Finding Influential People from Historical News Repository

Aayushee Gupta

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[1][2] (Example: Named Entity Recognition)

- Influential people identification has been done mostly in the field of social networking, marketing and diffusion research[3][4]

- 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

Scanned Image of Newspaper Page → Newspaper Page Text → Article Level Text

Data Preprocessing

Spelling Correction → Spell Corrected Articles

Development of People Gazetteer

Named Entity Recognition → People Gazetteer → Topic Detection

Detection of Influential People

Calculate Influential Person Index (IPI) → Rank Persons as 'Influential' across topics

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

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
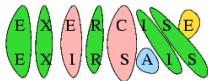
- String Edit distance algorithm for spelling correction:
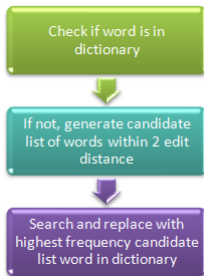
Spelling Correction Evaluation I

- 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

ORIGINAL TEXT: $W_1 W_2 \ldots W_{k-2} W_{k-1} W_k W_{k+1} W_{k+2} \ldots W_n$

Match with 2 words before and after W_k

SPELL CORRECTED TEXT: $W_1 W_2 \ldots W_{k-2} W_{k-1} \underline{W_k} W_{k+1} W_{k+2} \ldots W_n$

Check if W_k had been corrected

RAW OCR TEXT: $W_1 W_2 \ldots W_{k-2} W_{k-1} \underline{W_k} W_{k+1} W_{k+2} \ldots W_n$
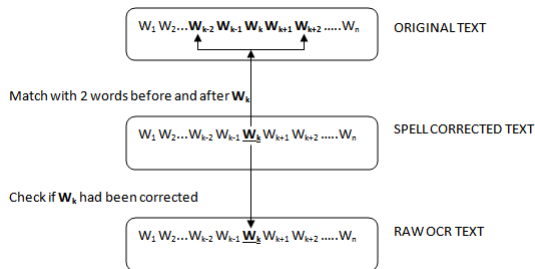
Match Found	Spell Corrected	Outcome
Y	Y	TP
Y	N	FN
N	Y	FP
N	N	TN

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

---

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

Example

Line text from 3 versions of a news article:

OcrLine= *Irnniluttry iiownlllllnu at tilchmond*
CorrectedLine= *Irnniluttry iiownlllllnu 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
iiownlllllnu	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 for spell correction = 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

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

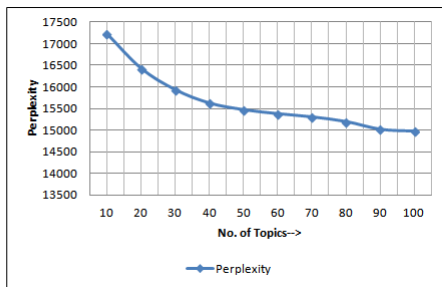| Number of Topics	Perplexity
10	17227
20	16431
30	15940
40	15634
50	15480
60	15390
70	15316
80	15208
90	15031
100	14988
200	14355

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.

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

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

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

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

  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:

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

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
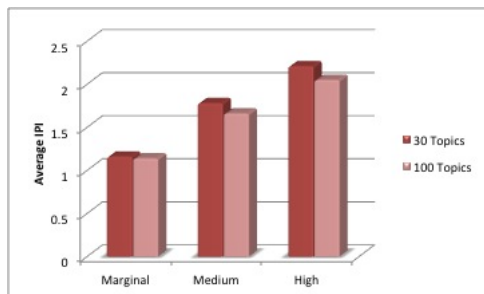

Figure : Comparison of the Average IPI for two ranked lists L_1 and L_2 using 30 and 100 topic LDA model respectively.

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 also 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
- 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, politics, government, shipping, etc. in lists L1 and L2 respectively

Agenda

Discussion and Future Work

- Spelling Correction algorithm needs to be improved in order to correct other OCR errors like New Line and Word Split and Join errors

- Lack of punctuation in the OCR dataset leads to high number of false positives during the process of PNER

- Person Named Entity Disambiguation : a hard problem to solve since persons with similar names can occur in multiple topic related articles in newspapers

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

- Heuristic-based parameter estimation of IPI can be replaced by an optimization approach such as multiple instance clustering to avoid choosing of parameter weights and biasing of results

📕 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.

📕 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.

References II

📕 CHA, M., HADDADI, H., BENEVENUTO, F., AND GUMMADI, P. K.
Measuring user influence in twitter: The million follower fallacy.
ICWSM 10 (2010), 10–17.

📄 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.