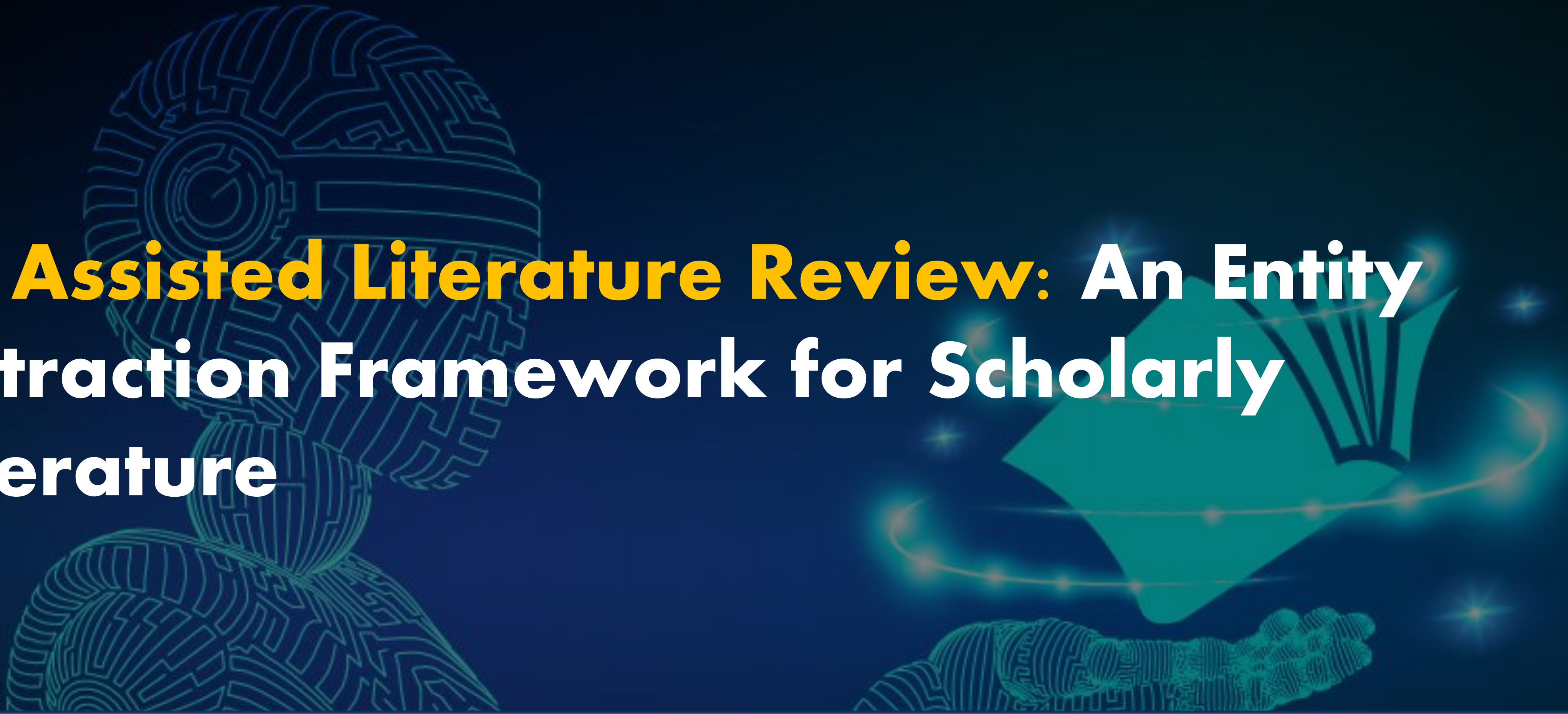
#semanticClimate
Transforming information into actionable knowledge

AI Assisted Literature Review: An Entity Extraction Framework for Scholarly Literature

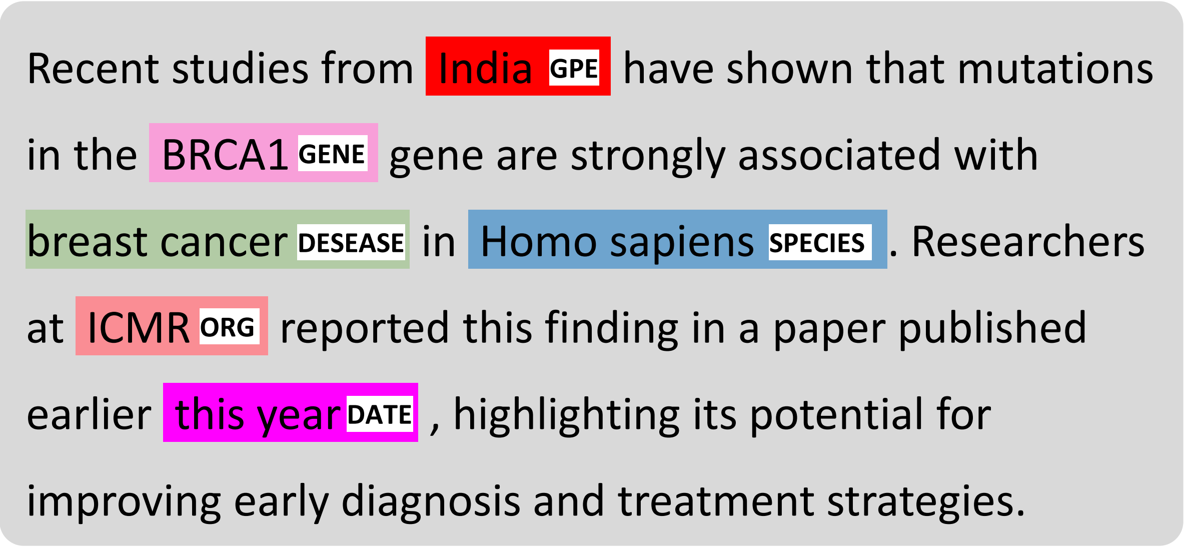


Presented by:
Moobashara Jawed

Named Entity Recognition (NER)

What is Entity Recognition?

Automated identification of key terms (entities) in text into categories like diseases, species or locations.



Recent studies from India GPE have shown that mutations in the BRCA1 GENE gene are strongly associated with breast cancer DESEASE in Homo sapiens SPECIES . Researchers at ICMR ORG reported this finding in a paper published earlier this year DATE , highlighting its potential for improving early diagnosis and treatment strategies.

■	GPE	■	GENE
■	DATE	■	DESEASE
■	ORG	■	Species

NER Workflow

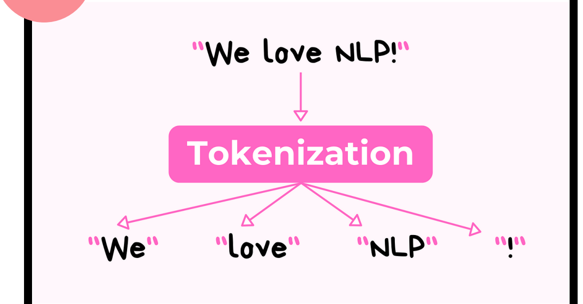
1

Automated article acquisition via **Pygetpapers**

2

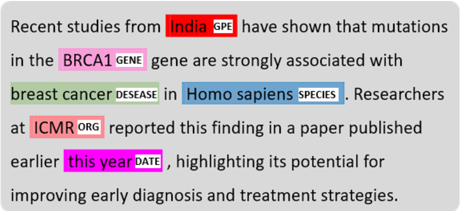
Document text extraction via PDF parsers

3



3. Text tokenization at sentence granularity

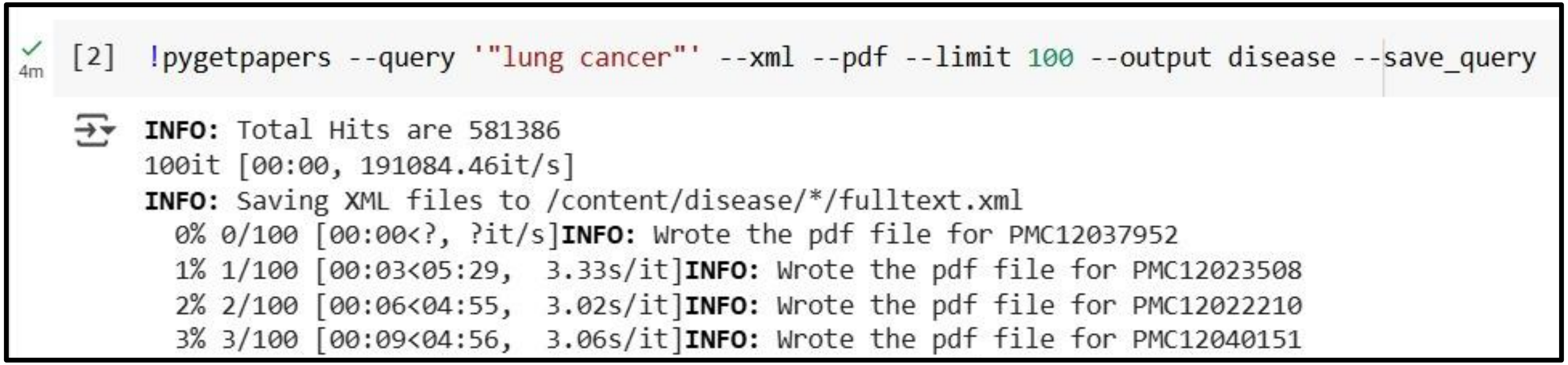
4



Recent studies from India GPE have shown that mutations in the BRCA1 GENE gene are strongly associated with breast cancer DESEASE in Homo sapiens SPECIES . Researchers at ICMR ORG reported this finding in a paper published earlier this year DATE , highlighting its potential for improving early diagnosis and treatment strategies.

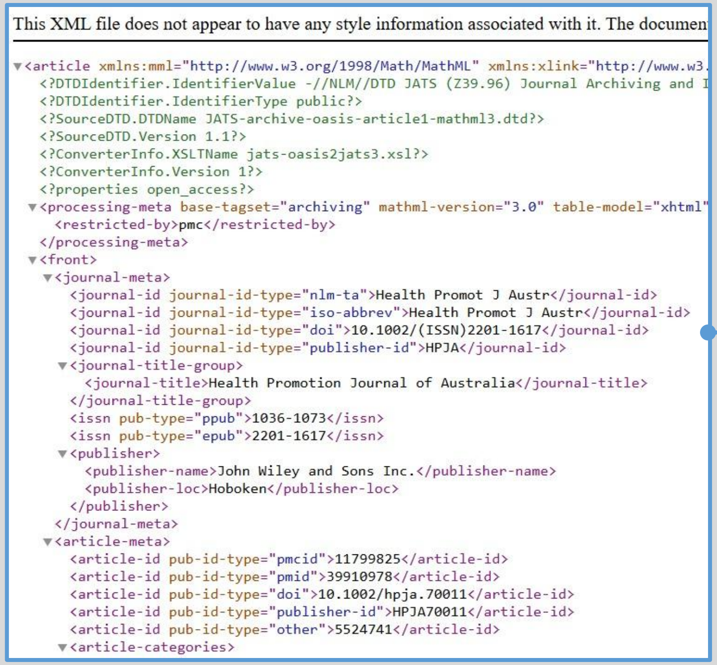
4. Named Entity Recognition (NER)

Step 1: Automated article acquisition via Pygetpapers`



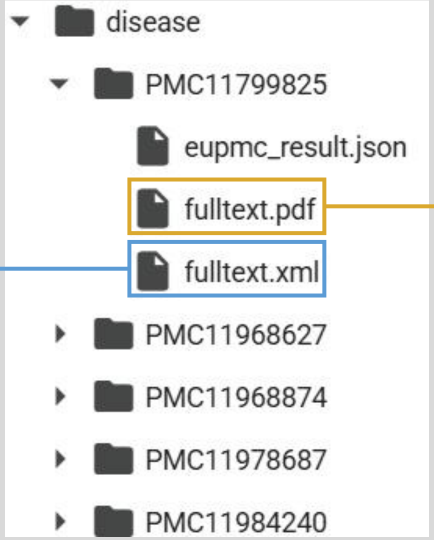
```
✓   [2]  !pygetpapers --query '"lung cancer"' --xml --pdf --limit 100 --output disease --save_query
4m

    INFO: Total Hits are 581386
    100it [00:00, 191084.46it/s]
    INFO: Saving XML files to /content/disease/*/fulltext.xml
      0% 0/100 [00:00<?, ?it/s]INFO: Wrote the pdf file for PMC12037952
      1% 1/100 [00:03<05:29,  3.33s/it]INFO: Wrote the pdf file for PMC12023508
      2% 2/100 [00:06<04:55,  3.02s/it]INFO: Wrote the pdf file for PMC12022210
      3% 3/100 [00:09<04:56,  3.06s/it]INFO: Wrote the pdf file for PMC12040151
```

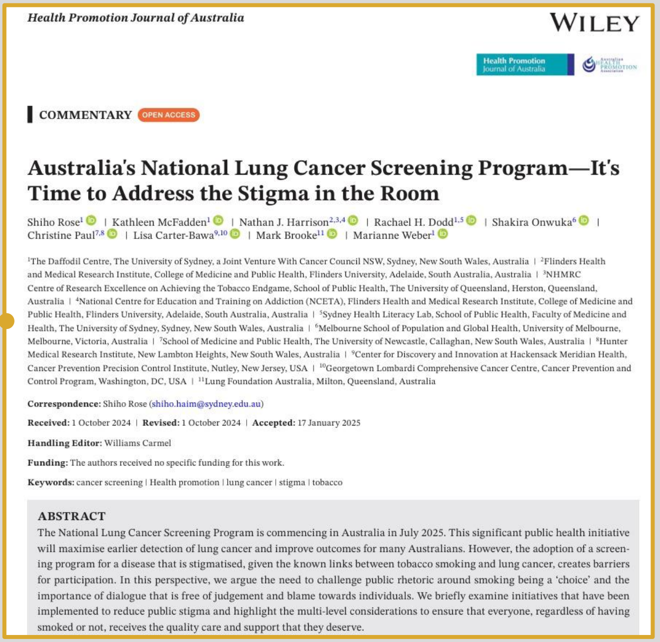


This XML file does not appear to have any style information associated with it. The documen

```
▼<article xmlns:mml="http://www.w3.org/1998/Math/MathML" xmlns:xlink="http://www.w3.
   <?DTDIdentifier.IdentifierValue -//NLM//DTD JATS (Z39.96) Journal Archiving and I
   <?DTDIdentifier.IdentifierType public?>
   <?SourceDTD.DTDName JATS-archive-oasis-article1-mathml3.dtd?>
   <?SourceDTD.Version 1.1?>
   <?ConverterInfo.XSLTName jats-oasis2jats3.xsl?>
   <?ConverterInfo.Version 1?>
   <?properties open_access?>
 ▼<processing-meta base-tagset="archiving" mathml-version="3.0" table-model="xhtml"
     <restricted-by>pmc</restricted-by>
   </processing-meta>
 ▼<front>
   ▼<journal-meta>
       <journal-id journal-id-type="nlm-ta">Health Promot J Austr</journal-id>
       <journal-id journal-id-type="iso-abbrev">Health Promot J Austr</journal-id>
       <journal-id journal-id-type="doi">10.1002/(ISSN)2201-1617</journal-id>
       <journal-id journal-id-type="publisher-id">HPJA</journal-id>
     ▼<journal-title-group>
         <journal-title>Health Promotion Journal of Australia</journal-title>
       </journal-title-group>
       <issn pub-type="ppub">1036-1073</issn>
       <issn pub-type="epub">2201-1617</issn>
     ▼<publisher>
         <publisher-name>John Wiley and Sons Inc.</publisher-name>
         <publisher-loc>Hoboken</publisher-loc>
       </publisher>
     </journal-meta>
   ▼<article-meta>
       <article-id pub-id-type="pmcid">11799825</article-id>
       <article-id pub-id-type="pmid">39910978</article-id>
       <article-id pub-id-type="doi">10.1002/hpja.70011</article-id>
       <article-id pub-id-type="publisher-id">HPJA70011</article-id>
       <article-id pub-id-type="other">5524741</article-id>
     ▼<article-categories>
```



```
▼ disease
    ▼ PMC11799825
          eupmc_result.json
          fulltext.pdf
          fulltext.xml
    ▶ PMC11968627
    ▶ PMC11968874
    ▶ PMC11978687
    ▶ PMC11984240
```



Health Promotion Journal of Australia

WILEY

COMMENTARY OPEN ACCESS

Australia's National Lung Cancer Screening Program—It's Time to Address the Stigma in the Room

Shiho Rose[1] | Kathleen McFadden[1] | Nathan J. Harrison[2,3,4] | Rachael H. Dodd[1,5] | Shakira Onwuka[6] | Christine Paul[7,8] | Lisa Carter-Bawa[9,10] | Mark Brooke[11] | Marianne Weber[1]

[1]The Daffodil Centre, The University of Sydney, a Joint Venture With Cancer Council NSW, Sydney, New South Wales, Australia | [2]Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Adelaide, South Australia, Australia | [3]NHMRC Centre of Research Excellence on Achieving the Tobacco Endgame, School of Public Health, The University of Queensland, Herston, Queensland, Australia | [4]National Centre for Education and Training on Addiction (NCETA), Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Adelaide, South Australia, Australia | [5]Sydney Health Literacy Lab, School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia | [6]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia | [7]School of Medicine and Public Health, The University of Newcastle, Callaghan, New South Wales, Australia | [8]Hunter Medical Research Institute, New Lambton Heights, New South Wales, Australia | [9]Center for Discovery and Innovation at Hackensack Meridian Health, Cancer Prevention Precision Control Institute, Nutley, New Jersey, USA | [10]Georgetown Lombardi Comprehensive Cancer Centre, Cancer Prevention and Control Program, Washington, DC, USA | [11]Lung Foundation Australia, Milton, Queensland, Australia

Correspondence: Shiho Rose (shiho.haim@sydney.edu.au)

Received: 1 October 2024 | **Revised:** 1 October 2024 | **Accepted:** 17 January 2025

Handling Editor: Williams Carmel

Funding: The authors received no specific funding for this work.

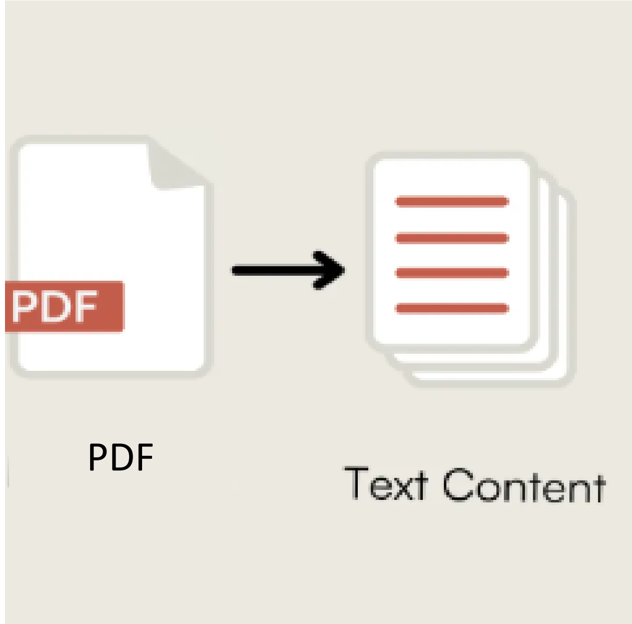
Keywords: cancer screening | Health promotion | lung cancer | stigma | tobacco

ABSTRACT

The National Lung Cancer Screening Program is commencing in Australia in July 2025. This significant public health initiative will maximise earlier detection of lung cancer and improve outcomes for many Australians. However, the adoption of a screening program for a disease that is stigmatised, given the known links between tobacco smoking and lung cancer, creates barriers for participation. In this perspective, we argue the need to challenge public rhetoric around smoking being a 'choice' and the importance of dialogue that is free of judgement and blame towards individuals. We briefly examine initiatives that have been implemented to reduce public stigma and highlight the multi-level considerations to ensure that everyone, regardless of having smoked or not, receives the quality care and support that they deserve.

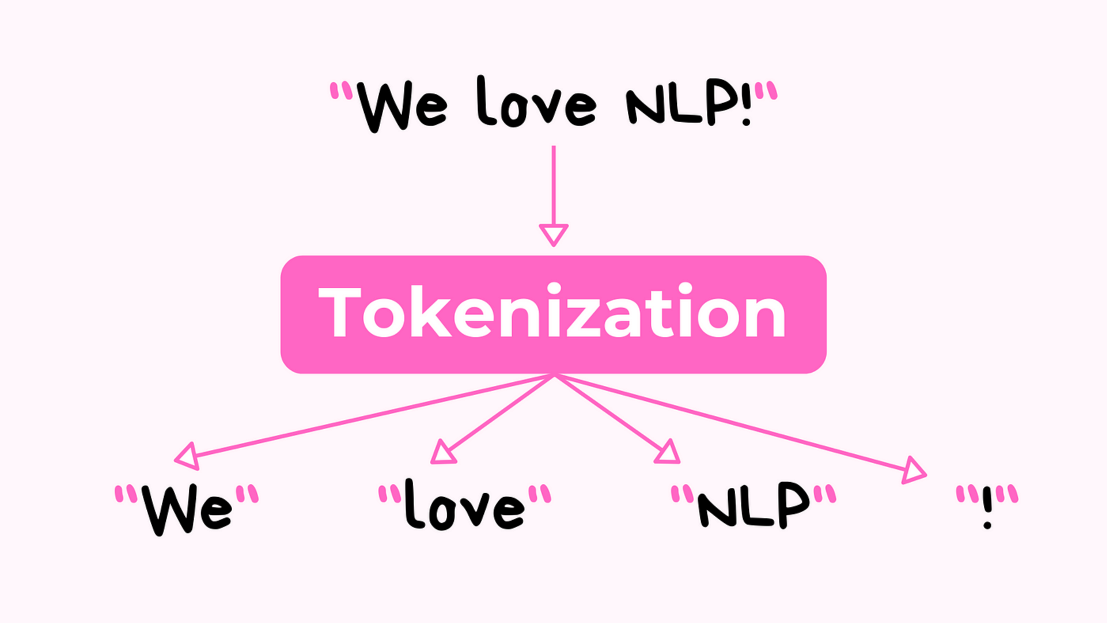
Step 2 & 3: PDF parsing and Text tokenization

2. Parsing is the process of extracting and structuring data from documents, such as PDFs.



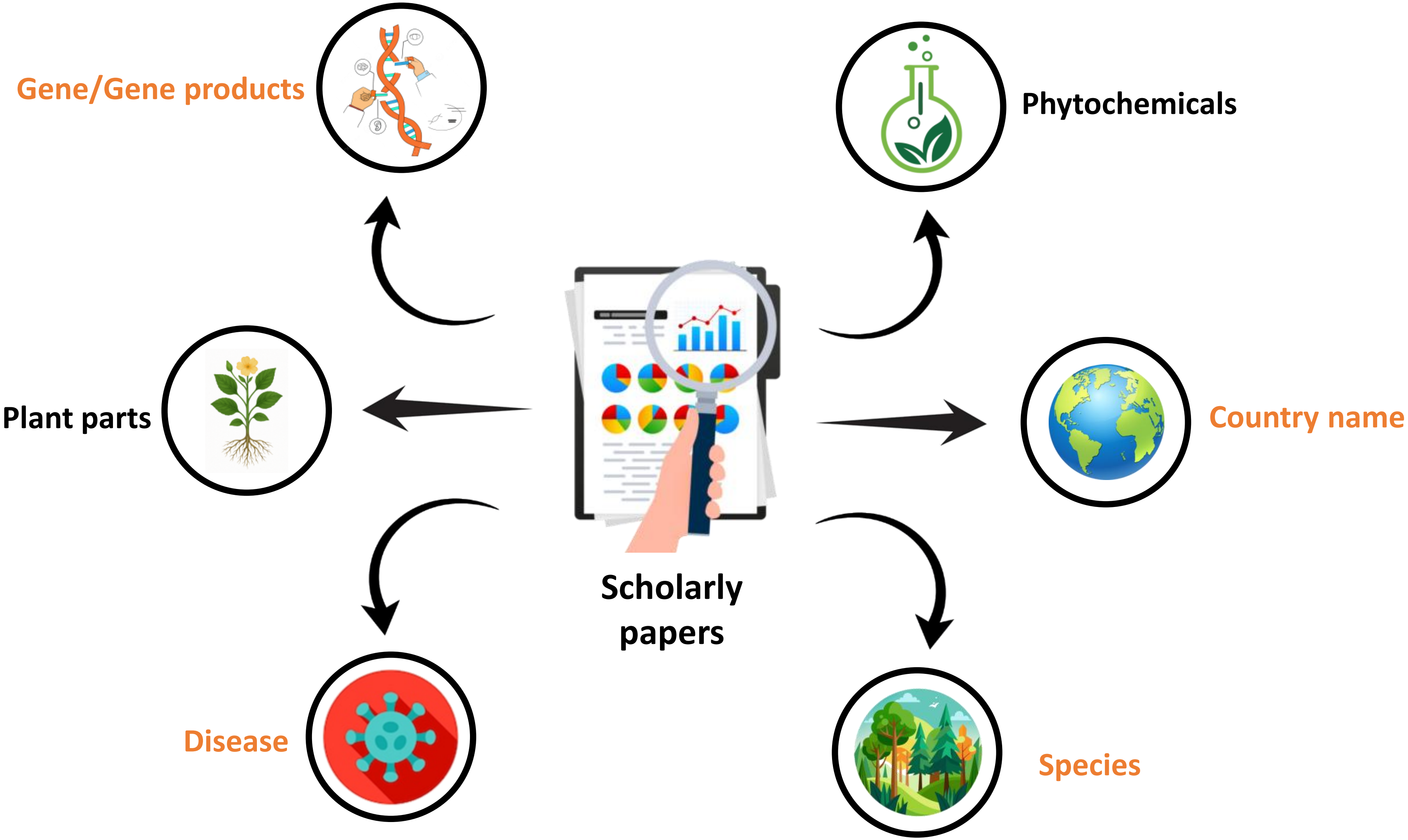
- ➢ PDF Parsing Packages: **fitz (from PyMuPDF)**
- ➢ **Fitz**: Extracts raw texts from pdfs

3. Tokenization is the process of splitting text into smaller units, such as sentences or words (tokens).



- ➢ Tokenization: **nltk.tokenize.sent_tokenize**
- ➢ Splits text into sentences

Step 4: Named Entity Recognition (NER)



Gene/Gene products

Phytochemicals

Plant parts

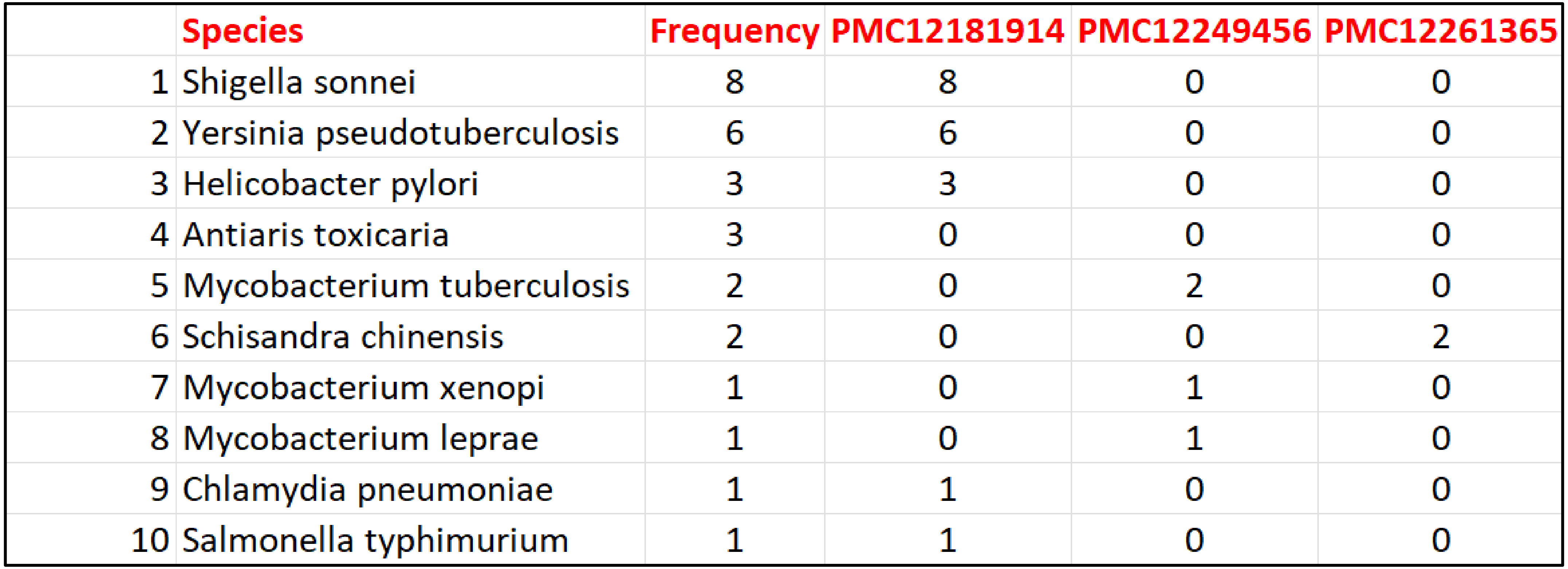
Country name

Scholarly papers

Disease

Species

Species Extraction

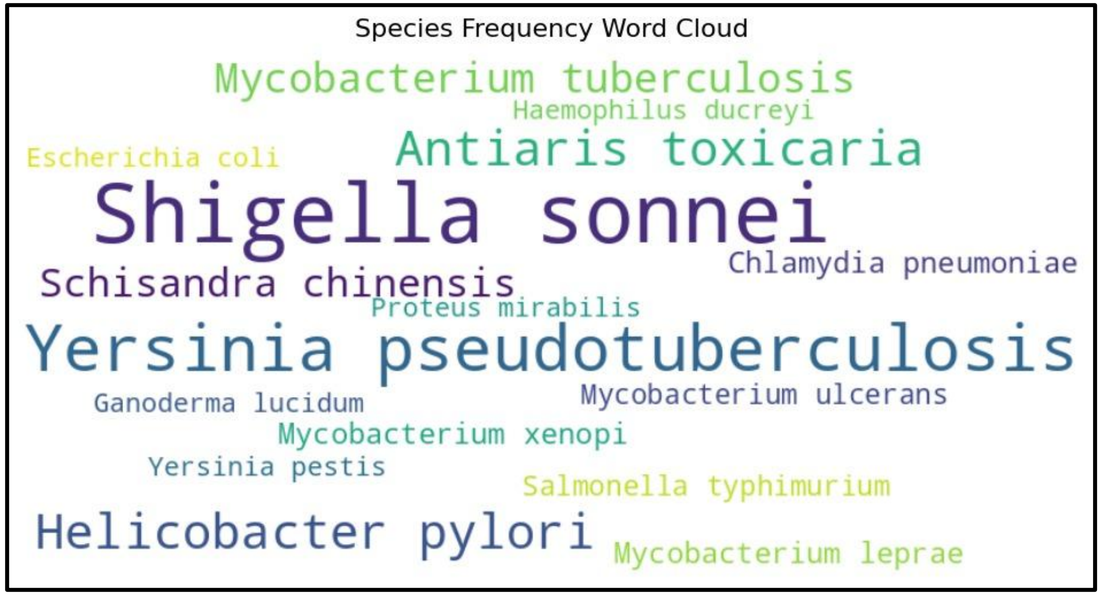


	Species	Frequency	PMC12181914	PMC12249456	PMC12261365
1	Shigella sonnei	8	8	0	0
2	Yersinia pseudotuberculosis	6	6	0	0
3	Helicobacter pylori	3	3	0	0
4	Antiaris toxicaria	3	0	0	0
5	Mycobacterium tuberculosis	2	0	2	0
6	Schisandra chinensis	2	0	0	2
7	Mycobacterium xenopi	1	0	1	0
8	Mycobacterium leprae	1	0	1	0
9	Chlamydia pneumoniae	1	1	0	0
10	Salmonella typhimurium	1	1	0	0

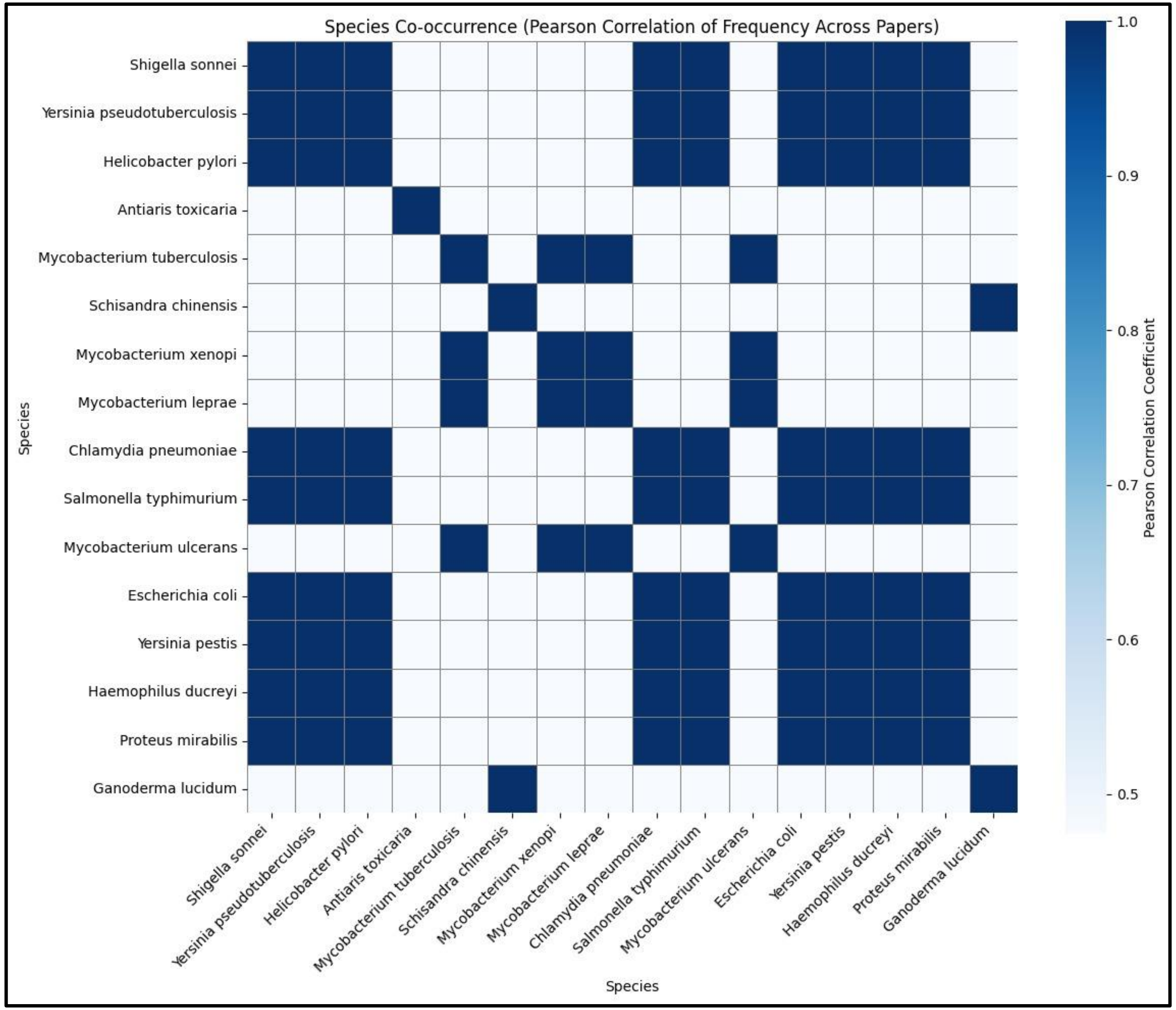
Frequency table showing the top 10 species across 3 PMCIDs out of the 50 papers

- NER Model: **scispaCy en_core_sci_md** (scientific text)
- Validation:
    - Regex for binomial nomenclature (Genus species)
    - **GBIF** API verification (EXACT species matches)

Species Extraction

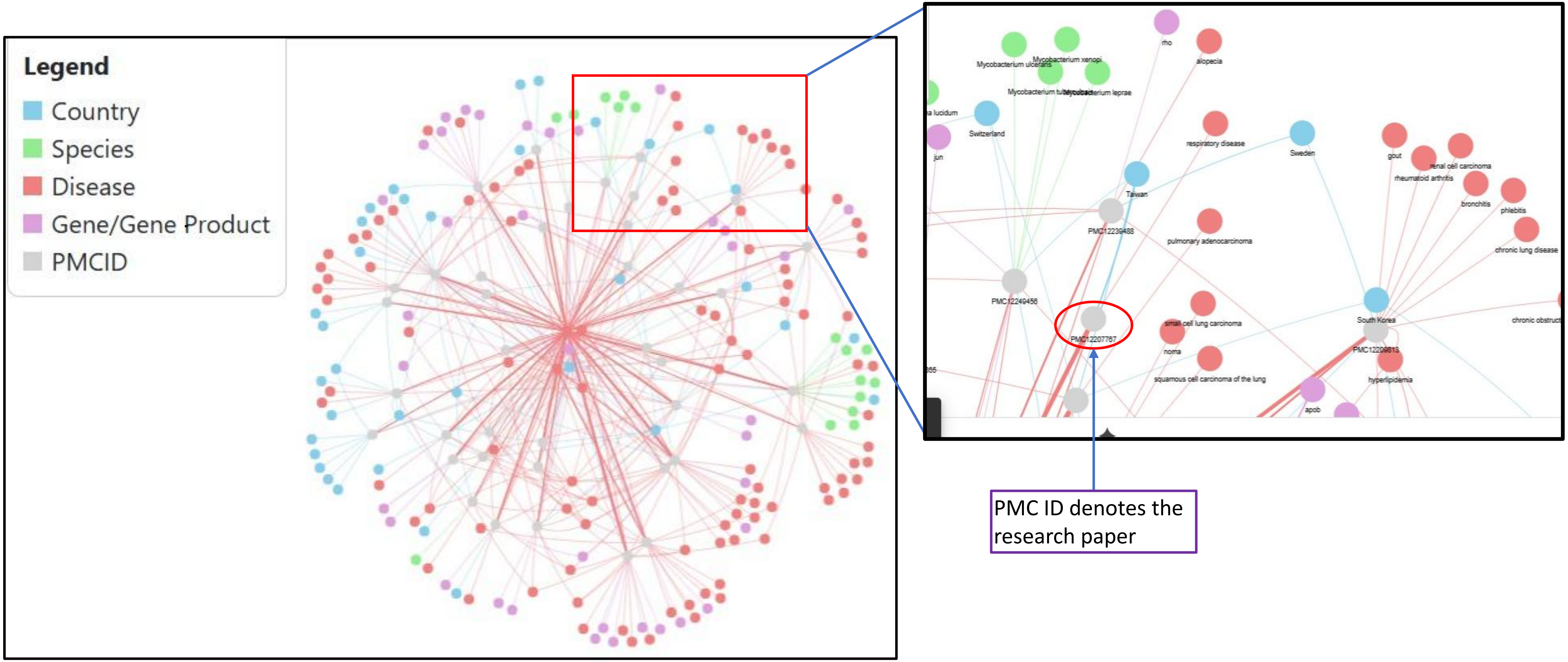


Word cloud of species obtained from the 50 papers



Co-occurrence plot of species for 50 papers

Network of the Extracted Entities



PMC ID denotes the research paper

THANK YOU

#semanticClimate

Transforming information into actionable knowledge