The Journal of Open Source Software

Lextract: A Python Pipeline for the Automated Extraction of European Commission Market Definitions

Shriyan S. Yamali [1]

1 Newark Charter High School, United States

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Lextract is a Python pipeline that automatically locates, downloads, and extracts relevant market definitions from the European Commission's merger and antitrust decision PDFs. Relevant market definitions establish the specific scope of competition legislation and identify the specific set of products in an area (Tsangaris, 2017), making them indispensable for economists, lawyers, and regulators when determining the effects of mergers and evaluating anticompetitive behavior (Patakyová, 2020). This pipeline has been designed for researchers and competition law experts who require a quick and scalable way to extract relevant market definitions from many cases at once. Considering that market definitions are highly sensitive and, to some extent, arbitrary, Lextract has been designed and implemented to extract definitions as accurately as possible, as a slight change in the language of a definition can drastically change its meaning.

Statement of need

Competition authorities routinely delineate the relevant market as a first step in merger and antitrust assessments. Scholars analyze this language to track precedent, analyze trends in the scope of definitions, and identify the evolution of market definitions (Robertson, 2019). Despite its significance, only one commercial product addressing this need exists: LexisNexis's Caselex Market Definitions Module which suffers from being proprietary, immutable, and inaccessible to many academics.

The Commission has published over 6,000 merger and antitrust decisions (Bernhardt & Dewenter, 2024; European Commission, 2025) and continues to add 280 more annually (Affeldt et al., 2021), each structured and formatted idiosyncratically, with inconsistent placement of definitions and headings that vary in language. As a result, deterministic approaches such as regex are brittle and ineffective (Wang et al., 2020), while manual extraction is slow and irreproducible at scale. This pipeline rectifies this issue by providing a simple, open source way to extract market definitions that does not require manual guidance nor rely upon inaccurate pattern-matching techniques.

General workflow

The general workflow for extracting market definitions is split into three sections and five steps. The first section involves the scraping of data, making use of regex: 1. A script processes an Excel file downloaded from the Commission's case search portal and extracts the links of decision documents and corresponding metadata (i.e., case number, year, policy area), saved in a plain text file. 2. Another script processes this file, and, using the decision document links, scrapes the decision text and converts it into a text corpus with the metadata, repeating this step for each link, while also sorting the corpus based on its length, with a breakpoint a 80,000 characters. It is during this process that decision documents without market definitions, identified by certain phrases or a page length less than three, are excluded.

The second section is responsible for the semantic extraction of market definitions: 3. Google Gemini is used to identify and extract only the section of the text corpus that contains the market definition section. 4. Afterwards, the process becomes more granular, with Gemini again being used, only this time to identify and isolate each individual market definition within those sections. Each definition is then tagged with a topic and saved in a structured JSON file, where each object contains all elements of the aforementioned metadata, a topic, and the market definition.

The third and final section improves the presentation of the data: 5. Each separate JSON file is cleaned to remove extraneous characters and then aggregated into a single file, which can then be used for research and analysis. By structuring the workflow this way, each processed case is consistently analyzed, reducing variability and improving accuracy. Lextract's code also maintains a high level of accuracy, substantiated by its comprehensive test suite with 94% code coverage.

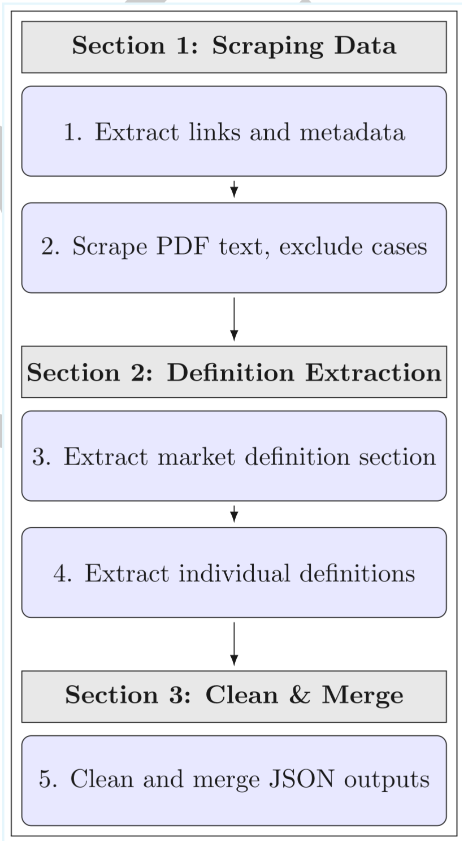

Figure 1: Workflow Diagram of Pipeline

Research applications

Lextract powers the database of JurisMercatus, an open source search interface that allows users to search for market definitions, leveraging the power of semantic search. The metadata provided by Lextract enables filtering by year, policy area, and case number. Further, this resource has the capability to support greater academic research and improve the accessibility of market definitions.

Limitations

It should be noted that this system, as with all systems, is not perfect and contains inaccuracies. First, with regards to step three, as a result of the previously mentioned fact that the heading used to identify the market definition section is inconsistently phrased, what constitutes the market definition is only heuristically defined, potentially leading to inaccuracies, especially when the language of decision texts deviates significantly from the expected pattern. Secondly, the quality and reliability of the extraction is limited to that of the input. In other words, should the input consist of missing pages or unconventional language, the model may be confused, leading to partial, hallucinated, or inaccurate results (Valentin et al., 2024). Additionally, though it is understood that decisions are adjudicated in many different languages, with the European Commission using multiple itself, the pipeline assumes that all decisions are provided in English and excludes all others, thereby limiting its application to other languages without at least a moderate amount of modification.

Lastly, while this pipeline makes use of Google Gemini, it is model-agnostic and, if properly refactored, could utilize any LLM, including commercially hosted models like OpenAI's or locally deployed ones such as LLaMA, Mistral, or DeepSeek. However, accuracy and consistency will vary significantly depending on model size and capabilities. Generally, smaller models, especially local ones without a sufficient context length or reasoning ability, will tend to hallucinate outputs, misidentify sections, or produce partial definitions (Sun et al., 2025).

Model Type	Accuracy	Context Length	Speed	Cost
Hosted L (eg. GPT-4o)	High	Very High	Moderate	High
Hosted S (eg. Gemini Flash)	Moderate	High	Fast	Moderate
Local L (eg. DeepSeek 67B)	Moderate	Medium	Slow	Low
Local S (eg. LLaMA 3-8B)	Low	Low	Moderate	Low

Table I compares the performance of different LLMs used to Extract Relevant Market Definitions. "L" = Large models (>30B parameters); "S" = Small models (<30B parameters).

Acknowledgements

I am grateful to Professor Thibault Schrepel for his invaluable advisement and guidance throughout the course of this project. This research received no specific grant from any funding agency, university, or commercial or not-for-profit sector.

References

Affeldt, P., Duso, T., & Szücs, F. (2021). 25 Years of European Merger Control. *International Journal of Industrial Organization*, *76*. https://doi.org/10.1016/j.ijindorg.2021.102720

Bernhardt, L., & Dewenter, R. (2024). The Impact of the More Economic Approach on EU Merger Decisions. *Stanford Computational Antitrust*. https://law.stanford.edu/wp-content/uploads/2024/12/Bernhardt-Dewenter.pdf

European Commission, D.-G. for C. (2025). *European Commission Competition Case Search*. https://competition-cases.ec.europa.eu/

Patakyová, M. T. (2020). Competition Law in Digital Era – How to Define the Relevant Market? *4th International Scientific Conference on Economics & Management*. https://doi.org/10.31410/EMAN.2020.171

Robertson, V. (2019). The Relevant Market in Competition Law: A Legal Concept. *Journal of Antitrust Enforcement*, *8*(2). https://doi.org/10.1093/jaenfo/jnz008

Sun, C., Li, Y., Wu, D., & Boule, B. (2025). *OnionEval: An Unified Evaluation of Fact-conflicting Hallucination for Small-Large Language Models*. https://doi.org/10.48550/arXiv.2501.12975

Tsangaris, P. (2017). Competition law enforcement. In *Capacity withdrawals in the electricity wholesale market: Between competition law and regulation* (pp. 37–103). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-55513-2_3

Valentin, S., Fu, J., Detommaso, G., Xu, S., Zappella, G., & Wang, B. (2024). *Cost-effective hallucination detection for LLMs*. https://www.amazon.science/publications/cost-effective-hallucination-detection-for-llms

Wang, P., Brown, C., Jennings, J. A., & Stolee, K. T. (2020). Proceedings of the 17th international conference on mining software repositories. *An Empirical Study on Regular Expression Bugs*, 103–113. https://doi.org/10.1145/3379597.3387464