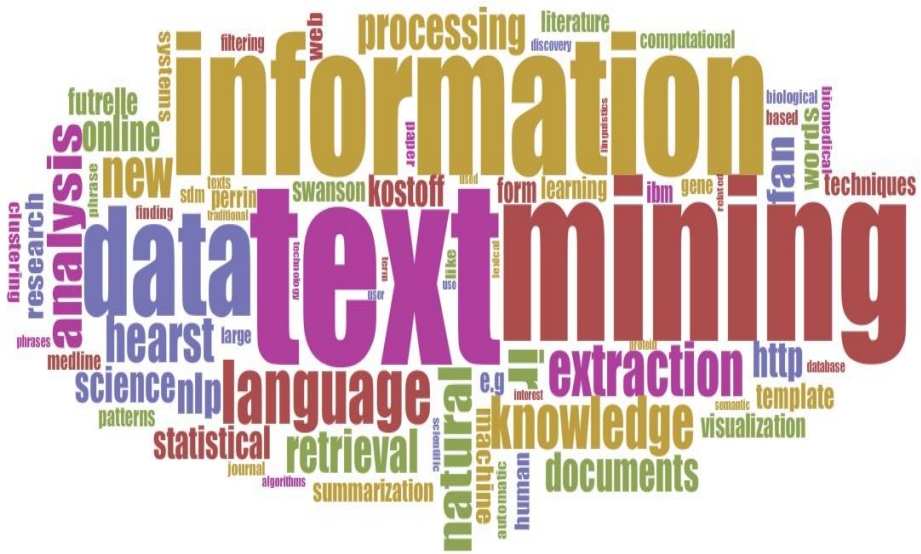
Citation Text Parsing and Text Mining



Martand Shardul
PG DBDA
CDAC

Text Mining in the area of research

"Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation"

— *Hearst, 2003*

- We are living in the era of information explosion
- Traditional online information repositories in the area of research are flooded with natural language based documents or un-structured data
- Scanning through thousands of growing volumes of research publication within stipulated time is beyond human capacity

Text Mining = Statistical NLP (structured data) + Data mining (pattern discovery)

— *Khandelwal, 2009*

Practice: **Text mining on citation data**

Un-structured data is transformed into data fields so that individual components within data can be labelled

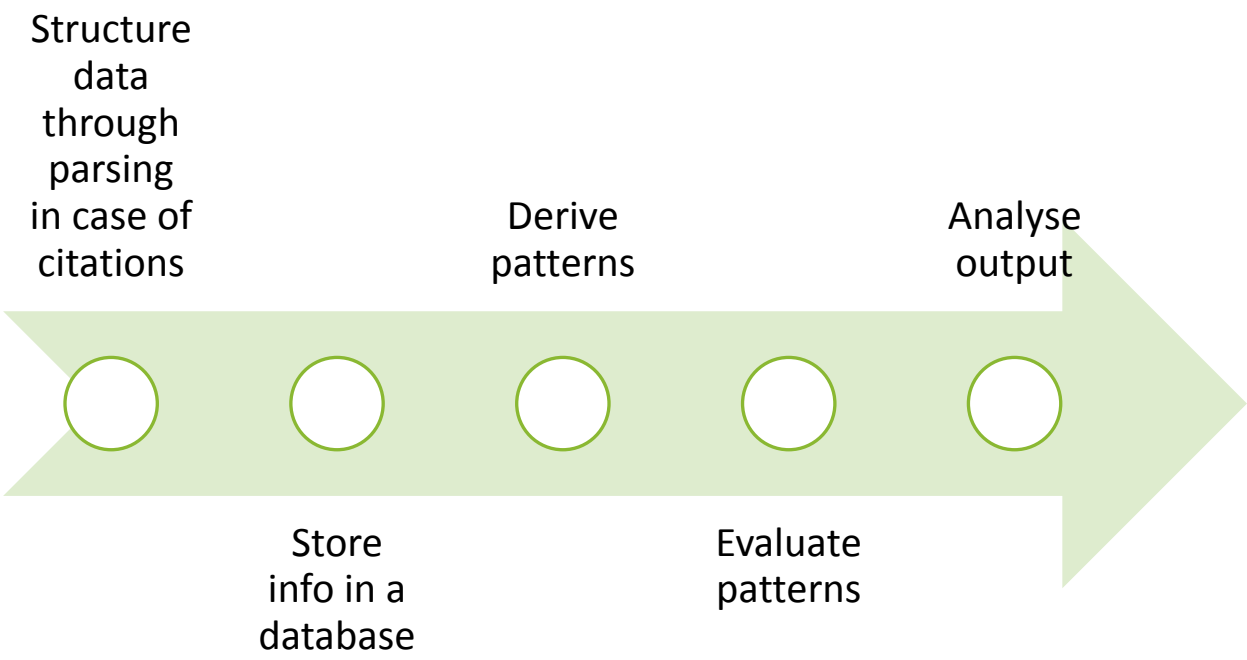
Processed information is stored into a database

Subsequently, patterns are derived from stored data

Results are evaluated and

Domain experts interpret findings

Note: Text mining is different form text analytics. The later might involve machine learning techniques. Unlike, text analytics, results of text mining might not directly feed into business intelligence applications



Project work: Approach

Domain: Climate Change

Sourcing data

1. Citations pertaining to publications of an eminent climate scientist, Dr V Ramanathan were sourced from Google Scholar and TERI Library
2. R package : “scholar” used by project team
3. Also used the R package “pdftools” which could only be used to read citations from PDF file
4. Both “scholar” and “PDFtools” had certain limitations
5. Java Apache PDFBox 2.0.0 API was used to import data from PDF file into text file format

Project work: Approach

Domain: Climate Change

Structuring data

1. “Title” , “Publication”, and “Author names” were individually analysed using Rstudio IDE and R language
2. Packages such as “tm”, “RefManageR”, “pdftools” were used
3. The use of RefManageR package required installation of an additional software “Poppler” which was not available for the R version available with the team
4. Common citation styles within the data were identified as Harvard style, Chicago style , APA style etc.
5. R based parser was developed by the project team to structure the citation dataset into structured data (Author name, Title, Citations, Year etc.)
6. Additionally, freely available software, “Publish or Perish 5” was used, where required

Project work: Approach

Domain: Climate Change

Preparing data

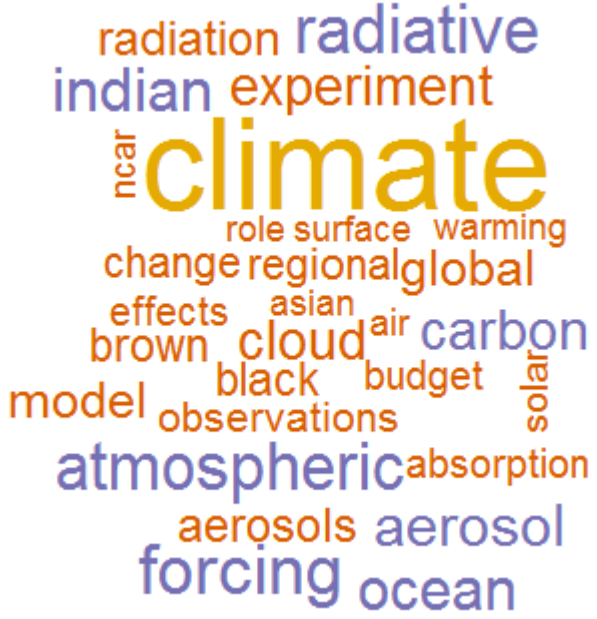
1. Data was imported into Rstudio
2. Further, whitespaces were stripped (where required), numbers and punctuations were also removed
3. R package “snowballc” was used for stemming data
4. Author names were imported into MySQL and all records that had author names like “ramanathan” were retrieved
5. Relevant records were identified

Project work: Approach

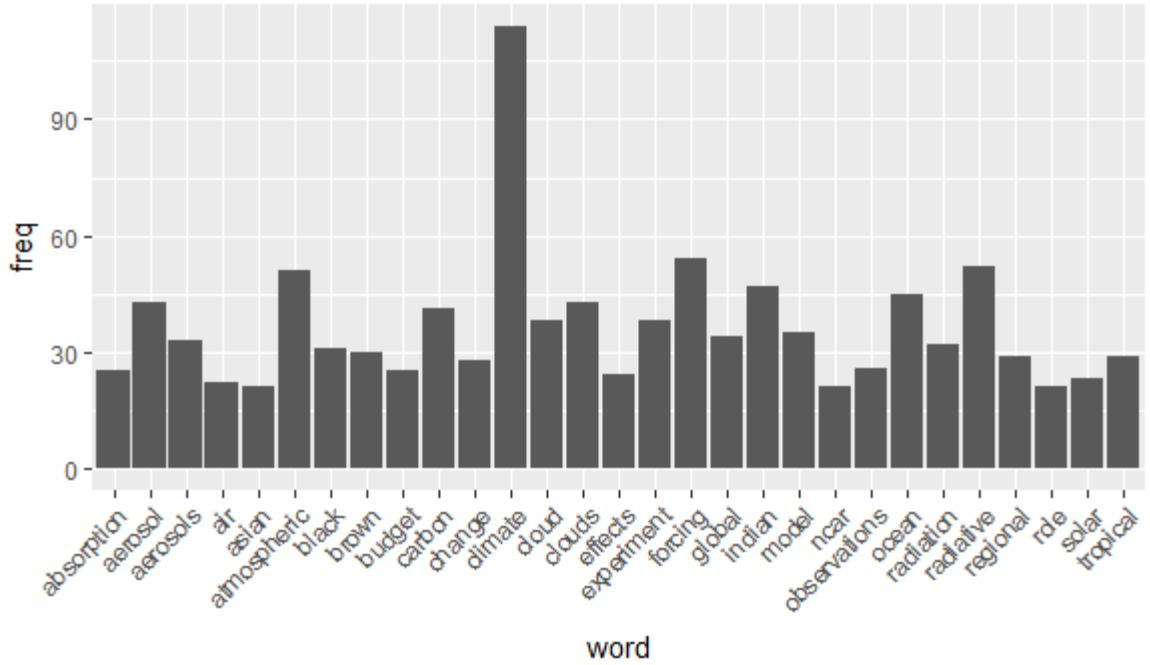
Domain: Climate Change

Exploring data: Text in Title ,Journal Name and Author names

Word cloud

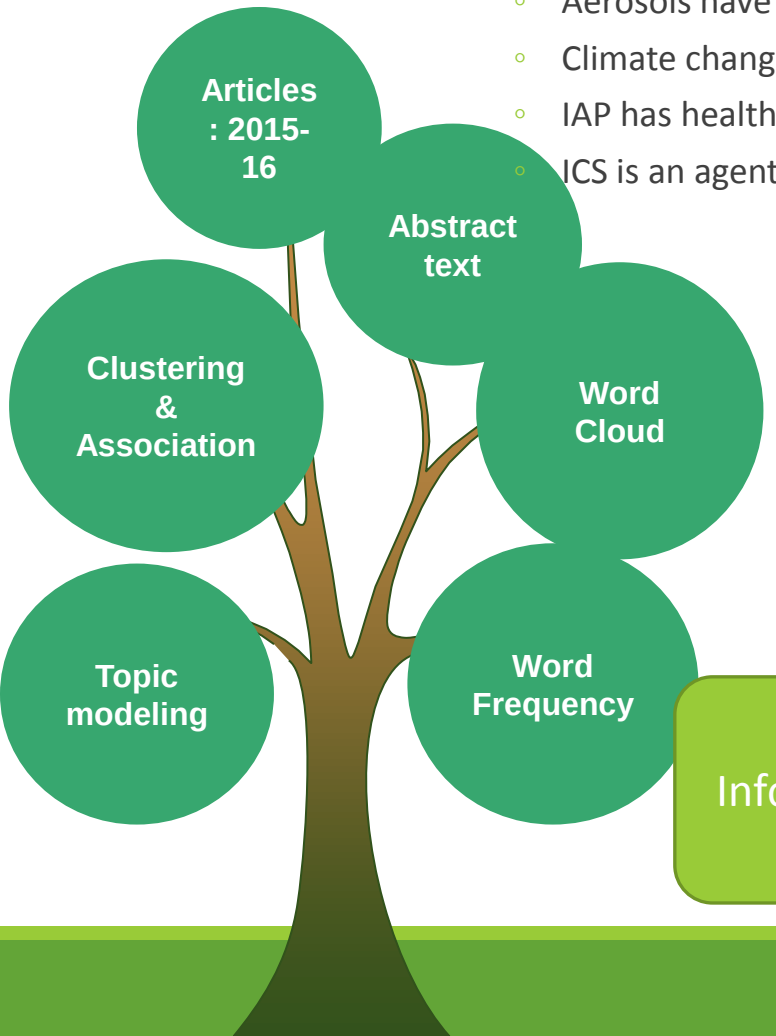


Word frequency



Most recent articles– What is being discussed and debated ?

- Aerosols have significant regional impact
- Climate change is a threat in South-Asia
- IAP has health impacts
- ICS is an agent for addressing aerosols from cookstove




```
> findAssocs(dtm, "climate", corlimit=0.5)
$climate
   global   general      high 
     1.00      0.98      0.98 
  include  reducing significant 
     0.98      0.98      0.98 
     time measurements       new 
     0.98      0.96      0.96 
  address aerosolcloud    affect 
     0.95      0.95      0.95 
 captured   carries challenges 
     0.95      0.95      0.95 
circulation component  directly 
     0.95      0.95      0.95 
  earth's    effect estimates 
     0.95      0.95      0.95 
    exist      fact  fidelity 
     0.95      0.95      0.95 
 findings finescale       gcm 
     0.95      0.95      0.95 
     gcms  hampered improving 
     0.95      0.95      0.95 
increasingly interaction    larger 
     0.95      0.95      0.95 
  limited measuring      need 
     0.95      0.95      0.95
```

Informing policy

Implementing energy access strategies

Design, development, commercialization and promotion of clean energy technology

Capacity building for energy access

Future work–
Text Analytics for classification of articles

Use of Machine Learning Techniques

- Identify articles pertaining to climate change

- Classify articles based on context

- Classify articles based major sectors that contribute to climate change

Thank You

Email: martand.cdac@gmail.com