Chair of Financial Management and Capital Markets
TUM School of Management
Technical University of Munich

Automated Text Extraction Pipeline for Firm Annual/Sustainability Reports

Keywords: Textual Analysis, Sustainability Reports, NLP, Biodiversity Risk

Project description

The aim of this IDP is to develop a robust and scalable pipeline for text extraction from annual and sustainability reports of listed European firms. This serves as the technical foundation for our ongoing research project on biodiversity and sustainability disclosure shared among Chair of Financial Management and Capital Markets (Prof. Dr. Christoph Kaserer) and the Center for Digital Transformation (Prof. Dr. Sebastian Müller).

The IDP project consists of three main components:

1. Build an automated extraction pipeline for sustainability reports

a. Collect sustainability reports

- Acquire and organise PDFs on annual/sustainability/ESG/CSR reports from standard data sources (e.g., LSEG, company websites).

b. Develop Python-based text parser (extraction) tools

- Extract raw text from PDFs using open-source libraries
- Handle complex layouts, including:
  - multi-column formats
  - tables
  - embedded images and charts
  - headers/footers and noisy metadata
- Ensure consistent sentence-level extraction while preserving document structure.

c. Clean and preprocess text

- Remove noise (tables, picture captions, page numbers, artefacts).
- Standardise encoding and apply basic NLP preprocessing (sentence correction, tokenisation, lemmatisation).
- Identify meaningful sections (using segmentation, heading detection, heuristics or embeddings).
- Output a clean, standardised sentence-level dataset for downstream NLP tasks.

2. Prepare data for advanced NLP analysis

- Create a consistent schema for storing text, metadata, and company–year identifiers.
- Store segmented text together with extracted metadata (e.g., industry, reporting framework, page references).
- Ensure smooth integration with downstream tasks (e.g., biodiversity-related sentence classification).
- Optional: apply embeddings or LLM-based methods for section classification or topic discovery.

3. Technical documentation and reproducibility

- Write concise and clean Python code.
- Provide clear and detailed documentation and a simple guide so that future researchers (PhD/postdocs) can reuse and scale the pipeline.

What we are looking for

- Strong IT programming skills
- Interest in text processing, NLP, or data engineering
- Ability to manage and structure large unclean datasets (PDFs, text files)
- Reliability, independence, and good project organisation
- Interest to learn something about finance and biodiversity risk

What we offer

- Kick-off session including introduction to relevant finance and/or business topics
- Experience with IDPs
- Open dialogue and support
- Potential for follow-up work
- Both single and group projects are possible

Interested?

Please send an e-mail with CV, academic transcript and your preference for this project to aida.cehajic@tum.de.

Questions?

In case of any (e.g. topic related) questions, please contact aida.cehajic@tum.de