A Machine Learning Approach to Quantitative Prosopography

Aayushee Gupta, Haimonti Dutta, Srikanta Bedathur, Lipika Dey

Abstract—Prosopography is an investigation of the common characteristics of a group of people in history, by a collective study of their lives. It involves a study of biographies to solve historical problems. If such biographies are unavailable, surviving documents and secondary biographical data are used. *Quantitative* prosopography involves analysis of information from a wide variety of sources about "ordinary people". In this paper, we present a machine learning framework for automatically designing a *people gazetteer* which forms the basis of quantitative prosopographical research. The gazetteer is learnt from noisy text of newspapers and is capable of identifying influential people from it. Our corpus comprises of 14020 articles from a local newspaper, "The Sun", published from New York in 1896. Some prominent people identified by our algorithm include Captain Donald Hankey (an English soldier), Dame Nellie Melba (an Australian operatic soprano), Hugh Allan (a Canadian shipping magnate) and Sir Hugh John McDonald (the first Prime Minister of Canada).

Index Terms—Gazetteer, Text Mining, Information Retrieval, OCR, Spelling Correction, Historical data, Influential people detection.

1 Introduction

HISTORICAL newspaper archives provide a wealth of information. They are of particular interest to historians [1], genealogists (e.g. Genealogy Bank[1], Ancestry[2]) and scholars. An important use of historical newspapers is for People Search [2], [3] – the process of finding information about a person and reconnecting them with others they are likely to know. The goal is to determine who knows whom and how. This is often achieved by studying biography. In historical groups, however, biographies may be largely untraceable. In such cases, secondary biographical information is studied by examination of the individual's experience and personal testimonies, some of which may be reported in newspaper articles. Identification of this group of individuals and studying the stories of their life is an important tool in the research historian's arsenal - called *prosopography*. It can be used to learn social structure such as analysis of the roles of a certain group of people, holders of titles, members of professional and occupational groups or economic classes. Quoting prosopographer Katharine Keats-Rohan,

> ...prosopography is about what the analysis of the sum of data about many individuals can tell us about the different types of connection between them, and hence about how they operated within and upon the institutions – social, political, legal, economic, intellectual – of their time.

The nature of prosopographical research has evolved over time. Lawrence Stone [4] discusses an "older" form of prosopography which was principally concerned with well-known social elites, many of whom were influential people. Their genealogies were well-researched, and social webs and kinship linking could be traced, allowing a prosopography of a "power elite" to emerge. This older prosopography can be contrasted with a newer form called *quantitative prosopography*, which studied much wider populations including "ordinary people".

In this paper, we present a framework to develop a *people gazetteer* which forms the basis of prosopographical research. The gazetteer is built from the text of historical newspapers subjected to Optical Character Recognition (OCR) and is capable of identifying influential people. Our paper has the following novel contributions: (1) **Development of the People Gazetteer** – an organized dictionary of people names and a list of newspaper articles in which the name occurs. (2) **Identification of Influential People**: we define an Influential Person Index (IPI) which helps identification and ranking of influential people.

To the best of our knowledge, the development of a framework for doing prosopographical research using machine learning has not been studied before. This exercise, however, opens up a wide range of possibilities

1. http://www.genealogybank.com/gbnk/
2. http://www.ancestry.com/

– for example, news articles related to the influential person can also be linked to a Wikipedia page entry to find out relevant details or build influential people networks that can learn about entities involved in historical events. Such applications can immensely help historians working on prosopography [5] and scholars in learning events related to historically significant people interactively.

Paper Organization: This paper is organized as follows: Section 2 discusses related work; the machine learning framework is discussed in Section 3; the characteristics of the data used for this research is presented in Section 4. Sections 5 and 6 present the development of the gazetteer and the influential people detection process; empirical results and discussions are presented in Section 7 and 8 and Section 9 concludes the paper.

2 Related Work

In this section, we review two types of related literature - digital humanities projects which build gazetteers from text and the process of identification of influential people from data.

2.1 Gazetteers for Digital Humanities Projects

Newspaper archives have been studied extensively for the design of search and retrieval algorithms ([6], [7], [8], [9]), summarization([10], [11], [12], [13], [14]), sentiment analysis ([15], [16], [17]), topic modeling ([18], [19], [20], [21], [22]), clustering([23]), classification ([24] and visualization([25], [26]). Historical newspaper archives from Chronicling America[3] have been used for topic modeling during historically significant time periods [27]. Newman et. al [28] use a combination of Statistical Topic Modeling and Named Entity Recognition for analyzing entities and topics from a news articles dataset. They also create networks based on the relationships among the entities. Lloyd et. al [29] discuss their approach for designing a news analysis system[4] where information about several types of entities can be searched. They perform temporal and spatial analysis and present time series popularity graphs based on the number of references and co-reference names for the entity.

Several digital humanities projects that have used machine learning and natural language processing techniques to learn from historic newspaper archives are relevant to this work – the libraries of Richmond and Tufts have examined the Richmond Times Dispatch during the civil war years for more than two decades and their work focuses on automatic identification and analysis of full OCR text in newspapers to provide advanced searching, browsing and visualization [30]. The focus of this work was on named entity extraction and ten categories prominent in these newspapers were studied including ship names, railroads, streets and organizations. In an earlier project at the universities, the Perseus project [31], [32], [33], a general system to extract dates and names from text was developed in order to detect significant events in document collections.

Developing gazetteers from news articles is a well established technique - different types of gazetteers are discussed under the General Architecture for Text Engineering (GATE[5]) framework. It defines a gazetteer as a set of lists containing names of entities (such as cities, organizations, days of the week, etc) which are used to find occurrences of these names in text. We use this definition to develop our People Gazetteer that finds person name entities from a news article repository and associates each unique person entity with the list of articles in which they occur.

Gazetteer lists are also discussed in [34] where they are used for learning name entity tagger using partial perceptron and aid in performing better NER compared to CRF based entity taggers. Zhang et. al [35] discuss automatic generation of gazetteer list by finding entities with similar type labels from Wikipedia articles. The evaluation is done over scientific domain of Archeology considering subject, temporal terms and location as named entities but no evaluation is presented for person entities. Allen et. al [5] describe an exploratory study for developing an interactive directory for the town of Norfolk, Nebraska for the years 1899 and 1900. Their work focuses on providing structured and richer information about the person entities by linking their occurrences with associated events described in historical newspapers. Their entity-based directory is similar to our people gazetteer although we do not restrict our problem to any specific town or significant era nor do we consider any specific town directory to begin with.

2.2 Influential People Detection

In prosopography, identification of the "social elite" plays an important role. Their experience and personal testimonies may be reported at length in newspaper articles.

In the context of machine learning and data mining, influential people detection has been mostly done in the field of social networks, marketing and diffusion research. Kempe et. al [36] present work on choosing

3. http://chroniclingamerica.loc.gov/
4. http://www.textmap.com
5. http://gate.ac.uk/sale/tao/splitch13.html

the most influential set of nodes in a social network in order to maximize user influence in the network. They consider spread of influence from an influential node cascading through a network which further influences other neighborhood nodes. In this research, we do not focus on the network formed by person entities. Lerman et. al [37] define popularity of a news story in terms of number of reader votes received by it. Popularity over time is based on voting history and the probability that a user in a list will vote. To identify influential bloggers, Agarwal et. al [38] quantify influence of each blogger by taking the maximum of the influence scores of each blog posted by the blogger. The influence score is calculated using the number of posts that refer to the blog, number of comments on the blog, number of other posts that the blog refers to and length of the blog. Influential blogger categories are also created based on the temporal patterns of blog posting.

Cha et. al [39] describe another set of measures for detection of top influential users on Twitter using number of retweets, mentions and followers for an individual. They perform ranking based on each measure separately and use Spearman's rank correlation coefficient to find correlation among ranks and effect of each measure contributing to a person's influence. The influence ranks of topmost influential users on Twitter are presented across various topics as well as time.

In all of the above, the goal is to measure influence or popularity – however, these cannot be directly adapted to the gazetteer or newspaper articles.

3 Machine Learning Framework for Prosopographical Research

Figure 1 presents the framework for machine learning to aid prosopographical research. It has the following components:

- **Data Gathering:** Prosopographical studies involve research on biographies of a group of people and is therefore severely limited by the quantity and quality of data accumulated about the past. Often in historical groups, a lot of information is available about some people, and almost nothing about some others. Studies are severely affected by lack of information and hence secondary sources of information are resorted to including demographic sources (such as parish registers), economic sources (such as deeds of sales), fiscal sources (such as tax lists), financial sources (such as city accounts), administrative sources (such as company records), religious sources (such as membership lists of fraternities), judicial sources (such as sentences), family archives and photographs, publicly available information (such as newspaper archives). The context of the research is sketched based on the available literature and has to be sufficient, relevant and easily accessible. Much debate has also gone into whether to use a single source or multiple sources. While some researchers favor verification from multiple sources, hypothesizing that a single source can lead to erroneous interpretation and one sided views of the past, others prefer a single source primarily due to the homogeneity of the data and ease of processing the data.
  For this research, the primary source are digitized newspaper archives. In order to make a newspaper available for searching on the Internet, the following processes [23] 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.
- **Data Pre-processing:** The images obtained from the OCR software are segmented to obtain article level data. Both manual and automatic segmentation procedures are used. Automatic segmentation primarily involves extraction of text from images by using the logical structure of the page to produce the informative units – the articles. Sometimes machine learning methods are used to label each image, then the page logical structure is constructed up from there by the detection of structuring entities such as horizontal and vertical separators, titles and text lines [40]. For our work, manual segmentation is resorted to. Following this, several preprocessing steps are applied on the text of the news articles including spelling correction and evaluation using a novel algorithm presented in [41].
- **Development of the People Gazetteer:** This component describes the process of development of people gazetteer which involves Named Entity Recognition (NER) and coreference resolution in order to find person entities. This is followed by topic detection using Latent Dirichlet Allocation (LDA) to find the primary topic(s) of news articles and link both to obtain an organized structure.
- **Detection of Influential People:** This compo-

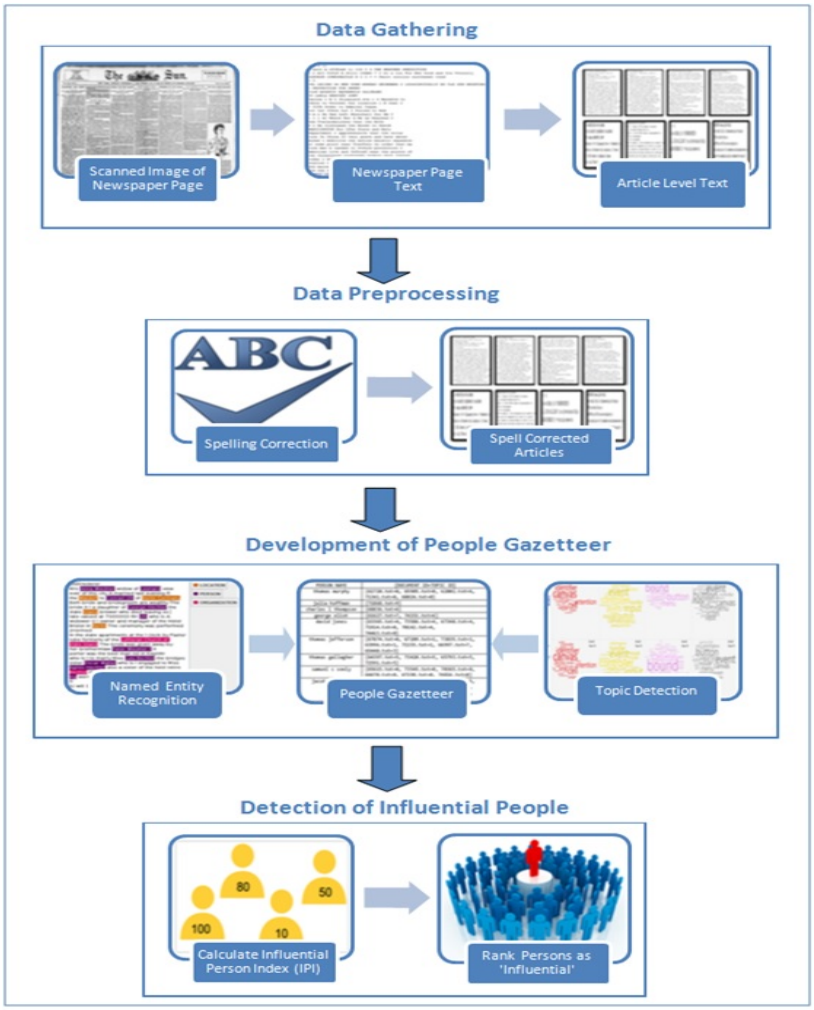


Fig. 1. Research Framework showing components of proposed solution

nent defines an **Influential Person Index** (IPI) that incorporates several criteria for identifying and ranking of influential people. Details about IPI, ranking and final results with some case studies are discussed in Section 7.

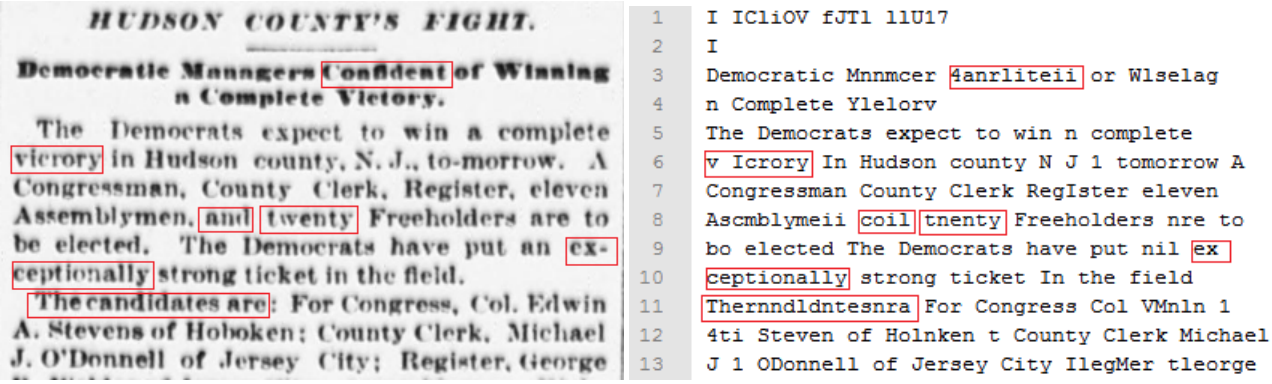
4 Dataset Description

Our prosopographical research is based on historical newspapers obtained from Chronicling America[6]. This is an initiative of the National Endowment for Humanities (NEH) and the Library of Congress (LC) whose goal is to develop an Internet-based, searchable database of U.S. newspapers(between 1836 and 1922) with descriptive information and select digitization of historic pages. Under this program, institutions such as libraries are receives an award to select and digitize approximately 100,000 newspaper pages representing that state's regional history, geographic coverage, and events of the particular time period being covered. The scanned newspaper holdings of the New York Public Library provides the source of prosopographical studies.

4.1 Characteristics

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. An individual OCR text article has at least one or more of the following types of spelling errors:

6. http://chroniclingamerica.loc.gov/



HUDSON COUNTY'S FIGHT.

Democratic Managers Confident of Winning a Complete Victory.

The Democrats expect to win a complete victory in Hudson county, N. J., to-morrow. A Congressman, County Clerk, Register, eleven Assemblymen, and twenty Freeholders are to be elected. The Democrats have put an exceptionally strong ticket in the field.

The candidates are: For Congress, Col. Edwin A. Stevens of Hoboken; County Clerk, Michael J. O'Donnell of Jersey City; Register, George

```
1  I ICliOV fJTl llU17
2  I
3  Democratic Mnnmcer 4anrliteii or Wlselag
4  n Complete Ylelorv
5  The Democrats expect to win n complete
6  v Icrory In Hudson county N J 1 tomorrow A
7  Congressman County Clerk RegIster eleven
8  Ascmblymeii coil tnenty Freeholders nre to
9  bo elected The Democrats have put nil ex
10 ceptionally strong ticket In the field
11 Thernndldntesnra For Congress Col VMnln 1
12 4ti Steven of Holnken t County Clerk Michael
13 J 1 ODonnell of Jersey City IlegMer tleorge
```

Fig. 2. Scanned Image of a Newspaper article (left) and its OCR raw text (right)

- **Real word errors** include words that are spelled correctly in the OCR text but still incorrect when compared to the original newspaper article image. For example: In Figure 2, the word "coil" has been correctly spelled in the OCR text but should have been "and" according to the original newspaper article.
- **Non-real word** errors include words that have been misspelled due to some insertion, deletion, substitution or transposition of characters from a word. For eg. In Figure 2, the word "tnenty" in the OCR text has a substitution error ('n' should have been 'w') which is actually "twenty" according to the original newspaper article.
- **Non-word errors** include words that have been spelled incorrectly and are a combination of alphabets and numerical characters. For example: In Figure 2, the word "4anrliteii" which is a combination of alphabets and number and should have been "confident" as per the original newspaper article.
- **New Line errors** include words that are separated by hyphens where part of a word is written on one text line and remaining part in the next line. For example: In Figure 2, the word "ex-ceptionally" where "ex" occurs on one line while "ceptionally" in the next and due to no punctuation in the text, they are treated as separate words in OCR text.
- **Word Split and Join errors** include words that either get split into one of more parts or some words in a sentence get joined to a make a single word. For example: In Figure 2, the word "Thernndldntesnra" in the OCR text is actually a combination of three words "The candidates are" while the words "v Icrory" are actually equivalent to a single word "victory" when compared with the original news article.

4.2 Statistics

Article level segmentation of text is available for only two months – since this requires human intervention. Articles of "The Sun" newspaper from November-December 1894 consisting of 14020 news articles are used in our study. A total of 8,403,844 tokens are generated from a bag-of-words extraction. The text from the articles do not have any punctuation and contain a large amount of garbled text containing above mentioned OCR errors.

4.3 Preprocessing

The garbled OCR text makes data preprocessing mandatory before application of any text mining algorithms. We,therefore, use edit distance algorithm based on Levenshtein distance to perform spelling correction on the OCR text articles. The algorithm is chosen because of its speed and ability to correct OCR errors compared to the n-gram approach [42]. Our edit distance algorithm also uses an enhanced person names dictionary for look up to give significance to personal names spelling correction in the dataset. The results of spelling correction and data preprocessing are presented in [43].

5 People Gazetteer

People Gazetteer as defined in Section 1 consists of tuples of person names along with list of documents in which they occur and their corresponding topics. It is developed as an organized structure that can facilitate the process of detection of influential persons from the dataset in an efficient and easy way. This section describes the 2-step process of construction of the People Gazetteer by a) Extraction of person

names from the news articles dataset using Named Entity Recognition in Section 5.1 and b) Assignment of topics to news articles using LDA topic detection in Section 5.2. Output of People gazetteer developed using these steps is presented in Section 5.3.

5.1 Person Named Entity Recognition (PNER)

5.1.1 Definition

NER (Named Entity Recognition) refers to classification of elements in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. It can also be considered as a sequence labeling problem that predicts label of each element of the text based on the adjacent element labels.

Person Named Entity Recognition (PNER) can be defined as the process of NER that marks up only person names that occur in the text. It is done in two phases: chunking ,i.e., segmentation of text for name detection followed by classification of the name by the type of entity it belongs to (person,organization,location,etc.).

PNER is required in this research so as to extract all person name entities occurring in the complete dataset and then identify influential person entities among them through development of the People Gazetteer. PNER aids in the development of People Gazetteer by first extracting a list of all person names occurring in the dataset followed by creation of a Person-Article list where each person is linked with the articles in which he/she occurs.

5.1.2 Methodology

While earlier NER models were rule-based or dictionary-based that use linguistic grammar based techniques, current studies use statistical sequential models that predict labels for sequential data using probabilistic techniques. They also require labeled training data with all of the entities of interest and their types. Statistical machine learning models like Hidden Markov Models or MaxEnt Markov Models follow a generative or discriminitave approach respectively. In the generative approach, a joint probability distribution over input and output variables is modeled assuming independent features and considering future observations into account while the discriminative approach models conditional distribution directly, without assuming independent features or considering future observations. The linear chain CRF (Conditional Random Field) sequence model for NER is a combination of generative and discriminative approach and is considered state of the art for performing NER ([44], [45], [46]). The model can be viewed as a conditionally trained finite state machine which is used to find the possible label sequence given an input sequence and learning. It combines features of discriminative and generative models by relaxing the assumption that features are independent and considers future observations into account during sequential labeling.

The Stanford CRF-NER[7] is used for PNER in this research. It can perform NER for 3 classes: Person, Organization and Location and is based on linear chain CRF sequence models. It is trained across several newspaper corpora and is fairly robust across multiple domains and performs best when compared to some other open source NER systems which is the reason it has been used for this study. As illustrated by Rodriquez et. al [47], Stanford NER gave overall the best performance across 2 OCR datasets, and was most effective for PNER when compared with 3 other open source NER systems.

Stanford NER recognizes a person's full name as separate names by default which is rectified by combining these multi-term entities into single person entities. For example, the person name "John Smith" is recognized as two separate person entities which we combine to form a single multi-term person entity. Person names tagged with "PERSON" category are stored while running NER on the dataset. Whenever a multi-term person name (number of terms in the person name must be greater than 1) occurs in a document, the person entity's name along with the document name is stored to obtain tuples of person names with their document occurrences in a Person-Article List. The Stanford NER takes 25 minutes to run on the complete news dataset of 14020 articles extracting a total of 36362 person entities.

We divide the people entities extracted into following categories so that separate analysis can be done for each category:

- **Marginally Influential**: This category includes all person entities with occurrence in less than 4 news articles. (36004 person entities)
- **Medium Influential**: This category includes all person entities with occurrence from 4 to 15 news articles. (344 person entities)
- **Highly Influential** : This category includes all person entities with occurrence in 16 or more news articles. (14 person entities)

Figure 3 shows the statistics for each of these categories of persons extracted from the dataset. These categories have been chosen manually simply based on the number of articles of occurrence of a person entity and do not directly lead to the conclusion of

7. http://nlp.stanford.edu/software/CRF-NER.shtml

a person entity with large number of articles being influential.

5.2 Topic Detection

Topic models are algorithms for discovering the main topics that occur across a large and otherwise unstructured collection of documents and can organize the collection according to the discovered topics. Here, a topic refers to a set of words which describe what any document is about. A topic model examines the set of documents and discovers based on the statistics of the words in each, what the topics might be and what each document's balance of topics is. Documents are considered as a mixture of topics and each topic a probability distribution over words. Topic detection is the process of identifying topics in a document collection using a topic model.

Topic detection is essential to this research in order to determine the topics of individual news articles that a person entity occurs in so that the person entity can be linked to the documents in which he/she occurs along with their respective topics.

5.2.1 Topic Detection Model

5.2.1.1 **:** Latent Dirichlet Allocation (LDA) Model

LDA is a generative probabilistic model in which each document is modeled as a finite mixture over an underlying set of topics and each topic, in turn, is modeled as an infinite mixture over an underlying set of topic probabilities [48]. In other words, documents exhibit multiple topics and each topic is a distribution over a fixed vocabulary. The LDA model can be briefly reviewed as follows:

Given an input corpus of D documents with K topics, each topic being a multinomial distribution over a vocabulary of W words, the documents are modeled by fitting parameters 'Φ' and 'Θ'. 'Φ' is a matrix of size $D \times K$ in which each row is a multinomial distribution of document d indicating the relative importance of words in topics. Θ is the matrix of size $W \times K$ with each column a multinomial distribution of topic j and corresponds to the relative importance of topics in documents.

Given the observed words x = x_{ij}, LDA inference is done by computing the posterior distribution over the latent topic assignments z = z_{ij}, the mixing proportions Θ_j and the topics Φ_k. The inferencing is either done using variational bayesian methods or Gibbs sampling which involves integration and sampling of latent variables. However, the simple LDA approach can take several days to run over a large corpora.

5.2.1.2 **:** Distributed LDA Model

The simple LDA method takes a long time for topic modeling which is why the distributed version suits large datasets such as ours. The data is partitioned across separate processors and inference is done in a parallel, distributed fashion.

The Approximate Distributed LDA (AD–LDA) model as proposed by [49] uses distributed computation where total dataset D is distributed equally among multiple P processors. Initialization involves data and parameters distribution to each processor and random assignment of topics so that each processor has its own copy of words x_p, topics z_p, word topic counts N_{wkp} and topic counts N_{kj_p}. The topic model inferencing then uses simultaneous local Gibbs sampling approach on each processor for a pre-decided number of iterations to reassign topic probabilities z_p, word topic N_{wkp} and topic counts N_{kj_p}. Global update is performed after each pass by using a reduce-scatter operation on word topic count N_{wkp} to get a single set of counts and obtain final topic assignments. The model requires user set parameters before inferencing such as number of processors/threads for parallel sampling of data, number of iterations of Gibbs sampling, number of topics and Dirichlet parameters.

5.2.2 Topic Models Evaluation

Different topic models can be evaluated using the metric of "Perplexity" which can be defined as how surprised a trained model is when given a held out test data. It has been used in [49] and [48] for evaluating the topic detection models under different parameter settings. Perplexity can be calculated using the following formula:

$$Perplexity = \exp(-\frac{\text{Log Likelihood of held-out test set}}{\text{Number of tokens in held-out test set}})$$

Here, held-out test set refers to the fact that complete dataset is split into two parts: one for training and the other for testing. The test set is taken as the held-out set for which perplexity is calculated. The document mixture is learned using the training data and log probability of the test data containing unseen documents is computed using the model developed.

Perplexity is a decreasing function of the log likelihood of the unseen documents as can be seen from its formula and lower the perplexity, better is the topic model.

5.2.3 Results

The AD-LDA model as described in [49] and implemented in the Mallet [50] toolkit (known as PLDA model) is used for topic detection over the complete

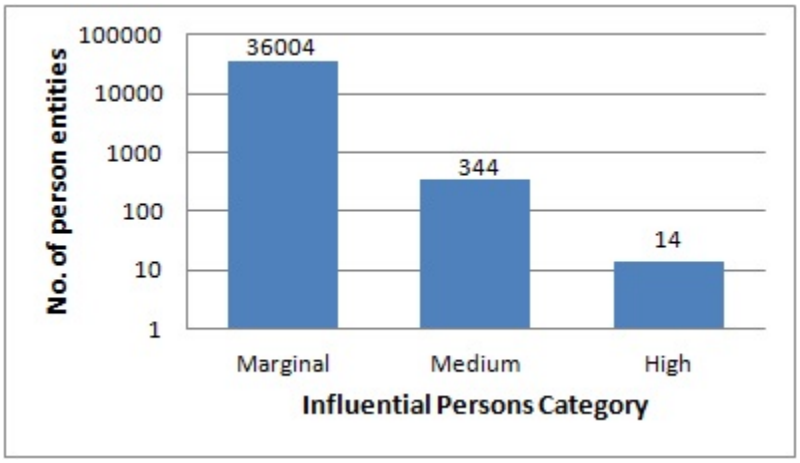


Fig. 3. Log Scale plot showing number of persons extracted for each influential person category after PNER from the dataset

dataset of 14020 news articles. Several topic models are first evaluated with different parameter settings in order to pre-decide the number of iterations, processors and topics for the final topic model to be used.

Perplexity is calculated by splitting the data into 90% for training and rest 10% for testing. Figure 4 shows the variation of the test perplexity versus the number of topics for one random $90 - 10$ split of the data[8]. The maximum number of words in each topic is set to 20, number of iterations 500 and the number of processors 4 for this experiment. It exhibits a decreasing perplexity with increase in number of topics. Typically, the number of topics should be chosen as high as possible in order to consider a better model with low perplexity but the model with high number of topics also takes longer to run on a large dataset. The number of topics is set to a value from where further increase in number of topics does not lead to a large decrease in perplexity. We choose the number of topics as 30 and 100 and demonstrate their effect on the influential people detection.

From the various topic models and parameter settings, the variability in perplexity with respect to the number of topics has been found to be much greater than the variability due to the number of processors or number of iterations. This is why two values of number of topics are experimented further while number of processors and number of iterations are kept fixed. The number of iterations of Gibbs sampling still need to be above the typical burn-in period of 200 which is why 500 is chosen as the parameter value for number of iterations. Number of threads/processors is similarly taken as 4 as least training time is obtained with this parameter value.

The two models from topic detection are thus used with following parameters:

1) **30 Topics LDA Model** : Number of topics = 30, Number of iterations = 500, Number of threads=4
2) **100 Topics LDA Model** : Number of topics = 100, Number of iterations = 500, Number of threads=4

The first model takes 7.5 minutes for training while the second one takes 8.6 minutes. Some of the topics words from the topic models can be easily identified to belong to the following topics: music performance, court events, elections and government and shipping.

Topic modeling gives as output, for each article in the dataset, a set of topics with their probability distribution score for the article. The topic with highest topic probability score is associated with each article in the dataset to obtain an Article-Topic List.

5.3 People Gazetteer Output

The procedure of development of people gazetteer can be seen in Figure 5.The list of articles obtained for each person entity after application of PNER (Person-Article List) and highest scoring topic assigned to each article during Topic Detection (Article-Topic List) are combined to obtain People Gazetteer. In each tuple of the gazetteer, a person entity gets associated with its list of articles in which it occurs and where each article is further associated with its corresponding highest scoring topic.

Two people gazetteers are finally developed, each corresponding to the two model settings of 30 Topics

8. We also vary the number of iterations from 100 to 500 and number of processors from 1 to 8 to study their effect on perplexity. However the number of topics is most influenced by perplexity and hence the other results are not presented here.

Number of Topics	Perplexity
10	17227
20	16431
30	15940
40	15634
50	15480
60	15390
70	15316
80	15208
90	15031
100	14988
200	14355

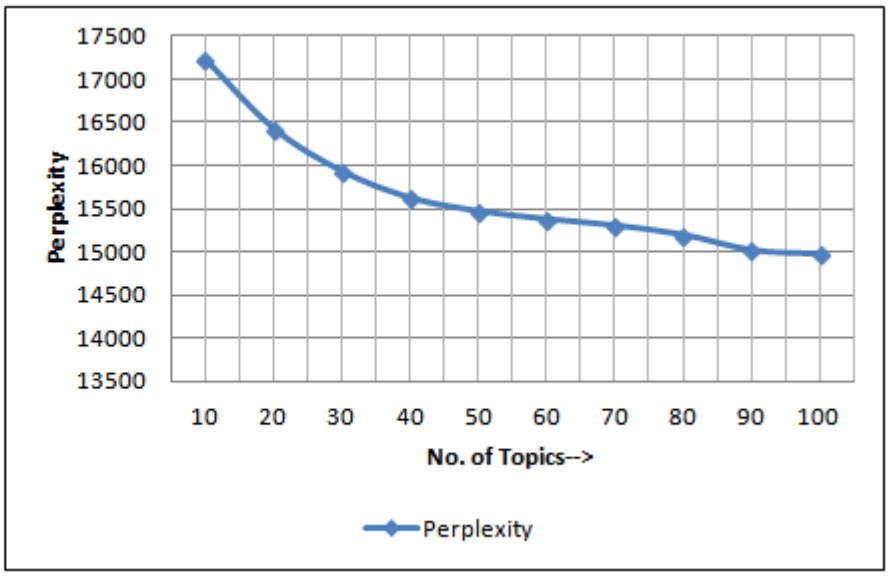


Fig. 4. Test Set Perplexity versus Number of Topics for a random $90 - 10$ split of the data. The maximum number of words in each topic is 20, number of iterations 500 and the number of processors 4 for this experiment.

Fig. 5. Procedure for development of People Gazetteer

PERSON ENTITY NAME	DOCUMENT LIST {DOCUMENT ID→DOCUMENT TOPIC}
Thomas Murphy	{61720.txt→16, 62002.txt→11, 65905.txt→19, 71341.txt→28, 68024.txt→16}
George Eliot	{74151.txt→5, 61627.txt→15}
Charles L Thompson	{68836.txt→9}
Thomas Jefferson	{67874.txt→19, 67209.txt→28, 63996.txt→6, 73835.txt→6, 71155.txt→6, 65440.txt→5, 66997.txt→20}
Jacob Schaefer	{70205.txt→21, 63936.txt→22, 68554.txt→21, 73420.txt→21, 74550.txt→21, 74922.txt→21, 64577.txt→21, 74759.txt→21, 67340.txt→0, 67924.txt→21}
Queen Victoria	{68231.txt→5, 74775.txt→5, 75097.txt→5, 72221.txt→2, 62731.txt→5, 62616.txt→17, 68368.txt→17}
Thomas Gallagher	{64397.txt→28, 65793.txt→21, 72591.txt→0, 73420.txt→21}
Samuel S Seely	{70365.txt→2, 64670.txt→23, 65615.txt→23, 67198.txt→19, 73545.txt→23, 74816.txt→16}
Matthew Parker	{64363.txt→11}
Daniel Frohman	{63704.txt→5, 66992.txt→25, 69668.txt→4, 68743.txt→5, 67554.txt→25, 67450.txt→5, 72274.txt→24, 69444.txt→4}

Fig. 6. Snapshot of People Gazetteer with Person names, Document list of occurrence and their corresponding Topic ID

LDA Model and 100 Topics LDA Model, respectively. Both gazetteer consist of a list of named entities along with the articles of their occurrence and topic associated with each article. A snapshot of the people gazetteer using 30 Topics LDA Model can be seen in Figure 6 where each person entity is followed by a document list consisting of a Document ID and its corresponding Topic ID. A similar people gazetteer is also obtained using 100 Topics LDA Model. Both People Gazetteers are further used in Section 6 for detecting and ranking influential person entities from them.

6 Influential People Detection

To measure influence in the newspaper environment and to compare and rank people as influential, we define an influence score measure called *"Influential Person Index" (IPI)* corresponding to each person entity in the people gazetteer. To calculate IPI for each person entity, we first define the *"Document Index" (DI)* to measure how each document in the person entity's associated list of documents affects his influence score. The choice of features for detecting a person entity as influential is motivated by following questions: Are frequently occurring persons in the newspaper influential? Does frequency mean occurrences of a person entity in a single article or across complete dataset? Do longer documents tend to talk about more important persons? Is a person entity occurring over multiple topics more influential or the one who is consistently talked about in similar topic articles?

Following subsections describe the features/parameters chosen for calculation of DI and IPI of a person entity followed by the complete algorithm for detection of influential persons:

6.1 Document Index (DI)

The Document Index (DI) of an article in the people gazetteer helps to measure a person's influence score. Following parameters are considered for the calculation of this index:

1) **Normalized Document Length** (NDL)
Document Length affects the influence score in the sense that a longer news article in which

a person entity occurs is deemed to be more important than a shorter one. It is defined as the number of tokens contained in a news article. Document Length is further normalized by dividing it with the maximum news article length (of 14020 articles in the dataset) to get Normalized Document Length as follows:

$$NDL = \frac{\text{Document Length}}{\text{Maximum Document Length in the dataset}}$$

2) **Normalized Term Frequency** (NTF)
Term Frequency (TF) accounts for the number of occurrences of a person's name in a news article. The TF of the person name affects a document's influence score as a higher number of occurrences in the document makes it more important. TF is further normalized and calculated as follows:

$$NTF = 1 + \log(\text{TF of person entity in current article})$$

The issues of co-reference resolution of person names (For Example, person entities such as "William Schmittberger","Captain Williams" are same but recognized as separate persons) and named entity disambiguation (Occurrence of different persons with similar name in news articles. For example, the person "John Smith" detected in two different articles might or might not be the same person) occur in our People Gazetteer which are not taken care of by PNER and need to be addressed separately. While the issue of co-reference can be still addressed by analyzing each news article, it is extremely hard to disambiguate among persons with similar names that can occur in multiple news articles with different topics. This is the reason coreference resolution is performed for the person entities obtained using PNER.The following section explains the approach to coreference resolution followed after development of People Gazetteer.

6.1.1 Coreference Resolution

The coreference resolution aims to find out all expressions that refer to the same entity in the text. Due to multiple references to a person in the text, coreference resolution been done keeping in mind that the number of articles of occurrence of a person entity, i.e.,TF is an important parameter in our study for determining an influential person.
Coreference resolution is performed using the Standford Deterministic Coreference Resolution System of the Stanford CoreNLP toolkit /footnote-http://nlp.stanford.edu/software/dcoref.shtml. It uses a multi-pass seive coreference resolution [51] which consists of 3 steps: mention detection which identifies clusters of all noun phrases, pronouns and named entity mentions, coreference resolution step which is a combination of ten independent seives applied from highest to lowest precision one by one sequentially but with global information sharing so that each seive builds on the previously clustered mentions followed by post processing which removes singleton mentions. Coreference Resolution was observed to be useful during this case study since the frequency of a person entity in a newspaper article changes when multiple references to the same person entity (coreferences) are found in an article. Figure 7 or 8 illustrates an example of coreference resolution when applied to an article text and its effect on term frequency for person entities. Initially, mentions are detected from the text followed by application of coreference resolution step which results in a chain of coference mentions/entities, which includes the mention itself along with other mentions that refer to it. The most representative mention out of each coference chain is found and checked whether it is a valid person entity in the People Gazetteer or not. If it is found to be a person entity, then the count of coreferences for that mention is considered its final term frequency in an article. It is also observed that due to lack of punctuation in the dataset, lots of meaningless coreference mentions are also detected. However, since the persons list is available from the People Gazetteer, frequencies of only those person entities are replaced through Coreference Resolution.

3) **Number of similar articles** (NSIM)
This parameter is used in calculation of the DI by finding articles of similar topic in the document list. Two documents are considered similar if they belong to the same topics. For a document d whose DI is to be calculated, we consider
SIM= Number of articles with the same topic as that of d in the document list of person entity.
This measure is normalized by dividing it with the number of total articles in the doc-

Text: Mr Eugene Kelly bead of the banking house of Eugene Kelly A Co Is dying at his home 33 West Fiftyfirst street He became ill on Dee 4 and was forced to lake to his bed

INITIAL MENTIONS EXTRACTED	COREFERRED MENTIONS (Most representative mention in bold)
Eugene Kelly	**Eugene Kelly**, Eugene Kelly, his, He, his
Mr Eugene Kelly bead of the banking house of Eugene Kelly A Co	**Mr Eugene Kelly bead of the banking house of Eugene Kelly A Co**
the banking house of Eugene Kelly A Co	**the banking house of Eugene Kelly A Co**
Eugene Kelly A Co	**Eugene Kelly A Co**
Eugene Kelly	
his home 33 West Fiftyfirst street He became ill on Dee 4 and was forced to lake to his bed	**his home 33 West Fiftyfirst street He became ill on Dee 4 and was forced to lake to his bed**
his	
33	**33**
He	
Dee 4	**Dee 4**
4	**4**
his bed	**his bed**
his	

MOST REPRESENTATIVE MENTION	IS MOST REPRESENTATIVE MENTION A PERSON NAMED ENTITY?	NEW TERM FREQUENCY OF PERSON NAMED ENTITY
Eugene Kelly	Yes	5
Mr Eugene Kelly bead of the banking house of Eugene Kelly A Co	No	
the banking house of Eugene Kelly A Co	No	
Eugene Kelly A Co	No	
his home 33 West Fiftyfirst street He became ill on Dee 4 and was forced to lake to his bed	No	
33	No	
Dee 4	No	
4	No	
his bed	No	

Fig. 7. Figure illustrating change in term frequency for person named entities on using coreference resolution on an article text

Text: Mr Eugene Kelly bead of the banking house of Eugene Kelly A Co Is dying at his home 33 West Fiftyfirst street He became ill on Dee 4 and was forced to lake to his bed

Mentions Extracted:
[“Mr Eugene Kelly bead of the banking house of Eugene Kelly A Co”], [“Eugene Kelly”], [“the banking house of Eugene Kelly A Co”], [“Eugene Kelly A Co”], [“Eugene Kelly”], [“his home 33 West Fiftyfirst street He became ill on Dee 4 and was forced to lake to his bed”], [“his”], [“33”], [“He”], [“Dee 4”], [“4”], [“his bed”], [“his”]

Coreferred Entity Chains with most representative entity in bold:
["**Mr Eugene Kelly bead of the banking house of Eugene Kelly A Co**"],
["**Eugene Kelly**", "Eugene Kelly", "his", "He", "his"],
["**the banking house of Eugene Kelly A Co**"],
["**Eugene Kelly A Co**"],
["**his home 33 West Fiftyfirst street He became ill on Dee 4 and was forced to lake to his bed**"],
["**33**"],
["**Dee 4**"],
["**4**"],
["**his bed**"]

Count of mentions for most representative entity which is also a Person Named Entity:
Eugene Kelly: 5

Fig. 8. Figure illustrating change in term frequency for person named entities on using coreference resolution on an article text

ument list of the person entity as follows:

$$NSIM = \frac{\text{SIM}}{\text{Total number of articles in the person's document list}}$$

NSIM can be said to be equivalent to the proportion of topic similar articles that any document d has.
This parameter takes into account the effect of a document's score on a person's IPI when there exist several other documents of the same topic in the person's list.

DI for each document is a function of the above mentioned parameters and is calculated using the following formula :

$$DI = w_a.NDL + w_b.NSIM + w_c.NTF$$

where, w_a,w_b and w_c are the weights associated with each of the parameters NDL, NSIM and NTF respectively.

DI is actually a heuristic measure of these three parameters where each of the parameters can be weighted as per dataset characteristics and user requirements. For example, a higher value to w_a and lower to w_b and w_c indicates documents with longer lengths are considered more important for influencing a person's IPI. On the other hand, a higher value to w_b and lower to w_a and w_c indicates a document with larger proportion of topic similar articles influences the person's IPI more suggesting assignment of high influence score to a person entity occurring repeatedly in a specific news topic.

6.2 Influential Person Index (IPI)

Once DI is calculated for each document in a person's list, an index is calculated for the person entity in order to measure its influence in the news dataset and calculate its influential score. The “Influential Person Index” defined for this purpose is calculated as follows:

$$IPI = maxDI(d_1, d_2, ..., d_n) + UniqT$$

where , $maxDI(d_1, d_2, ..., d_n)$ = Maximum Document Index of a document d_i in a person entity's list of n articles, and

$$UniqT = \frac{\text{Number of Unique Article Topics in a person entity's document list}}{\text{Total Number of Topics in the corpus}}$$

The parameter $UniqT$ is used to account for the fact that a single person entity can be talked about multiple news topics in the news articles and to include its effect on the person entity's influence score. It is normalized by dividing it with the total number of topics as obtained during topic detection on all 14020 articles.

Ranking is done across each person category of the people gazetteer to obtain top most influential persons. For this, IPI for each person entity across the person categories are sorted in decreasing order to obtain the most influential person entities with highest IPI at the top.

6.3 Procedure for finding influential persons

Algorithm 1 depicts the procedure for measuring influence and ranking of influential people from the gazetteer. It starts with calculation of DI for each news article in a person's document list by calculating the required parameters of NDL, NSIM and NTF which are assigned 0 values initially. The respective weights w_a,w_b,w_c are taken as inputs and multiplied with each parameter to get final DI score which is added to the list of DI scores $DIScoreList$. The list is sorted to find the maximum DI value among all news articles in the person's document list. The maximum DI score is then added to the UniqT parameter to get the final IPI for each person entity which are again stored and sorted to obtain a ranked list of influential person entities.

function CALCULATEIPI
Input: $PeopleGazetter(Persons, (DocList,$
Input: $TopicList))$, w_a,w_b,w_c
Result: Ranked list of Person Name and IPI
$NTF \leftarrow 0$, $NDL \leftarrow 0$, $NSIM \leftarrow 0$, $DI \leftarrow 0$, $UniqT \leftarrow 0$, $IPI \leftarrow 0$;
for *(String PersonName : Persons)* **do**
  for *(String doc : DocList)* **do**
    $NTF = 1 + \log(GetPersonTF(doc))$;
    $NDL = GetDocLength(doc)/GetMaxDocLength()$;
    $NSIM = GetTopicSimilarArticles(doc, DocList)$;
    $DI = w_a.NDL + w_b.NSIM + w_c.NTF$;
    $DIScoreList.add(DI)$;
  end
  $Sort(DIScoreList)$;
  $UniqT = GetUniqueTopics(Person, TopicList)$;
  $IPI = Max(DIScoreList) + UniqT$;
  $IPIScores.put(PersonName, IPI)$;
end
$Sort(IPIScores)$;
$PrintPersonNameandMaxIPI(IPIScores)$;
end function

Algorithm 1: Procedure to calculate IPI and rank person entities based on it

6.4 An alternative approach for detecting influential persons

A heuristics based approach for finding influential persons has been discussed in the previous section. An alternative approach involving clustering can also be used for detection of influential persons. One such multiple instance clustering algorithm is suggested in [52]. They suggest an algorithm called BAMIC which can be applied to our problem as well. The multiple instance clustering problem considers clustering objects that consist of sets of instances for clustering rather than single instance clustering. According to the BAMIC algorithm, a set of instances is represented by a bag object and k-medoids algorithm is used to cluster those bags. The k-medoids algorithm is adapted to use average Hausdorff distance to measure the similarity between instances of different bags. It averages the distance between each instance in one bag and its nearest instance in the other bag and partitions dataset into k disjoint groups each containing a set of bags. BAMIC is applied to MUSK 1 and MUSK 2 datasets available publically[9] which consist of 92 bags with 476 instances and 102 bags with 6598 instances, respectively and is used to test whether molecules are qualified to be used in a drug or not. This approach can be used to detect influential persons in our problem by clustering person entities into "influential" or "non-influential" considering each person entity of our people gazetteer as a bag with articles of their occurrence as the instances for each bag. The parameters used for calculating DI in the previous section: NDL,NTF and NSIM can be used as features associated with each article instance in a bag. Such a method can avoid choosing of parameter weights, biasing of results with respect to any specific parameter and decide which article plays a role in determining whether a person is influential or not. We tried to work with the open source version of the BAMIC algorithm to compare its results with the heuristic based approach suggested in this paper. But the clustering algorithm, due to its high complexity and the amount of data we worked

9. https://archive.ics.uci.edu/ml/datasets.html

Function Name	Description
GetPersonTF(doc)	Calculates TF of the person entity in document doc
GetDocLength(doc)	Calculates number of tokens in doc.
GetMaxDocLength()	Calculates maximum number of tokens in any document.
GetTopicSimilarArticles(doc,DocList)	Calculates normalized number of topic similar articles for doc in the $DocList$.
Sort(DIScoreList)	Sorts the $DIScoreList$
Max(DIScoreList)	Finds the maximum score from $DIScoreList$.
GetUniqueTopics(Person,TopicList)	Calculates normalized unique topics for $Person$ in its $TopicList$.
Sort(IPIScores)	Sorts the $IPIScores$ by IPI values.
PrintPersonNameandMaxIPI(IPIScores)	Prints $Person$ name with its IPI in decreasing order of IPI value.

TABLE 1
Description of the functions used in Algorithm 1

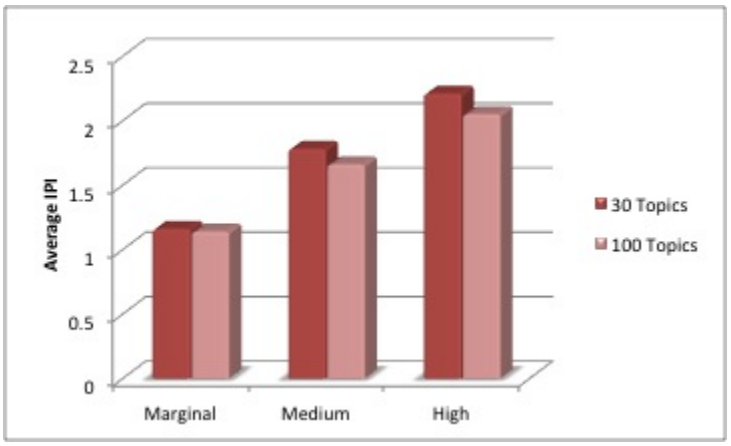


Fig. 9. Comparison of the Average IPI for two ranked lists L_1 and L_2 using 30 and 100 topics respectively.

with, the algorithm takes a very long time to give the results. Our dataset consisted of roughly 40000 person named entities on which multiple instance clustering was required and according to the estimation, it will take around 200 days to get the clusters of influential and influential persons. Due to unavailability of such a long time frame, we do not present the results of comparison between the two approaches for detection of influential persons. Since BAMIC has been used for smaller datasets in earlier studies, we believe if the BAMIC algorithm can be scaled for larger datasets, it can be applied to our scenario easily.

7 Results

Two ranked influential person lists, namely L1 and L2 are obtained after calculation of IPI from the people gazetteer (developed in Chapter 5) using 30 Topics and 100 Topics LDA Model respectively. The weights w_a, w_b and w_c are all set to 1 to give equal importance to each of the parameters during calculation of DI and IPI. The statistics obtained from both lists with respect to each person category of the people gazetteer are shown in Table 2. It can be clearly observed from the table that Highly Influential Persons occur in most number of news articles on an average and with highest average term frequency followed by Medium Influential and Marginal Influential Persons. Document Length need not always be too high for a person to be ranked higher as can be observed from the fact that average document length obtained for Marginally Influential People is high in spite of their Average IPI being low indicating that the varying number of similar articles for each document as well as its Term Frequency share also play an important part in measuring influence. Figure 9 shows the average IPI from the two ranked lists – it appears that the average IPI for highly influential people is more susceptible to changes in number of topics.

Person Category	Number of Person Entities	Average Number of Documents	Average Document Length	Average Term Frequency
Marginal	38066	1.04	2119.6	1.07
Medium	344	5.75	1976.3	6.68
High	16	22.8	2971.5	29.870

TABLE 2
Table illustrating average statistics for each Person Category of People Gazetteer across 2 Topic Models

The following sections present comparison between the ranked influential person lists L1 and L2, some case studies and evaluation results:

7.1 Comparison Across Ranked Influential Person Lists

The top 10 influential persons from List L1 and L2 detected from each of the people gazetteers are presented in Table 3 and 4 respectively. It can be clearly seen from both the tables that the person category labels assigned during development of people gazetteer do not hold true after detection of influential persons. This suggests that the highly influential category people which were defined as person entities with more than 16 articles in the dataset might not necessarily be the most influential. The top 10 influential persons in both tables are dominated by Medium and Marginal category persons having considerably less number of articles of occurrence. This indicates the fact that number of articles of occurrence has not been given priority while measuring influence of a person entity. The statistics for top 10 influential people from both the tables also suggest that none of the measures of NDL, NTF or NSIM can be alone used to say whether a person entity is influential since these value do not decrease or increase consistently although the NTF measure does contribute most to the IPI of any person.

The ranked influential lists L1 and L2 can be contrasted in terms of NSIM, UniqT and Topic Words since they vary across different number of topics and to see the effect of 30 and 100 Topics LDA Models on influential person detection. If NSIM (normalized number of topic similar articles) remains same in L1 and L2 during influential person detection from both the people gazetteers, then the same highest scoring article' DI is selected for calculation of IPI in both of them. This is why the parameters NDL (Normalized Document Length) and NTF (Normalized Term Frequency) remain same across both the lists. This can be seen for person like "capt creeten", "capt hankey", "aaron trow" and "mrs oakes" in Tables 3 and 4. But the value of UniqT for these persons decreases leading to decrease in their final IPI in the second table. This is because LDA model with higher number of topics (100) is used in this case due to which the proportion of unique topics becomes lower when NSIM does not change. However, when the NSIM (normalized number of topic similar articles) value changes because of change in number of topics, a different article with maximum DI score can get selected leading to change in the values of NDL, NTF, UniqT and the final IPI. This causes a shift in the ranking of influential persons across the two lists and can be seen when the rank of "alexander iii" in the first table moves from 9 to 4 in the second table. This indicates the fact that LDA Topic Model used affects the ranking of influential persons when number of topics are varied.

Wilcoxon signed rank paired test is also performed on the ranks of influential people across the two lists L1 and L2. This is done to test the hypothesis whether the differences in the ranking of person entities obtained using the 30 Topic and 100 Topic LDA models are due to chance or not. The null hypothesis for the test is:

H_0: the distribution of difference of ranks of the persons across L1 and L2 is symmetric about zero

On performing the normal distribution approximation for 36362 samples of person ranks from lists L1 and L2, the results are found to be significant for both one-tail and two-tail tests. For one-tail test, the p-value=0, z-score=129.7085544 and T-critical value=229859426 while for the two-tailed test, the p-value=0 and T-critical value= 229375179.6 The results lead to the conclusion that there is a significant difference between the ranks of influential persons across different topic models and number of topics chosen for topic modeling affects the ranking of person entities in this case.

7.2 Case Studies

Some of the topmost 10 influential person entities of lists L1 and L2 (Table 3 and 4) identified from each person category of the 2 people gazetteers are discussed below:

1) Highly Influential Category- This category as defined in Section 5.1.2 includes person entities influencing number of news articles greater than 16. However, only one person

Person Name	IPI	Number of Articles	Person Category	NDL	NTF	NSIM	TOPIC WORDS	UniqT	Rank
capt creeten	3.32	10	Medium	0.55	1.9	0.8	mr court police judge justice case yesterday street district	0.06	1
capt hankey	3.05	5	Medium	0.68	1.69	0.6	club game team play football half ball left college back	0.06	2
capt pinckney	2.93	3	Marginal	0.38	1.84	0.67	man ho men night back wa room left house told bad	0.03	3
john martin	2.89	14	Medium	0.55	1.6	0.57	mr court police judge justice case yesterday street district witness	0.16	4
ann arbor	2.87	44	High	0.19	1.77	0.63		0.26	5
john macdonald	2.85	3	Marginal	0.55	2.2	0	great people life man women good country world american part	0.1	6
aaron trow	2.81	1	Marginal	0.7	2.07	0	man ho men night back wa room left house told	0.03	6
mrs oakes	2.79	5	Medium	0.08	2.04	0.6	street mrs mr avenue wife house miss yesterday years home	0.06	7
alexander iii	2.71	31	High	0.24	2.04	0.25	great people life man women good country world american part	0.16	9
buenos ayres	2.7	6	Medium	0.49	1.47	0.67	white water indian black long found thu big dog time	0.06	10

TABLE 3
Table showing top 10 influential persons of List L1 detected from People Gazetteer with 30 Topics LDA model. Parameters NDL, NTF,NSIM and Topic Words belong to the maximum scoring DI in the person's document list.

entity ("alexander iii") from this category occurs in the top 10 influential persons. The entry for "alexander iii" has an IPI of 2.71 and 3.05 respectively in list L1 and L2 . The person entity occurs in 31 news articles with 5 and 7 different topics in each of the lists. The most common topic words associated with this person entity are: "emperor prince french alexander czar london nov government imperial russian" indicating the importance of this entity in government related news topics. The 100 Topic LDA model increases the IPI value of this entity because the NSIM value increases (more number of similar topic articles talk about this person) and a longer article gets maximum DI score resulting in a high IPI value and improvement in the ranking from rank 9 in the first table to rank 4 in the second. It is also observed that "ann arbor" occurring in 44 articles is ranked 5 in list L1 but this iis a false positive as it is actually a location and has been misrecognized in the PNER process as a person entity.

2) Medium Influential Category- The top 10 influential entities from Tables 3 and 4 contain the most number of person entities from this person category. The person entity"capt creeten" has been ranked as highest influential (Rank 1) across both the tables. It occurs in 10 news articles with 9 of them belonging to the same topic indicating the person influencing news articles of high topic similarity. Some of the most common topic words for this entity include " mr police witness committee capt asked captain money inspector paid" indicating the importance of this entity in a judicial or police related news topic. Several persons from this category like "mrs martin" , "mrs oakes" although identified among the top 10 influential persons but suffer from the problem of named entity disambiguation as it is hard to identify which exact person they refer to due to lack of first names.

3) Marginally Influential Category- Person entities belonging to this category have extremely low occurrence in news articles although the IPI of topmost influential entities belonging to this category are comparable to those in the other 2 categories. Several person entities occurring in low number of news articles like "aaron trow", "caleb morton", "john macdonald" belong to this category. These entities in

Person Name	IPI	Number of Articles	Person Category	NDL	NTF	NSIM	Topic Words	UniqT	Rank
capt creeten	3.28	10	Medium	0.55	1.90	0.8	mr police witness committee capt asked captain money inspector paid	0.06	1
mrs martin	3.21	8	Medium	0.20	2.36	0.62	mrs mr years wife home house ago woman city died	0.02	2
capt hankey	2.97	6	Medium	0.68	1.69	0.8	game team football play half line ball back yale eleven	0.02	3
alexander iii	3.05	31	High	0.49	2.04	0.45	emperor prince french alexander czar london nov government imperial russian	0.07	4
aaron trow	2.79	1	Marginal	0.70	2.07	0	day place long great water time feet found good men	0.01	5
john martin	2.78	14	Medium	0.55	1.6	0.57	mr police witness committee capt asked captain money inspector paid	0.05	6
john macdonald	2.77	3	Marginal	0.55	2.2	0	people american man great country men world life good english	0.02	7
mrs oakes	2.74	5	Medium	0.08	2.04	0.6	mrs mr years wife home house ago woman city died	0.02	7
ed kearney	2.63	7	Medium	0.16	1.6	0.85	won time race ran mile furlough half lo track fourth	0.01	9
caleb morton	2.61	1	Marginal	0.70	1.9	0	day place long great water time feet found good men	0.01	10

TABLE 4
Table showing top 10 influential persons of List L2 detected from People Gazetteer with 100 Topics LDA model. Parameters NDL, NTF,NSIM and Topic Words belong to the maximum scoring DI in the person's document list.

spite of occurring in very few articles (1 to 3) have high term frequency in those articles with comparatively longer article length indicating the importance of these entities with respect to the articles they occur in. Since each of the features has been given equal weight during the calculation of IPI, these person entities with high NDL and NTF have been identified among the top 10 influential persons.

7.3 Evaluation

Due to the unavailability of ground truth consisting of influential people in the newspaper archives from November-December 1894, there is no way to validate our results. To broadly evaluate our results, a simple web search query with the person entity's name in the context of 19th century was done on the Wikipedia website for the top 30 influential persons of Lists L1 and L2 detected from the people gazetteer with 30 Topics LDA and 100 Topics LDA Model respectively.

Among the top 30, 16 person entities from List L1 and 14 from List L2 were found to be influential and popular in the 19th century across topic categories like theatre, politics, government, shipping, etc. Some of these influential persons from Lists L1 and L2 found in Wikipedia are shown in Figure 10.

Most of the false positives although influential in other respects but were not influential 'person' entities which can attributed to the incorrect PNER (Person Named Entity Recognition) on noisy OCR data. False positives are obtained for person entities such as "mr got" which is not a person entity at all and for entities such as "ann arbor" and "van cortlandt" which are in fact locations but got recognized as highly influential person entities.

The ranked list of the top 30 influential persons with their IPI from Lists L1 and L2 can be seen in the Appendix (Table 5. 6) where evaluation result for each person entity is also presented.

Fig. 10. Some of the top 30 influential persons obtained from the dataset and also found on Wikipedia during evaluation

8 Discussion

- We used a linear combination of each of the parameters in calculation of DI and IPI and assigned equal values to the weights associated with each of them by not favoring any specific parameter. This is evident from the results which do not consistently favor any specific parameter. The parameters defined are based on heuristics and can be re-weighted according to user requirements or new parameters can be defined to do so.
- The parameters for calculation of DI and IPI can also be learned by performing regression analysis using a manually developed sample of topmost influential people and obtaining the complete list of ranked influential people based on the learned parameters.
- The NDL(Normalized Document Length) parameter defined for calculation of DI is normalized using the maximum length of any document in the dataset. However, there might exist other ways of normalization of Document Length like using total number of tokens in a person entity's document list or total number of tokens in the complete dataset which can be experimented with according to the dataset.
- The topmost influential people contain several false positives also which occur not due to the influence measures defined but due to other factors like Named Entity Disambiguation which is not addressed in this paper. Several location and organization names have been misrecognized as person entities after performing Spelling correction and PNER resulting in false detection of some highly influential entities like "van cortlandt", "ann arbor" , "sandy hook", etc.
- The choice of parameters for topic detection also affects the detection of influential people which is evident from the fact that we get different ranking of influential people for the two different LDA Topic model settings used.

9 Conclusion

The problem of finding influential people from historical OCR news repository has been studied in this research. In studying this novel problem, our main aim was to develop a complete solution framework for this problem and present insights from the results obtained. We made novel contributions to the problem solution by developing a people gazetteer for facilitating the process of influential people detection and finally defining parameters and measures in the

newspaper community to obtain the ranked list of influential people. Spelling correction algorithms with improved accuracy can certainly improve the influential persons results. Topic detection algorithms also need to be designed to enable them to deal with noisy OCR text in a better manner as some of the topics we obtained using LDA came out to be garbled and were difficult to understand in order to perform human-assigned manual labeling on them and use them further for finding similarity across articles. We didn't consider Named Entity Disambiguation into account while developing the people gazetteer for detection of influential people which is a difficult problem in itself since it is hard to disambiguate among persons with similar names that can occur in multiple topic related articles in newspapers. The problem presented in this paper requires research into better spelling correction, named entity recognition, topic detection algorithms and stricter measures of calculation of influence score and ranking of influential persons.

The parameters we defined for measuring influence scores of persons in news articles are based on heuristics and can be re-weighted according to user requirements or new parameters can be defined based on the characteristics of an OCR newspaper dataset making it an open research problem.

Non-heuristic based estimation for finding influential persons can also be done using optimization approaches such as unsupervised multiple instance clustering [53] [52] but they need to be adapted in order to be used in a large scale environment.

REFERENCES

[1] R. B. Allen and R. Sieczkiewicz, "How historians use historical newspapers," *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1–4, 2010.

[2] M. Bilenko, R. J. Mooney, W. W. Cohen, P. D. Ravikumar, and S. E. Fienberg, "Adaptive name matching in information integration," *IEEE Intelligent Systems*, vol. 18, no. 5, pp. 16–23, 2003.

[3] C. Friedman and R. Sideli, "Tolerating spelling errors during patient validation," *Comput. Biomed. Res.*, vol. 25, no. 5, pp. 486–509, Oct. 1992.

[4] L. Stone, "Prosopography," *Daedalus*, vol. 100, pp. 46–79, 1971.

[5] R. B. Allen, "Toward an interactive directory for norfolk, nebraska: 1899-1900," in *IFLA Newspaper and Genealogy Section Meeting, Singapore*, 2013.

[6] D. Shahaf and C. Guestrin, "Connecting two (or less) dots: Discovering structure in news articles," *ACM Transactions on Knowledge Discovery from Data*, 2011.

[7] E. Gabrilovich, S. Dumais, and E. Horvitz, "Newsjunkie: Providing personalized newsfeeds via analysis of information novelty," in *Proceedings of the 13th International Conference on World Wide Web*, 2004, pp. 482–490.

[8] O. Alonso, K. Berberich, S. J. Bedathur, and G. Weikum, "NEAT: news exploration along time," in *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*, 2010, p. 667.

[9] A. Khurdiya, L. Dey, N. Raj, and S. M. Haque, "Multi-perspective linking of news articles within a repository," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, ser. IJCAI'11, 2011, pp. 2281–2286.

[10] K. R. McKeown and D. R. Radev, "Generating summaries of multiple news articles," in *Proceedings, ACM Conference on Research and Development in Information Retrieval SIGIR'95*, Seattle, Washington, July 1995, pp. 74–82.

[11] J. Otterbacher, D. Radev, and O. Kareem, "News to Go: Hierarchical Text Summarization for Mobile Devices," in *29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, August 2006.

[12] D. R. Radev and K. R. McKeown, "Building a generation knowledge source using internet-accessible newswire," in *Proceedings, Fifth ACL Conference on Applied Natural Language Processing ANLP'97*, Washington, DC, April 1997, pp. 221–228. [Online]. Available: http://www.cs.columbia.edu/~radev/publication/anlp97

[13] D. R. Radev, S. Blair-Goldensohn, Z. Zhang, and R. Sundara Raghavan, "Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization," in *Demo Presentation, Human Language Technology Conference*, San Diego, CA, March 2001.

[14] D. R. Radev, J. Otterbacher, A. Winkel, and S. Blair-Goldensohn, "Newsinessence: Summarizing online news topics," Communications of the ACM, pp. 95–98, 2005.

[15] A. Balahur and R. Steinberger, "Rethinking sentiment analysis in the news: from theory to practice and back," *Proceeding of WOMSA*, vol. 9, 2009.

[16] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs." *ICWSM*, vol. 7, p. 21, 2007.

[17] J. Li and E. Hovy, "Sentiment analysis on the peoples daily," in *Proceedings of EMNLP*, 2014.

[18] B. Masand, G. Linoff, and D. Waltz, "Classifying news stories using memory based reasoning," in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '92, 1992, pp. 59–65.

[19] R. Nallapati, A. Feng, F. Peng, and J. Allan, "Event threading within news topics," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, ser. CIKM '04, 2004, pp. 446–453.

[20] D. R. Radev, "Topic shift detection - finding new information in threaded news," Columbia University, Tech. Rep. CUCS-026-99, 1999.

[21] C.-m. Au Yeung and A. Jatowt, "Studying how the past is remembered: towards computational history through large scale text mining," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1231–1240.

[22] A. Lee, H. Dutta, R. Passonneau, D. Waltz, and B. Taranto, "Topic identification from historic newspaper articles of the new york public library: A case study," in *5th Annual Machine Learning Symposium at the New York Academy of Sciences*, 2010.

[23] H. Dutta, R. J. Passonneau, A. Lee, A. Radeva, B. Xie, D. L. Waltz, and B. Taranto, "Learning parameters of the k-means algorithm from subjective human annotation." in *FLAIRS Conference*, 2011.

[24] H. Dutta and W. Chan, "Using community structure detection to rank annotators when ground truth is subjective,"

in *NIPS Workshop on Human Computation for Science and Computational Sustainability*, 2012, pp. 1–4.

[25] A. J. Torget, R. Mihalcea, J. Christensen, and G. McGhee, "Mapping texts: Combining text-mining and geo-visualization to unlock the research potential of historical newspapers," 2011.

[26] H. Southall, P. Aucott, and J. Westwood, "Pastplace–the global gazetteer from the people who brought you'a vision of britain through time'," in *UK Archives Discovery Forum*, 2014.

[27] T.-I. Yang, A. J. Torget, and R. Mihalcea, "Topic modeling on historical newspapers," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, 2011, pp. 96–104.

[28] D. Newman, C. Chemudugunta, P. Smyth, and M. Steyvers, "Analyzing entities and topics in news articles using statistical topic models," in *Intelligence and Security Informatics*. Springer, 2006, pp. 93–104.

[29] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large-scale news analysis," in *String Processing and Information Retrieval*. Springer, 2005, pp. 161–166.

[30] G. Crane and A. Jones, "The challenge of virginia banks: an evaluation of named entity analysis in a 19th-century newspaper collection," in *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2006, pp. 31–40.

[31] D. A. Smith, "Detecting and browsing events in unstructured text," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 73–80.

[32] ——, "Detecting events with date and place information in unstructured text," in *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2002, pp. 191–196.

[33] D. A. Smith and G. Crane, "Disambiguating geographic names in a historical digital library," in *Research and Advanced Technology for Digital Libraries*. Springer, 2001, pp. 127–136.

[34] A. Carlson, S. Gaffney, and F. Vasile, "Learning a named entity tagger from gazetteers with the partial perceptron." in *AAAI Spring Symposium: Learning by Reading and Learning to Read*, 2009, pp. 7–13.

[35] Z. Zhang and J. Iria, "A novel approach to automatic gazetteer generation using wikipedia," in *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*. Association for Computational Linguistics, 2009, pp. 1–9.

[36] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[37] K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 621–630.

[38] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proceedings of the 2008 international conference on web search and data mining*. ACM, 2008, pp. 207–218.

[39] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy." *ICWSM*, vol. 10, pp. 10–17, 2010.

[40] T. Palfray, D. Hebert, S. Nicolas, P. Tranouez, and T. Paquet, "Logical segmentation for article extraction in digitized old newspapers," *CoRR*, vol. abs/1210.0999, 2012. [Online]. Available: http://arxiv.org/pdf/1210.0999.pdf

[41] A. Gupta, "Finding influential people from a historical news repository," Master's thesis, IIIT-Delhi, 2014.

[42] I. Chattopadhyaya, K. Sirchabesan, and K. Seal, "A fast generative spell corrector based on edit distance," in *Advances in Information Retrieval*. Springer, 2013, pp. 404–410.

[43] A. Gupta and H. Dutta, "Finding influential people from a historical news repository," 2014.

[44] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. Association for Computational Linguistics, 2003, pp. 188–191.

[45] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

[46] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Machine Learning*, vol. 4, no. 4, pp. 267–373, 2011.

[47] K. J. Rodriquez, M. Bryant, T. Blanke, and M. Luszczynska, "Comparison of named entity recognition tools for raw ocr text," in *Proceedings of KONVENS*, 2012, pp. 410–414.

[48] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[49] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," *The Journal of Machine Learning Research*, vol. 10, pp. 1801–1828, 2009.

[50] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[51] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky, "Deterministic coreference resolution based on entity-centric, precision-ranked rules," *Computational Linguistics*, vol. 39, no. 4, pp. 885–916, 2013.

[52] M.-L. Zhang and Z.-H. Zhou, "Multi-instance clustering with applications to multi-instance prediction," *Applied Intelligence*, vol. 31, no. 1, pp. 47–68, 2009.

[53] D. Zhang, F. Wang, L. Si, and T. Li, "M3ic: Maximum margin multiple instance clustering." in *IJCAI*, vol. 9, 2009, pp. 1339–1344.

Person Name	IPI	Whether found on Wikipedia	Comments
capt creeten	3.380151	no	spelled incorrectly;capt creedon
capt hankey	3.022371	yes	
capt pinckney	2.933288	yes	
john macdonald	2.854389	yes	
john martin	2.827969	yes	
aaron trow	2.814171	yes	fictional character
mrs oakes	2.791536	no	false positive
buenos ayres	2.767399	no	location name
alexander iii	2.742552	yes	
mr got	2.736363	no	false positive
mrs martin	2.719383	no	false positive
ann arbor	2.681657	no	location name
caleb morton	2.63808	no	fictional character
anthony comstock	2.633381	yes	
toledo ann arbor	2.610495	no	location name
john thompson	2.609841	yes	
nat lead	2.594452	no	false positive
ed kearney	2.543152	yes	name of horse
van cortlandt	2.533131	no	location
louis philippe	2.523525	yes	
mrs talboys	2.522888	yes	fictional character
jim hooker	2.500915	yes	false positive
marie clavero	2.497384	no	false positive
father watson	2.450817	no	false positive
james mccutcheon	2.431448	no	part of an organization name
hugh allan	2.4287	yes	
william i	2.4222	yes	
marie antoinette	2.40731	yes	
schmitt berger	2.396639	no	spelled incorrectly;max f schmittberger
jacob schaefer	2.392976	yes	

TABLE 5
Table representing top 30 influential person entities detected from people gazetteer with 30 Topics LDA Model along with evaluation results and comments.

Person Name	IPI	Whether found on Wikipedia	Comments
capt creeten	3.333485	no	spelled incorrectly; capt creedon
mrs martin	3.23105	no	false positive
alexander iii	3.090361	yes	
capt hankey	2.975704	yes	
aaron trow	2.790838	yes	
john macdonald	2.774389	no	
mrs oakes	2.744869	no	false positive
john martin	2.711302	yes	
ed kearney	2.629342	yes	name of horse
caleb morton	2.614746	no	fictional character
john ward	2.57499	yes	
nat lead	2.571118	no	false positive
mrs talboys	2.499555	yes	fictional character
buenos ayres	2.490502	no	location
van cortlandt	2.490169	no	location
john thompson	2.482063	yes	
louis philippe	2.476858	yes	
marie clavero	2.474051	no	false positive
hardy fox	2.449248	no	
mme melba	2.415785	yes	
charles weisman	2.405938	no	false positive
hugh allan	2.405367	yes	
mr got	2.389697	no	false positive
schmitt berger	2.373305	no	spelled incorrectly
phil king	2.363644	yes	
henry a meyer	2.350396	yes	
north orlich	2.348236	no	false positive
james mccutcheon	2.338115	no	part of organization name
gen porter	2.330658	yes	
miller hageman	2.327831	no	

TABLE 6
Table representing top 30 influential person entities detected from people gazetteer with 100 Topics LDA Model along with evaluation results and comments.