*Machine Learning and Content Analytics*

*Delta Team*

*Katsoula Eleftheria – p2822416*

*Papanikolaou Georgios – p2822423*

Intelligent Lead Generation

Project Scope and Business Need

This project was developed to address a widespread need across industries and sectors: the ability to efficiently discover and connect with relevant organizations, professionals, and content sources based on thematic and geographic criteria. In today’s information-rich environment, traditional search methods often fall short-either returning generic results or failing to surface niche entities and context-specific contacts.

The solution presented here combines **machine learning**, **natural language processing**, and **web scraping** to build a personalized discovery engine tailored for strategic outreach, research, and intelligence gathering. Whether the goal is to identify potential collaborators, suppliers, thought leaders, or investment targets, the tool enables users to generate curated lists of organizations and contact information aligned with their interests.

This approach is especially valuable to a wide range of business users, including:

* **Business development professionals** seeking new partnerships or clients
* **Market analysts** mapping competitive landscapes or emerging trends
* **Innovation scouts** identifying early-stage ventures or technology providers
* **Consultants and advisors** conducting due diligence or opportunity assessments
* **Investors and venture capitalists** sourcing deal flow across regions
* **Policy researchers and think tanks** exploring sectoral activity by geography
* **Sales and outreach teams** building targeted prospect lists for campaigns

By transforming the abstract task of “lead generation” into a **targeted and context-aware sourcing workflow**, this platform empowers professionals to navigate complex digital landscapes with precision, speed, and relevance.

Technical Workflow and Result Generation

The project is built around a modular pipeline of intelligent agents, each responsible for a distinct phase of the search and discovery process:

1. **Synonym Expansion** The SmartSynonymAgent uses a generative language model (Mistral-7B) to produce 12 alternative search phrases for a given topic. This ensures semantic diversity and improves the chances of uncovering relevant results that may not use the exact keyword.
2. **Keyword Construction** The KeywordAgent combines these synonyms with intent phrases (e.g., "contact us", "about us") and country-specific demonyms (e.g., "Greek", "Hellenic") to generate targeted search queries. It also incorporates domain filters (e.g., .gr, .com) to localize results.
3. **Web Search Execution** The SerpApiSearchAgent sends these queries to SerpApi, retrieving organic Google search results. It filters out duplicate domains using base URL logic and collects titles and links for each result.
4. **Result Filtering** The FilterAgent removes entries from known aggregators like LinkedIn, Crunchbase, and Wikipedia to ensure the final list contains primary sources—typically official company websites.
5. **Contact Extraction** The ContactPageFinder visits each homepage and scrapes the HTML to extract the site title and any visible email addresses using regex. This step transforms raw URLs into meaningful contact entries.
6. **Personalized Orchestration** The PersonalizedSearchAgent ties everything together. It uses the user’s profile (name, company, LinkedIn) to personalize the search—excluding their own company from results—and returns a deduplicated, filtered list of organizations with contact info.
7. **Final Output** The results are compiled into a structured directory, displayed in an interactive table, and saved as a CSV file for offline use or further analysis.

User Experience Through the UI

The user interface is built with ipywidgets, offering a clean and intuitive experience directly within the Jupyter notebook:

* **Input Fields** Users enter their first name, surname, company, LinkedIn profile, topic of interest, and select a country from a dropdown menu. They also specify the number of results they want.
* **Run Button** A single click initiates the entire pipeline, with progress feedback shown in real time.
* **Output Area**
  + Displays the number of results fetched and deduplicated
  + Shows a progress bar while contact info is being scraped
  + Presents the final directory in a styled HTML table with clickable homepage links and visible email addresses
  + Automatically saves the results to a CSV file named directory\_results.csv

The UI is designed for **simplicity and clarity**, allowing users to focus on insights rather than technical complexity. It transforms a sophisticated backend into a seamless front-end experience suitable for both technical and non-technical users.