Thoughts, Words and Actions: Understanding the Growth of an Ideological Movement

Garth Griffin

May 13, 2011

1 Introduction

Political change can seem sudden and unexpected, but when viewed through the lens of history, revolutionary political movements often turn out to be the mere tip of an ideological groundswell. An ideological movement, born out of disparate individuals who share a common goal, represents the coalescence of shared ideas into a unified force. In light of the recent unrest in the Middle East, it is clear that understanding this process of unification is sociologically and politically relevant today. Horace Seldon, a Boston-based historian and National Park Ranger, leads a group of historical experts who are studying the growth of the abolitionist movement in 19th-century America. The abolitionist movement called for the emancipation of the slaves, and was ultimately successful with Lincoln's emancipation proclamation. By the end, the abolitionist movement was a recognized political group, but it did not begin that way. It began with a few individuals who held what were then radical ideas about the plight of the colored population in America. Seldon's group is exploring how those individuals and ideas came together to form a powerful, unified political movement.

Of particular interest to Seldon's group is William Lloyd Garrison, a prominent, radical abolitionst. Garrison published a newspaper called *The Liberator* from 1831 to 1865 [9]. Garrison played an active role in the shaping of the abolitionist movement, using *The Liberator* as his primary platform for speaking. In addition to editorials by Garrison, the paper contains pieces written by supporters of abolition as well as reprints of pieces by Garrison's opponents. The temporal regularity, topical relevance, and depth of information of *The Liberator* make it an ideal primary source for the work of Seldon's group. With 34 years of publication at 52 issues per year and four pages per issue, *The Liberator* contains a total of 7,072 pages. Assimilating that quantity of information is a daunting task even for a group of domain experts.

To distill knowledge and insight from this large-scale textual corpus, we turn to the field of computational information retrieval. Our goal is to use automated information retrieval techniques to support the domain experts in their analysis of the text. The starting point of our project is a corpus of scanned images of the pages of *The Liberator*. We first use optical character recognition to convert the images into machine-readable text. This already provides some advantage to the historians by enabling basic text searching. Next, we use a language model to extract named entities from the text, from which we build an index of where the entities are mentioned over the course of the publication of the newspaper, which we call an *entity mention index*. This index can then be the focus of a variety of analytic approaches. Because we are working with domain experts who may not have expert technical knowledge and skills, we focus on data visualization as an output medium for the analysis. The entity mention index provides a back-end that drives visual representations of the text corpus. These visualizations can then be used by the domain experts to more quickly and easily explore the text corpus in pursuit of their research agenda.

Seldon's historical research of the abolitionist

movement is an ongoing project. Since our contribution to that effort is in collaboration with the domain experts, we are continuing to develop and refine these computational tools to better suit their needs. In this paper, we describe our computational pipeline for building the entity mention index, evaluate the robustness of the pipeline on noisy input data, present preliminary analysis results as evidence for the efficacy of the approach, and highlight some of the avenues that are the focus for the future of the project. We do not attest that our results or process are conclusive or final. Rather, we hope to provide a sense of scope for the project and to supply a proof of concept for our analysis pipeline and our overall approach.

2 Dataset

Our corpus is a collection of scanned pages from Garrison's newspaper *The Liberator*, as images. The available image corpus comprises the full run of the newspaper, spanning 34 years with 52 issues per year and four pages per issue. Due to limited computational resources, it was necessary to develop our analysis pipeline on a small subset of the full corpus. This subset consists of a few years from the beginning of the paper's publication plus a few years from the end. This enables some longitudinal analysis of the text, although it is not as useful as having the complete publication. We plan to expand in scope to cover the full publication run.

3 Related Work

Our work relies on two foundational information retrieval techniques, optical character recognition (OCR) and named entity recogition (NER). We also draw heavily on existing visualization research. Both OCR and NER technologies are active areas of ongoing research, and there are some papers that have particular relevance for this project. Impedovo et al. describe a variety of OCR techniques in a survey paper [5]. Mantas provides a survey of OCR techniques broken down according to the various methodologies [6]. These papers give an overview of the technology behind the character recognition component of our project.

A synopsis of the NER technologies underlying this work is presented by Nadeau et al. in [8]. The particular tool we employ draws on the hidden markov model NER techniques discussed by Zhou et al. in [14]. In addition, the lack of hand-corrected OCR data necessitates an NER module that is especially robust to noisy input. Miller et al. present an approach to this problem in [7].

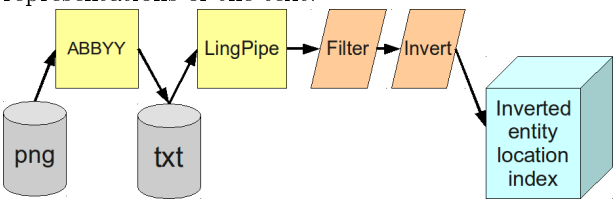
We use visualization of the data to support analysis and insight. The problem of visualizing textual data remains an active interdisciplinary research area. Wise proposes an ecological approach to this problem, where many different visualizations are used to support various strategies for analysis [12]. This tactic is an important component of our overall approach. In [13], Wise et al. discuss the use of spatial techniques in textual analysis. This work underlies much of the visualization that we employ.

Though our dataset is currently somewhat limited in size, there are nonetheless challenges with respect to scalability of the visualization. Efficient approaches for large-scale text corpora are discussed by Grobelnik et al. in [3]. A crowd-sourced approach to visualization could also be applied to our task, such as has been done by Viegas et al. with ManyEyes [11]. Their work provides an excellent pipeline for a variety of text visualizations. The three areas of OCR, NER, and visualization represent the technological foundations of our research, and the works we mention in this section represent aspects of these fields that are of particular relevance to our project.

4 Approach

Our dataset consists of a set of images. In order to perform text mining, we must first convert the images into text. We use optical character recognition technology to accomplish this. Next, we extract the named entities from the resulting text. When an entity is recognized, we record the issue date and sentence of the occurrence. This generates a set of entities tagged with textual location information. To handle the noisy OCR, we employ a number of fil-

Figure 1: Overview of the textual analysis portion of the project. The images in our corpus are converted to text using the ABBYY FineReader OCR tool [1]. We extract named entities from this text corpus using the HMM entity extractor in the LingPipe toolkit. The resulting set of entities is filtered and normalized to reduce noise. This is used to construct an inverted index, which for each normalized entity stores the locations in the textual corpus where the entity is found. This index is used to drive interactive visual representations of the text.



tering and normalization steps. We then build an inverted index of the normalized entities. Finally, this index is used as a back-end to drive visual representations of the data, which support analysis by human experts. Figure 1 shows a high-level diagram of the various stages of textual analysis. Our approach is described in more detail below.

4.1 Optical Character Recognition

In order to perform automated text information retrieval, it is first necessary to convert the scanned images of the newspaper corpus into a form that could be understood as text by a computer, a process known as optical character recognition (OCR). To accomplish this, we used the commercial product ABBYY FineReader [1]. This product was chosen over several other OCR tools on the basis of recognition accuracy, as determined by manual evaluation of a few pages from the corpus. The output of this stage of the process was a collection of simple text files. Some layout information is extracted as well, but we do not assume that the layouts are preserved completely and instead treat the text files as loosely-ordered bags of words.

4.2 Named Entity Recognition

Having obtained textual versions of the scanned pages, a variety of computational information retrieval techniques can be applied to extract useful information from the corpus. We elected to begin with named entity recognition, as the process is relatively straightforward and the results can be readily used as a precursor to other automated analysis techniques such as topic modeling or network analysis. Various toolkits provide support for NER. We chose to use LingPipe [2], an open-source text processing toolkit written in Java.

It is desirable for the NER process to preserve the locations of the entity occurrences with regard to the order of the newspaper issue releases, as this provides temporal information that can be used for co-occurrence mapping and investigation of change over time. In addition to recording the issues in which an entity is mentioned, recording the locations of entity mentions within a particular issue allows for the future possibility of intra-issue co-occurrence mapping. This could be used for issue-level information extraction, such as finding information about particular articles or recurring sections of the paper. Consequently, before extracting the named entities we separate the text into sentences using the sentence chunker provided by LingPipe. It is unlikely that the sentence chunking is perfectly correct, but it does provide some information about where in the issue an entity was mentioned.

Once an issue is chunked into sentences, we extract the entities using LingPipe's entity recognizer. This recognizer is an unsupervised recognizer that uses a hidden markov model combined with a precomputed language model to extract a set of probable named entities from a text. The fact that the process does not use a gazette combined with the noise introduced in the OCR phase means that the resulting set of entities contains some error. We observed that many times the output from the recognizer would contain names that a human reader would immediately understand as references to the same entity but that the program had recognized as being different entities. For example, the entity recognizer might conclude that "New Haven", "NewHaven", " New Haven.

", and "the New Haven" are four separate entites, but it is clear to a human eye that they all refer to the name "New Haven". To address this, we perform normalization of the extracted entities. The normalization removes punctuation and capitals and trims leading and trailing spaces. It also removes articles such as "a" and "the" if they occur at the beginning of the entity. Normalization of the "New Haven" examples above would map all four to the single entity "new haven". It is worth noting that the normalization process does not detect plurals, and as it is rule-based there are likely to be some errors that are not corrected. However, it provides significant improvements over the raw LingPipe output.

After the entities have been normalized, we perform a second pass that removes entities matching a stoplist. The stoplist contains a number of words that were observed to commonly occur and that are clearly not entities, by virtue of not being the correct part of speech, being obvious OCR errors, and so on. For example, the entity recognizer incorrectly labeled the word "and" as an entity in a number of cases. This stoplist was created by hand and would need to be adjusted if the corpus was changed. The stoplist removes many entries that hold no informational value, reducing computational load and analytic clutter in subsequent stages.

4.3 Index Construction

We construct an inverted index using the normalized entities as keys. For each entity, we store a list of issues where that entity is mentioned, and for each of those issues, we store a list of the sentences from that issue in which the entity is mentioned. An inverted index is a common approach in information retrieval. It is significantly more efficient than storing the full matrix of entities and locations, because that matrix will be very sparse. The inverted index provides fast boolean querying on entities, and enables efficient computation of relative frequency of entities. By storing both issue-level and sentence-level positional information, the index can operate at two different levels of granularity. Applications using the data structure can determine at run-time whether to use full granularity at the level of sentences or to remain at the higher issue level to take advantage of faster query time. For convenience, the index is implemented in a serializable class that can be written and read either as a binary object or as a human-readable character-separated file format. This makes porting the index to different environments relatively straightforward. We use the index as a back-end for data visualizations, but it could also be used as a base for other information retrieval techniques.

4.4 Interactive Visualization

We aim to support domain experts performing analysis of the corpus. These experts may not have technical backgrounds, so we present the data in the form of an interactive visualization. This provides an intuitive and powerful channel for exploration of the corpus. We have so far created two visualizations based on the entity mention index. The first visualization is a social network graph that uses a force-directed layout. The layout was intended to be adjustable in real-time by the user, but this proved intractable for larger datasets, so we are now pursuing a static network visualization that can be rendered offline. The second visualization is an interactive matrix of entities versus time. The visualization displays a binary matrix of entity mentions, where each column corresponds to an issue of the newspaper and each row corresponds to an entity. The rows can be reordered interactively according to a variety of criteria. These criteria can be global features, such as total number of hits or by entity name alphabetically, they can be local features, such as similarity to one particular entity. We provide a temporal similarity sorting feature. This is calculated by representing entity mentions as a binary vector with a one for each issue where the entity is mentioned and zero otherwise. We use the Hamming distance metric between two such vectors to compare the similarity of entities. This supports finding clusters of entities that are mentioned concurrently over time. We present preliminary results using the matrix visualization in Section 6. We intend to incorporate additional visualization methods according to the needs of the domain experts.

Table 1: Total entity mentions in ground truth versus OCR text.

	Total entity mentions
Ground truth	1720
OCR	1498

5 Evaluation

It is to be expected that there will be some error introduced by the OCR process. As there is not an objective ground truth against which the visualizations can be compared, it is meaningless to measure the accuracy at the final stage of the pipeline. However, rather than measuring the accuracy of the OCR directly, it is desirable to characterise the extent to which the OCR errors impact the NER and index building. To accomplish this, we compare the output of our NER module on two versions of the same set of issues from the corpus, one being the text generated using OCR and the other a hand-corrected version. The issues were taken from 1861 because hand-curated text was already available for *The Liberator* during that time.

Table 1 shows the total entity mentions found in each corpus. There is a 13% difference between the two counts, indicating that the errors introduced by OCR slightly decrease the number of recognized entities overall. If all the recognized entities were correct, this would be a manageable discrepancy. However, as can be seen in Table 2, there are significant differences between the sets of entities that were recognized for the OCR versus ground truth text. Of the 264 entities recognized in the OCR text, 28% were not in the ground truth set of recognized entities, and 54% of the 293 entities in the ground truth text were missing from the set of entities recognized in the OCR text. These discrepancies indicate that OCR errors significantly degrade the quality of the entity recognition. However, investigation of the resulting entity sets revealed a number of correspondent pairs between the missing and extra entities. For example, the entity "lucretia mott" was found in the ground truth entity set but missing from the OCR entity set, and the OCR entity set included "luorotia mott", which was not found in the ground truth set. It is exceedingly likely that these two entities occur at the same places in the text but the OCR incorrectly identified the name. Finding pairs such as that could significantly reduce the impact of OCR errors on the overall pipeline. A possible approach to this problem using Levenshtein distance is discussed in Section 7. Overall, we find that the impact of OCR errors is significant, but conclude that there may be large improvements to be gained from simple solutions.

Figure 2: Binary matrix of entity mentions in *The Liberator* during its first three years of publication, 1831-1833. Gray squares denote mentions, rows are unique entities, and columns are issues sorted by date. The rows are sorted by frequency of occurrence of the entity during the visible time range. The most frequent entities are "united states", "philadelphia", "boston", and "massachusetts", which is unsurprising considering that Garrison was publishing *The Liberator* from Boston with help from friends in Philadelphia.

6 Analysis Results

As a proof of concept for the efficacy of our pipeline, we performed a few short analyses using the matrix

Table 2: Unique entity mentions in ground truth versus OCR text

	Found in ground truth (total 293)	Not in ground truth
Found in OCR (total 264)	190	74 (28%)
Not found in OCR	103 (54%)	

Figure 3: Binary matrix of entity mentions in *The Liberator* during its first three years of publication, 1831-1833. Gray squares denote mentions, rows are unique entities, and columns are issues sorted by date. The rows are sorted by similarity to the entity "colonization society". The row for the entity "wilberforce" is highlighted, and its proximity to the top of the matrix indicates a relationship between "wilberforce" and "colonization society". This is significant because the abolitionist Wilberforce advocated the formation of slave colonies as a means of emancipation, a stance with which Garrison strongly disagreed [10].

visualization described in Section 4.4. In general, we present analyses supporting well-known historical facts, rather than endeavoring to introduce revelatory historical interpretations. The aim of these preliminary results is to support the claim that our analysis pipeline can be used effectively to mine the textual corpus. More in-depth analysis will be directed by the domain experts.

We can use the matrix visualization to focus on a particular time period of interest. Figures 2 and 3 show two displays of the first three years of publication of *The Liberator*. In the former, we sort the entities by overall frequency, and in the latter, we find a cluster of terms related to the entity "wilberforce". Similarly, Figure 4 displays the entities mentioned in *The Liberator* during the Civil War years. By sorting the rows according to similarity to the row for the entity "lincoln", Garrison's coverage of President Lincoln becomes apparent. This temporally windowed analysis of the first three years and the Civil War years shows how the visualization can be used to gain a sense of what is being discussed during a particular time period.

The matrix visualization also provides an intuitive way to compare two temporally disparate segments of publication in order to highlight long-term changes. Figure 6 shows the first three years and last three years of publication juxtaposed against one another. Because each entire row is a single entity, sorting according to frequency in only one of the separate time segments visually emphasizes entities that are notably different between the two time periods. The above example illustrates this with the appearance of warlike terminology during the Civil War years. These analytical results showcase some of the ways in which our pipeline supports exploration of the text corpus. Future development will be guided by the

Figure 4: Binary matrix of entity mentions in *The Liberator* during the Civil War years, 1861-1865. Gray squares denote mentions, rows are unique entities, and columns are issues sorted by date. The rows are sorted by similarity to the entity "lincoln", which is highlighted at the top of the matrix. The second and third rows visually stand out as having a clear similarity to the entity "lincoln". Those rows correspond to the entities "room no" and "abraham lincoln". The relation of the latter is obvious, though the connection with the former is unknown. By looking horizontally at these first three rows of the matrix, it is clear that Garrison took serious note of Lincoln in 1862 and continued to write about him until the end of the paper. This is likely related to the events surrounding the Emancipation Proclamation.

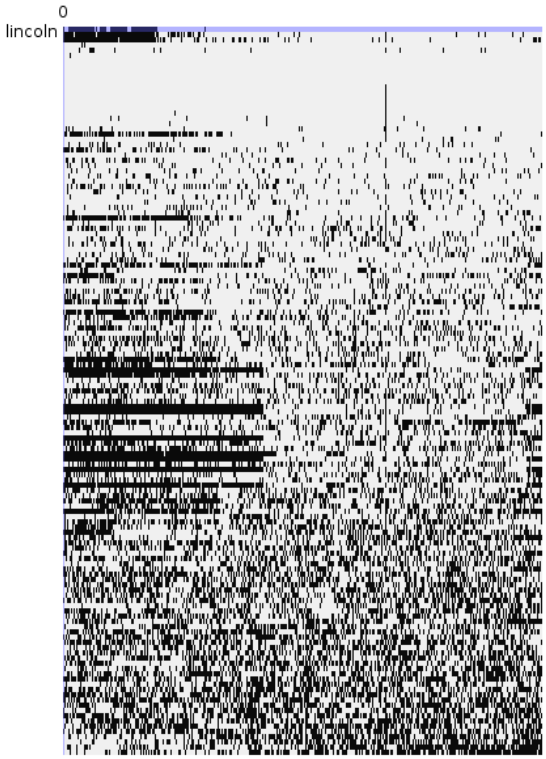


Figure 5: Binary matrix of entity mentions in *The Liberator* from two three-year periods. Gray squares denote mention, each row is a unique entity, and each column is an issue. The left half comprises the first three years of publication of the newspaper, 1831-1833, and the right half comprises the last three years of publication, 1863-1865. Each row corresponds to a unique entity that was mentioned frequently during both three-year periods. The rows are sorted alphabetically by the name of the entity. The point at which the first three years end and the last three years begin is quite visually obvious. This shows that a significant change in the discourse of *The Liberator* is reflected by an easily observable change in the visual representation, lending credence to the claim that this visualization can support intuitive exploration of the text corpus.

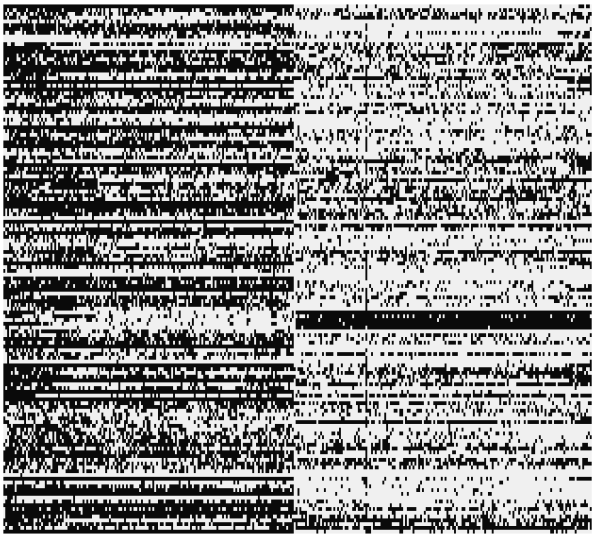
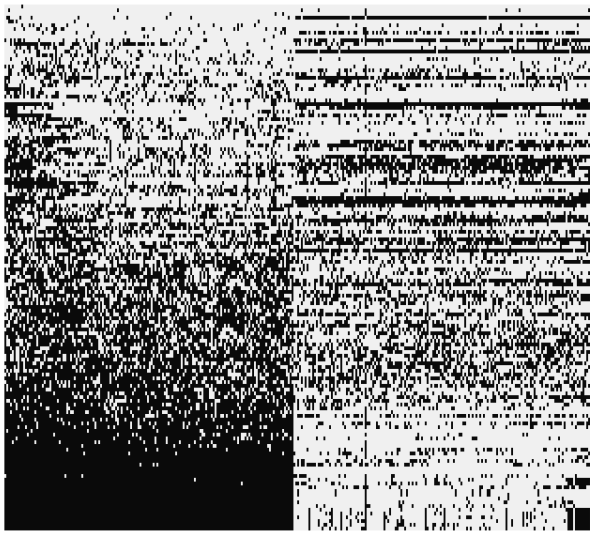


Figure 6: Binary matrix of entity mentions in *The Liberator* from two three-year periods. Gray squares denote mention, each row is a unique entity, and each column is an issue. The left half comprises the first three years of publication of the newspaper, 1831-1833, and the right half comprises the last three years of publication, 1863-1865. Each row corresponds to a unique entity that was mentioned frequently during both three-year periods. The rows are sorted according to the frequency of the entity in the left half. This highlights changes between the beginning and the end of the publication. For example, there are a number of rows near the bottom of the matrix that are almost all black on the left and almost all gray on the right. This indicates an entity that was mentioned rarely between 1831 and 1833 and was mentioned frequently between 1863 and 1865. Many of those rows correspond to entities with military connotations, such as "martial power takes", "has power to order", and "commander of". This reflects Garrison's coverage of the Civil War.



specific needs of Seldon's historical research group.

7 Future Work

This project is still in the preliminary stages. While we have a workable entity mention index, there are a number of ways in which the indexing pipeline could be improved, in particular by addressing the weaknesses of the current OCR and NER modules. So far, the system has been used for only minimal analysis of the text. There are a number of other avenues that could be pursued both in terms of more sophisticated computational text mining and in terms of visualization. Lastly, it remains of high importance to expand the project to cover the full corpus of scanned pages rather than the subset that is currently covered.

As evidenced by the comparative evaluation with hand-curated data, noisy OCR has a significant impact on the results of the NER despite efforts at normalization. In particular, we observed many instances where infrequent entities that were incorrectly read during the OCR phase had only a few letters difference from a more frequently-occurring correct entity. To address this, we plan to implement a Levenshtein distance normalization function. The procedure would map uncommon entities that differ by only a few letters from a frequent entity onto that frequent entity. This can be parameterized for a variable level of sensitivity by adjusting the maximum edit distance at which keys are normalized to one another.

Another way to improve the index is through the use of entity metadata. During the creation of the index, the context surrounding entity mentions could be used to predict descriptive tags for each entity. This could be done in a supervised fashion, for instance by starting with trained data for tags such as *person* or *place*. Alternatively, an unsupervised approach could find associations between entities based on context. Entity metadata such as this would then provide another layer of information in the index that could be used for analysis and visualization.

There is also additional structural information in the text corpus that is currently not retained by the index. Because the corpus is a newspaper, each is-

sue has internal structure in the form of articles and sections. This structure could provide a level of granularity in the index between the issue level and sentence level. There might also be recurring internal structures across issues in the text. These internal structures could be used to draw out contextual information associated with the recognized entities.

Thus far, actual analysis of the corpus has been fairly simplistic. The matrix visualization that formed the basis for most of the results presented in this paper provides only minimal abstraction from the index. We intend to apply a number of other visualization techniques to the index in order to gain different perspectives on the text. We hope to take advantage of existing visualization toolkits and products, such as Prefuse [4]and ManyEyes [11]. Leveraging existing visualization technology will provide a variety of representations that can then be adapted and specialized based on feedback from the domain experts.

Another promising avenue is network analysis, where entities are linked based on coreferences in the text. This approach was the focus of our first visualization attempt, but was discarded due to the computational demands of rerendering the network in real time. Future work in this area will be oriented towards offline network-based analysis, including techniques such as network backboning, to discover important communities in the graph. The above-mentioned improvements to the index and these additional analytic approaches are the primary areas of ongoing development in the project.

8 Conclusion

In this paper, we describe a computational pipeline for computational analysis of Garrison's *The Liberator*. We use OCR to convert our image corpus into textual data, and then use NER to build an entity mention index. We assess the robustness of our approach, and present preliminary analysis results using our pipeline as evidence for its efficacy. We discuss the future directions of the work, and offer possible solutions to some of the weaknesses that we found in our current implementation. The results shown in this paper represent a proof of concept for our analysis pipeline and serve to validate our overall approach. The motivation for the project stems from an ongoing historical research effort by a group of domain experts who seek to understand the growth of the abolitionist movement in 19th-century America, and our work continues in collaboration with this group.

References

[1] ABBYY. Abbyy finereader. `http://www.abbyy.com`.

[2] Alias-i. Lingpipe. `http://alias-i.com/lingpipe/`.

[3] M. Grobelnik and D. Mladenic. Efficient visualization of large text corpora. In *Proceedings of the seventh TELRI seminar. Dubrovnik, Croatia*, 2002.

[4] J. Heer, S.K. Card, and J.A. Landay. Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430. ACM, 2005.

[5] S. Impedovo, L. Ottaviano, and S. Occhinegro. Optical character recognition-a survey. *Issues*, 1(2):1–24, 1991.

[6] J. Mantas. An overview of character recognition methodologies. *Pattern recognition*, 19(6):425–430, 1986.

[7] David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. Named entity extraction from noisy input: speech and ocr. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, pages 316–324, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[8] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30:3–26(24), January 2007.

[9] Horace Seldon. The liberator files. `http://www.theliberatorfiles.com`.

[10] Horace Seldon. William wilberforce, on colonization. `http://www.theliberatorfiles.com/william-wilberforce-on-colonization/`, May 2011.

[11] F.B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. Manyeyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, pages 1121–1128, 2007.

[12] J.A. Wise. The ecological approach to text visualization. *Journal of the American Society for Information Science*, 50(13):1224–1233, 1999.

[13] J.A. Wise, J.J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *infovis*, page 51. Published by the IEEE Computer Society, 1995.

[14] GuoDong Zhou and Jian Su. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 473–480, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.