Visualizing Document Image Collections using Image-based Word Clouds

Tomas Wilkinson and Anders Brun

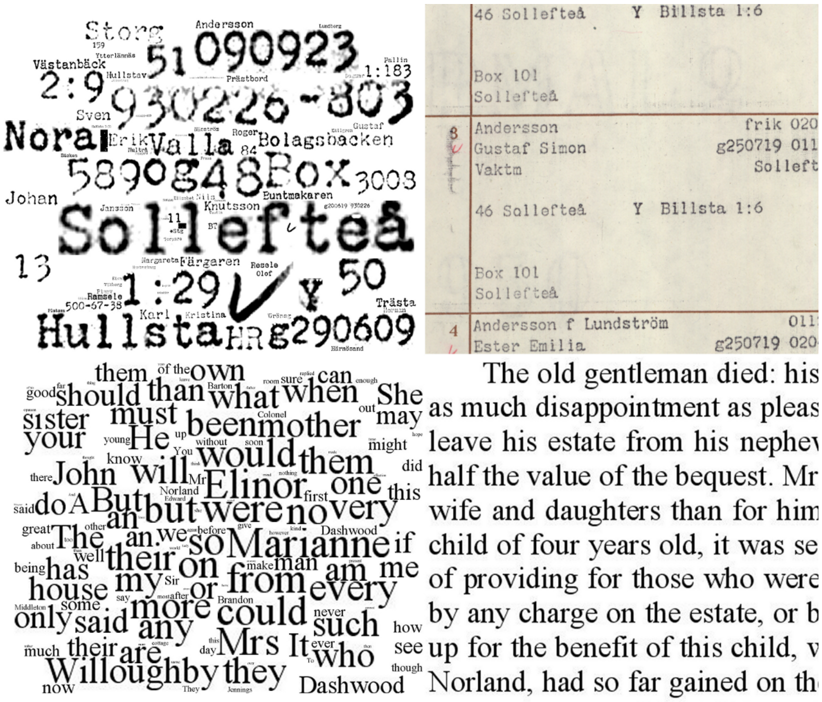
Department of Information Technology
Uppsala University
{tomas.wilkinson, anders.brun}@it.uu.se

Abstract. In this paper, we introduce image-based word clouds as a novel tool for a quick and aesthetic overviews of common words in collections of digitized text manuscripts. While OCR can be used to enable summaries and search functionality to printed modern text, historical and handwritten documents remains a challenge. By segmenting and counting word images, without applying manual transcription or OCR, we have developed a method that can produce word or tag clouds from document collections. Our new tool is not limited to any specific kind of text. We make further contributions in ways of stop-word removal, class based feature weighting and visualization. An evaluation of the proposed tool includes comparisons with ground truth word clouds on handwritten marriage licenses from the 17th century and the George Washington database of handwritten letters, from the 18th century. Our experiments show that image-based word clouds capture the same information, albeit approximately, as the regular word clouds based on text data.

1 Introduction

In the last decade, word clouds have evolved to become a common tool on blogs and websites. They have the power to compile a large amount of text into a single image that grants the viewer an overview of the text's contents from which properties of the text may be inferred, such as genres or important keywords. In libraries around the world today, there are many collections of old handwritten manuscripts that few people alive today have read or are able to read. The number of experts able to read these manuscripts are few and far between and hence heavily outnumbered by the amount of texts. Because of this, selecting which manuscripts to study can be an overwhelming process. Access to a qualified guess as to what a book contains would help alleviate this problem. Image-based word clouds would allow an expert to get a feel for what a manuscript contains prior to actually working with the material.

Text-based word clouds can be described as doing basic histogram processing (i.e., counting word occurrences) and can be seen as one of the most basic examples of data mining. Word clouds scales very well with text size, in principle it allows you to reduce all the text on the entire Internet to a single



46 Sollefteå Y Billsta 1:6

Box 101
Sollefteå

8 Andersson frik 020
Gustaf Simon g250719 011
Vaktm Solleft

46 Sollefteå Y Billsta 1:6

Box 101
Sollefteå

4 Andersson f Lundström 011
Ester Emilia g250719 020

The old gentleman died: his
as much disappointment as pleas
leave his estate from his nephev
half the value of the bequest. Mr
wife and daughters than for him
child of four years old, it was se
of providing for those who were
by any charge on the estate, or b
up for the benefit of this child, v
Norland, had so far gained on th

Fig. 1: Two image-based word clouds and their respective source material. The top left image depicts an image-based word cloud created from Swedish tax registries from the mid 20th century, shown to the top right. The bottom left image shows an image-based word cloud created from the book ”Sense and Sensibility” and a sample image is shown to the bottom right.

image. The typical word cloud processing is as follows. Given some text as input, the word occurrences are counted. With the help of a dictionary, common and uninformative words (called stop words) are removed. The final step is then to place the informative words on a canvas and rescale the size of the words so that the word size is proportionate to the word count. The placement can be done in many different ways, spanning from something simple like a circle or spiral to complex shapes like elephants or guitars. Two different examples of image-based word clouds can be seen in Figure 1.

2 Image-based Word Clouds

The purpose of image-based word clouds is to try and capture the same information contained in regular word clouds but using only image data, and without doing any recognition. Some of the processing steps have an exact analogue to the text version, e.g., clustering corresponds to calculating the histogram, whereas other steps are new, such as word segmentation and feature extraction. The layout step is similar to the text-based word cloud except for the fact that the images that are laid out are bitmaps, not vectorial. In some cases, this can cause severe pixelation in the images due to excessive rescaling of small words. The processing pipeline consists of five stages, visualized in Figure 2.

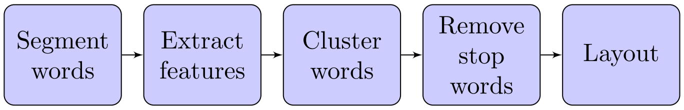


Fig. 2: The processing pipeline for image-based word cloud.

2.1 Word Segmentation

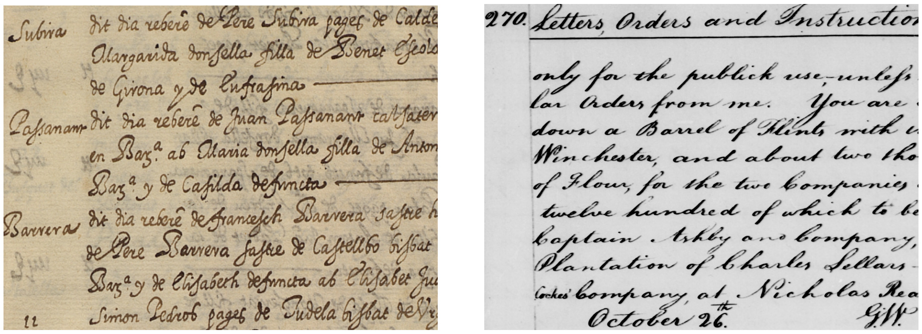
In our current pipeline for image based word clouds, word segmentation is a necessary step. Using a word segmentation method adapted from [1], we have repeatedly created word clouds for both good and poor quality typewritten text, see Figure 1. For handwritten material in general, we have so far only been able to find stable word segmentations by manual tuning. For this reason, this paper does not present a solution to the word segmentation problem for handwritten sources in general and it remains a topic for future research. Instead, in our quantitative experiments, we have used challenging handwritten datasets where the word segmentation has been known beforehand.

2.2 Feature Extraction

For each word image, a set of features is extracted using the approach described in [2]. First, a simple pre-processing step is applied wherein the images are padded with 8 pixels and resized to a fixed size (56×160 in this case). The image is then split into cells of size 8×8 pixels and for each cell, HOG [3] (31 dimensional) and Local Binary Pattern (LBP) [4] (58 dimensional) descriptors are extracted, using the VlFeat library [5]. The HOG and LBP features are then normalized so they have unit l^2 norm and then concatenated. The resulting feature vector, $\mathbf{v} \in \Re^k$, where $k = (31 + 58) \cdot 7 \cdot 20 = 12460$.

After extracting HOG and LBP features for each image in the dataset, choose a random subset of your data set of size $n = 3750$ and call it $\mathbf{M} \in \Re^{n \times k}$. Define a partition P as the interval $[1, ..., n]$ split into groups of size $p = 15$ (the values for n and p are adopted from [2]). Then, project all of your feature vectors onto M, i.e., $\mathbf{u} = \mathbf{M}\mathbf{v}$. Finally, split each $\mathbf{u}$ using the partition P and do max pooling. That is, for each group of size 15 keep the max value, giving you the final vector $\mathbf{x} \in \Re^{250}$.

Given the extracted features, their optimal weighting is unknown. Yet the weighting will be of great importance for the following unsupervised clustering



Subira dit dia rebere de Pere Subira pages de Calde
Margarida donsella filla de Benet Esco
de Girona y de Eufrasina
Passanant dit dia rebere de Juan Passanant calsater
en Barz.^{a} ab Maria donsella filla de Anton
Barz.^{a} y de Cafilda defuncta
Barrera dit dia rebere de Francesch Barrera sastre h
de Pere Barrera sastre de Castellbo bisbat
Barz.^{a} y de Elisabeth defuncta ab Elisabet Ju
11 Simon Pedrol pages de Tudela bisbat de V

270. Letters, Orders and Instruction
only for the publick use–unless
lar Orders from me. You are
down a Barrel of Flints with
Winchester, and about two tho
of Flour, for the two Companies
twelve hundred of which to be
Captain Ashby and Company,
Plantation of Charles Sellars –
Cockes' Company, at Nicholas Rea
October 26th. G.W.

(a) Marriage License dataset (b) George Washington dataset

Fig. 3: Sample images of the three datasets used for the experiments, going from easiest to most difficult.

step. Some feature dimensions are correlated, leading to an emphasis of certain word characteristics, while others contain noise. One recent approach in query-by-string word spotting is to use Canonical Correlation Analysis (CCA) [6], which aligns and de-correlates the feature dimensions in a query string feature space and a word image feature space. In our work here, we use a related but different method to learn an optimal representation suitable for unsupervised clustering, which has previously been described in [7]. In this method, we perform CCA between image features and their class labels, encoded using the one-hot encoding. The gain from this reweighting scheme is two-fold; we reduce the feature dimension to 100 from 250, and to put our performance gain into perspective, get about 10% higher Mean Average Precision on the Washington dataset compared to [2], using only 20% of the available data as training data. In particular, this method does not require training on fully transcribed words. In the training step of this method, each word image needs a label, but there is no need to submit any string data representing a transcription of the word.

2.3 Clustering

The clustering step in the image-based word cloud method corresponds to counting the occurrences of each word in a given text. In the ideal case, each cluster would correspond to a unique word and only contain instances of that word, i.e., be completely homogeneous. However, in the general case the number of unique words in a text is unknown and therefore also the number of clusters to use. There are two ways to get around this problem. One approach is to use algorithms where the number of clusters is not a parameter [8]. However, they often have other sensitivity parameters that more or less correspond to confining the number of clusters to an interval. Another approach would be to try and estimate the number of clusters somehow. This has been done before in the clustering of word images using Heaps' Law [9]. For a given natural language, Heaps' Law provides an estimate of the number of unique words in a document given its length. However, in an environment where the language is constrained, e.g., scanned forms, Heaps' Law might not perform as well.

Using Heaps' Law to estimate the number of clusters unlocks the use of many powerful clustering algorithms. The algorithms that have shown previous success [10] and also performs the best in our experiments are Hierarchical Agglomerative Clustering (HAC) [11] techniques with different linkage-criteria. In our case, we found that average linking together with cosine distance measure used in [2] gave the best result.

2.4 Stop Word Removal

This step has an exact analogue with the regular word cloud, though in the text case, it is very simply solved by using a dictionary of words to remove for each language. The image case is more complex. The goal is to remove images of words that are frequently occurring and bear no information. For English, these include common words that occur often in text such as "and", "the", and "of". For the image case, words that occur often should correspond to largest clusters from the clustering. Related to this are some early results in information retrieval that suggest that one can split the types of words in a text into three categories.

1. Frequently recurring stop words, corresponding to the largest clusters. These words hold little to no information.
2. Sporadically occurring words, noncrucial to describe the content of the text. These correspond to the smallest clusters.
3. Moderately recurring words that correspond to the middle sized clusters are the words that are informative and make up the content of a text.

This property of text can be described by Zipf's Law [12]. It states that for a given text, the plot of term frequencies exhibits a distribution where the kth most frequent term has a frequency of k/f_0, where f_0 is the most frequent term. Figure 4 shows a plot of the term frequencies for the Washington dataset along with Zipf's Law. The sporadically occurring words typically cause no trouble for word clouds since they are usually so small you can barely see them, if they are even in the image. The stop words on the other hand, cause problems. A simple, approximate way to remove them is to discard the largest clusters before proceeding to the next stage. This is not a perfect solution and therefore the efficacy of this may vary from quite high to relatively low depending on the quality of the clustering and the text that you are working with. To choose an appropriate number of clusters to discard, we look at what words the largest clusters contain and manually choose the cutoff. However, this method will not eliminate all stop words, as stop word clusters smaller than the largest informative word cluster will not be discarded.

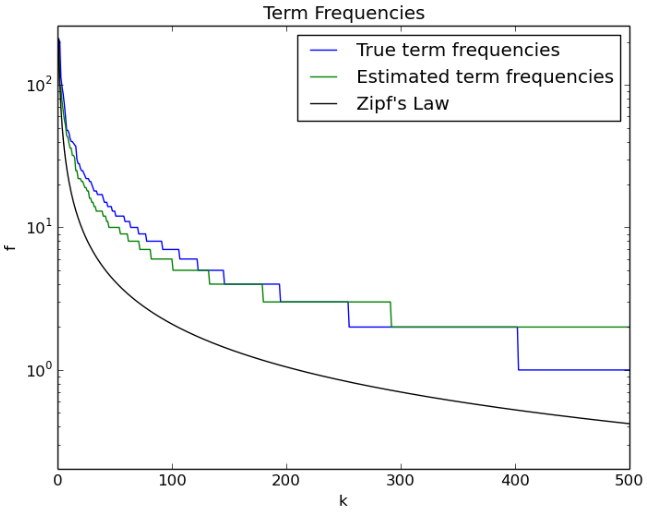
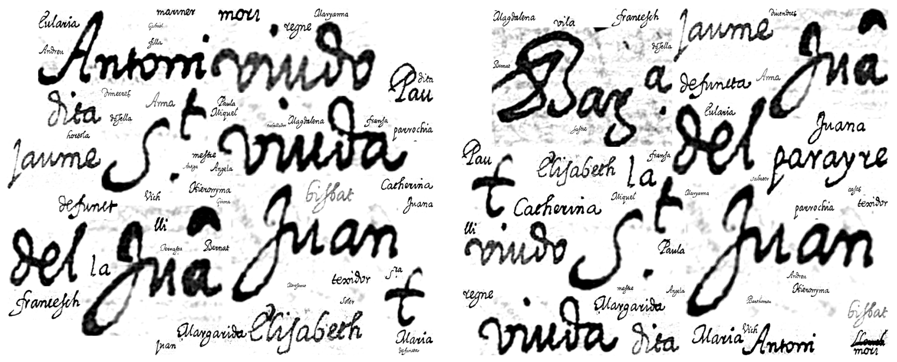


Fig. 4: Term frequencies using ground truth data, estimated term frequencies (cluster sizes), and Zipf's Law in the Washington dataset. The frequencies for words 500 to 1204 are one.

2.5 Layout

The layout procedure that is used in this paper can be described as a greedy search over a set of coordinates generated in a specific shape, where an image is placed in the first place that it fits. More concretely, for each cluster, the image closest to its cluster's centroid is selected as a representative for that cluster. They are then sorted with respect to the sizes of the clusters in descending order. First, an empty canvas is created, then starting with the image representing the largest cluster, the image is (non-linearly) scaled according to the cluster size, then we iterate over the set of coordinates and place the image at the first position that is available. There are many possible scaling functions. We have used $c^\alpha+1$, where c is the cluster size for that word and α is a tunable parameter. Typically $\alpha = 0.3$ works well. This procedure is repeated for all images until there is no more room on the canvas. To increase the density of the canvas, in addition to scaling the images according to their cluster size, they are also scaled down when placing an image rescaled by cluster size fails to find an empty spot in the set of coordinates, due to overlap with another image. While this does create a false sense of proportion, we feel that it improves the final look of the word cloud. However, more sophisticated layouts may eliminate the need for this step. The final step is some basic contrast enhancement to remove small variations in the background. This is done using gray level transforms with the parameters manually chosen for each dataset.



(a) (b)

Fig. 5: Word clouds of the BH2M dataset. Left: the estimated word cloud. Right: the true word cloud, where the ground truth clustering has been used. For both word clouds, Zipfs law has been applied to remove stop words.

3 Experiments

To evaluate the image-based word clouds, we perform experiments on two handwritten datasets manuscripts from the 17th and 18th century for which transcriptions are available, allowing us to create ground truth word clouds using correct counts as a comparison. The first dataset is the Barcelona Historical Handwritten Marriages database (BH2M) [13]. It is comprised of a book of marriage licenses, 174 pages long, written between 1617 and 1619 by a single author in old Catalan. However, in interest of keeping computation time and memory requirements low, we only make use of the first 79 pages, or 26087 words. Since the dataset is comprised of a large amount of marriage licenses, and each license follows a kind of grammar (essentially they are historical forms), the language is quite repetitive. Therefore, Heaps' Law does not necessarily hold. As it turns out, this is not such a big problem. Using Heaps Law works quite well, with parameters estimated on running English text.

The second dataset is the George Washington letters [14]. This is a well known and widely used dataset for word spotting [2]. It consists of 20 pages, written during the mid 18th century by George Washington and his secretary, which correspond to 4860 words.

To quantitatively evaluate the estimated word clouds, we measure the cosine similarity between them and their ground truth counterparts. The information present in a word cloud is essentially a histogram, or bag-of-words model, and is therefore represented as a high dimensional vector, with as many dimensions as there are unique words in the word clouds. The cosine similarity measure calculates the angle between these two vectors and produces a number $s \in [0, 1]$, where 1 means that the word clouds contain the same information and 0 means that they contain completely different information. The cosine measure can be defined as

$$s = \frac{\mathbf{u^T} \cdot \mathbf{v}}{\sqrt{\mathbf{u^T} \cdot \mathbf{u}}\sqrt{\mathbf{v^T} \cdot \mathbf{v}}} \tag{1}$$

where u and v are two bag-of-words vector representations of word clouds. Another measure that is simpler but slightly more intuitive is the ratio between the number of words present in the two clouds and the max of the number of words in each cloud, i.e. the overlap percentage defined as

$$o = \frac{M}{max(O, P)} \tag{2}$$

where M is the number of words present in both clouds, and O and P is the number of words rendered in their respective clouds.

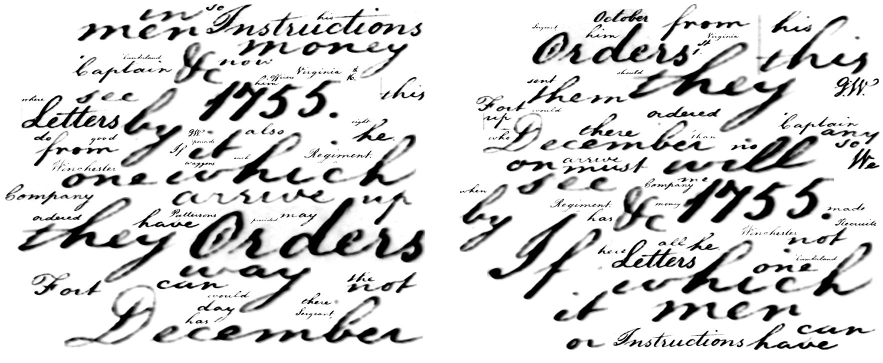
3.1 Marriage Licenses

The purpose of this experiment is to show the that even though the assumptions of Heaps' Law and Zipf's Law do not necessarily hold for this type of data, results are still interpretable and satisfactory. As mentioned above, Heaps' Law tuned for English seems to work well for old Catalan. We use 10% of the data, or 2608 words, as training data for the CCA ,and we use the remaining data to generate the word clouds. Heaps' Law estimates that there are 3356 clusters, which is around 57% too many. However, the estimated and the ground truth cloud look quite similar, see Figure 5. A lot of the words correspond to common names like "Juan" and "Antoni".

3.2 George Washington Letters

For the second experiment, we create a word cloud of a more difficult dataset, even though Heaps' Law should work much better. The main difficulty stems from the fact that it is a lot smaller than the BH2M dataset, resulting in less data for the CCA feature weighting. Using 20%, or 972 words, as training data for CCA works sufficiently well. We estimate the number of clusters to be 1190, which is about 14% too many. The generated clouds can be seen in Figure 6. By inspecting Figure 6a, one can guess a genre or topic for the text. Words like "Captain" and "Orders" indicates some sort of military text.

The results from the two experiments, as well as the Sense and Sensibility dataset can be seen in Table 1, the cosine similarity and overlap is presented for the three datasets. The quantitative results reinforce the notion that similar information is captured by the two word clouds.



(a) (b)

Fig. 6: Word clouds for the Washington dataset. Left: the estimated word cloud. Right: the true word cloud. For both word clouds, Zipfs law has been applied to remove stop words.

Table 1: Results from word cloud evaluation.

Dataset	cosine similarity	overlap
Sense and Sensibility	0.98	0.93 (94 of 101)
BH2M	0.88	0.76 (56 of 74)
Washington	0.76	0.64 (38 of 59)

4 Conclusion

In this paper, we have presented a novel framework for generating image-based word clouds. We have quantitatively and qualitatively shown that image-based word clouds generate word clouds that look similar to ordinary word clouds. Meaningful visualizations of large document collections can be generated using this framework, despite the lack of a proper stop word list and without advanced recognition algorithms that decode the text letter by letter. From a theoretical point of view, this algorithm is conceptually close to earlier work in word spotting. However, from an application point of view, the introduction of image based word clouds could potentially open up completely new ways for users to experience large document collections.

As of right now, the word segmentation algorithm is the limiting factor in this framework. Better word segmentation for handwritten documents and segmentation free approaches to image-based word clouds are interesting directions for future research. Interactive tools that allow users to quickly correct, merge clusters and remove obvious stop words manually, are also possible extensions. The current evaluation focuses on comparing the quality of the clustering compared to the true word counts. Another interesting comparison for the future would be to compare the word clouds to fully text-based word clouds, i.e., using a dictionary for stop word removal.

Acknowledgment

This project is a part of q2b, From quill to bytes, a framework program sponsored by the Swedish Research Council (Dnr 2012-5743) and Uppsala university. The work is done in part as a collaboration with the Swedish Museum of Natural History (Naturhistoriska riksmuseet). We would also like to thank Alicia Fornés and the Document Analysis group of the Computer Vision Center at Universitat Autnoma de Barcelona for access to the BH2M dataset.

References

1. Zagoris, K., Pratikakis, I., Antonacopoulos, A., Gatos, B., Papamarkos, N.: Handwritten and machine printed text separation in document images using the bag of visual words paradigm. In: ICFHR. (2012) 103–108
2. Kovalchuk, A., Wolf, L., Dershowitz, N.: A simple and fast word spotting method. In: Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. (2014) 3–8
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1., IEEE (2005) 886–893
4. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on **24** (2002) 971–987
5. Vedaldi, A., Fulkerson, B.: Vlfeat: An open and portable library of computer vision algorithms. In: Proceedings of the international conference on Multimedia, ACM (2010) 1469–1472
6. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Handwritten word spotting with corrected attributes. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1017–1024
7. Johansson, B.: On classification: simultaneously reducing dimensionality and finding automatic representation using canonical correlation. (2001)
8. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315** (2007) 972–976
9. Heaps, H.S.: Information Retrieval: Computational and Theoretical Aspects. Academic Press, Inc., Orlando, FL, USA (1978)
10. Rath, T.M., Manmatha, R.: Word spotting for historical documents. International Journal of Document Analysis and Recognition (IJDAR) **9** (2007) 139–152
11. Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R.: The elements of statistical learning. Volume 2. Springer (2009)
12. Zipf, G.K.: Human behavior and the principle of least effort. (1949)
13. Fernández-Mota, D., Almazán, J., Cirera, N., Fornés, A., Lladós, J.: Bh2m: The barcelona historical, handwritten marriages database. In: Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE (2014) 256–261
14. Lavrenko, V., Rath, T.M., Manmatha, R.: Holistic word recognition for handwritten historical documents. In: Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on, IEEE (2004) 278–287