Framework for the Analysis of Big Historic Text Collection

Project Lead :
Dan Costa Baciu

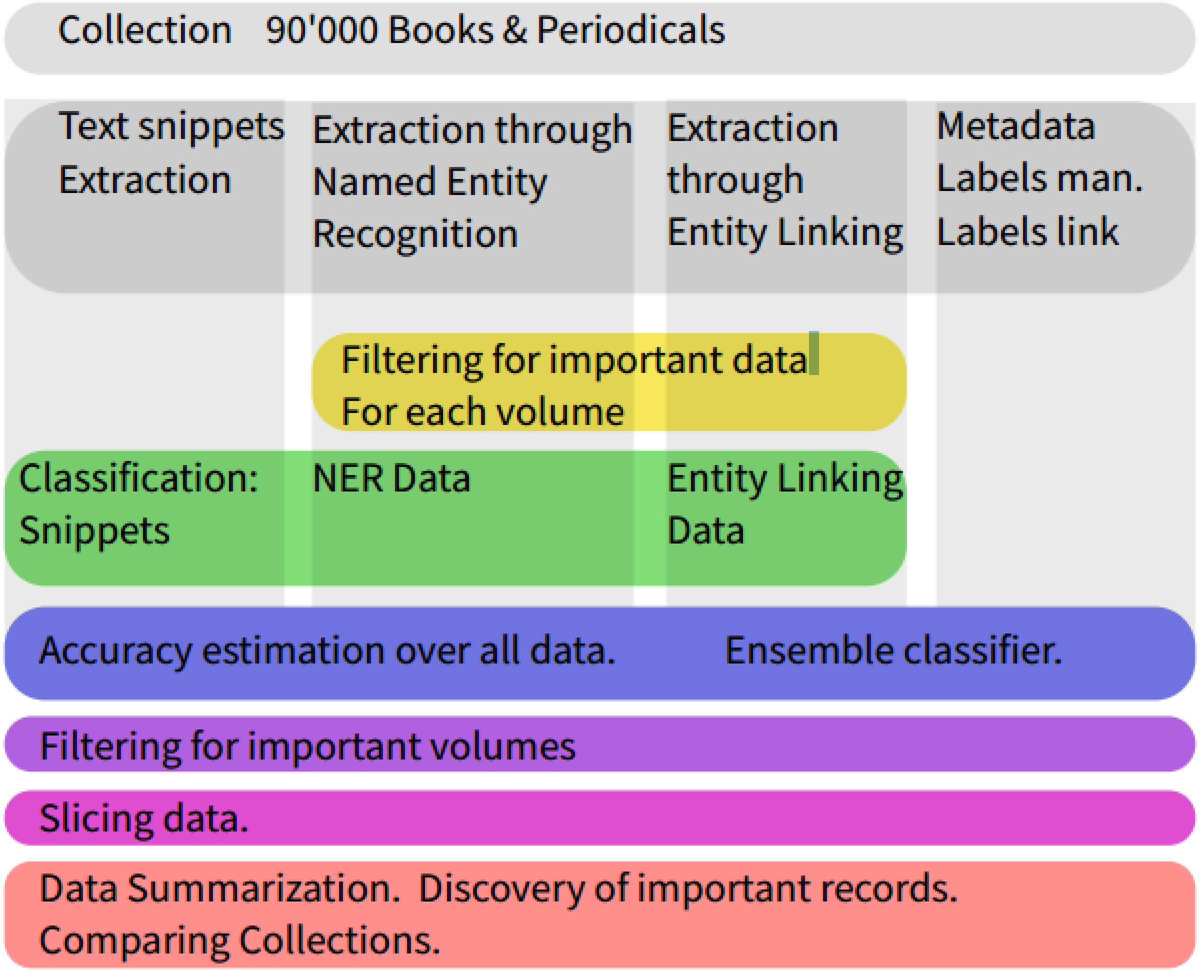
Team members:
Chandra Kumar Basavaraju
Harsha Keragodu Shivappa
Khushboo Mrugender Shah
Mayuri Kadam
Rima Sukhadia
Shruthi Naik

Introduction and Goal of the Project

- “Slice and Dice the Data”.
- Discovery of important records.
- Comparing collections.
- Data summarization .

Necessary steps:

- Structuring the text data.
- Running algorithms.
- Summary and Discovery.



Collection 90'000 Books & Periodicals

Text snippets Extraction

Extraction through Named Entity Recognition

Extraction through Entity Linking

Metadata Labels man. Labels link

Filtering for important data For each volume

Classification: Snippets

NER Data

Entity Linking Data

Accuracy estimation over all data. Ensemble classifier.

Filtering for important volumes

Slicing data.

Data Summarization. Discovery of important records. Comparing Collections.

Data and Approaches:

- Data: For a collection of books in architectural and cultural history
  1. Structured data extracted through NER
  2. Structured data extracted through entity linking
  3. Extracted text snippets
  4. Metadata

- Approaches :
  1. Data summarization
  2. Filtering - Using results from entity linking, for books and collections
  3. Classification - Using results from Entity Linking, Text snippets and NER
  4. Inverted Indices

Experiments and Results:

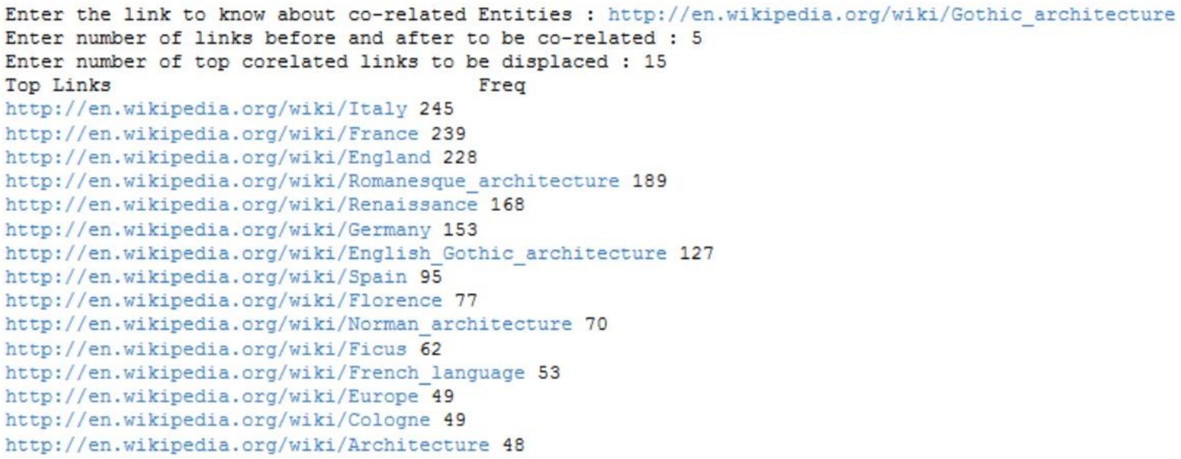
Snippet Classification Results (bigram model)		
Method	Parameters	Accuracy
Logistic Regression	(default parameters)	0.84688
Linear Support Vector Machine	C=0.1	0.84619
Neural Network	Logistic, (5,10), alpha=0.01, early_stopping=True, max_iter=100, solver="lbfgs", shuffle=True	0.86623

Number of features : 60
Number of instances : 1450

Results for Classification based on Entity Links(Bag Of Words)		
Method	Parameters	Accuracy
Random Forest	Class_weight="balanced", random_state=35	0.82849

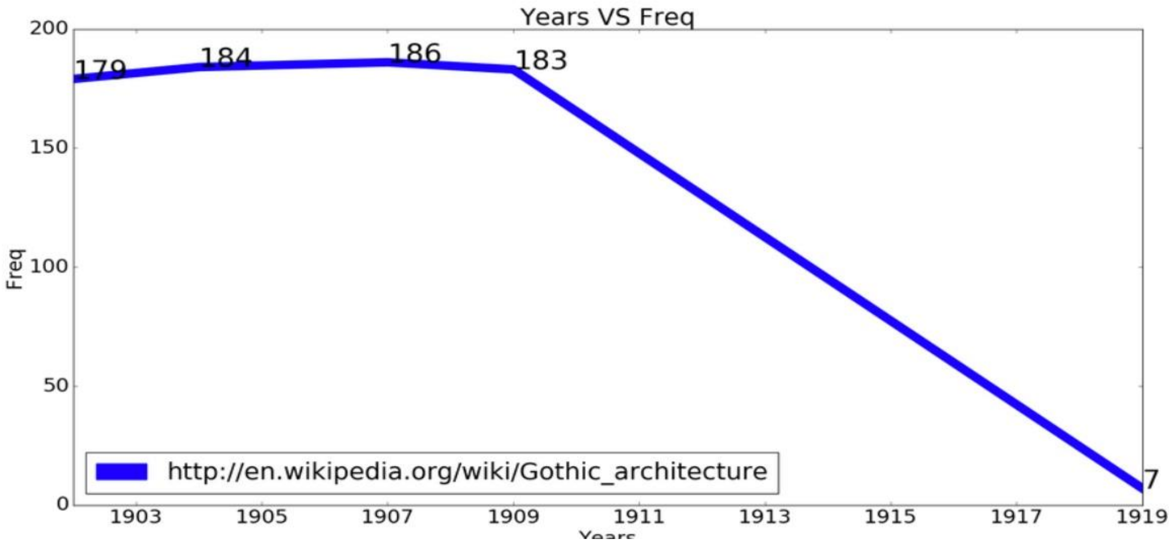
Number of features : 202624
Number of instances : 58651

2.



```
Enter the link to know about co-related Entities : http://en.wikipedia.org/wiki/Gothic_architecture
Enter number of links before and after to be co-related : 5
Enter number of top corelated links to be displaced : 15
Top Links                                       Freq
http://en.wikipedia.org/wiki/Italy 245
http://en.wikipedia.org/wiki/France 239
http://en.wikipedia.org/wiki/England 228
http://en.wikipedia.org/wiki/Romanesque_architecture 189
http://en.wikipedia.org/wiki/Renaissance 168
http://en.wikipedia.org/wiki/Germany 153
http://en.wikipedia.org/wiki/English_Gothic_architecture 127
http://en.wikipedia.org/wiki/Spain 95
http://en.wikipedia.org/wiki/Florence 77
http://en.wikipedia.org/wiki/Norman_architecture 70
http://en.wikipedia.org/wiki/Ficus 62
http://en.wikipedia.org/wiki/French_language 53
http://en.wikipedia.org/wiki/Europe 49
http://en.wikipedia.org/wiki/Cologne 49
http://en.wikipedia.org/wiki/Architecture 48
```

3.



Analysis and Conclusion:

- Mining for important topics within a book.
- Mining for important books on a particular topic.
- Important historic sources found.
- Analyzing the commonalities and differences in different collections.

CHICAGO SCHOOLS:

AUTHORS, AUDIENCES, & HISTORY

NOV 17-18

2ND INTERNATIONAL GRADUATE STUDENT SYMPOSIUM

PHD PROGRAM IN ARCHITECTURE

IMAGE: GTA ARCHIVES / ETH ZURICH, (ESTATE SIGFRIED GIEDION)

CALL FOR PAPERS

ARCH.IIT.EDU/STUDY/PHD

SYMPOSIA@YAHOO.COM

IIT ARCHITECTURE
CHICAGO
S. R. CROWN HALL

College of Architecture
ILLINOIS INSTITUTE OF TECHNOLOGY