Death 'n Curse: visualizing Tarantino's vulgarity

150022854

April 21, 2016

Abstract

The dataset I chose is *Swearing and dieing in Tarantino's movies*. This paper describes the visualisation generated from such dataset: first by introducing the data, then by describing the ideation process followed and the end concept chosen. Consequently, its implementation in Processing will be discussed, followed by the insights gathered from such visualisation. Finally, a critique on its limitations will be addressed.

1 The Dataset

The data used is (mainly) a collection of counts of death occurrences and swearwords pronounced in Tarantino's films. The attributes available were divided in two sections.

- First Section:
  - Movie: name of film. Categorical
  - type: type of occurrence: word or death. Categorical
  - word: exact word pronounced. Empty for death. Categorical
  - minutes in: exact time in minutes of the occurrence. Quantitative
- Second Section:
  - Movie: name of film. Categorical
  - run time: total running time of film. Quantitative
  - release year: year of the film commercial release. Quantitative
  - Budget in Million: total budget of film. Quantitative
  - Box Office in Million: total revenues for the movie. Quantitative
  - description: long string of text providing the description to the movie.

In addition, I derived a new attribute called *Percentile*: it represent occurrence of a curse word/death expressed in percentile of a movie run time. In this way, all the occurrences were normalized in a scale [0, 1] to allow for comparison between movies with different running times. Moreover, external data was gathered from Lexicon regarding the most frequent curse words used on U.S. Facebook interactions.

The questions that my visualisation allows to answer are:

1. What are the most common swear words in Tarantino movies? Did they change depending on the film?
2. Are there (absolute or relative) temporal pattern in which swearing and deaths occur in Tarantino movies?
3. Is there a relation between popularity of a Tarantino movie (based on the box office), its budget and the amount of violence and profane language?

2 Ideation Process

In order to generate the visualisation, I followed the "Five Design Sheet Methodology" [1]. In this way, I was able to first generate ideas and subsequently converge to a final design through the systematic application of the theory studied in class and summarized in Munzner's textbook [5] (See attached sketches). Nevertheless (unfortunately) I was forced to take into account implementation constraints even during the selection process due to my limited experience with Java code, which influenced and limited the possible design space. Here, I will describe the most relevant alternatives that were not chosen, dividing them between the question they would have answered.

Question 1: the first design was based on the proportions of words: in this visualisation, the number of appearances of a word would have encoded the importance of it: *e.g.* if *Fuck* represented 60% of all swearwords, it would be encoded by 60 repetitions of it: see Figure 1 for an example. However, such design was rejected because it did not provide the exact counts but just the relative proportions. A second alternative was to encode word counts in the sizes of words, *i.e.* the more a word is used in Tarantino's, the greater its size. Nevertheless, it was ultimately not chosen for the same reasons as above: it did not provide enough information.

Queston 2: the first concept for this question was similar to the one ultimately chosen. The source of such idea can be found at blprnt.blg. Horizontal position would have encoded *Percentile* while a low saturation mark each occurrence of swear word/death in each movie. Therefore, the result would have been a single line composed by marks aggregating all movies: overlapping marks would have increased the saturation of that spot, identifying more common "swearing times". However, this idea was rejected due to its limitation by

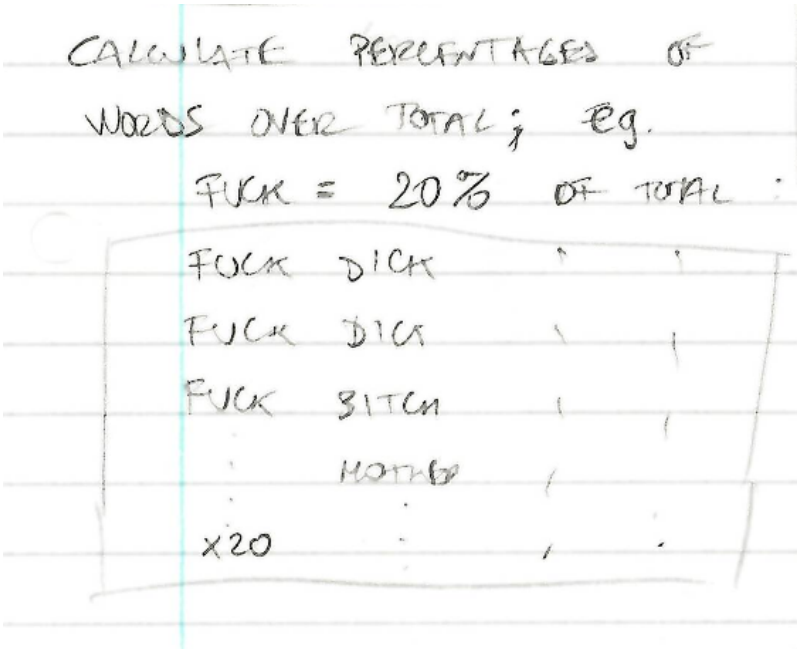

Figure 1: Words in proportion: every words appears as many times as its percentage total in all movies

showing just the aggregate of all movies and because saturation has a perceptual limitation for the length characteristic: see Figure 2 for an example. The second idea was based on a line chart, where vertical position encoded counts, horizontal encoded percentiles and hue was used for movie titles. This design was not chosen for two reasons: firstly, I wanted to experiment with a new (relative to my experience) visualisation encoding and secondly because it would have required an interaction (*e.g.* a filter) based on movie title to better convey information: in fact, 8 different lines would have probably rendered the visualisation quite ineffective.

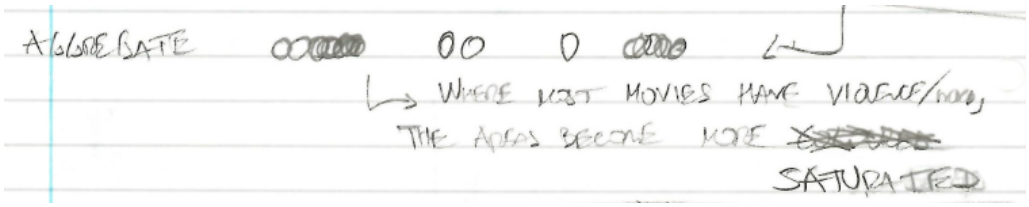

Figure 2: Saturation line: the more frequent are occurrences at a certain percentile, the more saturated will be the mark at that percentile

Question 3: for the last question the alternative concept would have been a fusion of a scatterplot with two labels atop each data-point to show the most used curse words for that particular movie (see Figure 3). The encoding would have been: horizontal position for revenues and vertical position for budget. However, no other clue would have been provided with respect to the numerical word distribution, their relative importance and

so on. Therefore, it would not have given a complete answer to the question (*i.e.* no clue of the relationship between vulgarity and revenues would have been present).

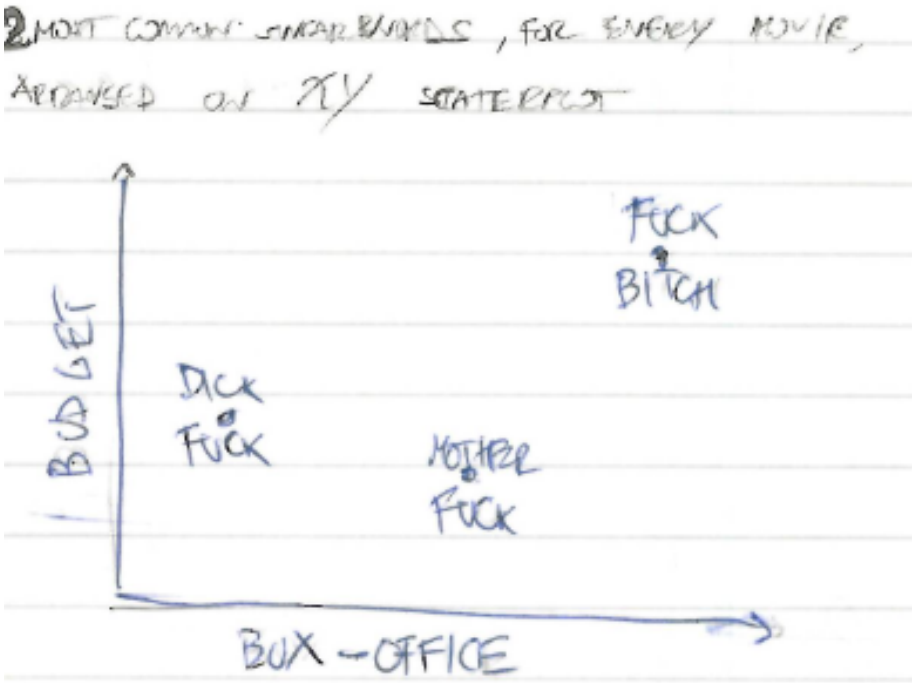

Figure 3: Scatterplots with most frequent words

3 Description of Visualisation Concept

The visualisation is composed of three views: *Tarantino's violence & swear pattern*, *Tarantino's most common swearwords* and *Do violence and profanity make a blockbuster?*. To navigate between the views, the user must press a *Next* button positioned in each one of them. The first visualisation is based on a series of vertical line marks, representing each single occurrence of curse word or death. Their horizontal position encodes the percentile of movie time in which they happen. Vertical position is applied to differentiate between movie titles, which are arranged in increasing order of time (from oldest to newest). Hue does not encode any value: a plain colour scheme was used (black and white) to cope with the width size of marks, which therefore required maximum saturation and luminance to ensure contrast with the background. Other combinations of colour were tried, but none of them provided enough visibility as black & white, even though it is not the best choice from the aesthetic point of view. See Figure 4 for the screenshot of the visualisation.

The second page can be accessed (as the third) by clicking on the *Next* button. This view is composed by a bar-chart: obliviously, a line mark is used with size (variation in length) to encode word counts, while horizontal position represents the different word

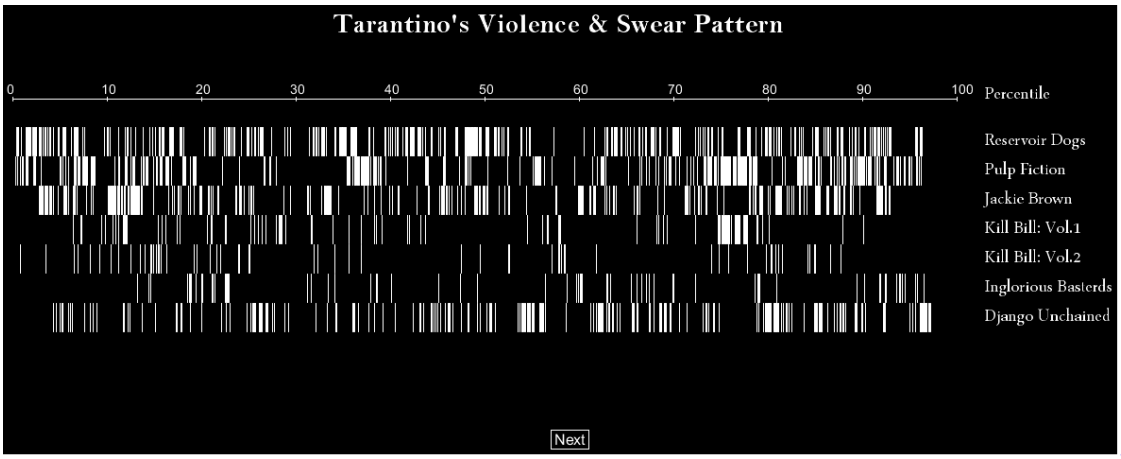

Figure 4: First view of the visualisation: the swearing & death pattern in Tarantino's movies. Note the *Next* button in the lower part.

labels. Note that not all used words are represented: only the first 13 are shown to avoid overcrowding of the visualisation space. In addition, a list is provided which shows the most common curse words used on U.S. Facebook. This allows a comparison between Tarantino's vocabulary and the everyday one. The choice of colour scheme was guided again by the size of the mark: in this case, the bars are quite big and require therefore limited saturation. For this reason the colour used is composed of hue N 230 with 80% of brightness and 60% of saturation. Unfortunately, I was unable to change the colour of the axis: indeed, their luminance should be decreased towards black, to avoid interaction with the background. From this main chart, which provides an overview of the most used curse words in Tarantino's filmography, the user can subset the data (by clicking on the *By Movie* button) and visualize the barchart for every single film in isolation. I also included a *Back* button to allow for playback: this ensure a more fluid interaction with the visualisation. See Figure 5 for the screenshot of the aggregate barchart.

The last view is instead based on a bubbleplot with vertical position encoding counts of curse words/deaths, horizontal position representing box-office revenues and area marks with size as visual variables (*i.e.* variation in area) to encode budget. The colour scheme was again based on similar consideration as before, in order to cope with its interaction with size. In this case, the presence of both small and big marks forced the choice in favour of a compromise for saturation. The colour chosen has hue N 230, with both saturation and brightness at 80%. Also in this case I was unable to modify the colour for the axis and the labels. See Figure 6 for a representation of this last view. By clicking on *Next*, the first view reappears.

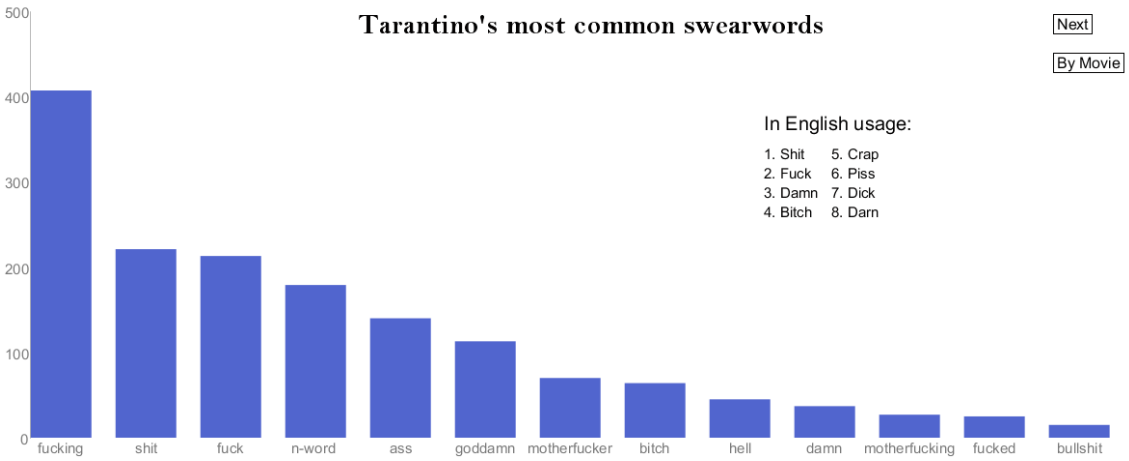

Figure 5: Second view: the barchart with the frequency of swear words. Note the two buttons in the higher left corner.

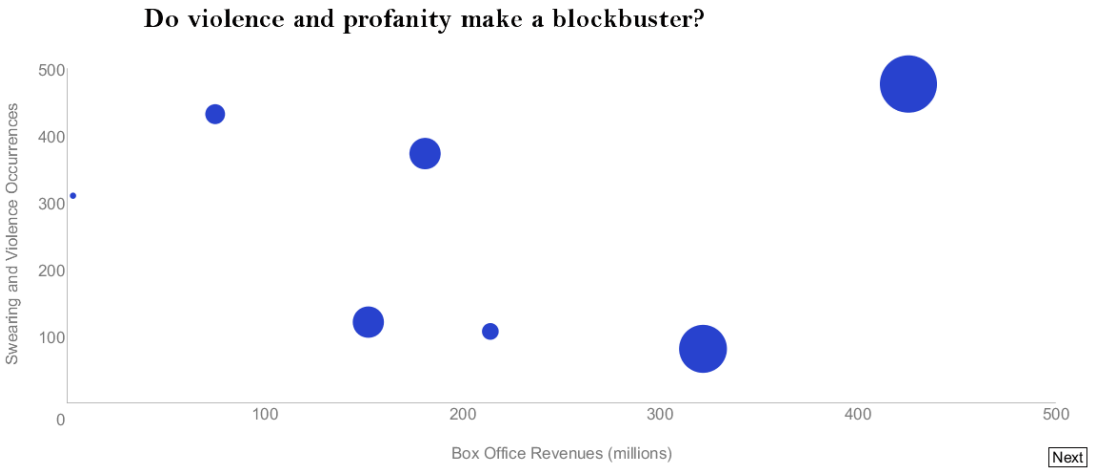

Figure 6: Third view: the bubbleplot

4 Implementation

To construct such visualisation, the Java based Processing language [4] was used together with R [3] for data manipulation. Unfortunately, my limited programming experience constituted a significant obstacle for implementation, especially with respect to interactivity. Indeed, many aspects of the algorithm require improvements. First of all, the loading of data is not executed as a single table/file but through a sequence of multiple, already prepared datasets. For sure, such method is highly inefficient, and would likely influence latency in a negative way if bigger datasets were used.

To implement the first view I mainly used a series of for loops while for the other two views, I applied the methods found in the giCentre[1] library. All three views are enclosed

[1] org.gicentre.utils.stat

by the *draw* method which allows for interactivity through a button. The button was devised thanks to the source code provided as an example on the Processing webpage. I extended it in order to allow the user to navigate between different visualisations and also to filter the barchart data in the third view.

5 Insights from the Visualisation

It is possible to divide the insights gained depending on the three different views. From the first view, the user can see that no pattern seems present: violence and cursing are distributed quite randomly both within and between movies. Nevertheless, an absolute pattern is quite visible: violence and cursing seem to be diminishing during time: *Kill Bill: Vol. 1/2* and *Inglorious Basterds* show significantly less marks. However, *Django Unchained* presents an inversion to this trend. Note that no movie has violence after around the 97% percentile: this is due to the end titles, which are included in run times.

From the second view, the insight is straightforward: the most common swearwords in Tarantino's movies are *fucking* and *shit*. Moreover, through a comparison with the most common curse words on Facebook, we see that Tarantino choice is different from normal usage. In addition, by clicking on *By Movie* button, it is possible to see how the most used curse words vary significantly within movies. For example, the *Kill Bill* saga presents a similar pattern of curse words, which is quite different from other movies.

Finally, the answer to the third question is negative and it's provided by the third view. It seems that the best predictor of box-office revenues is budget: we see that the rightmost (hence more profitable) movies a bigger mark size (hence a higher budget). On the contrary, the second data-point has a higher swear words count than the penultimate, but, at the same time, also lower revenues.

6 Limitations

My visualisation falls on two aspect: interactivity and aesthetic. The former could solved by implementing a filter to the curse words barchart, instead of forcing the user to pass through all movies. A pop-up window activated by mouse click, could also be added to the bubbleplot, allowing the user to receive more information on the movie. In addition, the circular path that the user is forced to follow could be avoided by jumping directly to the view of interest. Secondly, I recognize that the aesthetic value is pretty simple, given that the visualisation does not incorporate any new and catchy encoding paradigm. On the contrary, it uses standard data visualisation techniques which fail to captivate the user.

References

[1] Roberts JC, Headleand C, Ritsos PD, "Sketching designs using the Five Design-Sheet methodology", *Visualization and Computer Graphics*, IEEE Transactions, 22(1), 419-428 (2016).

[2] Heer J, Bostock M, Ogievetsky V, "A tour through the visualization zoo", *Communications of the ACM* 53-59 (2010)

[3] R Core Team, "R: A Language and Environment for Statistical Computing", *R Foundation for Statistical Computing*, Vienna (2015)

[4] Fry B, Reas C, "Processing", *Processing Foundation* (2012)

[5] Munzner T, "Visualization analysis and design", *CRC Press, Taylor Francis Group*, Boca Raton (2015)