AIM Reply

Micro-Internship Task

Dates: 8th December – 12th December 2025

Context

Welcome to AIM Reply!

Throughout this micro-internship, you will act as junior analysts helping AIM Reply explore how artificial intelligence can transform the way organisations analyse large patent portfolios.

Patents contain valuable information about innovation trends, competitive positioning, and strategic opportunities. However, manually reviewing thousands of patents is time-consuming and prone to bias. Your task is to use modern AI techniques, including text embeddings, unsupervised clustering and LLMs to automatically discover patterns and themes within patent data.

Your Client: Nestlé

Nestlé is one of the world's largest food and beverage companies, with a diverse portfolio spanning confectionery, snacks, beverages, dairy, pet care, and more. To stay competitive, Nestlé needs to understand:

1. What innovations are we developing internally? (their own patent portfolio)

2. What are our competitors working on? (competitive intelligence)

Your analysis will provide actionable insights to Nestlé's innovation and strategy teams.

Task Assignments

Intern 1: Internal Portfolio Analysis

Question: What is Nestlé good at? Where are we innovating?

You will receive a dataset of approximately 1,000 patents filed by Nestlé across various domains (food technology, packaging, manufacturing, etc.).

Your goal is to:

- Discover the main innovation themes within Nestlé's portfolio

- Identify clusters of related patents
- Provide strategic insights: What are Nestlé's core innovation strengths? Are there emerging areas of focus?

Intern 2: Competitive Intelligence Analysis

Question: What are our competitors doing? Where should we be paying attention?

You will receive a dataset of approximately 1,000 patents filed by three major competitors:

- PepsiCo (snacks, beverages)
- Mars (confectionery, pet care)
- General Mills (cereals, snacks)

Your goal is to:

- Discover the main innovation themes across competitors
- Identify clusters of related patents
- Provide strategic insights: What are competitors investing in? Are there white space opportunities for Nestlé?

Your Dataset

Each intern will receive a CSV file containing approximately 1,000 patent records with the following fields:

Field	Description
publication_number	Unique identifier for the patent
title	Patent title
abstract	Brief summary of the invention
claims	Legal claims defining the scope of the patent
ipc_classification	International Patent Classification codes
publication_date	Date the patent was published
assignee	Company or entity that owns the patent
legal_status	Current status (e.g., active, expired, pending)

Task

Step 1: Understand and explore your dataset

Learn what's inside the patent data:

- Get uv set up as your package manager (https://docs.astral.sh/uv) and install project dependencies with `uv sync`

- Examine the structure
- Understand the columns
- Review text fields

This may include:

- Loading the dataset into Python (e.g. with Pandas)
- Check for missing values, inconsistencies, etc.
- Cleaning and preparing the data for further analyses

Clean this dataset and produce a short *Data Summary* for your presentation.

Step 2: Use Azure OpenAI API

Generate text embeddings and learn how to interact with modern LLM endpoints.

You will:

- Review the API key and endpoint structure
- Write a simple Python script to test the connection (you can use `openai-quickstart/example_embeddings.py ` as a starting point)
- Generate embeddings using an appropriate model
- Store them in your dataset and verify their shape and behaviour
- Optionally run simple similarity checks to see how embeddings behave

This step is essential for clustering and semantic analysis.

Step 3: Build an unsupervised clustering method

Using your embeddings you will:

- Choose a clustering method (the scikit-learn documentation has a few good starting points)
- Apply the method to your embedding vectors
- Inspect the output and review cluster characteristics (how many clusters were formed, are they balanced in size, do they make sense when you check which patents have been clustered?)

Once clusters are generated, use an LLM to:

- Summarise the patents within each cluster
- Generate a clear, human-readable cluster label
- Describe any key themes or topics
- **Note**: You may wish to view and run `openai-quickstart/example_llm.py` for how to call the Azure OpenAI API endpoint

Step 4: Build a Streamlit dashboard

Streamlit is a Python framework for quickly building interactive web apps. You'll use it to make your analysis accessible and engaging. Here are a few suggestions for your dashboard, but feel free to get creative and add your own features!

- An overview page with summary statistics, sample patents and exploratory visuals
- A clustering page with scatterplots, cluster summaries and labels, and filtering options

You may also, if you have time, wish to add a patent detail view, or a keyword search function, etc. Think about your client and what they might find useful to see.

Helpful Resources:

- Streamlit Documentation
- Streamlit Gallery for inspiration

Step 5: Present your findings

On Friday, you will give a short slide presentation showcasing your work

Your presentation should include

- The project goal
- A dataset summary
- Your methodology
- Embedding and clustering explanation
- Insights per cluster
- Any recommendations you may have to the fictional client
- A personal reflection on your experience at AIM

Learning outcomes

By the end of the week, you will have gained practical experience with

- Working with real-world messy text data
- Making API calls using Azure OpenAI
- Understanding embeddings and why they matter
- Applying basic unsupervised ML
- Using LLMs for automatic cluster naming and summarisation

- Building a simple Streamlit web app
- Communicating technical insights to business audiences

Timeline

The AIM Reply team will always be available should you need help, both in our daily standup sessions at 11am, and on teams.

Monday - On your first day we will introduce the team members, give a tour of the office and help you to settle in before introducing this task.

Tuesday - On Tuesday our CEO David Semach will give you a more detailed introduction into AIM Reply and our goals.

Wednesday - Wednesday we have a mid-week checkpoint to discuss your progress, any blockers and next steps

Thursday – Thursday will be an opportunity to work on your final presentation and catch up with team members.

Friday - Friday we have a meeting scheduled for your final presentations and wrap up of the micro-internship

Best of luck!