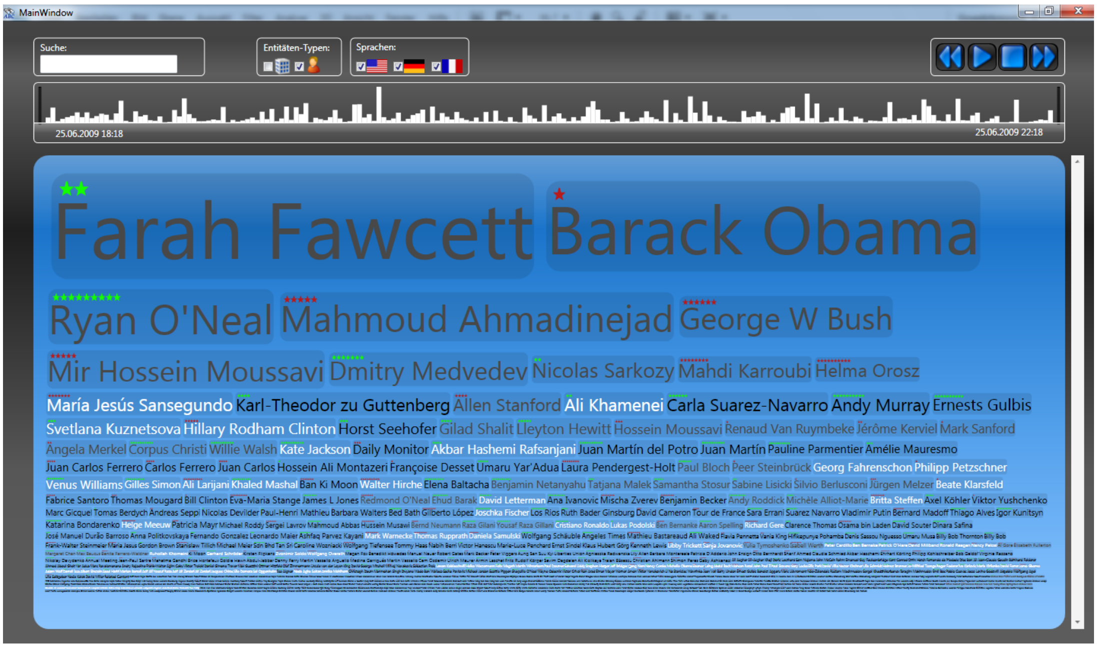
[VCPK09] Vuillemot R., Clement T., Plaisant C., Kumar A.: What's Being Said Near "Martha"? Exploring Name Entities in Literary Text Collections. In *IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST)* (Oct. 2009), pp. 107–114. 2

[VW08] Viégas F. B., Wattenberg M.: Timelines tag clouds and the case for vernacular visualization. *interactions 15*, 4 (2008), 49–52. 1

[VWF09] Viégas F. B., Wattenberg M., Feinberg J.: Participatory visualization with wordle. *IEEE Trans. Vis. Comput. Graph. 15*, 6 (2009), 1137–1144. 1, 2, 4

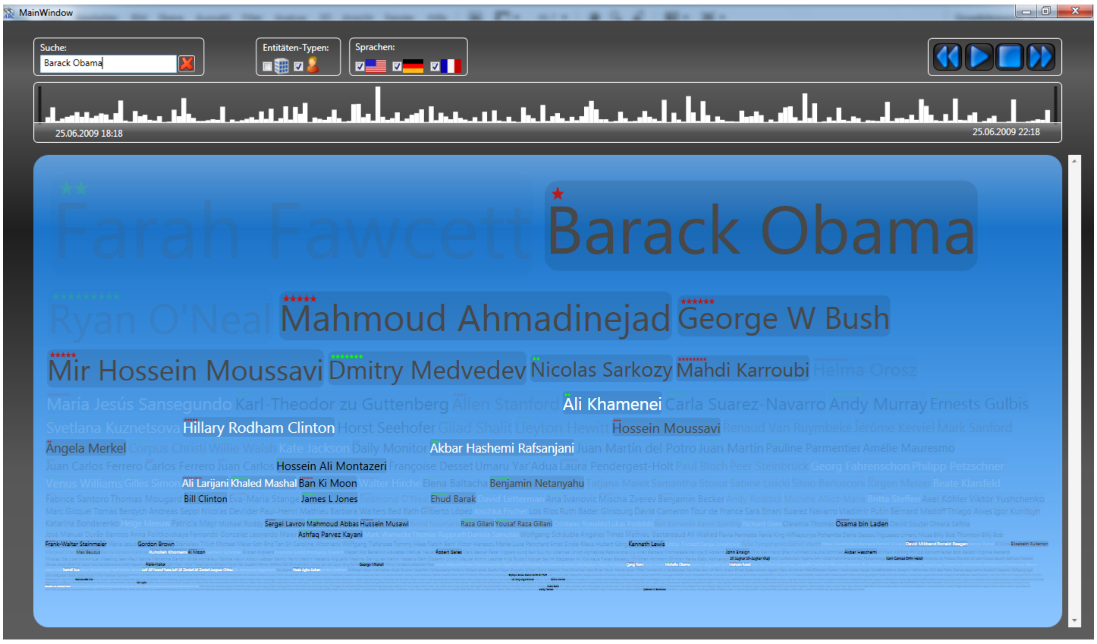
[VWvH*07] Viegas F. B., Wattenberg M., van Ham F., Kriss J., McKeon M.: Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics 13* (2007), 1121–1128. 1

[WDL] wordle, http://www.wordle.net/, april 23rd, 2010. 1


MainWindow

Suche:

Figure 3: *Visualization of the June 25th 18:18h to June 25th 22:18h, with applied persons filter.*


MainWindow

Suche: Barack Obama

Figure 4: *Visualization of the June 25th 18:18h to June 25th 22:18h, with applied persons filter and search filter "Barack Obama"*

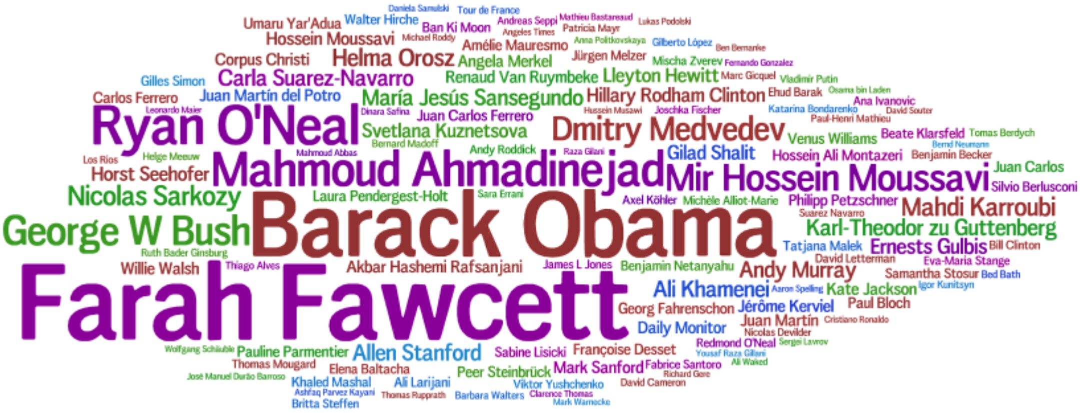


Figure 5: *Wordle visualization of the June 25th 18:18h to June 25th 22:18h, with applied persons filter.*

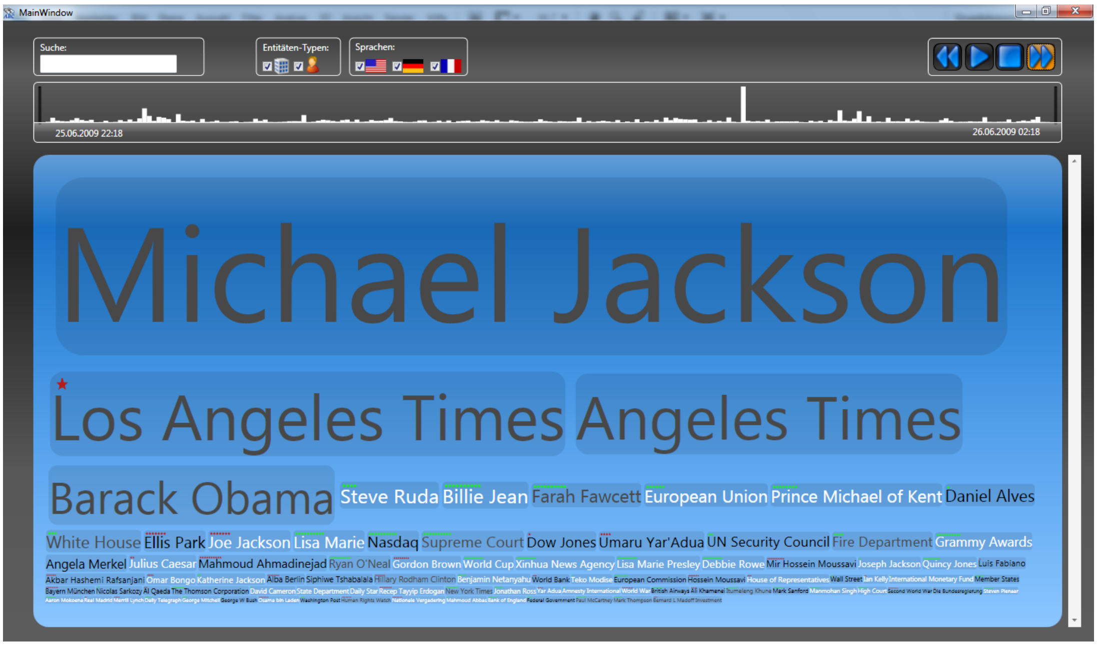


Figure 6: *Visualization of the June 25th 22:18h to June 26th 02:18h: Shows the easy detection of the highest frequency like "Micheal Jackson" with word size and position. The comparison between the entities can be seen with "Barack Obama" and "Steve Ruda".*

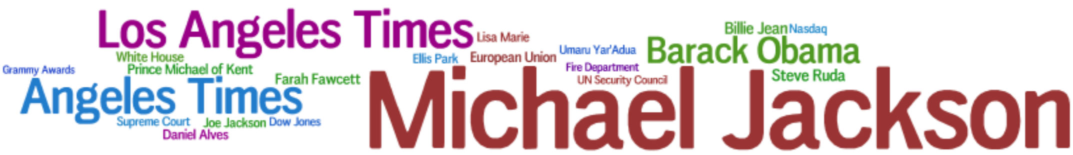


Figure 7: *Wordle visualization of the June 25th 22:18h to June 26th 00:18h: Aesthetically nice but the functional advantages like the detecting of the ranking order is not possible.*