PAPERMATCH

AI-ENHANCED PLATFORM FOR CURATED ACADEMIC AND MEDIA RESEARCH INSIGHTS

PROPOSAL PRESENTATION

Alvin Wong Ann Ying, Bertrand TAN

Scope

- Business & Technical Problem Statements
- Methodology & System Architecture
  - Datasets
  - System Framework
  - Document Workstream
  - Query Workstream
  - Knowledge Workstream
- Project Deliverables

Business problem statement

Navigating the Information Overload in Research

Exponential Growth of Academic Publications

- The number of scholarly articles published annually has surged, making it difficult for researchers to keep up with the latest developments.
- Researchers struggle to efficiently filter and extract relevant insights from vast databases such as ACM Digital Library, IEEE Xplore, and arXiv.

Fragmented Research Dissemination

- Research findings are scattered across multiple platforms with varying accessibility, making discovery inefficient.
- The lack of structured categorization and summarization tools slows down knowledge acquisition and innovation.

The Role of Media in Research Awareness

- Media framing and narratives significantly influence public and academic discourse, shaping research trends and funding priorities.
- Researchers often miss how their work is portrayed in mainstream media, affecting public perception and policy impact.

Business Need for PaperMatch

PaperMatch provides a **unified AI-driven platform** to help researchers efficiently discover, analyze, and contextualize academic findings. By integrating **cutting-edge NLP, thematic clustering, and media analytics**, PaperMatch enables faster knowledge translation, ensuring research has a tangible impact on academia, industry, and society.

Technical problem statement

Technical Challenges in Research Aggregation and Analysis

Automating Research Summarization

- Processing large volumes of academic papers requires advanced **NLP techniques** to extract key insights without losing critical details.
- Current summarization methods (e.g., See, Liu, and Manning, 2017) need to be optimized for domain-specific technical content.

Scalable Thematic Clustering

- Organizing research papers into relevant clusters demands efficient **unsupervised learning** models (e.g., Aggarwal & Reddy, 2013).
- Traditional keyword-based categorization often fails to capture nuanced topic relationships and interdisciplinary connections.

Integrating Media and Research Contexts

- AI-driven **media aggregation** must effectively filter, extract, and summarize news articles relevant to academic research.
- Sentiment and framing analysis are needed to assess **how media narratives influence scientific discourse**.

Bridging Research with Practical Applications

- Mapping research findings to **real-world implementations** requires linking academic papers with **code repositories (e.g., GitHub)**.
- Challenges include maintaining **up-to-date recommendations** and ensuring relevance to evolving industry needs.

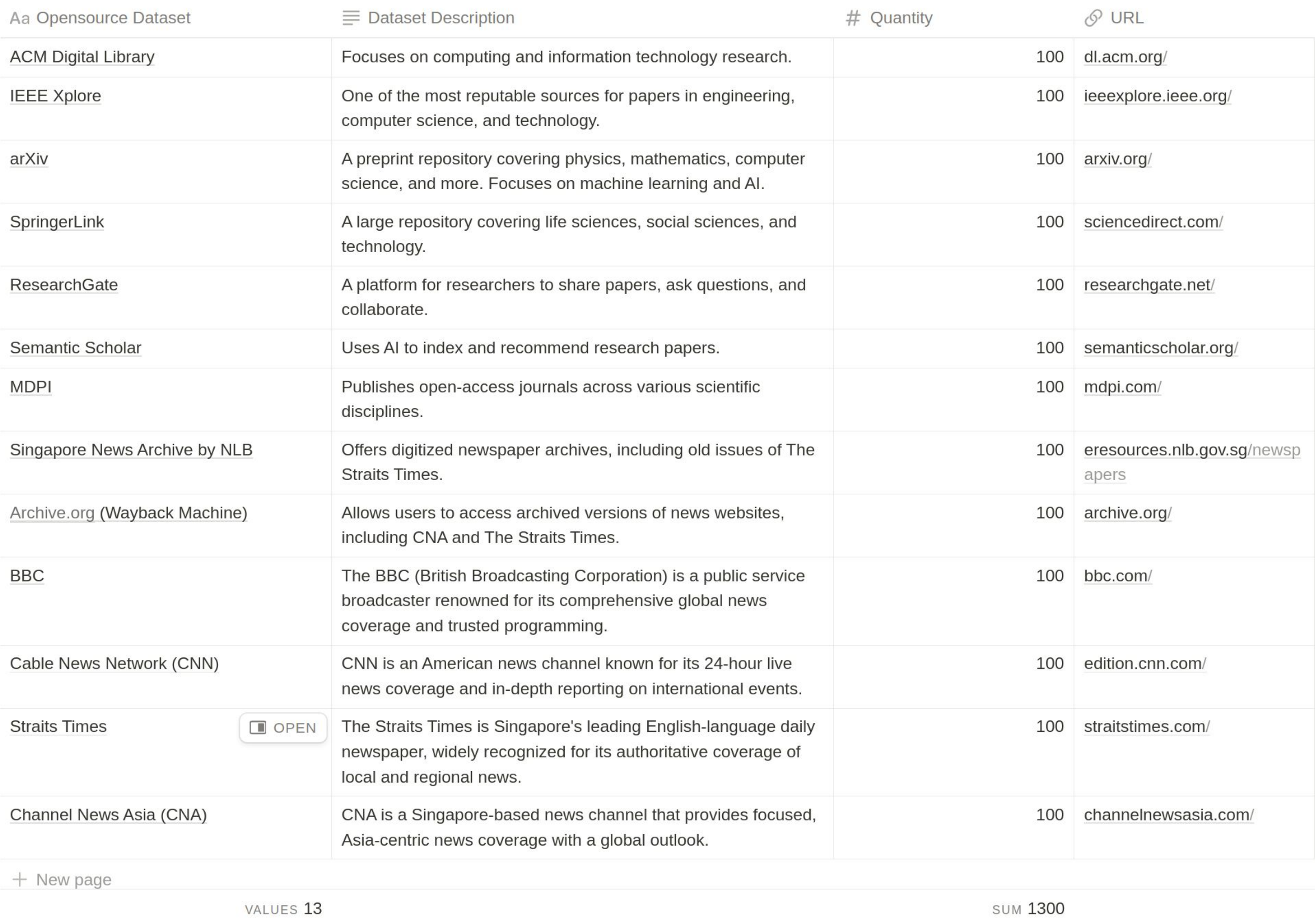
How PaperMatch Solves These Challenges

PaperMatch employs a **multi-layered AI approach** to tackle these technical barriers:

- **NLP-enhanced summarization** to condense academic papers into digestible insights.
- **Advanced clustering techniques** to organize research into actionable themes.
- **AI-driven news aggregation** to integrate media narratives with scholarly work.
- **Dynamic research-to-application mapping** to facilitate real-world impact.

Dataset

- Scrapping Open-source datasets from various sources:

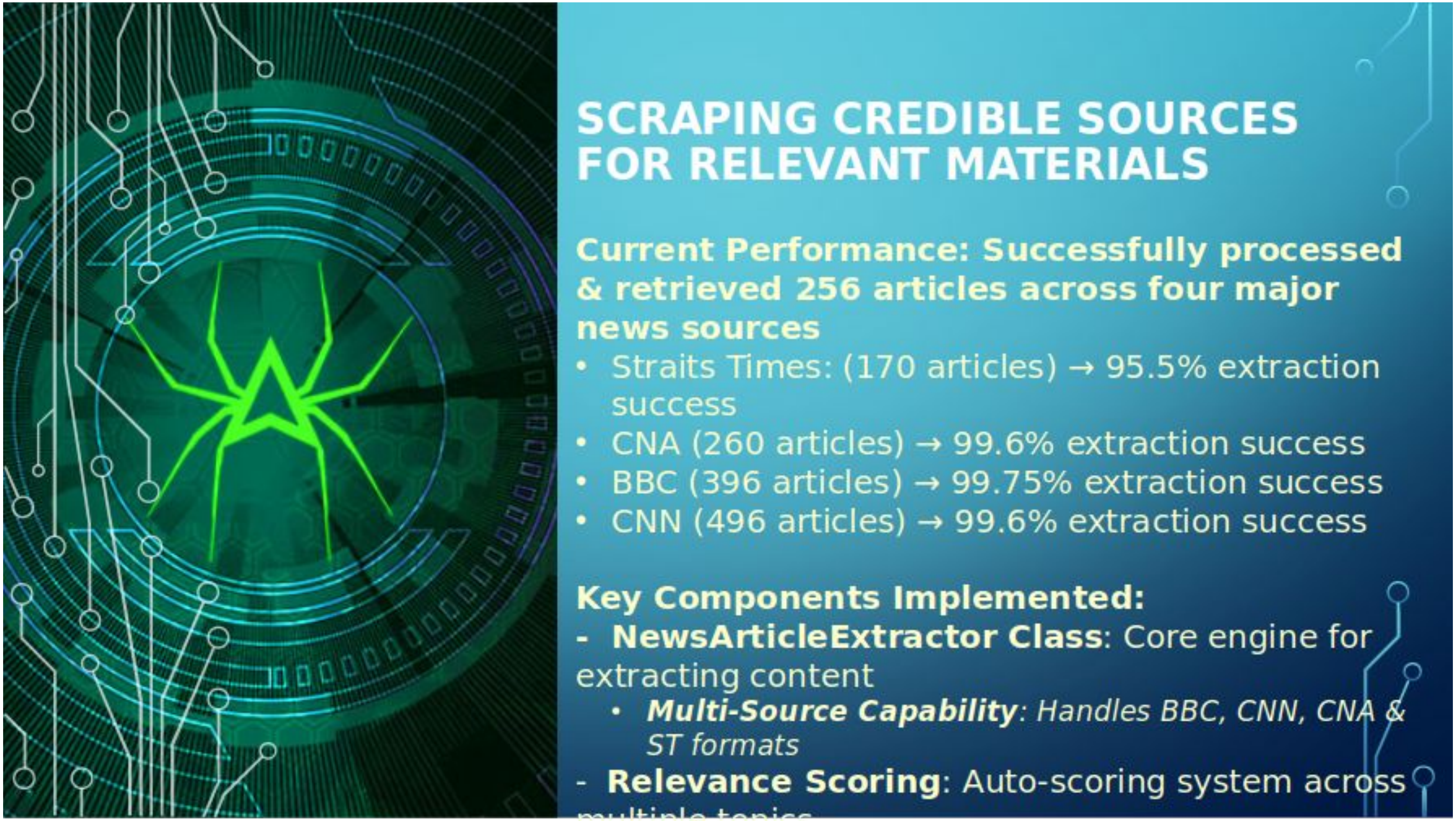


Aa Opensource Dataset	Dataset Description	# Quantity	URL
ACM Digital Library	Focuses on computing and information technology research.	100	dl.acm.org/
IEEE Xplore	One of the most reputable sources for papers in engineering, computer science, and technology.	100	ieeexplore.ieee.org/
arXiv	A preprint repository covering physics, mathematics, computer science, and more. Focuses on machine learning and AI.	100	arxiv.org/
SpringerLink	A large repository covering life sciences, social sciences, and technology.	100	sciencedirect.com/
ResearchGate	A platform for researchers to share papers, ask questions, and collaborate.	100	researchgate.net/
Semantic Scholar	Uses AI to index and recommend research papers.	100	semanticscholar.org/
MDPI	Publishes open-access journals across various scientific disciplines.	100	mdpi.com/
Singapore News Archive by NLB	Offers digitized newspaper archives, including old issues of The Straits Times.	100	eresources.nlb.gov.sg/newspapers
Archive.org (Wayback Machine)	Allows users to access archived versions of news websites, including CNA and The Straits Times.	100	archive.org/
BBC	The BBC (British Broadcasting Corporation) is a public service broadcaster renowned for its comprehensive global news coverage and trusted programming.	100	bbc.com/
Cable News Network (CNN)	CNN is an American news channel known for its 24-hour live news coverage and in-depth reporting on international events.	100	edition.cnn.com/
Straits Times OPEN	The Straits Times is Singapore's leading English-language daily newspaper, widely recognized for its authoritative coverage of local and regional news.	100	straitstimes.com/
Channel News Asia (CNA)	CNA is a Singapore-based news channel that provides focused, Asia-centric news coverage with a global outlook.	100	channelnewsasia.com/

+ New page

VALUES 13 SUM 1300

Dataset



SCRAPING CREDIBLE SOURCES FOR RELEVANT MATERIALS

Current Performance: Successfully processed & retrieved 256 articles across four major news sources

- Straits Times: (170 articles) → 95.5% extraction success
- CNA (260 articles) → 99.6% extraction success
- BBC (396 articles) → 99.75% extraction success
- CNN (496 articles) → 99.6% extraction success

Key Components Implemented:

- **NewsArticleExtractor Class**: Core engine for extracting content
  - ***Multi-Source Capability***: *Handles BBC, CNN, CNA & ST formats*
- **Relevance Scoring**: Auto-scoring system across multiple topics
- **Standardized Output Format**:
  - *YYYYMMDD_SourceTopicR[1-5]_Clean-Title.txt*

Dataset

SOURCE-SPECIFIC EXTRACTION

Tailored Extraction Methods

- Straits Times Processing:

- Skip articles with broken links leading back to main landing page
- Skip articles that are unrelated advertorials inserted by publisher’s algorithm
- Paywall detection and archive-based retrieval (Disabled)

- CNA Processing:

- Newsletter termination markers detection
- Exclusion of recommended/related articles with article content

- BBC Processing:

- Special handling for paragraphing class patterns (javascript)
- Context-aware heading hierarchy preservation
- Content deduplication for overlapping sections

- CNN Processing:

- Exclusion filtering for "Up Next" and "Most Read" sections
- Contextual article boundary detection

Dataset

TECHNICAL CHALLENGES & SOLUTIONS

Overcoming Web Content Extraction Barriers

Challenge 1: Varied Behaviour of Publishers' Search Function

- Problem:
  - *Some publishers' search algorithm operate using 'AND' while others use 'OR' function*
  - *ST search page uses javascript to display results (no pagination); limit: 20 results*
- Solution: Use of different search keywords for different publishers to maximise search results
- Result: Managed to complete trawl of relevant results for each publisher

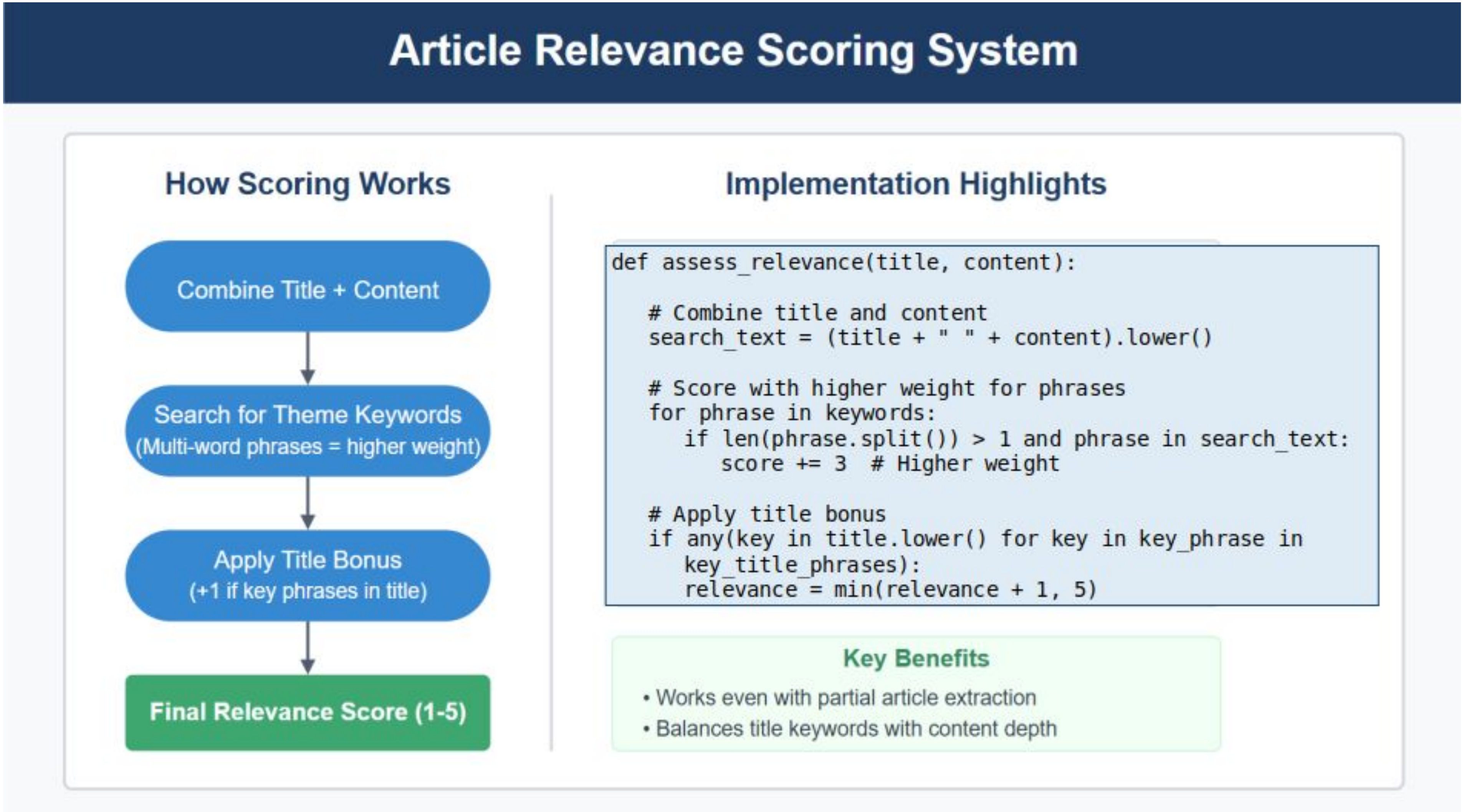
Challenge 2: Inconsistent HTML Structures

- Problem: Each news source uses different CSS selectors and DOM structures
- Solution: Implemented multi-selector cascading approach with 5+ fallbacks per source
- Result: Achieved 95%+ extraction success rate across varied HTML structures

Challenge 3: Content Duplication

- Problem: Multiple similar sections in modern news sites
- Solution:
  - *Implemented intelligent deduplication to detect article boundaries*
  - *Created excluded section detection for "Related Articles" & Avertisments*
- Result: Clean, non-redundant content for LLM processing

Dataset



Article Relevance Scoring System

How Scoring Works

Combine Title + Content

Search for Theme Keywords
(Multi-word phrases = higher weight)

Apply Title Bonus
(+1 if key phrases in title)

Final Relevance Score (1-5)

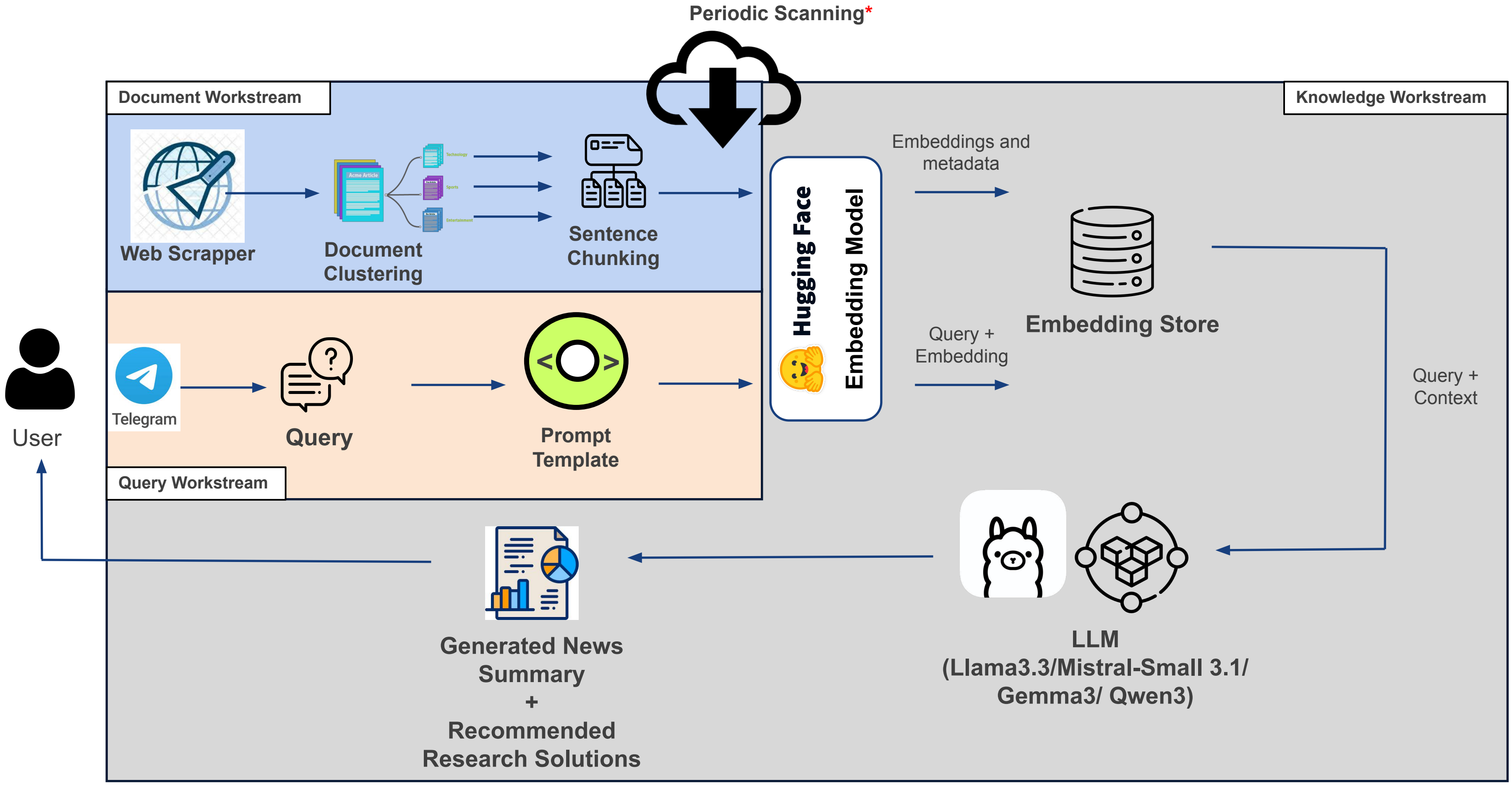
Implementation Highlights

```
def assess_relevance(title, content):

   # Combine title and content
   search_text = (title + " " + content).lower()

   # Score with higher weight for phrases
   for phrase in keywords:
      if len(phrase.split()) > 1 and phrase in search_text:
         score += 3  # Higher weight

   # Apply title bonus
   if any(key in title.lower() for key in key_phrase in
      key_title_phrases):
      relevance = min(relevance + 1, 5)
```

Key Benefits

- Works even with partial article extraction
- Balances title keywords with content depth

System Framework for PaperMatch



Periodic Scanning*

Document Workstream

Web Scrapper

Document Clustering

Sentence Chunking

Query Workstream

User

Telegram

Query

Prompt Template

Knowledge Workstream

Hugging Face Embedding Model

Embeddings and metadata

Query + Embedding

Embedding Store

Query + Context

LLM (Llama3.3/Mistral-Small 3.1/ Gemma3/ Qwen3)

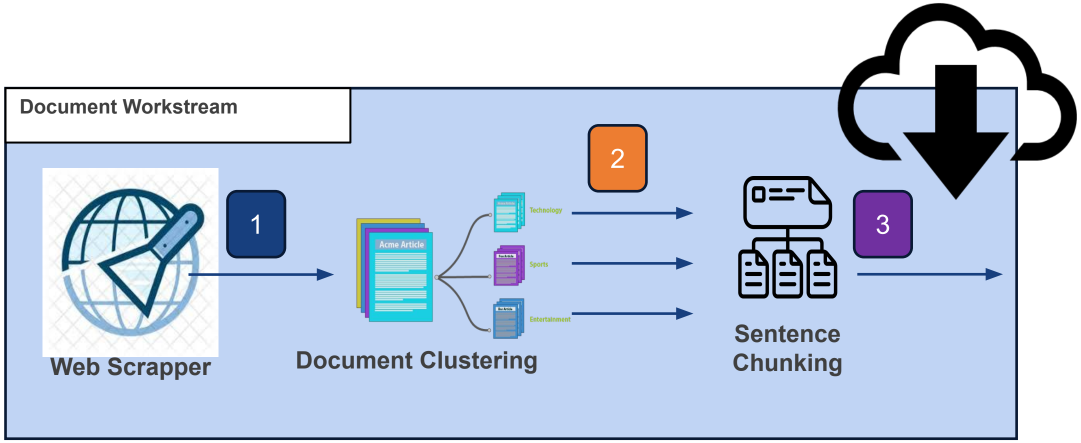
Generated News Summary + Recommended Research Solutions

*Knowledge Database should be updated periodically. However, we will exclude this feature for this project scope

Document Workstream

The **Document Workstream** focuses on automating the retrieval, processing, and categorization of academic research papers and news articles. This pipeline ensures seamless integration of both scholarly and media perspectives, providing researchers with structured insights.

Periodic Scanning*



Web Scraper for Relevant URLs
A web scraping module will be developed to identify and collect URLs from selected research publication repositories and news sources.

Key Features:

- **Targeted Domains:**
- Research: ACM Digital Library, IEEE Xplore, arXiv, Papers with Code.
- News: Major scientific news outlets (e.g., Nature, Science, TechCrunch).
- **Dynamic URL Extraction:**
- Uses **BeautifulSoup/Selenium/Scrapy** to navigate webpages and extract links to relevant research papers and news articles.
- Identifies patterns in URL structures to filter out irrelevant pages.
- **Metadata Collection:**
- Extracts essential details (e.g., title, author, publication date, abstract) to facilitate later categorization.
- **Duplicate Filtering:**
- Implements a hash-based mechanism to remove duplicate URLs before processing.

Document Download and Clustering
Once URLs are collected, the next step involves downloading documents and categorizing them into research or news groups.

Download Function:

- **Automated PDF/HTML Fetching:**
- Downloads full-text documents using requests-based methods or APIs (where available).
- Handles authentication and rate-limiting for paywalled content (if applicable).
- **File Format Standardization:**
- Converts various formats (PDF, HTML, TXT) into a unified text format for further processing.

Clustering Methodology:

- **Categorization of News Content:**
- Implements **unsupervised learning techniques (e.g., k-means, DBSCAN, or hierarchical clustering)** to group news articles based on themes.
- Uses **TF-IDF and topic modeling (LDA, BERT-based embeddings)** to detect key topics and separate articles into categories (e.g., AI research, policy changes, industry trends).
- **Academic vs. Media Differentiation:**
- Uses **rule-based classifiers and NLP models** to distinguish between scientific papers and news reports.
- Assigns confidence scores to ensure accuracy.

Sentence Chunking and Content Extraction
After downloading, the documents undergo text extraction and segmentation for structured analysis.

Sentence Chunking:

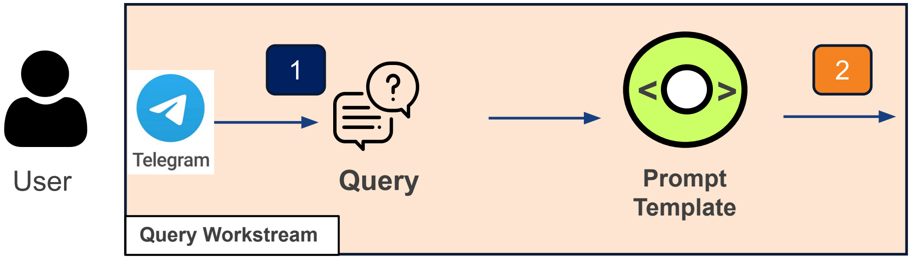
- **Text Segmentation:**
- Breaks long academic papers and news articles into meaningful **sentence-level or paragraph-level** chunks.
- Uses **NLTK, SpaCy, or Hugging Face models** to segment text while preserving context.
- **Entity Recognition:**
- Extracts key entities (e.g., research topics, organizations, authors, keywords) using **Named Entity Recognition (NER)**.

Content Extraction:

- **Key Information Identification:**
- Extracts **abstracts, conclusions, research methodologies** from academic papers.
- Identifies **headlines, quotes, and key takeaways** from news articles.
- **Summarization Pipeline:**
- Utilizes **state-of-the-art NLP models (e.g., BART, T5, Pegasus)** to generate concise summaries for each document.

Query Workstream

The **Query Workstream** enables users to search for relevant academic research and media coverage using natural language queries. This pipeline processes user queries efficiently, retrieves relevant documents, and provides structured insights based on predefined templates.



User Query Input

The user initiates the search by entering a natural language query into the system.

Example Query:

"Provide me the latest trends about deepfake and recommended research solutions on publication forums."

Key Features:

- **Intuitive Search Interface:**
- Supports **free-text queries** with natural language understanding (NLU).
- Users can specify **research domains, publication types, and media sources** for more refined searches. *(optional)*
- **Query Preprocessing:**
- Uses **tokenization, lemmatization, and stop-word removal** to clean and standardize the input.
- Identifies key entities (e.g., "deepfake," "research solutions") using **Named Entity Recognition (NER)**.

Prompt Template for Query Processing

To ensure effective search results, the system utilizes predefined **prompt templates** for structuring the query before execution.

Prompt Template Format:

[User Query] → [Structured Prompt for Retrieval System]

Example transformation for the query:

Input:

"Provide me the latest trends about deepfake and recommended research solutions on publication forums."

Processed Prompt Template:

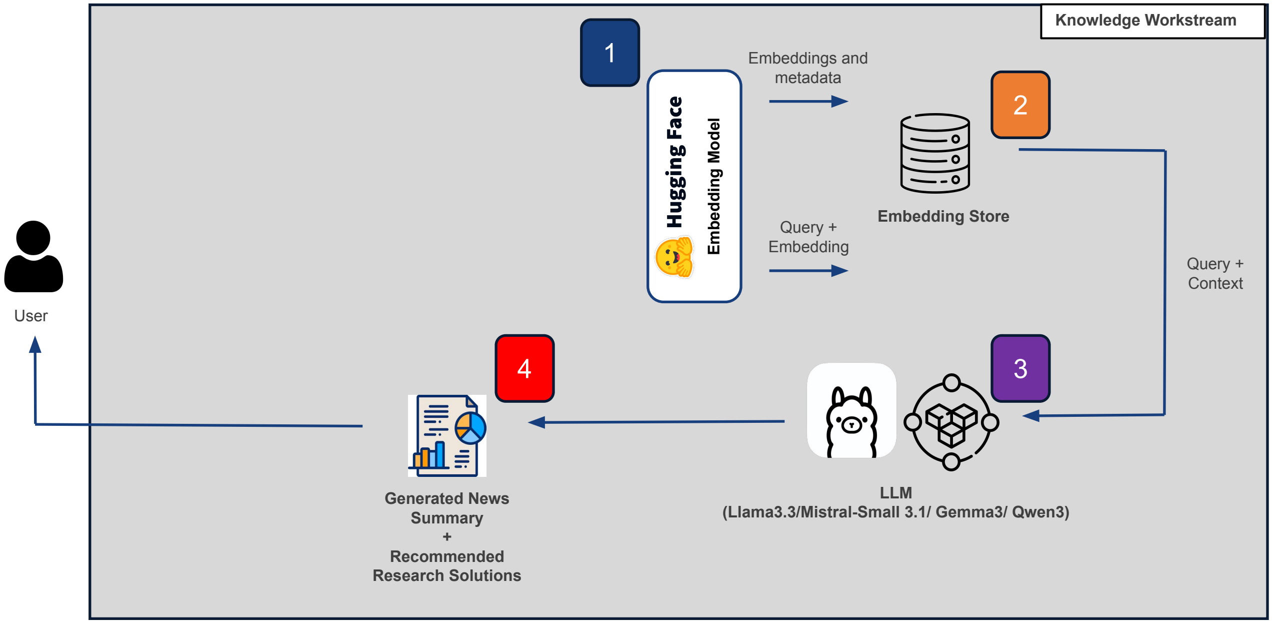
"Retrieve recent publications from IEEE Xplore, ACM Digital Library, and arXiv that discuss 'deepfake' in the last 12 months. Identify research solutions related to detection, mitigation, and ethical considerations. Provide key findings, methodologies, and direct links to papers."

Key Processing Steps:

- **Contextual Expansion:**
- Enhances the user query by adding structured retrieval conditions (e.g., time filters, research topics).
- **Query Routing:**
- Determines whether the query should be directed towards **academic sources, media sources, or both**.
- **Integration with Retrieval System:**
- Translates the prompt into API calls or database queries for retrieving relevant documents.

Knowledge Workstream

The **Knowledge Workstream** is responsible for transforming the structured outputs from the **Document Workstream** and **Query Workstream** into meaningful, retrievable knowledge representations. This is achieved through embeddings, vector databases, summarization via LLMs, and structured reporting for decision-makers.



Generate Embedding Pairs
To facilitate **efficient retrieval and semantic search**, we convert processed documents (from academic publications and news articles) into **vector embeddings** using a **Hugging Face embedding model** (e.g., sentence-transformers/all-mpnet-base-v2).
Processing Steps:

- Extract **title, abstract, key insights, and metadata** from **academic publications**.
- Extract **headlines, key narratives, and sentiment** from **news articles**.
- Use a **pre-trained transformer model** to generate embeddings for each document.
- Generate **paired embeddings** to allow cross-referencing between **academic** and **media** insights.

Store Embeddings in a Vector Database
Once embeddings are generated, they are stored in a **vector database** or **MongoDB (in JSON format)** for efficient retrieval.
Options for Storage:

- **Pinecone** / **FAISS** for high-performance vector searches.
- **MongoDB** for structured storage in **JSON format**, enabling document retrieval based on metadata.

Summarized Outcomes Using LLMs
To generate actionable insights, we use **LLMs such as Llama v3, Mistral, or Deepseek** to summarize trends and recommend publications & GitHub links.
Processing Steps:

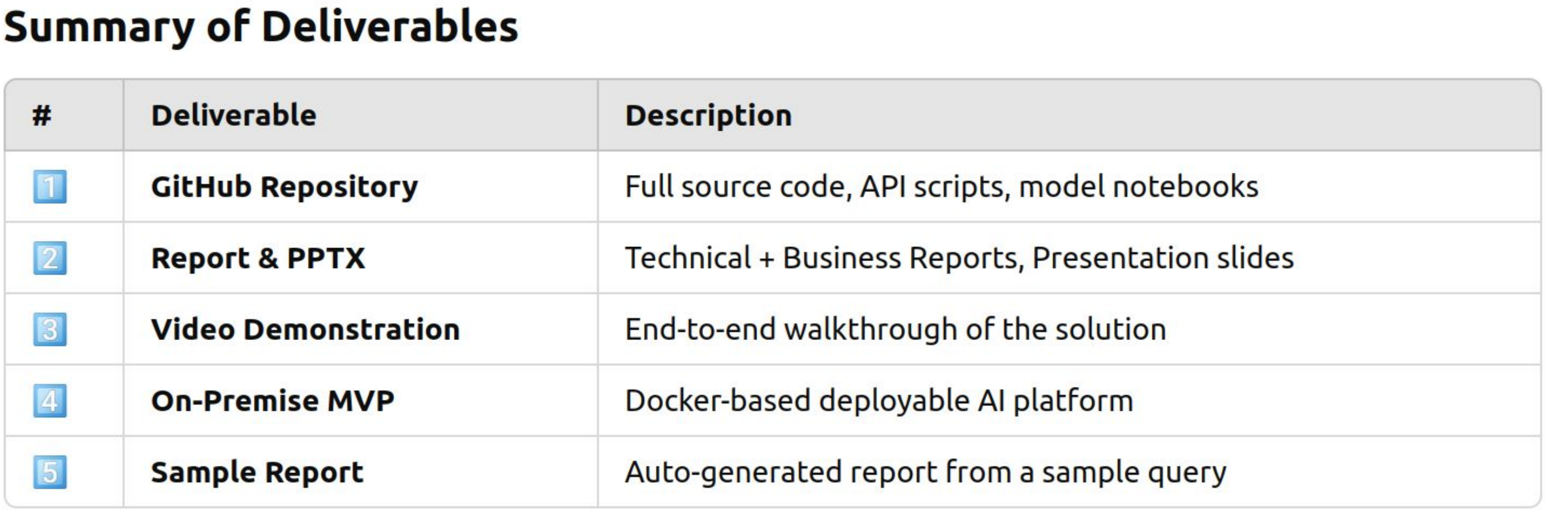
- Retrieve **top research papers** and **news articles** based on query relevance.
- Use **LLM prompting** to generate **concise summaries** of key insights.
- Extract **recommended research directions** and **potential solutions**.

PDF Report Generation for Senior Management
To facilitate decision-making, the system compiles insights into a **PDF report** summarizing the top **5 news trends** and **5 latest research directions**.
Processing Steps:

- Extract **top 5 news trends** and **top 5 research directions**.
- Format insights into a structured **management report**.
- Convert content into a **PDF report using Python libraries** (e.g., reportlab or pdfkit).

Project deliverables

The project deliverables encompass all critical components required to **demonstrate, document, and deploy** the solution, ensuring seamless integration and stakeholder alignment.

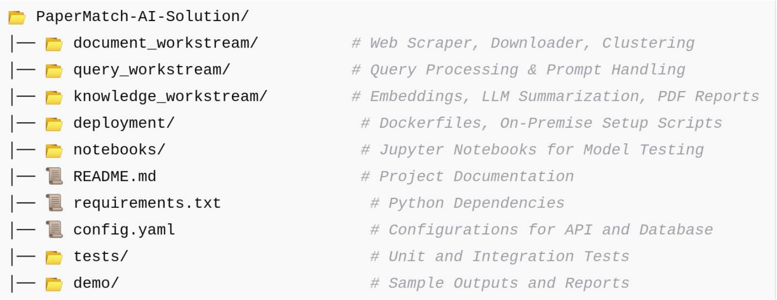


Summary of Deliverables

#	Deliverable	Description
1️⃣	**GitHub Repository**	Full source code, API scripts, model notebooks
2️⃣	**Report & PPTX**	Technical + Business Reports, Presentation slides
3️⃣	**Video Demonstration**	End-to-end walkthrough of the solution
4️⃣	**On-Premise MVP**	Docker-based deployable AI platform
5️⃣	**Sample Report**	Auto-generated report from a sample query

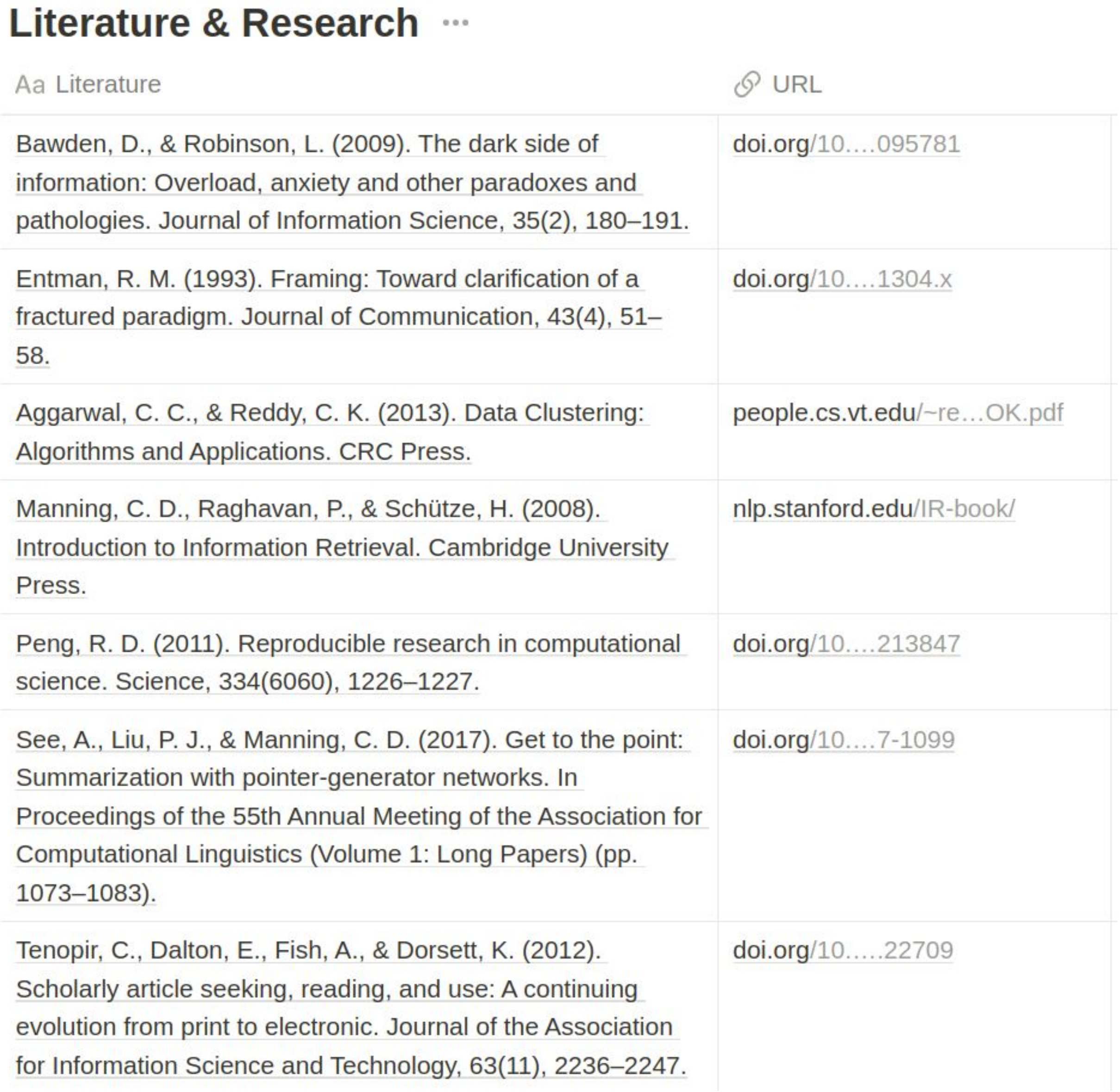
Final Outcome: A fully functional **end-to-end AI system** for **automated research retrieval, media trend analysis, and executive reporting**, ready for **enterprise use**.

Sample GitHub Repository



```
📂 PaperMatch-AI-Solution/
|— 📂 document_workstream/       # Web Scraper, Downloader, Clustering
|— 📂 query_workstream/          # Query Processing & Prompt Handling
|— 📂 knowledge_workstream/      # Embeddings, LLM Summarization, PDF Reports
|— 📂 deployment/                 # Dockerfiles, On-Premise Setup Scripts
|— 📂 notebooks/                  # Jupyter Notebooks for Model Testing
|— 📜 README.md                   # Project Documentation
|— 📜 requirements.txt             # Python Dependencies
|— 📜 config.yaml                  # Configurations for API and Database
|— 📂 tests/                       # Unit and Integration Tests
|— 📂 demo/                        # Sample Outputs and Reports
```

References



Literature & Research ···

Aa Literature	URL
Bawden, D., & Robinson, L. (2009). The dark side of information: Overload, anxiety and other paradoxes and pathologies. Journal of Information Science, 35(2), 180–191.	doi.org/10.…095781
Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of Communication, 43(4), 51–58.	doi.org/10.…1304.x
Aggarwal, C. C., & Reddy, C. K. (2013). Data Clustering: Algorithms and Applications. CRC Press.	people.cs.vt.edu/~re…OK.pdf
Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.	nlp.stanford.edu/IR-book/
Peng, R. D. (2011). Reproducible research in computational science. Science, 334(6060), 1226–1227.	doi.org/10.…213847
See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1073–1083).	doi.org/10.…7-1099
Tenopir, C., Dalton, E., Fish, A., & Dorsett, K. (2012). Scholarly article seeking, reading, and use: A continuing evolution from print to electronic. Journal of the Association for Information Science and Technology, 63(11), 2236–2247.	doi.org/10.….22709

References

Annex A

RELEVANT KEYWORDS BY THEME [1/3]

Cybercrime & Digital Fraud

Cybercrime Keywords - cyber attack, malware, ransomware, phishing, data breach, hacking, cybercrime, cyber security, cyber criminal, dark web, cyber fraud, identity theft, cyber espionage, botnet, DDoS attack, cyber warfare, computer virus, data theft, online fraud, cryptocurrency crime

Cyber-Tactics Keywords - zero-day exploit, social engineering, encryption, keylogger, backdoor, brute force attack, SQL injection, man-in-the-middle, password cracking, spyware, trojan horse, rootkit, cryptojacking, extortion

Forensic Science & Criminal Investigation

Forensic Techniques Keywords - DNA analysis, fingerprint analysis, ballistics, toxicology, forensic pathology, crime scene investigation, digital forensics, blood pattern analysis, forensic anthropology, trace evidence, forensic entomology, forensic psychology, autopsy, serology, chain of custody, forensic odontology

Scientific Methods Keywords - chromatography, spectroscopy, PCR amplification, mass spectrometry, microscopy, luminol test, substance identification, comparative analysis, facial reconstruction, voice analysis, handwriting analysis, geographic profiling

References

RELEVANT KEYWORDS BY THEME [2/3]

Annex A

Medical Fraud & Malpractice

Healthcare Fraud Keywords - medical fraud, healthcare fraud, insurance fraud, billing fraud, medicare fraud, medicaid fraud, upcoding, phantom billing, patient brokering, kickback scheme, medical identity theft, prescription fraud, fraudulent diagnosis, unnecessary procedure, health insurance fraud, pharmaceutical fraud

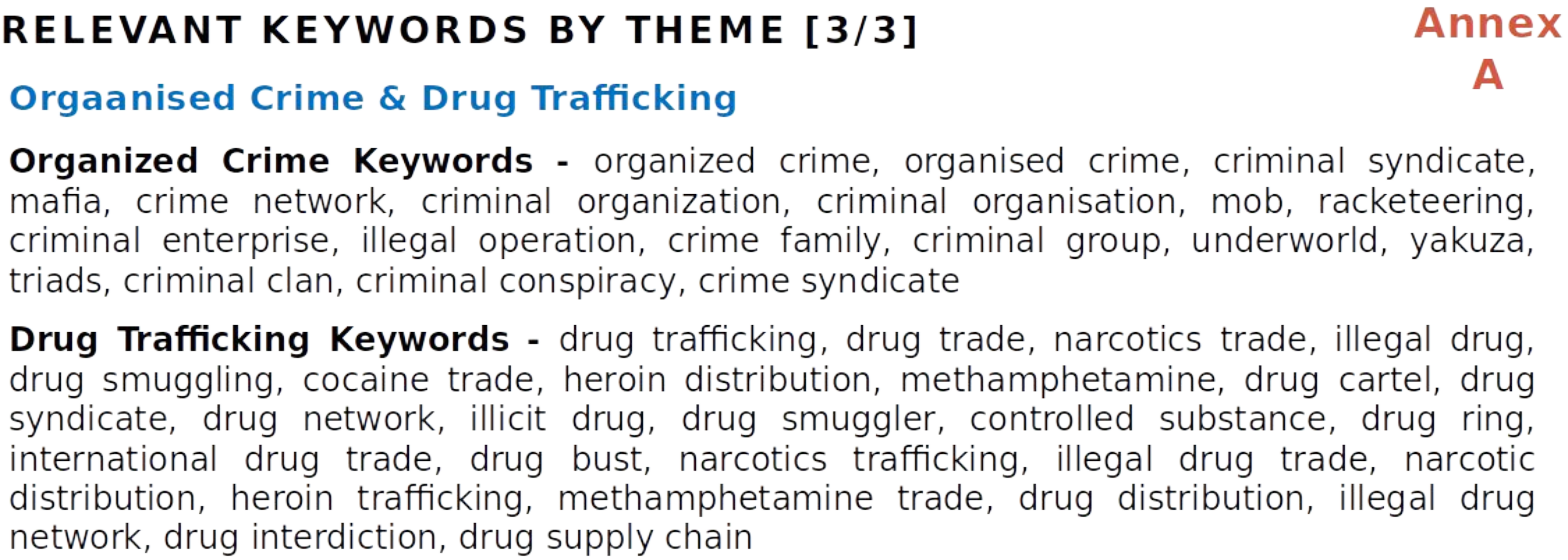
Medical Ethics Keywords - malpractice, unethical treatment, experimental treatment, informed consent violation, patient exploitation, medical negligence, falsified credentials, medical license fraud, counterfeit medicine, unlicensed practice, medical data manipulation, clinical trial fraud

Misinformation & Fake News

Misinformation Keywords - fake news, disinformation, misinformation, propaganda, conspiracy theory, information warfare, social media manipulation, deepfake, fact-checking, media literacy, information disorder, filter bubble, echo chamber, algorithmic bias, synthetic media, information operations, coordinated inauthentic behavior, influence operation, astroturfing, computational propaganda

Source Credibility Keywords - source verification, media bias, journalistic integrity, information source, primary source, secondary source, citation needed, unverified claim, anonymous source, information provenance, source attribution, fact versus opinion, media literacy, source criticism

References



RELEVANT KEYWORDS BY THEME [3/3]

Annex A

Orgaanised Crime & Drug Trafficking

Organized Crime Keywords - organized crime, organised crime, criminal syndicate, mafia, crime network, criminal organization, criminal organisation, mob, racketeering, criminal enterprise, illegal operation, crime family, criminal group, underworld, yakuza, triads, criminal clan, criminal conspiracy, crime syndicate

Drug Trafficking Keywords - drug trafficking, drug trade, narcotics trade, illegal drug, drug smuggling, cocaine trade, heroin distribution, methamphetamine, drug cartel, drug syndicate, drug network, illicit drug, drug smuggler, controlled substance, drug ring, international drug trade, drug bust, narcotics trafficking, illegal drug trade, narcotic distribution, heroin trafficking, methamphetamine trade, drug distribution, illegal drug network, drug interdiction, drug supply chain