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

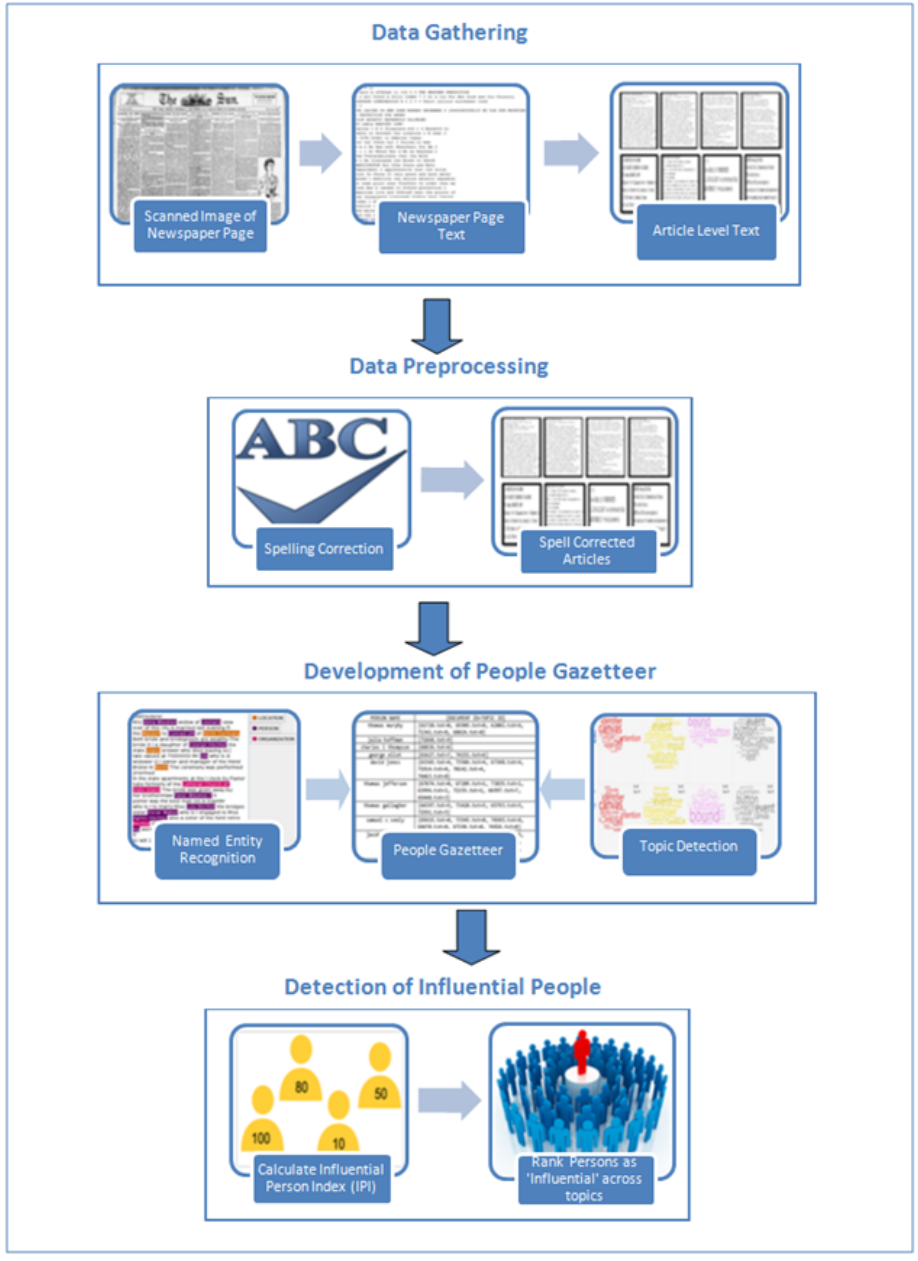
Agenda

Related Work

- GATE gazetteers 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 (Example: Named Entity Recognition)
- Influential people identification has been done mostly in the field of social networking, marketing and diffusion research
- Number of votes, tweets, comments, followers, etc are common parameters used for defining influence but no applicable to the newspaper environment

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

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

  Based on levenshtein distance and can correct errors due to substitution, insertion and deletion of at most 2 letters in a word

Spelling Correction Algorithm

- "Edit distance" corresponds to the minimum number of insertion, deletion and substitution required to transform one string into another
- Precompiled dictionary is used to search for candidate list of words within edit distance 2 from the word to be corrected
- Word correction is done by replacement with the highest frequency word and lowest edit distance among the candidate list of words
- 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 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 automatic evaluation of word by word post spelling correction on OCR dataset
- Algorithm uses an n-gram words approach by considering a window of k words before and after each word in the original text for each word in the corrected text
- Each word in the corrected text is marked as a True Positive, True Negative, False Positive, False Negative based on whether the spelling had been corrected and if a match is found in the k words window of the original text

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 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

Development of People Gazetteer

Two step process:

- Person Named Entity Recognition (PNER)- Extraction of person names from the news articles dataset using Named Entity Recognition
- 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

People Gazetteer Sample Output

PERSON ENTITY NAME	DOCUMENTLIST {DOCUMENTID→DOCUMENTTOPIC}
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}

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 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

Calculation of IPI

- 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

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
- 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 to avoid analyzing unnecessary person names
- 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