Finding Influential People from Historical News Repository

Aayushee Gupta

aayushee1230@iiitd.ac.in

Indraprastha Institute of Information Technology

June 19, 2014

Agenda

Motivation

- People Search - an important practical application of historical newspapers to find information about people and track the timeline of news articles related to them

- Finding influential people from historic newspaper archives- a novel problem

Agenda

Problem Description

Aim: To find "influential" people from historical news OCR archives where an influential person can be defined as:
"A person whose actions and opinions strongly influence a course of events"

Problem Description

Aim: To find "influential" people from historical news OCR archives where an influential person can be defined as:
"A person whose actions and opinions strongly influence a course of events"

Divided into subproblems:

Problem Description

Aim: To find "influential" people from historical news OCR archives where an influential person can be defined as:
"A person whose actions and opinions strongly influence a course of events"

Divided into subproblems:

- Spell Correction and Cleaning of OCR text

Problem Description

Aim: To find "influential" people from historical news OCR archives where an influential person can be defined as:
"A person whose actions and opinions strongly influence a course of events"

Divided into subproblems:

- Spell Correction and Cleaning of OCR text
- Development of a People Gazetteer-an organized structure to ease the process of identification of influential people

Problem Description

Aim: To find "influential" people from historical news OCR archives where an influential person can be defined as:
"A person whose actions and opinions strongly influence a course of events"

Divided into subproblems:

- Spell Correction and Cleaning of OCR text
- Development of a People Gazetteer-an organized structure to ease the process of identification of influential people
- Influential People Identification-define the criteria to identify and rank people as "influential".

Agenda

Novel Contribution

- A novel Spell Correction Evaluation (SCE) algorithm for measuring performance of Spelling Correction
- Development of People Gazetteer - an organized dictionary of people names and a list of news articles of their occurrence along with the corresponding topic label of each article which can be used to identify and rank influential people
- Define an Influential Person Index (IPI) and metrics for its calculation to identify and rank influential people from the People Gazetteer

Agenda

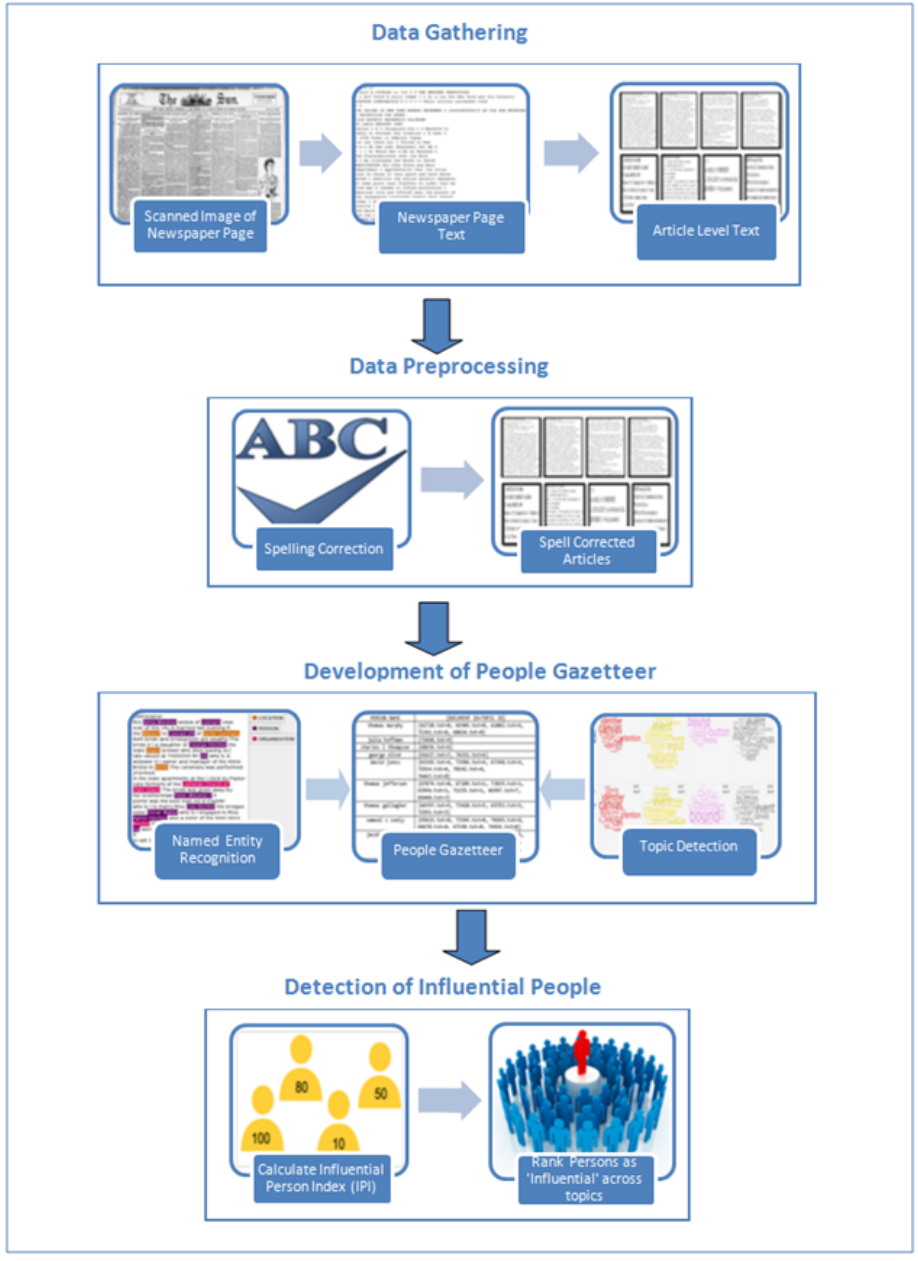
Related Work

- GATE gazetteers[1] define gazetteers as set of lists containing names of entities such as cities, organizations, days of week, etc
- Gazetteers have been used as a processing resource to find occurrence of entity names in text[2][5] (Example: Named Entity Recognition)
- Influential people identification has been done mostly in the field of social networking, marketing and diffusion research[3][1]
- Number of votes, tweets, comments, followers, etc are common parameters used for defining influence but not applicable to the newspaper environment

[1] http://gate.ac.uk/sale/tao/splitch13.html

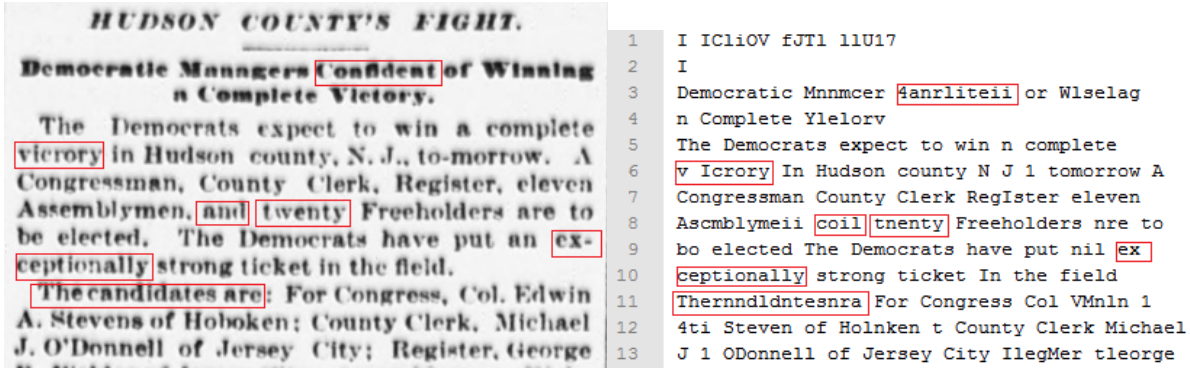
Agenda

Solution Framework



Data Gathering

- **Data Source** : Chronicling America - provides scanned OCR newspaper pages of American newspapers published between 1836 and 1922
- **Data Statistics** : 14020 news articles of "The Sun" newspaper published between November-December 1894 consisting of 8 million tokens
- **Data Characteristics** : News articles consist of one or more OCR errors of the types- Real word, Non-real word, Non-word, Word Split and Join and New line errors



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

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

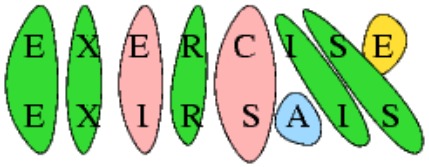
Figure : Scanned newspaper image and its corresponding OCR text

Data Preprocessing

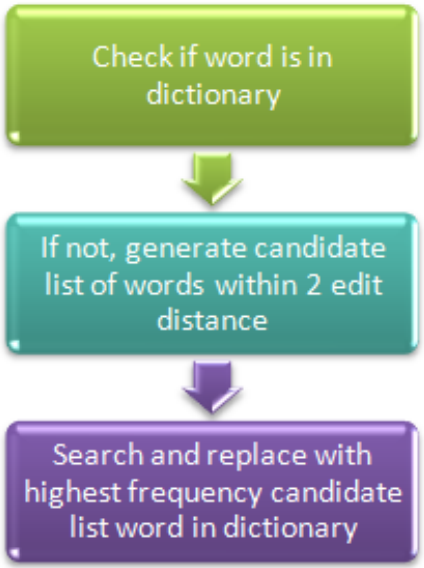
- Required to deal with OCR errors in the news articles
- Edit distance algorithm used for spelling correction of non-real and non-word OCR errors using precompiled dictionary for look-up
- Person name correction is improved by enhancing dictionary with people names by running Stanford NER-CRF parser on subsets of the ClueWeb12 dataset available as a part of TREC 2013 Crowdsourcing track

Spelling Correction Algorithm

- "Edit distance" corresponds to the minimum number of insertion, deletion and substitution required to transform one string into another



- String Edit distance algorithm for spelling correction:



Spelling Correction Evaluation I

- Required to measure the performance of spelling correction

- Evaluation Parameters:

1. Accuracy : measures the percentage of actual errors that get corrected in the OCR text after spelling correction and defined as follows:

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN}$$

where,
TP=Number of True Positives,
TN=Number of True Negatives,
FP=Number of False Positives,
FN=Number of False Negatives.

2. Time taken to run Spelling Correction Algorithm

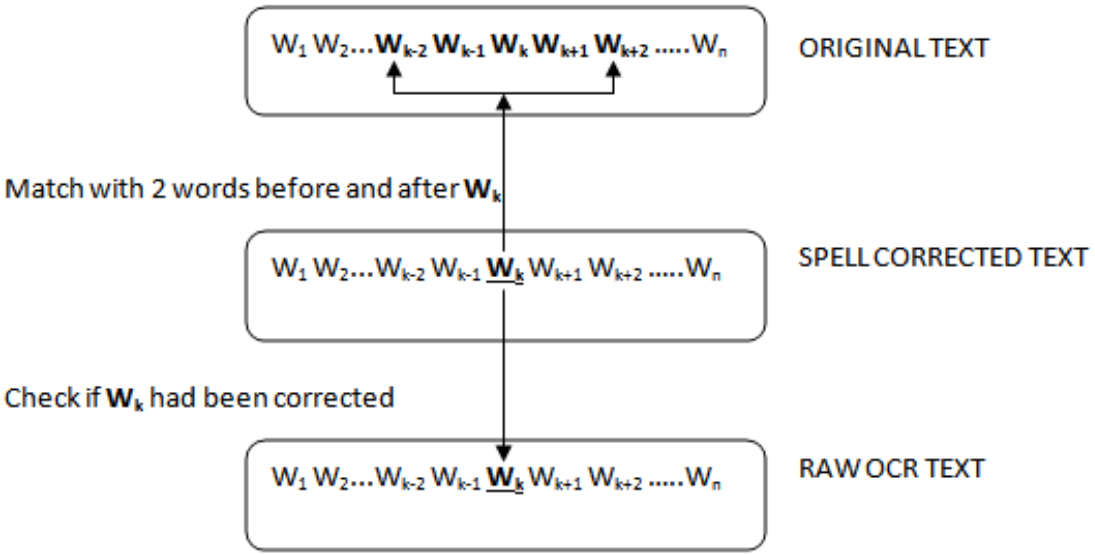
Spelling Correction Evaluation II

3. Person Names Detection Rate (PNDR) : defined as the ratio of person names recognized through Named Entity Recognition (NER) before spelling correction process and the total number of person names recognized in the original newspaper articles

$$PNDR = \frac{\text{Person Names recognized before/after spelling correction}}{\text{Person Names recognized in original newspaper articles}}$$

Spelling Correction Evaluation (SCE) Algorithm

- Word by word correspondence between corrected and original dataset not possible because of Word Split and Join errors in OCR dataset
- SCE algorithm performs word by word automatic evaluation on post spell corrected OCR dataset using an n-word grams approach



Match Found	Spell Corrected	Outcome
Y	Y	TP
Y	N	TN
N	Y	FP
N	N	FN

Figure : Schematic diagram for alignment of spell corrected article text with original article text for a word W_k

Example

Line text from 3 versions of a news article:

OcrLine= *Irnniluttry iiownllllnu at tilchmond*
CorrectedLine= *Irnniluttry iiownllllnu at Richmond*
OriginalLine= *Grand jury now sitting at Richmond*

Word in Corrected Line	**Corresponding Word Window in Original Line**	**Result**
Irnniluttry	Grand jury now	FN
iiownllllnu	Grand jury now sitting	FN
at	Grand jury now sitting at	TN
Richmond	sitting at Richmond	TP

Spelling Correction Evaluation Results

- SCE algorithm tested on 50 spell corrected articles using 3 versions of each article: Original text, Raw OCR text and Spell Corrected text

Accuracy : 72.7%
Time taken : 9 seconds on average per article

PNDR from Raw OCR : 72.5%
PNDR from spell corrected text : 63.3%
PNDR from spell corrected text with enhanced people names dictionary : 82.5%

- Results indicate less accuracy due to large number of incorrected OCR errors and usefulness of person names dictionary for spell correction and person names detection

Discussion

- Spelling Correction accuracy can be improved by correcting other OCR errors like New Line and Word Split and Join errors
- Choice of a dictionary for the edit distance algorithm affects the results of spelling correction and PNDR using enhanced person names dictionary
- SCE algorithm can be used to compare among multiple spell correction algorithms and decide which one suits the dataset better and gives best accuracy

Development of People Gazetteer

- People Gazetteer: an organized structure developed to ease the process of influential people identification
- Two step process:
  1. Person Named Entity Recognition (PNER)- Extraction of person names from the news articles dataset using Named Entity Recognition
  2. Topic Detection-Assignment of topics to news articles using LDA model

PNER

- Stanford CRF-NER used for the process of Named Entity Recognition
- Only multi-term person entities are analyzed
  Example: Henry is ignored while Henry Smith is stored
- Inverted Index is created to link person entities with the list of news articles in which they occur
- 38426 person entities recognized from 14020 news articles

Person Categories

Extracted person entities are divided into following categories for separate analysis of each category:

Person Category	**Number of news articles**	**Statistics from dataset**
Marginally Influential	less than 4	38066
Medium Influential	4 to 15	344
Highly Influential	more than 15	16

Topic Detection

- Topic : a set of words which describe what any document is about
- Topic Detection: use topic modelling algorithm for examining a set of documents and discover main topics occurring across the documents as well as the balance of topics in each document based on the statistics of the words in the complete document set
- LDA : generative probabilistic model in which documents exhibit multiple topics and each topic is a distribution over a fixed vocabulary
- AD-LDA : approximate distributed LDA model that uses distributed computation on multiple processors to infer document topics and is faster than the simple LDA approach

Topic Model Evaluation

- Evaluation required to decide parameters to be used for topic modeling
- Evaluation measure: Perplexity
  1. It indicates how surprised a trained model is when given a held out test data and calculated as follows:

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

  2. It is a decreasing function of the log likelihood of the unseen documents
  3. Lower the perplexity, better is the topic model

Topic Model Evaluation Results I

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

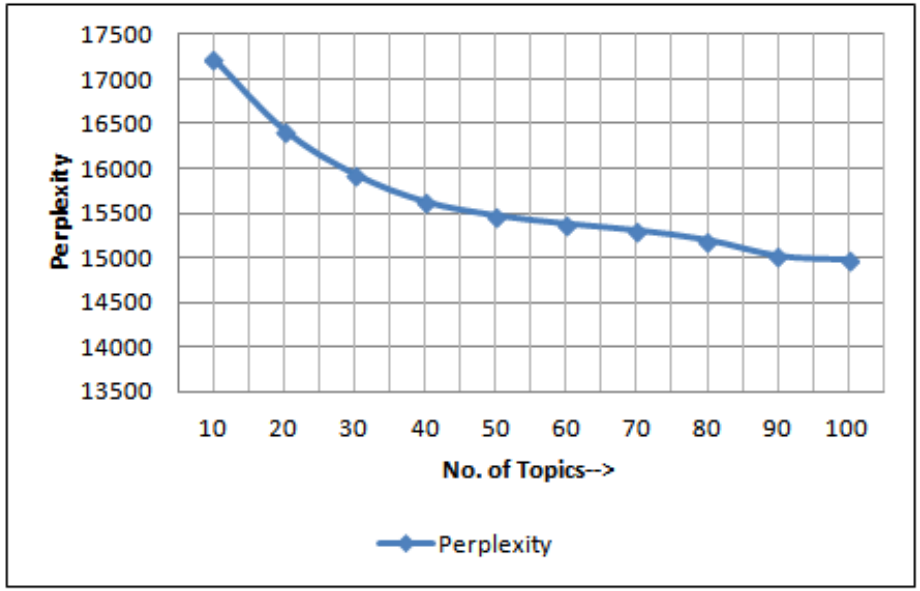


Figure : Test Set Perplexity versus Number of Topics for a random $90 - 10$ split of the data using AD-LDA. The maximum number of words in each topic is 20, number of iterations 500 and the number of processors 4 for this experiment.

Topic Model Evaluation Results II

- Variability in perplexity with respect to the number of topics was found to be much greater than the variability due to the number of processors or number of iterations
- Two models finalized from topic detection and used for developing people gazetteer:
  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

Output of People Gazetteer

- Person Name + Document List + Document Topic => 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}

Figure : Snapshot of the people gazetteer using 30 Topics LDA Model where each person entity is associated with a list consisting of a text Document ID and its corresponding Topic ID.

Discussion I

- Lack of punctuation in the OCR dataset leads to high number of false positives as well as missing recognition of several person entities during PNER

Example

"They gave money to Ronn Collector A Augustus Healy Speaker has been appointed.." leading to the recognition of person entity "Ronn Collector A Augustus Healy Speaker"

- Person Named Entity Disambiguation is required to differentiate among persons with similar names in news articles

Discussion II

- Co-reference Resolution of person names needs to be performed to link multiple ways in which a single person is addressed

Example

"William Schmittberger", "Captain William" recognized as separate person entities in an article

- The LDA topic detection model is also not geared to be used on OCR dataset directly since it recognizes some topics having completely meaningless words

Example

1. air ran ur fur ui full tt al tl late mr ant liar art lay told met ti tr
2. la lu ot lo tu au tb ta ha tea day al aa ut ar uu wa tt te

Identification of Influential People

- Define Document Index (DI) to measure effect of each document on a person's influence score
- Calculate Influential Person Index (IPI) for each person entity based on the maximum DI in their document list
- Rank person entities in the people gazetteer in decreasing order of IPI

Document Index I

- Measures how each document in the person entity's associated list of documents affects his influence score
- Parameters for estimating DI:
  1. Normalized Document Length (NDL) : normalized number of tokens contained in a news article

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

  2. Normalized Term Frequency (NTF) : normalized number of occurrences of a person's name in a news article

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

  3. Number of similar articles (NSIM) : proportion of topic similar articles for a news article in a person's list

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

Document Index II

- Formula for estimating DI:

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

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

- DI is a heuristic measure of these three parameters where each of the parameters can be weighted as per dataset characteristics and user requirements

Influential Person Index

- An index calculated for each person entity in order to measure its influence in the news dataset and calculate its influential score
- Formula for calculation of IPI:

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

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

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

Agenda

Comparison Across Influential Person Lists I

Two ranked influential person lists, namely L_1 and L_2 are obtained after calculation of IPI from each people gazetteer using 30 Topics and 100 Topics LDA Model respectively

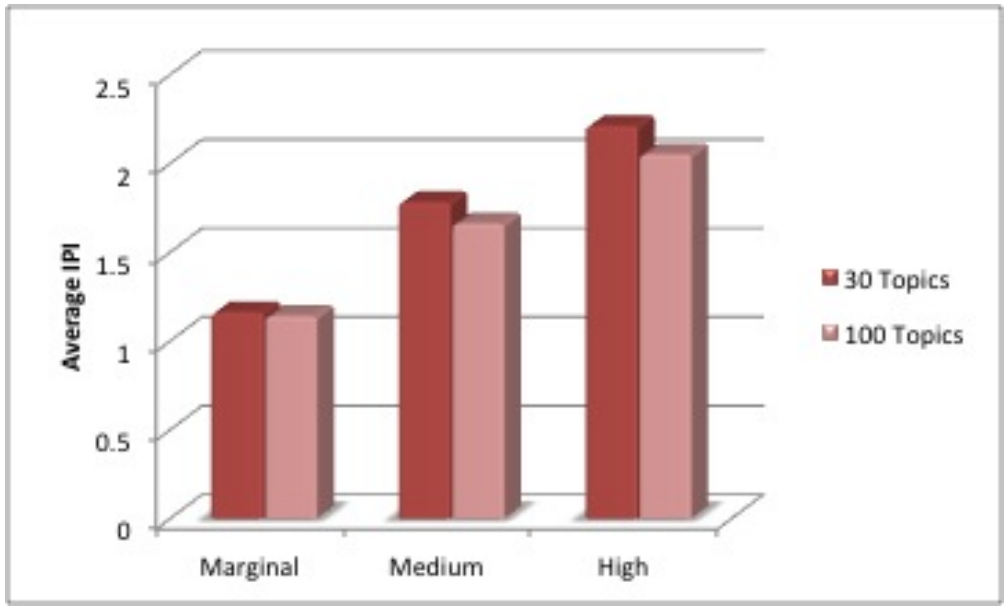


Figure : Comparison of the Average IPI for two ranked lists L_1 and L_2 using 30 and 100 topic LDA model respectively.

Comparison Across Influential Person Lists II

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

Table : Average statistics for each Person Category of People Gazetteer across 2 Topic Models

- Results indicate the fact that highly influential category people are more susceptible to change in number of topics
- Number of articles of occurrence for a person entity cannot be used uniquely to determine a person as "influential"

Person Name	IPI	Number of Articles	Person Category	NDL	NTF	NSIM	TOPIC WORDS	UniqT	Rank
capt creeten	3.38	10	Medium	0.56	1.95	0.8	mr court police judge justice case yesterday street district	0.06	1
capt hankey	3.02	6	Medium	0.68	1.6	0.66	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

Person Name	IPI	Number of Articles	Person Category	NDL	NTF	NSIM	Topic Words	UniqT	Rank
capt creeten	3.33	10	Medium	0.56	1.95	0.8	mr police witness committee capt asked captain money inspector paid	0.02	1
mrs martin	3.23	8	Medium	0.20	2.38	0.5	mrs mr years wife home house ago woman city died	0.02	2
alexander iii	3.09	31	High	0.49	2.04	0.48	emperor prince french alexander czar london nov government imperial russian	0.07	3

Table : Comparison of top 3 influential persons for two ranked lists L_1 and L_2 using 30 and 100 topic LDA model respectively.

Evaluation

- No gold standard data available regarding influential persons in the newspaper archives from November-December 1894
- Evaluation done through Wikipedia search of top 30 influential persons of ranked lists L1 and L2

Evaluation Results

16 and 14 out of the top 30 person entities were found to be influential and popular in the 19th century across topic categories like theatre, sports, government, shipping, etc. in lists L1 and L2 respectively

Influential Persons

Alexander iii
(Tsar of Russia)

Capt Hankey
(English Soldier)

John Thompson
(4th PM of Canada)

John Macdonald
(1st PM of Canada)

Jacob Schaefer
(Carom Billiards Player)

Anthony Comstock
(US Postal Inspector)

Mme Melba (Australian
operatic soprano)

Hugh Allan(Canadian
shipping magnate)

Figure : Some of the top 30 influential persons found during evaluation from Wikipedia

Discussion

- Linear combination of parameters used in calculation of DI and IPI
- Parameters for calculation of DI and IPI can also be learned by performing regression analysis
- Effect of different ways of normalizing parameters on ranking of influential people needs to be analyzed
- The topmost influential people contain several false positives due to OCR errors and PNER issues
  Example : “van cortlandt”, “ann arbor” , “sandy hook”,“mrs martins” , “mrs oakes”

Agenda

Conclusion and Future Work

- Proposed a novel approach and highlighted challenges for finding influential people from historical OCR news repository
- Non-heuristic estimation for finding influential persons possible through optimization approaches such as unsupervised multiple instance clustering[4]

Conclusion and Future Work

- Proposed a novel approach and highlighted challenges for finding influential people from historical OCR news repository
- Non-heuristic estimation for finding influential persons possible through optimization approaches such as unsupervised multiple instance clustering[4]
- Algorithm is a combination of Constrained Concave-Convex Procedure and Cutting Plane method with faster convergence

Conclusion and Future Work

- Proposed a novel approach and highlighted challenges for finding influential people from historical OCR news repository
- Non-heuristic estimation for finding influential persons possible through optimization approaches such as unsupervised multiple instance clustering[4]
- Algorithm is a combination of Constrained Concave-Convex Procedure and Cutting Plane method with faster convergence
- Cluster person entities into “influential” or “non-influential” by considering each person entity as a bag with articles of their occurrence as the instances for each bag

Thank You!

Questions??

References

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

[2] Carlson, A., Gaffney, S., and Vasile, F.
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.

[3] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K.
Measuring user influence in twitter: The million follower fallacy.
ICWSM 10 (2010), 10–17.

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

[5] Zhang, Z., and Iria, J.
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* (2009), Association for Computational Linguistics, pp. 1–9.

Algorithm 1 MatchWordGrams function of SCE Algorithm for measuring accuracy

```
function MatchWordGrams(OcrLine, CorrectedLine, OriginalLine, jstart, jend, i)
    for (int j=jstart; j<jend; j++) do
        if ((CorrectedLine[i].equals(OriginalLine[j]))&&(!(OcrLine[i].equals(CorrectedLine[i])))) then
            tp = tp + 1  flag0=false  return tp
        end
        else if ((CorrectedLine[i].equals(OriginalLine[j]))&&(OcrLine[i].equals(CorrectedLine[i]))) then
            tn = tn + 1  flag1=false  return tn
        end
    end
    if (!(OcrLine[i].equals(CorrectedLine[i]))&&flag0==true) then
        fp = fp + 1  return fp
    end
    else if ((OcrLine[i].equals(CorrectedLine[i])) && flag1==true) then
        fn = fn + 1  return fn
    end
```

TOPIC ID	TOPIC WORDS
1	total ii won club score night ran furlough alleys tournament time mile fourth rolled curling scores race national game
2	la lu ot lo tu au tb ta ha tea day al aa ut ar uu wa tt te
3	iii lie tin nail tn lit hut ill ii nn thu tu anti thin inn hit lu lo nut
4	line street feet point western easterly northerly feel southerly distance place distant lo fret hue beginning laid early felt
5	opera theatre music company week play stage evening night performance concert mme audience manager season de orchestra house miss
6	great people life man women good country world american part ot ha made la years make long place bad
7	election mr party republican state district vote democratic county senator elected city committee mayor political candidate majority york democrats
8	time ho work tn men city bo lie anti day thin long thu made part ago lot york make
9	st room av sun wife board front lo december rent lot november sunday ht west ar house private si
10	dr book st story books cloth author cure free work york blood illustrated remedy goods medical library health price
11	church dr father funeral school st college sunday year rev catholic pastor services late service held society holy clock
12	horse race class horses won racing years prize record year show ring track mile money jockey trotting trotter ran
13	cent year week pf market total net stock today central st ft lit sales short cotton ohio lot month
14	white water indian black long found thu big dog time ground wild tree killed birds bird day great lake
15	price black silk goods prices ladies worth dress fine white full tea quality style wool made fancy cloth fur

Table : Topic ID and words obtained from the 30 Topics LDA model.

TOPIC ID	TOPIC WORDS
16	street mrs mr avenue wife house miss yesterday years home woman night ago husband found died daughter children mother
17	war american government army chinese japanese china japan foreign united nov emperor states prince minister military french port navy
18	feet north minutes avenue boundary seconds degrees west york minute degree point east south feel city angle county laid
19	man ho men night back wa room left house told bad door found turned place ran lie front morning
20	water feet building boat company car train road fire miles railroad island work line city great river built bridge
21	club game team play football half ball left college back yale played harvard line eleven men match yacht field
22	ii iii ill lit ll si ti il im vi st iv ft mi li till lull lui oil
23	bank money national gold amount notes banks hank business treasury account cent paid bonds note currency company stock estate
24	mr john william york henry charles james club city ii george dec dr thomas smith jr brooklyn van held
25	piano st rooms car york daily chicago city sunday upright parlor furnished broadway hotel av west train brooklyn monthly
26	york daily steamship nov directed letter dec fur orleans al steamer walls letters close australia china japan city london
27	mr court police judge justice case yesterday street district witness jury charge asked attorney trial arrested lawyer told office
28	mr law present made public year state committee president secretary bill report states con tin united number meeting york
29	air ran ur fur ui full tt al tl late mr ant liar art lay told met ti tr
30	company york trust bonds city cent railroad mortgage interest wall bond stock street st central january coupon committee jan

Table : Topic ID and words obtained from the 30 Topics LDA model.

Person Name	IPI	Number of Articles	Person Category	NDL	NTF	NSIM	TOPIC WORDS	UniqT	Rank
capt creeten	3.38	10	Medium	0.56	1.95	0.8	mr court police judge justice case yesterday street district	0.06	1
capt hankey	3.02	6	Medium	0.68	1.6	0.66	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 macdonald	2.85	3	Marginal	0.55	2.2	0	great people life man women good country world american part	0.1	4
john martin	2.82	12	Medium	0.56	1.6	0.5	mr court police judge justice case yesterday street district witness	0.17	5
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
buenos ayres	2.76	6	Medium	0.43	1.6	0.67	white water indian black long found thu big dog time	0.06	8
alexander iii	2.74	31	High	0.24	2.04	0.29	great people life man women good country world american part	0.16	9
mr got	2.73	3	Marginal	0.56	1.47	0.67	mr court police judge justice case yesterday street district witness	0.03	10

Table : 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.

Person Name	IPI	Number of Articles	Person Category	NDL	NTF	NSIM	Topic Words	UniqT	Rank
capt creeten	3.33	10	Medium	0.56	1.95	0.8	mr police witness committee capt asked captain money inspector paid	0.02	1
mrs martin	3.23	8	Medium	0.20	2.38	0.5	mrs mr years wife home house ago woman city died	0.02	2
alexander iii	3.09	31	High	0.49	2.04	0.48	emperor prince french alexander czar london nov government imperial russian	0.07	3
capt hankey	2.97	6	Medium	0.68	1.6	0.66	game team football play half line ball back yale eleven	0.02	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 macdonald	2.77	3	Marginal	0.55	2.2	0	people american man great country men world life good english	0.02	6
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
john martin	2.71	12	Medium	0.56	1.6	0.5	mr police witness committee capt asked captain money inspector paid	0.05	8
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 : 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.