Learning Parameters of the K-Means Algorithm from Subjective Human Annotation

Haimonti Dutta, Rebecca J. Passonneau, Austin Lee, Axinia Radeva, Boyi Xie and David Waltz
Center for Computational Learning Systems
Columbia University
New York, NY 10115
{haimonti@ccls, becky@cs, alee@ccls,
axinia@ccls, xie@cs, waltz@ccls}.columbia.edu

Barbara Taranto
New York Public Library Labs
5th Avenue and 42nd Street
New York, New York 10018
barbarataranto@gmail.com

Abstract

The New York Public Library is participating in the *Chronicling America* initiative to develop an online searchable database of historically significant newspaper articles. Microfilm copies of the papers are scanned and high resolution OCR software is run on them. The text from the OCR provides a wealth of data and opinion for researchers and historians. However, the categorization of articles provided by the OCR engine is rudimentary and a large number of the articles are labeled "*editorial*" without further categorization. To provide a more refined grouping of articles, unsupervised machine learning algorithms (such as K-Means) are being investigated. The K-Means algorithm requires tuning of parameters such as the number of clusters and mechanism of seeding to ensure that the search is not prone to being caught in a local minima. We designed a pilot study to observe whether humans are adept at finding sub-categories. The subjective labels provided by humans are used as a guide to compare performance of the automated clustering techniques. In addition, seeds provided by annotators are carefully incorporated into a semi-supervised K-Means algorithm (Seeded K-Means); empirical results indicate that this helps to improve performance and provides an intuitive sub-categorization of the articles labeled "editorial" by the OCR engine.

1 Introduction

Chronicling America[1] is an initiative of the National Endowment for Humanities (NEH) and the Library of Congress (LC) whose goal is to develop an online, searchable database of historically significant newspapers between 1836 and 1922. The New York Public Library (NYPL) is part of this initiative and has scanned 200,000 newspaper pages published between 1890 and 1920 from microfilm.

In order to make a newspaper available for searching on the Internet, the following processes must take place: (1) the microfilm copy or paper original is scanned; (2) master and Web image files are generated; (3) metadata is assigned for each page to improve the search capability of the newspaper; (4) OCR software is run over high resolution images to create searchable full text and (5) OCR text, images, and metadata are imported into a digital library software program. The scanned newspaper holdings of the NYPL offers a wealth of data and opinion for researchers and historians. The goal of our research project is to enable users of this historical archive including scholars (genealogists, geologists, marine biologists investigating oil spills in the New York area) and other library patrons to efficiently search for articles of interest to them.

The newspaper titles and digitized pages available through the Chronicling America website can be searched using the OpenSearch protocol[2]. Unfortunately, the current search facilities are rudimentary and irrelevant documents are often more highly ranked than relevant ones. The newspapers are scanned on a page-by-page basis and article level segmentation is poor or non-existent; the OCR scanning process is far from perfect and the documents generated from it contains a large amount of garbled text. In a bid to serve its patrons better, the New York Public Library employed human annotators to clean headlines of articles and text, but the process of manually reading all the old newspapers article-by-article and cleaning them soon became very expensive.

In addition, categorization of article level data using the OCR software was not very successful; for instance, an attempt to categorize articles in the edition of *The Sun* newspaper published on November 4th, 1894 resulted in 338 articles classified as editorial, 32 unclassified[3], 10 sports, 23 advertising, 5 commercial, 3 birth-related announcements and 2 reviews. There was no easy mechanism to do fine-grained categorization of editorial articles – thus articles dealing with elections and governmental appointments, crime and public health were all labeled "editorial".

In this paper, we describe an automated technique of categorizing articles using an unsupervised learning algorithm - the K-Means algorithm. To ensure that it converges and produces satisfying results, the parameters of the algorithm need to be set correctly. This includes the choice of the number of groups in the data set (**K**) and a mechanism for the selection of the seeds. We performed a pilot study to show that humans are adept at navigating ambiguous and hierarchical situations and therefore integrating their wisdom into a

[1] http://chroniclingamerica.loc.gov/

[2] http://www.opensearch.org/Home

[3] These were later identified as banners of the newspaper.

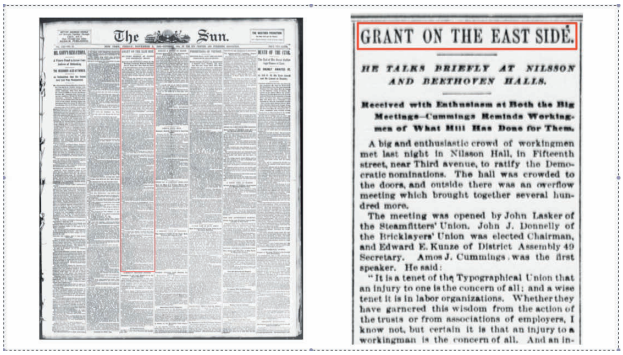
learning algorithm helps the automated technique considerably. Furthermore, if the unsupervised learning algorithm is guided by appropriate choice of seeds, the results are much more intuitive.

This paper is organized as follows: Section 2 presents related work; Section 3 discusses the data available from the NYPL archives and methods of pre-processing and cleaning applied on it; Section 4 describes the human annotation task and Section 5 provides mechanisms to improve K-Means using human-insight. Finally Section 6 concludes the paper and discusses future work.

2 Related Work

In many machine learning tasks, there is a large supply of unlabeled data but insufficient labeled data to learn from. Online document repositories (such as JSTOR, IEEE, ACM, Google digital libraries) provide good examples of domains where unlabeled data is available in surplus. Unsupervised machine learning algorithms (such as clustering) can be used to automatically learn from these large archives. In recent years, *topic models* (Blei, Ng, and Jordan 2003; Blei 2004) have been used extensively for finding useful structures in otherwise unstructured document collections. These are probabilistic models used for uncovering the underlying semantic structure of a document collection based on hierarchical Bayesian analysis of original texts. However, these models assume an underlying distribution of topics and do not always keep the browsing needs of a human user in mind. An alternative is to use one of the oldest and most commonly used clustering algorithms such as the K-means algorithm (Lloyd 1957), (MacQueen 1967). While relatively easy to use, parameters of this algorithm need to be tuned carefully; cluster labels obtained by running the algorithm need to be compared to prior knowledge or "ground truth".

Iterative clustering techniques (such as K-Means) are sensitive to the choice of initial starting points (seeds) and the number of clusters they learn (K). A common technique is to seed at random by arbitrarily creating K partitions and choosing the mean of each partition as seeds. Forgy (Forgy 1965) proposed a variant that chooses K instances at random as seeds, then assigns the remaining instances to the cluster represented by the nearest seed. MacQueen (MacQueen 1967) recalculates the centroids after the assignment of instances to the cluster represented by the nearest seed. Kaufman and Rousseeuw (Kaufman and Rousseeuw 1990) propose an elaborate mechanism of seed selection: the first seed is the instance that is most central in the data; the rest of the representatives are selected by choosing instances that promise to be closer to more of the remaining instances. In K-Means++ (Arthur and Vassilvitskii 2007), centers are chosen at random from the data points, but weighted according to their squared distance from the closest center already chosen. By augmenting K-Means using this simple, randomized seeding technique, K-Means++ is θ(log K) competitive with the optimal clustering. Bradley and Fayyad (Bradley and Fayyad 1998) propose refining the initial seeds by taking into account the modes of the underlying distribution. This refined initial seed enables the iterative algorithm to converge to a better local minimum.


The Sun.

GRANT ON THE EAST SIDE.

HE TALKS BRIEFLY AT NILSSON AND BEETHOVEN HALLS.

Received with Enthusiasm at Both the Big Meetings—Cummings Reminds Workingmen of What Hill Has Done for Them.

A big and enthusiastic crowd of workingmen met last night in Nilsson Hall, in Fifteenth street, near Third avenue, to ratify the Democratic nominations. The hall was crowded to the doors, and outside there was an overflow meeting which brought together several hundred more.

The meeting was opened by John Lasker of the Steamfitters' Union. John J. Donnelly of the Bricklayers' Union was elected Chairman, and Edward E. Kunze of District Assembly 49 Secretary. Amos J. Cummings was the first speaker. He said:

"It is a tenet of the Typographical Union that an injury to one is the concern of all; and a wise tenet it is in labor organizations. Whether they have garnered this wisdom from the action of the trusts or from associations of employers, I know not, but certain it is that an injury to a workingman is the concern of all. And an in-

Figure 1: (Left) A newspaper page from the NYPL archive. The red-border shows an article from the newspaper, zoomed in on the right hand figure.

To efficiently estimate the number of clusters in the data Pelleg and Moore (Pelleg and Moore 2000) search the space of cluster locations and number of clusters using the Bayesian Information Criterion (BIC) or Akaike Information Criterion. The G-Means algorithm (Hamerly and Elkan 2003) is based on a statistical hypothesis test that subsets of data follow the Gaussian distribution and the PG-Means algorithm (Feng and Hamerly 2007) extends this algorithm further by learning the number of clusters from a classical Gaussian Mixture Model. Tibshirani et al. (Tibshirani, Walther, and Hastie 2001) propose the use of the "gap statistic" for estimating the number of clusters in the data, comparing the change in within-cluster dispersion with that expected under an appropriate reference null distribution.

In semi-supervised clustering algorithms, labeled data has been used to provide iterative feedback (Cohn, Caruana, and McCallum 2003) and conditional distributions in auxiliary space (Sinkkonen and Kaski 2002). Seeding mechanisms for semi-supervised clustering have been studied in (Basu, Banerjee, and Mooney 2002), (Wagstaff et al. 2001). Klein et al. (Klein, Kamvar, and Manning 2002) were able to show that by allowing instance-level constraints to have space-level inductive constraints, improved methods of clustering can be obtained with very limited supervisory information.

3 The Data

Figure 1 shows a scanned newspaper (*The Sun*, November 2, 1894) from the NYPL archive and an article from this paper. The historical newspaper archive contains two types of XML files: (1) **Page-Level XMLs:** For each page of a newspaper, there is an XML file that contains metadata about the page and the text in it. Among other things, the metadata presents information about the OCR software used and alternate suggestions for words scanned whenever possible[4]. (2) **Issue-Level XMLs:** The issue-level XMLs (illustrated in Table 1) provide the following information about articles:

[4]We found that its primary selections are usually better than their alternatives.

<dmdSec ID="artModsBib_1_3"
<mdWrap MDTYPE="MODS" LABEL="
Article metadata">
<xmlData>
<mods:mods>
<mods:detail type="**headline**">
<mods:text>**Grant on the East Side**</mods:text>
</mods:detail>
<mods:detail type="**classification**">
<mods:text>**article/opinion-editorial**</mods:text>
< /mods:detail>
<mods:detail type="pageIdentifier">
<mods:text>pageModsBib1< /mods:text>
< /mods:detail>
< /mods:mods>
< /xmlData>
< /mdWrap>
< /dmdSec>

Table 1: A segment of the issue-level XML file illustrating the OCR Classification (as "article/editorial") for the article and its headline.

(a) *Headlines cleaned by humans* which are of much higher quality than the text produced by the OCR software. (b) *Article segmentation information:* Each newspaper article is represented as a collection of one or more text blocks and their pixel coordinates are available. This helps to determine where one article ends and the next one begins and is particularly useful when an article spans more than one page. (c) *High-level categorization* of the articles produced by the OCR software. We have access to only a subset of the NYPL archive[5] – issues of *The Sun* newspaper dating from November 1, 1894 to December 31, 1894. For experiments used in this paper, we used only one randomly chosen newspaper (November 2nd, 1894 issue of *The Sun*). Figure 2 shows all the categories found by the OCR software for this issue. Articles in the "editorial/opinion" and "sports" categories contain statistically significant amounts of text - the remaining 28 articles in the newspaper are not included in our study.

Pre-processing: We first preprocess the documents to reduce dimensionality and have clean data to learn from. For each article, a bag-of-words representation and tf-idf weights are obtained. Stop words such as "the", "and", etc. are removed from the set of words. Terms of length three or less and words that contain digits or repeated characters (e.g. "paaa" and "ornnn") are also removed. After applying the above noise reduction techniques, the dimensionality of the feature space is 3210.

4 A Pilot Study Involving Human Annotators

A pilot study was conducted to test whether the category labeled "article/editorial" by the OCR software could be further broken down to more meaningful sub-categories. Six annotators were recruited to determine the number of natural categories found in a random sample of twenty-five articles, and the divergence across annotators. The articles (all labeled article/editorial by the OCR software) were selected from the November 2nd, 1894 issue of *The Sun* newspaper. All the annotators were given the same set of articles to work with. They were asked to skim the articles first and group them into obvious and intuitive categories and focusing on the "bigger picture". The defined categories had be described in 5 - 10 words and preferably had to include words from the articles. Finally, they were interviewed with the following set of questions:

1. What was the strategy you used for coming up with the categories?
2. Were there any documents that you found difficult to assign to categories?
3. Did you find any part of the study particularly difficult or ambiguous? If so, describe the problem you faced.
4. How long did it take you to complete the study?
5. If you had the opportunity to change anything with this study, what would it be?

While there are many other interesting research questions that can be investigated with human annotated data, the focus was on determining a meaningful number of subcategories for the "article/editorial" category; thus reaction times, self-consistency among annotators were not emphasized.

Interpreting Results from the Pilot Study: Table 2 shows the number of categories found by the annotators. It so happened that the November 2nd, 1894 newspaper was published immediately after general elections; thus a lot of articles in this issue had to do with politics and elections. This

[5]These have been substantially cleaned by humans

Category	Article counts
Editorial/Opinion	154
Sports	6
Advertising	9
Commercial/Legal/Public notices	7
Birth/Death/Wedding	2
Unclassified	10
Total	188

Figure 2: Top-level categories of articles from OCR for the November 2nd, 1894 issue of *The Sun* newspaper.

ID	Number of sub-categories found
Annotator 1	8
Annotator 2	13
Annotator 3	13
Annotator 4	9
Annotator 5	10
Annotator 6	13

Table 2: Sub-Categories found by humans in the random sample of 25 articles labeled "article/editorial" by the OCR software.

is also reflected in the random sample used for the categorization task – annotators unanimously agreed that seven of the twenty five articles used for the study belong to the category "politics/elections/governmental appointments". Three of the annotators found hierarchies among this category such as "politics/ballot, politics/election, politics/nomination, politics/war, politics/social, politics/entertainment, politics/gossip". This accounted for the increased number of total categories they listed. Since the instructions explicitly mentioned focusing on the "bigger picture" and not drilling down to very fine-grained categories, these were merged together to form the category "politics". Annotators also agreed unanimously on one article belonging to the category "medicine, public-health and safety". This article presented a report on a new diphtheria remedy and announced the arrival of fresh serum from Germany which was tried on two cases in Philadelphia. Although slightly tricky, annotators merged together articles that contained arts, biographies, book reviews and the like into one category called "arts/human interest". Creating a homogeneous category for these articles was not easy due to the wide variety of articles. Articles pertaining to "death" and "marriage announcements" were binned into separate categories. There was no agreement among annotators on eleven articles – for example, an article with a headline "President Cleveland goes hunting for squirrels" was labeled as belonging to the following categories: human interest/politics/sports/entertainment/social. All of these eleven articles had a much higher level of ambiguity and there was no agreement among annotators. Since we did not have categories pre-defined for the annotation task and chose rather to let annotators come up with appropriate categories by themselves, computing agreement on these articles was not straight forward.

In essence, **six**[6] sub-categories for the "article/editorial" OCR category were found by human annotators and are illustrated in the Table 3. It must be noted that in this application, it is hard to obtain "ground truth" or a "gold standard" which can be used for further labeling. Consequently, we are forced to rely on the subjective opinion of annotators who sometimes disagree on labels. There is considerable interest in the research community on whether this subjective labeling at low cost is indeed useful for machine learning algorithms (Raykar et al. 2009; Hsueh, Melville, and Sindhwani 2009).

The interview section of the annotation task indicated that small or singleton categories lead to less agreement among humans; these outliers do not fit into a larger category easily and this raised confusion and difficulty in categorization. Thus, learning from more examples of similar kind was the norm.

Many of the annotators based the initial decision of the number of categories by reading the headlines of the articles and making notes; a feedback loop was almost always involved where annotators refined the initial estimates based on more careful and thorough reading of the articles. Finally, since it was not clearly indicated whether an article is allowed to belong to multiple categories, this question was raised by several annotators. The time recorded by annotators indicates that it took anywhere between 45 mins - 2 hrs to complete the task.

[6]This is the value of K chosen for our experiments in later sections.

ID	Category	Article counts
1	politics, elections, governmental appointments	7
2	medicine, public health and safety	1
3	death	3
4	arts, human interest, entertainment	2
5	marriage	1
6	Other	11

Table 3: Sub-Categories formed by humans in the random sample of 25 articles.

5 Refining Parameters of the K-Means Algorithm

The K-Means Algorithm: One of the oldest and most commonly used clustering algorithms is the K-means algorithm (Lloyd 1957), (MacQueen 1967). Assume we are given an integer K and a set of N data points $X \subset \mathbb{R}^d$; the goal is to partition X into K clusters, $K < N$. This can be achieved by choosing K centroids $C_1, C_2, \cdots, C_K$ so as to minimize the potential function $\phi = \sum_{x \in X} \min_{c \in C} \mathit{Dist}[x - c]$, where *Dist* represents a distance function (such as squared euclidean, L1 norm, cosine metric etc.). The basic steps of the algorithm are as follows: Arbitrarily choose initial K centroids $C_1, C_2, \cdots, C_K$ from X; for each $i \in \{1, 2, \cdots K\}$ set the cluster $\mathcal{C}_i$ to be the set of all points in X that are closer to centroid C_i than they are to centroid $C_j, \forall j \neq i$; for each $i \in \{1, 2, \cdots K\}$ set the cluster centroid $C_i = \frac{1}{|\mathcal{C}_i|} \sum_{x \in \mathcal{C}_i} x$; the last two steps are repeated until the process stabilizes and there are no new cluster assignments.

Choice of Parameters: There is much debate on how to choose a suitable K appropriate for the data set. For our experiments we relied on human annotators to come to a consensus regarding the choice of an appropriate K as described in Section 4. The other parameter that warrants some discussion is the choice of initial seeds; we have used two different seeding mechanisms in our experiments: (a) randomly chosen seeds which do not use information about clusters that humans produced (b) a semi-supervised K-Means algorithm called **Seeded K-Means** (Basu, Banerjee, and Mooney 2002). This algorithm assumes that there exists $S \subseteq X$, called the "seed set" on which *supervision* is provided by annotators; thus, for each $x_i \in S$ the annotator indicates which cluster it seeds; there is at least one seed point x_i per cluster. Once appropriate parameters have been set, the labels from K-Means are compared with those *suggested* as "ground-truth" by human annotators. Note that all articles where annotators did not agree on labels were designated to a category called "Other".

Testing the validity of clusters: In order to measure the quality of the clusters produced by the K-means algorithm,

Seeding Algorithm	Mean	Std over 10 trials
Random Sampling	0.19428	$\pm$0.100840554
Semi-supervised	0.25825	$\pm$0.074990418

Table 4: Mutual Information over 10 trials using Random vs Semi-supervised Seeding techniques.

we compare the clusters they produce to human annotated data marking each instance as one of the six categories illustrated in Table 3. This procedure allows us to quantitatively measure how useful the cluster labels are when compared to the annotated class labels. The external cluster-validity measure used in this work was first suggested by Dom (Dom October 2001) and is equivalent to mutual information when cluster labels and class labels are exactly the same. Let each data set D have n instances $O_1, O_2, \cdots, O_n$ and we want to partition it into K clusters. Let $K = \{1, 2, \cdots 6\}$ be the set of cluster labels and $C = \{1, 2, \cdots, 6\}$ be the expert annotated class labels assigned to the objects in D. Consider a two-dimensional contingency table, $\mathcal{H} = h(c, k)$ where $h(c, k)$ represents the number of objects labeled class c are assigned to cluster k by the algorithm. Then, if there is a perfect clustering $\mathcal{H}$ is a square matrix with only one non-zero element per row / column. The marginals are defined as $h(c) = \sum_k h(c, k)$ and $h(k) = \sum_c h(c, k)$. Since in our experiments the number of clusters are known apriori, the cluster-validity measure is essentially is the empirical mutual information $\hat{I}(C, K) = \hat{H}(C) - \hat{H}(C|K)$, where $\hat{H}(C) = -\sum_{c=1}^{|C|} \frac{h(c)}{n} log \frac{h(c)}{n}$ and $\hat{H}(C|K) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{h(c,k)}{n} log \frac{h(c,k)}{h(k)}$.

Empirical Evaluation: There are 25 articles used in the pilot study; a bag-of-words representation and tf-idf weights are obtained for these articles. Each article has 3210 features and one of possible six labels provided by human annotators. The K-Means algorithm with $K = 6$ is run over 10 trials using both the random seeding and semi-supervised seeding. For semi-supervised seeding, one representative article from each category provided by human annotators is randomly selected for creating the seed; however care is taken to ensure that all six categories are represented by at least one seed. In each trial, the labels obtained after clustering are tested against the "ground-truth" generated by annotators and mutual information is recorded. The average and standard deviation of mutual information obtained over all trials is presented in the Table 4. Clearly, using Seeded K-Means with semi-supervision from annotators is more robust than the random seeding mechanism.

In another experiment, we used the results from the pilot study to annotate unlabeled articles. We applied the Seeded K-Means algorithm with seeds suggested by annotators, on the remaining articles of the November 2nd, 1894 issue of *The Sun* newspaper that were not included in the pilot study. At least one representative article from each category was randomly selected from clusters found by humans for creating the seed and care is taken to ensure that all six categories are represented. We ran the Seeded K-Means algorithm ten times on the unlabeled articles; for each run, the number of clusters is fixed at six, the cosine distance metric is used to compare similarity between instances and the same technique (randomly choose one of the representative documents of a category as the centroid) is used for generating seeds. The labels obtained from each run can be considered as produced by an **automated annotator**. Since each automated annotator only provides labels between 1 and 6 we are able to use Krippendorff's alpha[7] to measure inter-annotator agreement between them. It is seen that there is a very low agreement (α=0.316) when 200 resamplings are used for calculating two-tailed 1% confidence intervals. To illustrate this point further, we closely examined the labels provided by two representative automated annotators as shown in the confusion matrix illustrated in Table 5. For these two automated annotators, there is a complete agreement on subcategories for 20.7% of the articles used for blind testing; 61.9% of articles labeled "death" and 33% of articles labeled "Medicine" are correctly labeled. While these results are encouraging, there seems to be confusion in distinguishing between the "election" and "human interest" categories. It is worthwhile to note that humans also found it difficult to assign articles to the "human interest" category and thus this task appears to be significantly harder. An interesting direction for future work is to use other mechanisms of finding representative seeds to be used with the Seeded K-Means algorithm. One such approach is to identify a centroid of the human clusters by calculating the cosine distance of each pair of documents in each human cluster, estimate the mean and then find the document closest to the mean as the seed.

6 Conclusion and Future Work

The New York Public Library has an archive of over 200,000 historical newspapers published between 1890 and 1920 which have been subjected to OCR and are currently stored in an online database making them accessible to patrons. Unfortunately search facilities on this database are rudimentary; newspapers are scanned on a page-by-page basis and article level segmentation is almost non-existent; the OCR scanning process introduces a lot of garbled text. In a bid to make these archives more accessible to the general public, text mining algorithms are being considered for categorization of articles. The OCR software provides a rough categorization, but a large chunk of the articles are labeled "article/editorial" without division into fine-grained categories. Thus, articles dealing with medicine and crime are deemed to belong to the same category; this makes search and retrieval of articles difficult. We designed a pilot study to observe if humans were able to find coherent categories in a small subset of articles in a newspaper; these sub-categories discovered served as "ground-truth" against which labels learnt from unsupervised clustering algorithms are compared. Our results indicate that the presence of small and noisy clusters in the data made it difficult to find an agreement in the optimal choice of K between human annotators and the automated technique. This raises questions about what is the best way to quantify closeness of the automated

[7]We used the implementation available from `http://ron.artstein.org/resources/`

	Elections	Medicine	Other	Death	Human Interest	Marriage	Total
Elections	0	1	2	0	23	0	26
Medicine	6	7	0	3	1	4	21
Other	1	4	4	2	2	9	22
Death	7	0	0	13	0	1	21
Human Interest	2	16	0	5	1	1	25
Marriage	3	0	14	0	0	3	20

Table 5: Confusion Matrix generated by two runs of Seeded K-Means on blind test data formed by articles of the newspaper not considered for the pilot study.

method to the human clustering. Future work also involves analysis of more sophisticated seeding techniques, use of non-parametric algorithms for clustering and design of experiments for incorporating "wisdom of crowds" into machine learning algorithms.

7 Acknowledgements

This work is supported by funding from the National Endowment for Humanities, Grant No: NEH HD-51153-10. The authors would like to thank Mariya Riskova for maintenance of code and generating feature vectors from articles.

References

Arthur, D., and Vassilvitskii, S. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027 – 1035.

Basu, S.; Banerjee, A.; and Mooney, R. J. 2002. Semi-supervised clustering by seeding. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, 27–34.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

Blei, D. 2004. *Probabilistic Models of Text and Images*. Ph.D. Dissertation, U.C. Berkeley, Division of Computer Science.

Bradley, P. S., and Fayyad, U. M. 1998. Refining initial points for K-means clustering. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, 91 – 99.

Cohn, D.; Caruana, R.; and McCallum, A. 2003. Semi-supervised clustering with user feedback. Technical Report TR2003-1892, Cornell University.

Dom, B. E. October - 2001. An information-theoretic external cluster-validity measure. *IBM Research Technical Report* RJ - 10219.

Feng, Y., and Hamerly, G. 2007. G.: Pg-means: learning the number of clusters in data. In *Advances in Neural Information Processing Systems 19*, 393–400. MIT Press.

Forgy, E. 1965. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* 21.

Hamerly, G., and Elkan, C. 2003. Learning the k in k-means. In *Advances in Neural Information Processing Systems*, volume 17.

Hsueh, P.-Y.; Melville, P.; and Sindhwani, V. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, HLT '09, 27–35.

Kaufman, L., and Rousseeuw, P. J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, Canada.

Klein, D.; Kamvar, S. D.; and Manning, C. D. 2002. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, 307–314.

Lloyd, S. 1957. Least squares quantization in pcm. In *Bell Telephone Laboratories Paper*.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of the 5th Berkeley Symposium*, 281 – 297.

Pelleg, D., and Moore, A. 2000. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, 727–734. San Francisco: Morgan Kaufmann.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Jerebko, A.; Florin, C.; Valadez, G. H.; Bogoni, L.; and Moy, L. 2009. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Sinkkonen, J., and Kaski, S. 2002. Clustering based on conditional distributions in an auxiliary space. *Neural Comput.* 14:217–239.

Tibshirani, R.; Walther, G.; and Hastie, T. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal Of The Royal Statistical Society Series B* 63(2):411–423.

Wagstaff, K.; Cardie, C.; Rogers, S.; and Schroedl, S. 2001. Constrained k-means clustering with background knowledge. In *Proceedings of the International Conference on Machine Learning (ICML)*, 577–584. Morgan Kaufmann.