Advanced Data Mining
CS 522
Final Project Report

Text Mining Algorithms for Systematic Reading of Massive Text Records measuring the Cultural Impact of Historic Chicago High-Rise

AUTHOR	**ARNAB MUKHOPADHYAY**
CWID	**A20353463**
DATE	**05/7/2016**

Table of Contents

1. INTRODUCTION

Studying big topics in cultural and architectural history is time expensive due to the high about of text information that needs to be processed by the scholars. In this project, we explore the usability of a computer to systematically read the text and offer objective insights into the text. Various data mining techniques such as network analysis, bag of words, power iteration are tested on datasets from journals and books.

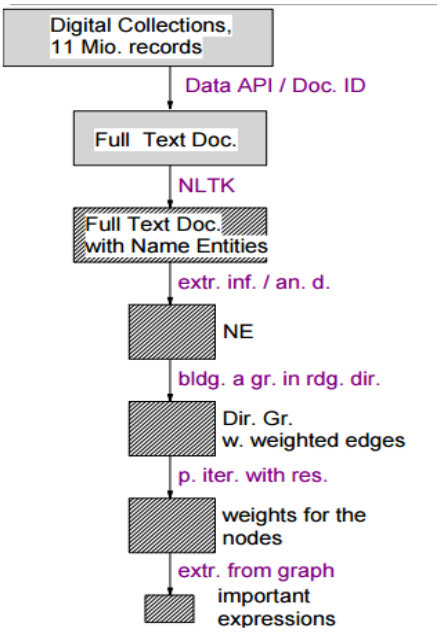


Fig. Overall flow of our project

2. DATA

2.1 Data collection

Our source of data is JStore.We have selected around 150 documents in pdf format. All the documents are related to architecture.

2.2 Data Pre-processing

We have manually converted the pdf files to .txt type in Year_DocId.txt format. Then use it as the input of the Name Entity Recognition(NER).

2.2.1 Name Entity Recognition

- **What is NER (Name Entity Recognition)?**

  Named entities are "atomic elements in text" belonging to "predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc." Named entity recognition (NER) is the task of identifying such named entities.

- **Process Flow:**

  The goal of a named entity recognition (NER) system is to identify all textual mentions of the named entities. We are using the following Named Entity.

  - PERSON, FACILITY, ORGANIZATION, GSE, GSP, LOCATION

  The below figure explains the architecture for a simple information extraction system.

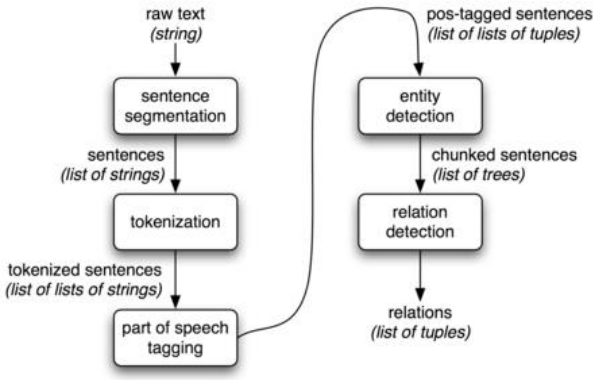


Fig. Information Extraction System with NER

- **Implementation of NER in our project:**

NER is the first data pre-processing step in our project. The output of NER is the main source of data for our experiments. But NER contains many irrelevant data which is not important for us. So as a remedy we have created a list of data (cross data) to filter the output of NER and to use it as the source data in our code. Below is the sample result from our NER code:

Source

The Correlation of Literature with Architecture Author(s): Francis Shoemaker Source: College Art Journal, Vol. 9, No. 2 (Winter, 1949-1950), pp. 181-186 Published by: College Art Association Stable URL: http://www.jstor.org/stable/772993 Accessed: 01-05-2016 23:14 UTC Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://about.jstor.org/terms JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org. College Art Association, Taylor & Francis, Ltd. are collaborating with JSTOR to digitize, preserve and extend access to College Art Journal This content downloaded from 216.47.136.20 on Sun, 01 May 2016 23:14:25 UTC All use subject to http://about.jstor.org/terms THE CORRELATION OF LITERATURE WITH ARCHITECTURE' By Francis Shoemaker AMONG writers and critics of architecture, I am not alone in the fact that I am not a professional architect and can make no claim for special knowledge in this field. But for the past six months I have lived in a house which Mrs. Shoemaker and I helped to design and build. Some interesting things have happened in that time. Two five-year-old girls came to the door un-invited and said, "This is a pretty house; may we come in ?" A Columbia University professor said, "This house implies a whole new way of looking at life."

(lp0 (lp1 (VPERSON p2 VFrancis Shoemaker p3 tp4 a(VORGANIZATION p5 VCollege Art Journal p6 tp7 aa(lp8 (VORGANIZATION p9 VCollege Art Association p10 tp11 aa(lp12 (VORGANIZATION p13 VCollege Art Association , Taylor & Francis , Ltd. p14 tp15 a(VORGANIZATION p16 VCollege Art Journal p17 tp18 a(VPERSON p19 VFrancis Shoemaker p20 tp21 a(VPERSON p22 VShoemaker p23 tp24 a(VORGANIZATION p25 VColumbia University p26 tp27 aa.

2.3 Cross Data

Cross Data is used as a remedy of the irrelevant text output from NER. The issue of irrelevant text output occurred because of the uncleaned source text inputs. In NER we have tried to clean it as much as possible but it was not possible to clean it completely due to the high volume of the text input. Hence we created two lists of required entities to compare it with the NER output and filter it.

Two lists that we have used are – Architect names and Organisations names. We have created it manually. Hence there is always a scope of improvement in filtering as we can add more names gradually with time.

- **Main Source Data:**

  It is the main source data of our experiments (Bag of Words, N-Gram, Network Graph, Power iteration, etc).

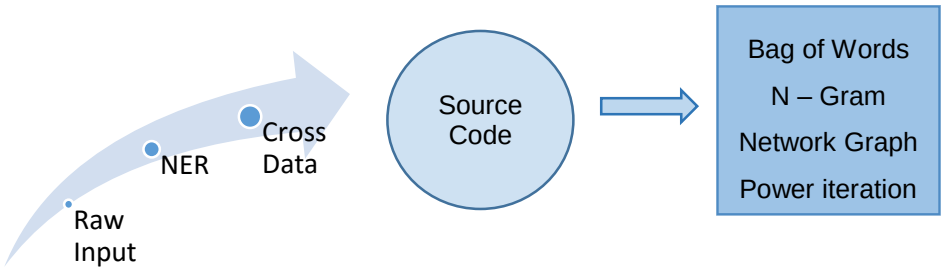


Fig. Source data to our code

3. BAG OF WORDS

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

Here we get the list of the words with frequencies which is useful to find out the most frequent or important terms in the documents.

In our project we have used this approach to compare with the result of power iteration of the entities. We have plotted a graph on year vs count of edges where one set of data is the filtered entities after power iteration and other set of data is the bag of words without filtration (Discussed later)

5. NETWORK GRAPH

A graph in this context is made up of vertices, nodes, or points which are connected by edges or lines. To convert the sentences into a graph/network, we map each named entity to a vertex, and each sentence containing more than one entity to an edge. A node is also weighted by the co-occurrence frequency, which is then changed to mutual information for clustering.

The vertices in this graph are person names, facility, organization, location, gpe, gsp. For each sentence containing two names, we add an edge between two vertices. If such a node already exists, we increment the edge weight.

- **Implementation in our project:**

  We have taken the output of the cross data as the input of to the graphs. As per our design we have created one graph per document. Below is the graph of one such document.

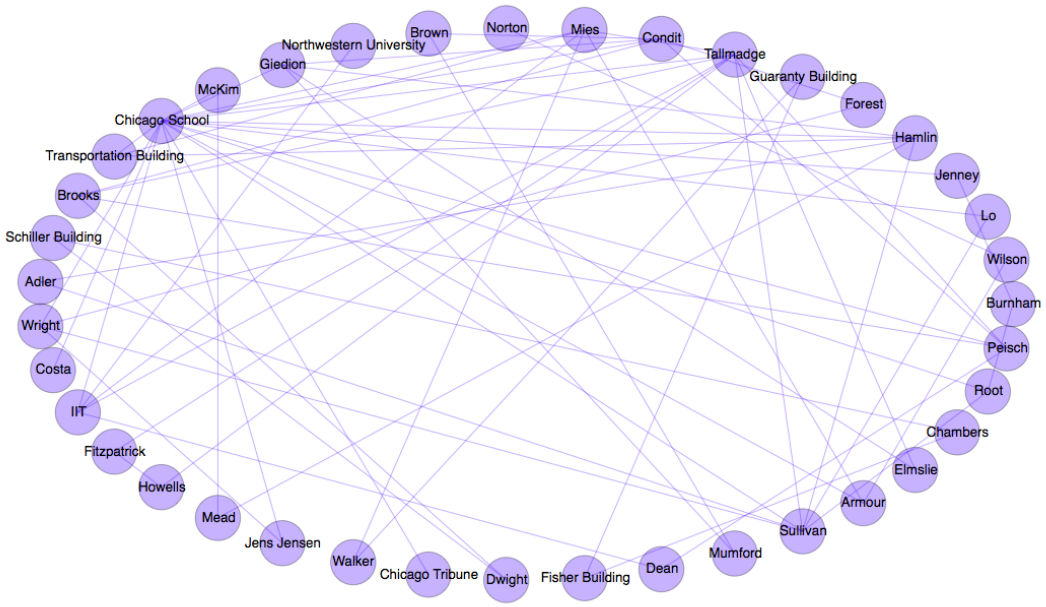


Fig. Network graph of a document

The above graph is without any weights. We have used the weighted graph (edges weight) for the power iteration to find out the most important entities per document.

- **Sample result of graph of a document:**

  [(u'University of California Press', u'Society of Architectural'), (u'Society of Architectural', u'University of California Press'), (u'University of California Press', u'CHICAGO'), (u'CHICAGO', u'Chicago Area'), (u'Chicago Area', u'Carl W. Condit'), (u'Carl W. Condit', u'Condit'), (u'Condit', u'AMERICA'), (u'AMERICA', u'Jens Jensen'), (u'Jens Jensen', u'Leonard K. Eaton'), (u'Leonard K. Eaton', u'Jens Jensen'), (u'Jens Jensen', u'Jensen'), (u'Jensen', u'Leonard Eaton'), (u'Leonard Eaton', u'Jensen'), (u'Jensen', u'Derek Hill'), (u'Derek Hill', u'Oleg Grabar'), (u'Oleg Grabar', u'Turkey'), (u'Turkey', u'Persia'), (u'Persia', u'Afghanistan'), (u'Afghanistan', u'Asia'), (u'Asia', u'Derek Hill'), (u'Derek Hill', u'Oleg Grabar'), (u'Oleg Grabar', u'University of Michigan'), (u'University of Michigan', u'Harry S. Ransom'), (u'Harry S. Ransom', u'John Lyon Reid'), (u'John Lyon Reid', u"O' Neil Ford"), (u"O' Neil Ford", u'Victor Gruen'), (u'Victor Gruen', u'I. M. Pei'), (u'I. M. Pei', u'Vernon De Mars'), (u'Vernon De Mars', u'Pietro Belluschi'), (u'Pietro Belluschi', u'Charles M. Goodman'), (u'Charles M. Goodman', u'CHICAGO'), (u'CHICAGO', u'Chicago'), (u'Chicago', u'London'), (u'London', u'LOUIS SULLIVAN'), (u'LOUIS SULLIVAN', u'Maurice English Chicago'), (u'Maurice English Chicago',

u'Nietzsche'), (u'Nietzsche', u'Spencer'), (u'Spencer', u'William James'), (u'William James', u'Sullivan'), (u'Sullivan', u'Sullivan'), (u'Sullivan', u'Jeremiah'), (u'Jeremiah', u'Sullivan'), (u'Sullivan', u'CARL W. CONDIT'), (u'CARL W. CONDIT', u'University Place'), (u'University Place', u'Evanston'), (u'Evanston', u'Illinois'), (u'Illinois', u'Avery Memorial Architectural Library'), (u'Avery Memorial Architectural Library', u'Columbia University AVERY'), (u'Columbia University AVERY', u'Mariners Museum'), (u'Mariners Museum', u'Newport News'), (u'Newport News', u'Virginia'), (u'Virginia', u'Roux'), (u'Roux', u'Bard'), (u'Bard', u'Jacobsen'), (u'Jacobsen', u'KUNSTHISTORISCHEN INSTI'), (u'KUNSTHISTORISCHEN INSTI', u'Lincoln Street'), (u'Lincoln Street', u'Boston'), (u'Boston', u'Massachusetts')]

Here each tuple contains two entities of a document which are connected with each other with an edge.

6. POWER ITERATION

It is a simple iterative scheme used over a graph to find the most important nodes. Here we have used this to find out the most important entities per document. We have filtered the result using it.

- Top 30% entities per document has been taken after power iteration.
- As an output we get a vector of entities with the frequencies.
- It is the most important step to find out the entities which will be considered for creating important expression for our main analysis.
- **Sample result of power iteration:**

```
[ 0.0025     0.03372599  0.01486464  0.01371787  0.01470602  0.01371787
  0.01573456  0.01371787  0.03504788  0.01326055  0.01374591  0.01270534
  0.01395792  0.01009222  0.01493314  0.01403488  0.01314152  0.01201255
  0.01317441  0.0103261   0.01613238  0.01493314  0.01425079  0.01371787
  0.00941397  0.01265888  0.01298959  0.01493314  0.02281135  0.01377237
  0.01791832  0.01374591  0.0160381   0.03610186  0.01395792  0.05107487
  0.0243253   0.02181518  0.01403487  0.02348996  0.01486464  0.01374591
  0.01270534  0.01107705  0.02215604  0.01123239  0.01201255  0.01143064
  0.01430041  0.01425079  0.01508428  0.01508428  0.01374591  0.01470602
  0.04964778  0.02669348  0.00941397  0.01314152  0.01430041  0.01317441]
```

This is a list of frequencies of entities of a document(1950_772993.txt) after power iterations.

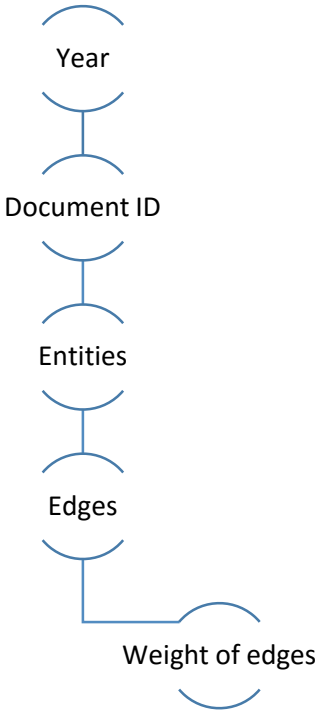
Below are the top 30% entities of the above list displayed in (entity, value) pair:

{u'Plymouth': 0.015084275975823402, u'Jimmy Herf': 0.017918318510055542, u'Wisconsin': 0.02181517519056797, u'Great Lakes Basin Region': 0.022156041115522385, u'Chicago': 0.03372599184513092, u'Dante': 0.01603809744119644, u'College Art Journal': 0.03610185533761978, u'Winwold': 0.016132377088069916, u'College Art Association': 0.05107486620545387, u'National Council of Teachers': 0.049647778272628784, u'A. Richards': 0.02669348008930683, u'Hotel Statler': 0.015734555199742317, u'Harcourt Brace': 0.024325301870703697, u'New York': 0.015084280632436275, u'Buffalo': 0.02348995767533779, u'Manhattan': 0.03504788130521774, u'Williams': 0.014933143742382526, u'Columbia University': 0.022811345756053925}

7. STRUCTURING THE DATA OUTPUT

In this part we have created a structure with dictionary to show the overall structure of output in a hierarchical format. By traversing it we can easily relate which entity is connected with which entity of a document and which all documents are created in a particular year.

The hierarchical format that we have used is given below:




```
{
'1979':
   {
    'doc1ID':
       {
           'Entity1': [('C',20), ('B',10)],
           'Entity2': [('D',20), ('B',10)],
        }
       },
 '1908':
    {
     'doc3ID':
        {
            'Entity1': [('C',20), ('B',10)],
             'Entity2': [('D',20), ('B',10)],
         },
        'doc2ID':
          {
              'Entity1': [('C',20), ('B',10)],
               'Entity2': [('D',20), ('B',10)],
           }
        }
 }
```

Fig: Hierarchical structured output

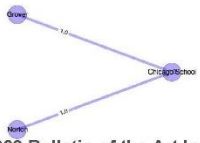
- **Sample Result:**

{'1964': {'988187': {u'Nietzsche': {u'Spencer': 1.0, u'Maurice English Chicago': 1.0}, u'University Place': {u'CARL W. CONDIT': 1.0, u'Evanston': 1.0}, u'Oleg Grabar': {u'Derek Hill': 2.0, u'Turkey': 1.0, u'University of Michigan': 1.0}, u'Jensen': {u'Derek Hill': 1.0, u'Leonard Eaton': 1.0, u'Jens Jensen': 1.0}, u'Avery Memorial Architectural Library': {u'Columbia University AVERY': 1.0, u'Illinois': 1.0}, u'Sullivan': {u'Sullivan': 1.0, u'CARL W. CONDIT': 1.0, u'Jeremiah': 1.0, u'William James': 1.0}, u'Boston': {u'Lincoln Street': 1.0, u'Massachusetts': 1.0}, u'Chicago Area': {u'Carl W. Condit': 1.0, u'CHICAGO': 1.0}, u'Columbia University AVERY': {u'Mariners Museum': 1.0, u'Avery Memorial Architectural Library': 1.0}, u'Leonard Eaton': {u'Jensen': 1.0}, u'KUNSTHISTORISCHEN INSTI': {u'Lincoln Street': 1.0, u'Jacobsen': 1.0}, u'Persia': {u'Turkey': 1.0, u'Afghanistan': 1.0}, u'CARL W. CONDIT': {u'Sullivan': 1.0, u'University Place': 1.0}, u'Illinois': {u'Evanston': 1.0, u'Avery Memorial Architectural Library': 1.0}, u'Jeremiah': {u'Sullivan': 1.0}}}

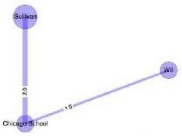
- ✓ From the above result we can understand that document id:988187 is present in year 1964 and 'University Place' entity is related to 'CARL W. CONDIT' and 'Evanston' which has equal weight of 1 which implies both entities are equally important to 'University Place' entity.
- ✓ If one of the edges has more weight, it implies more importance of that edge node to the parent entity.
- ✓ This structure creates a nice hierarchy to understand the importance of the terms in each document over the timeline.

8. PLOTTING IMPORTANT EXPRESSIONS

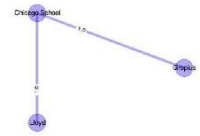
Using the above structured data, we have plotted graph with weighted entities and edges. These are the top most 30% entities which has been considered to form the important expressions.



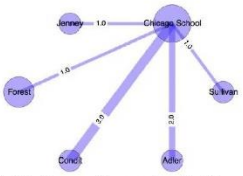
1909 Bulletin of the Art Insitute of Chicago, Art School matters.



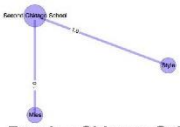
1920, Art Institute, Chicago Room exhibit (Sullivan as Chicago School)



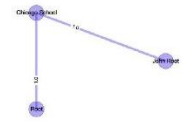
1959 Buchard, Gropius at Harvard ACSA



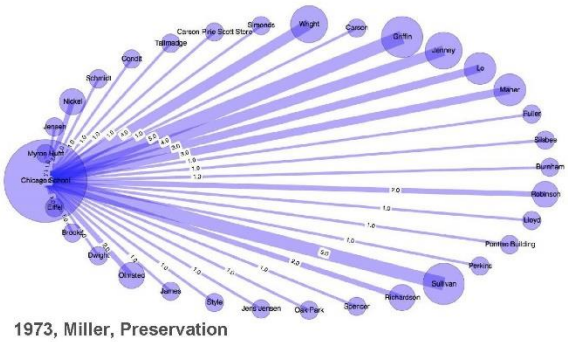
1970, Hoffman, Essay by Sullivan



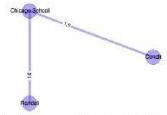
1966, Brooks, Chicago School



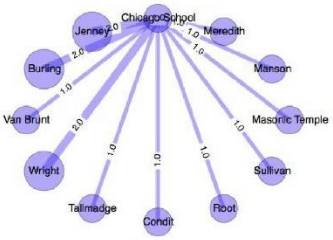
1972 Hoffmann John Wellborn Root (Back Matter book announcement)



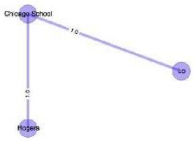
1973, Miller, Preservation



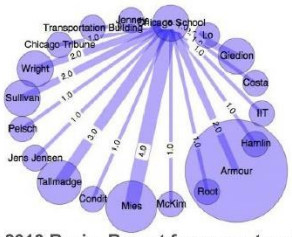
1982 MIT Press, Back Matter. (Bibliographic Notes)



1985, Clausen, Wright and the Chicago School quest for light



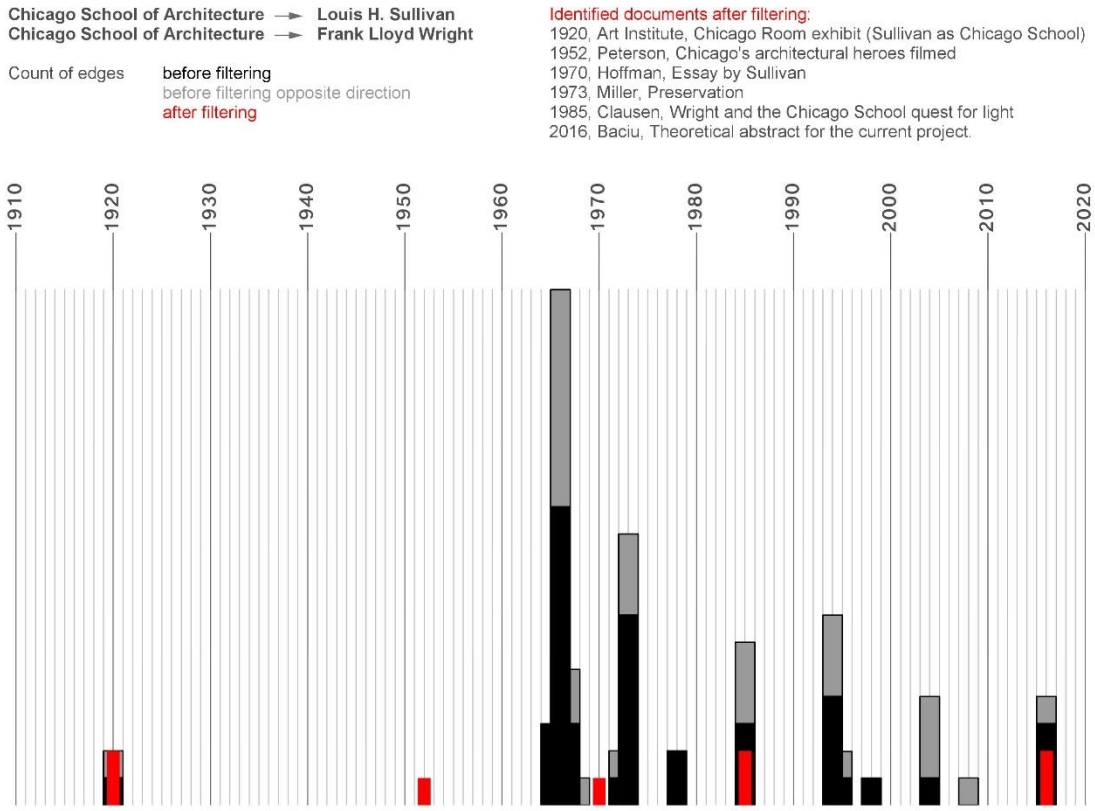
2008 Marie Frank, Pure Design



2016 Baciu, Report for current project

Here the bigger nodes indicate higher frequency and thicker edges indicate higher weight.

9. COMPARISIONS OF RESULTS FROM NER AND POWER ITERATION



- **Analysis:**

1. Here we compared the result of power iterated data with the out of bag of words model. The output of the power iteration is filtered and the output of bag of words is not filtered. Red bar denotes the filtered data and black bar denotes the non-filtered data.
2. Above we can see both the filtered data and not filtered entities has same weight in 1920.It implies the book was present in that time for both the approach.
3. Between 1960 and 1970 the non-filtered entities have highest weight but filtered entities are absent which denotes only the bag of word approach indicates the presence of books at that time.
4. In 1952 and 1970 only filtered entity is present which indicates the power iteration approach identified documents after filtering.

10. OVERALL ANALYSIS

1. NER runs good in cleaned data but performs poor in uncleaned data. In our experiment firstly we started experiments on uncleaned data but it generates lots of irrelevant text from NER. Next we cleaned the data which gave better results from NER.
2. To get better result and analysis we created cross data with manual list of entities which is used to filter the output of NER.After this filtration we got better result.
3. We have seen the filtered data of power iteration gives better results than the data of bag of words approach. We have found some interesting results for power iteration filtered data over the non-filtered data. It found new records which were not found by the bag of words approach. (Section 9)

11. CONCLUSION

Presently we have run all the experiments on limited dataset. We have found some interesting approach to relate the entities between different documents. In future we have plan to apply these approaches to bigger data set by which we can extract important information from a library of architectural journals and books.

12. REFERENCE

Baciu, Dan Costa & Nadine Kahnt. “Tour Eiffel, Ein Rückblick verknüpft mit einer Big Data-Analyse,” Phoenix 3, June 2015, p.56-61.

Cheng, Xiao and Dan Roth. “Relational Inference for Wikification.” EMNLP (2013).

Ratinov, Lev, Dan Roth, D. Downey and M. Anderson. “Local and Global Algorithms for Disambiguation to Wikipedia.” ACL (2011).

Bonta, Juan Pablo. American Architects and Texts A Computer-Aided Analysis of Literature; Electronic Companion to American Architects and Texts. Cambridge: MIT Press, 1996.

Kuter Williamson, Roxane. American Architects and the Mechanics of Fame. Austin: University of Texas Press, 1991.

Baciu, Dan Costa. “Sigfried Giedion. Historiography and History of Reception on a Global Stage” Proceedings of ar(t)chitecture, the International Conference at the Technion Israel, Faculty of Architecture. Haifa, forthcoming 2016.

Link of our code: https://bitbucket.org/onlinesocialnetworkanalysis/project/src