Automating Secondary Research

Comparative Insights from KPMG and PwC

Himanshu

April 23, 2025

Phone: 9306983260
Email: ydvhimanshu22@gmail.com

GitHub Repository: https://github.com/22Himanshu/KPMG_-_PwC_article_resarch_automation

1 Objective

The goal of this project is to **automate secondary research** by scraping and analyzing articles from the websites of **KPMG** and **PwC**. The purpose is to identify and compare themes, industries, and insights using AI/ML methods and visualize these findings effectively.

2 Approach Overview

This project was divided into 3 main phases:

A. Web Scraping

- **Source Websites:** KPMG Insights & PwC Insights pages
- **Scraped Fields:** Title, Article URL, Date Published, Full Text Content.
- **Output Format:** JSON

B. Data Analysis & Thematic Extraction

- **Industry Classification:** LLM-based zero-shot classification
- **Theme Extraction:** Using Sentence Transformers (GoogleAIEmbedder)
- **Clustering:** KMeans clustering to group articles by themes
- **Traceability:** Article links retained for cluster-level attribution

C. Visualization & Reporting

- **Visualization Tool:** Matplotlib & Seaborn
- **Deliverables:** Theme clusters, industry comparison, keyword clouds

3 Tools, Models & Libraries Used

Purpose	Tools / Libraries / Models Used
Web Scraping	`requests`, `BeautifulSoup`, `Reader(r.jina/ai)`
Embeddings & LLMs	`sentence-transformers`
Clustering & NLP	`sklearn (KMeans)`, `KeyBERT`, `TfidfVectorizer`
Visualization	`matplotlib`, `seaborn`, `wordcloud`
Classification	Google Gemini-2.0 flash api call based zero-shot classification (inferred)

4 Key Industries Covered

The articles were classified into the following industries:

- Finance & Banking
- Tax & Regulation
- Data & Artificial Intelligence
- Supply Chain & Trade (includes tariffs)
- Technology & Innovation

Major themes revolved around **regulatory changes**, **tax policies**, **AI adoption**, and **financial innovations**.

5 Key Themes Identified

Articles were clustered using embeddings and KMeans into the following themes:

Cluster	Theme	Common Keywords
1	Regulatory & Tax Reforms	tax, policy, compliance, regulation, legislation
2	Artificial Intelligence & Data	AI, data, automation, innovation, machine learning
3	Financial Markets & Economy	banking, investment, economic trends, risk, fintech
4	Trade & Tariffs	tariffs, supply chain, global trade, imports

6 Visualizations

- **Bar Graphs** – Showcasing frequent terms (e.g., word counts).

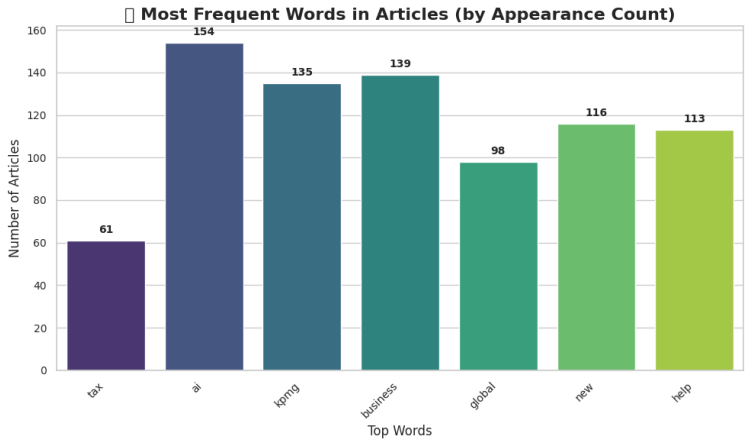

Figure 1: Bar graph illustrating frequently used words in KPMG articles.

- **Clusters Plots** – Visualizing clustered articles using dimensionality reduction (UMAP).

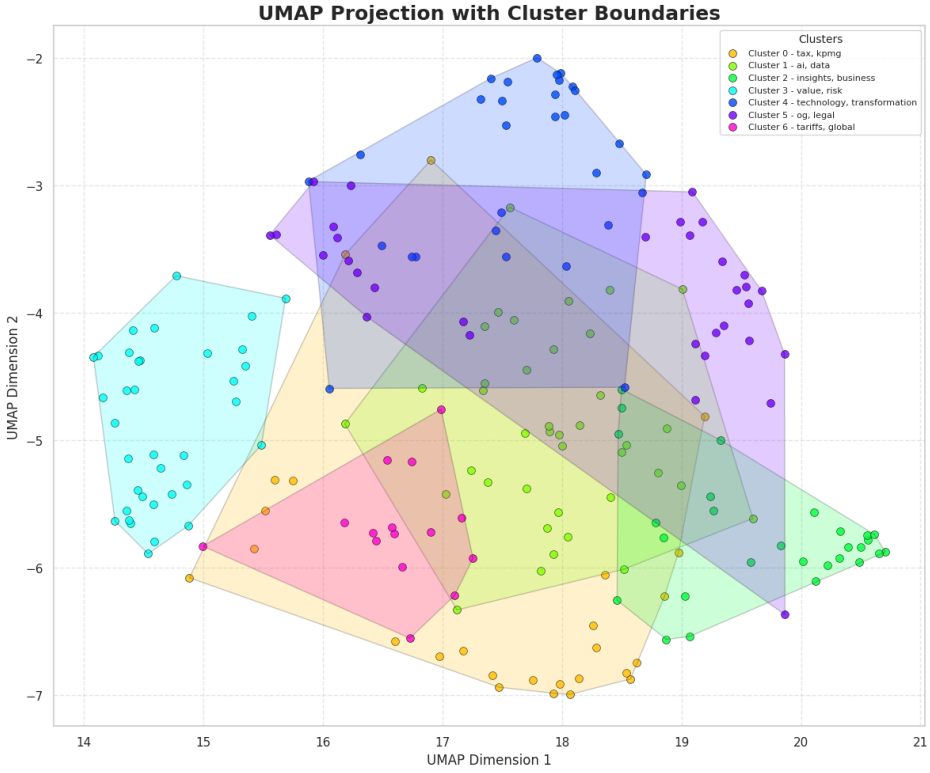

Figure 2: UMAP scatter plot showing the clustering of different article groups based on semantic similarity.

- **Word Clouds** – Highlighting prominent keywords and themes.

Figure 3: Word cloud generated from KPMG article text, visualizing key terms.

7 Deliverables Summary

- `AI_web_scraper.ipynb` – Scraping and basic cleaning
- `KPMG_PWC_data_visualization.ipynb` – Clustering, classification, visuals
- `data.json` – Raw and processed structured data
- **This Report** – Contains methodology, insights, visuals

8 Future Work

There is considerable potential to expand this project further:

- **Use of Paid LLM APIs:** Integrating advanced commercial language models (e.g. Gemini-Pro) can improve classification, especially for long, nuanced articles.
- **Sentiment Analysis:** Extract sentiment polarity for articles to understand tone or outlook across clusters.
- **Firm Comparison Expansion:** Extend analysis to other firms like Deloitte, EY for broader industry insights.
- **Web App Deployment:** Deploy insights via a lightweight Streamlit/Dash interface.

Conclusion

This project automated the process of gathering, analyzing, and visualizing insights from leading consulting firms using AI/ML. The results provide a scalable way to derive actionable intelligence from secondary research.