The different sentiment values of books
based on gender, century and location.

Text Mining and Collective Intelligence
20/12/2016
Elise Mol

Introduction

Whether or not male and female writing differs has been extensively researched, and although there is a discussion on how this should be measured, the overall opinion in the field is that there is a difference. The aim of the conducted research is not to find an answer to this debated question, but rather to shed light on one specific aspect of this topic. One of the cornerstones of many of the theories about the proposed difference is that where female authors use more emotion words in their writing, male authors use more articles (Newman et al., 2016; Yu, 2013; Schler, 2005). In this research the amount of emotion words will be equated as to the amount of expression of sentiment in a text. This could relate to positive, negative or neutral sentiment, although this research will focus on the positive sentiment. By applying sentiment mining on texts written by female authors and male authors, previous literature would predict a higher sentiment score for texts with female authors, and this is what will be tested. Furthermore the influence of time and location on the difference in male and female writing will be discussed.

Literature Review

That emotion words are indeed an indicator of female writing was found by Koppel et al. and by Newman et al. (2002; 2016). Koppel et al. found that it was possible to automatically classify the author of a text as either male or female by applying machine learning. He found many indicators of male or female writing and among those, that female authors seem to use more negations, pronouns, and certain prepositions, which are all indicators of a higher sentiment value (2002). By using this pattern, and many others, Koppel et al. were able to correctly classify 80 percent of the unseen texts from a genre constricted subset from the BNC. That this subset was genre constricted was to avoid that the differences in writing based on genre would interfere with the differences caused by gender. Our research will also use this tactic. All documents will be of the same category, namely novels.

Newman et al conducted a similar study but instead of classifying on 200 to 300 documents they expanded their research by training on, and classifying 14000 documents. They found there is indeed a difference between male and female writing, and they found the most important features for classifying a text with a female author were related to sentiment (2016). The method Newman et al. used to measure whether there was a difference between male and female writing was by using the Cohen d. The Cohen d, a value that is widely used in this field, is the way in which to

measure the effect. Although the original recommendation by Cohen was that a value of 0.2 for d should indicate a small effect, 0.5 should indicate a medium effect and 0.8 should indicate a large effect, Newman et al. edited these values because of the specifics of their research (0.1 was an indicator of a small effect) (Cohen, 1988).

Another study that utilised the Cohen d measurement was a study by Bei Yu (2013). His research was related to the proposed difference between male and female writing, but focused especially on this difference in a professional setting; Congress. He hypothesised the difference between male and female speech would differ less as Congress is a male dominated environment and a formal setting. This is indeed what he found, but that there was not as much difference did not mean there was no difference at all. In fact, it was found that female legislators used more emotion words and other text attributes that are considered to be feminine compared to the male legislators.

Data set

The data used to conduct this research was obtained from the Gutenberg project. Originally the way in which this data would be gathered was by using a python Library to interface with Project Gutenberg, but this method turned out to be ill suited. As web scraping is not allowed by Gutenberg, and there is no API, the data was stored locally after being downloaded by hand from the site. The contents of these downloaded files were books. The books were selected in one of the following categories: British female novel writers from the 18th century, British female novel writers from the 19th century, British male novel writers from the 18th century, British male novel writers from the 19th century, American male novel writers from the 19th century, American female novel writers from the 19th century. In each of these categories around 20 authors were selected, based on the combination of the following sources (19th-century English novelists, 2016; 18th-century English novelists, 2016; 19th-century American novelists, 2016) and the availability of the books on Gutenberg. The authors were selected in alphabetical order until either the amount of books reached twenty or there were no more authors available in this category. All the writers were novel writers in order to avoid confusion between the differences in genre or gender as mentioned in the previous section.

The original format of the data was a text file filled with Unicode. The data was then transformed and made to be roughly the same length to avoid bias.

Method

The aim of the research is to compute the difference between male and female writing and to research whether this difference changes per location or century. The way in which this is achieved is by transforming a directory of text files into a dictionary with the mean and standard deviation of the positive sentiment. As each dictionary represents one of the previously mentioned categories, a Cohen d score can then be computed to measure the differences between these categories.

In the original directory, all the files are text files of books, and to be able to compute the sentiment for each text file, the files need to be transformed to fit the type that is allowed for the HTTP POST belonging to the nltk sentiment miner [text-processing.com/demo/sentiment/]. This type is a string with no more than 50000 characters. To obtain these strings, each file in the directory is reduced to exclusively the text of the author, and from that text random sentences are selected until the amount of characters is almost 50000, on which the sentiment will be computed. As this is done for all the files in the directory, a list is obtained with n strings of less than 50000 characters, where n is the amount of authors in this category. When posting all these individual strings to the nltk sentiment miner, a json object is returned. Now a list of json objects is available, with n items. This json object can be interpreted as a nested dictionary where the first key is *probability* and the second is *label*. The probability is the aim of this research, and we do not work with the label as it can also be obtained from the probability. The value for this probability is again a dictionary with three keys. The first key is *pos*, the second *neg* and the third is *neutral*.

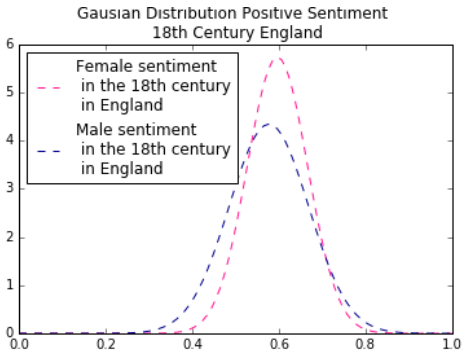
This list of json objects is then transformed into a single dictionary with six values. These values are the mean of each of the probabilities of the sentiment, *pos, neg* or *neutral*, and the standard deviation of these values in the list. This process is repeated 4 iterations to avoid skewed data. This skewed data can arise as the random selection of sentences can cause for unreliable means and standard deviations.

This dictionary of means and standard deviations can then be computed for all directories, where each directory represents one of the following categories mentioned above. For each of these categories, a Gaussian distribution can be plotted that corresponds to the positive sentiment, negative sentiment or neutral sentiment. As this research is focused on the differences in positive sentiment, this visualisation will be made for those that fit in that category.

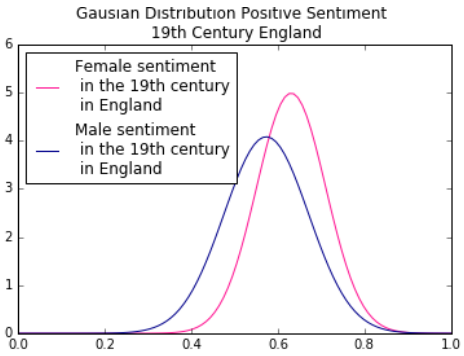
These can then be compared, and a Cohen score can be computed to measure the difference. These results can be seen in the following figures.

Results

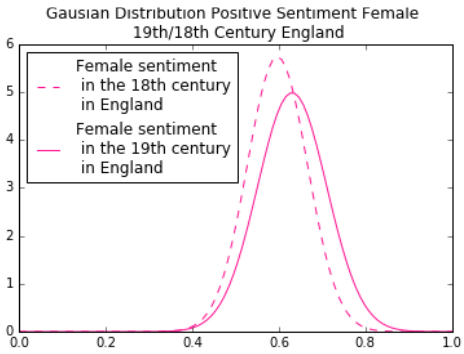
d = -0.242784527854



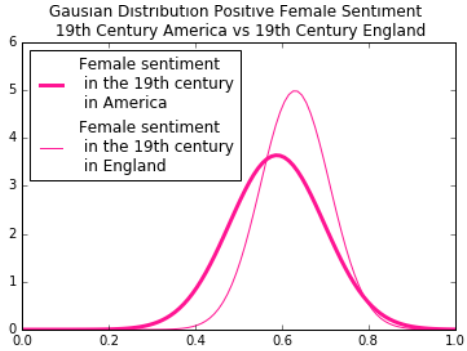
d = -0.642291481163



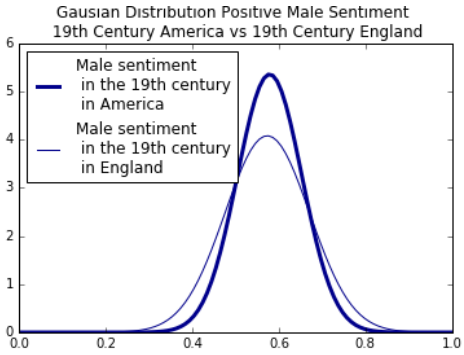
d = -0.450871518459



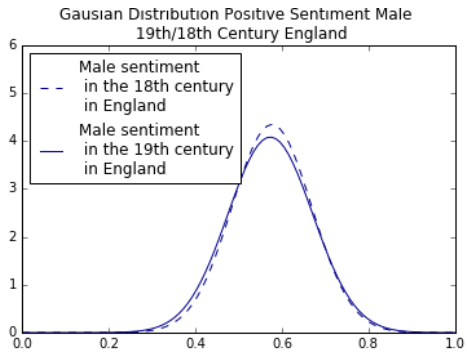
d = -0.440891084784



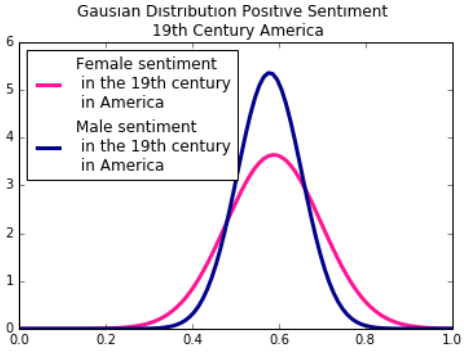
d = -0.440891084784



d = -0.440891084784



d = -0.103187554972



Conclusion

As can be deducted from the Cohen scores represented above the figures, there is a small difference between male and female writing in England in the 18th century, but a much larger (0.71) difference in the 19th century. The difference between male and female positive sentiment in 19th century America is especially low, and not significant at all as its effect is below 0.2. Comparison between centuries shows that the female sentiment in England from the 18th to 19th century has an almost medium effect, where the male positive sentiment has not changed, indicated by a score of 0.004.

Comparison between locations shows quite a difference.

The amount of positive sentiment in the 19th century for English female writers is higher than for their colleagues in America, with a medium value for d. Where the location made a medium difference for female writers, this is not the case for their male colleagues.

It is hard to base reliable statistics on such a small subset of works, and future researchers are encouraged to continue this line of study. Suggestions for improving this research are: working with more indicators of difference, not just sentiment mining, and improving the sentiment mining as the nltk sentiment miner is based on current use of English, and might therefore not be suitable to the 18th or 19th century. Furthermore, the use of a larger subset then 20 authors per category is encouraged and a more subsets in this data based on genre, compared to only one genre; novels.

References

Argamona, S., Koppelb, M., Finec, J., & Shimonib, A. R. (2003). Gender, Genre, and Writing Style in Formal Written Texts. To appear in Text, 23, 3.

Biber, D., & Burges, J. (2000). Historical change in the language use of women and men: gender differences in dramatic dialogue. Journal of English Linguistics, 28(1), 21-37.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd edn. London, UK: Routledge Academic. Dahlerup,

Koppel, M., Argamon, S., & Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. Literary and Linguistic Computing, 17(4), 401-412.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. Discourse Processes, 45(3), 211-236.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006, March). Effects of Age and Gender on Blogging. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (Vol. 6, pp. 199-205).

19th-century English novelists. In Wikipedia. Retrieved December 20, 2016, from https://en.wikipedia.org/wiki/Category:19th-century_English_novelists

18th-century English novelists. In Wikipedia. Retrieved December 20, 2016, from https://en.wikipedia.org/wiki/Category:18th-century_English_novelists

19th-century American novelists. In Wikipedia. Retrieved December 20, 2016, from https://en.wikipedia.org/wiki/Category:19th-century_American_novelists