Lextract: An Python pipeline for the automated extraction of European Commission market definitions

Shriyan S. Yamali [1]

1 Newark Charter High School, United States

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 product in an area (Tsangaris, 2017), making them indispensable for economists, lawyers, and regulators when determining the effects of mergers and evaluating anticompetitive behavior (Patakyová1, 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 are accurately as possible, as a slight change in the language of a definition can drastically change its meaning. The process is split into five main steps: (1) fetch links, (2) exclude irrelevant decisions PDFs and scrape the text of the rest, (3) extract the market definitions section, (4) extract individual market definitions, and (5) clean and combine JSON files. This modular design allows for the scalable, reproducible extraction of market definitions with vast applications.

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 annually (Affeldt et al., 2021), each formatted idiosyncratically and lengthy. Headings vary in language and definition placement within decisions is inconsistent. As a result, deterministic approaches such as regex are brittle and ineffective, 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 relies upon inaccurate patter 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 form the Commission's case search portal and extracts the links of

40 decision documents and corresponding metadata (i.e. case number, year, policy area), saved in
41 a plain text file. 2. Another script processes this file, and, using the decision document links,
42 scrapes the decision text and convert them into a text corpus with the metadata, repeating
43 this step for each link, while also sorting the corpus based on its length, with a break point a
44 80,000 characters. It is during this process that decision documents without market definitions,
45 identified by certain phrases or a page length less than three, are excluded.

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

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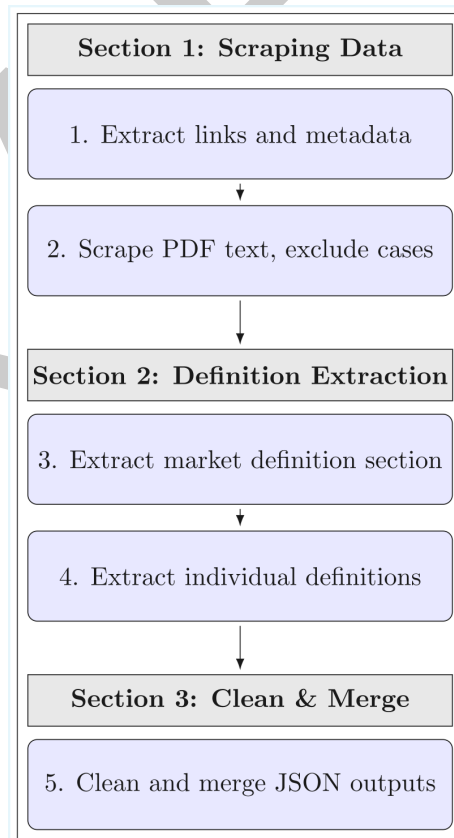58 coverage.

Figure 1: Workflow Diagram of Pipeline

59 Research applications

60 Lextract powers the database of JurisMercatus, an open source search intercace that allows
61 users to search for market definitions, leveraging the power of semantic search. The metadata

⁶² provided by Lextract enables filtering by year, policy area, and case number. Further, this
⁶³ resource has the capability to support greater academic research and improve the accessability
⁶⁴ of market definitions.

Limitations

⁶⁶ It should be noted that this system, as with all systems, is not perfect and contains inaccuracies.
⁶⁷ First, with regards to step three, as a result of the previously mentioned fact that the heading
⁶⁸ used to identify the market definition section is inconsistently phrased, what constitutes the
⁶⁹ market definition is only heuristically defined, potentially leading to inaccuracies, especially
⁷⁰ when the language of decision texts deviate significantly from expected pattern. Secondly, the
⁷¹ quality and reliability of the extraction is limited to that of the input. In other words, should
⁷² the input consist of missing pages or unconventional language, the model may be confused,
⁷³ leading to partial, hallucinated, or inaccurate results (Valentin et al., 2024). Additionally,
⁷⁴ though it is understood that decisions are adjudicated in many different languages, with the
⁷⁵ European Commission using multiple itself, the pipeline assumes that all decisions are provided
⁷⁶ in English and excludes all others, thereby limiting its application to other languages, without
⁷⁷ at least a moderate amount of modification.

⁷⁸ Lastly, while this pipeline makes use of Google Gemini, it is model-agnostic and, if properly
⁷⁹ refactored, could utilize any LLM, including commercially hosted models like OpenAI's or locally
⁸⁰ deployed ones such as LLaMA, Mistral, or DeepSeek. However, accuracy and consistency will
⁸¹ vary significantly depending on model size and capabilities. Generally, smaller models, especially
⁸² local ones without a sufficient context length or reasoning ability, will tend to hallucinate
⁸³ outputs, misidentify sections, or produce partial definitions (**?**).

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

⁸⁴ Table I compares the performance of different LLMs used to Extract Definitions. "L" = Large
⁸⁵ models (typically >30B parameters); "S" = Small models (typically <30B parameters).

Acknowledgements

References

⁸⁹ Affeldt, P., Duso, T., & Szücs, F. (2021). 25 Years of European Merger Control. *International*
⁹⁰ *Journal of Industrial Organization*, *76*. https://doi.org/10.1016/j.ijindorg.2021.102720

⁹¹ Bernhardt, L., & Dewenter, R. (2024). The Impact of the More Economic Approach on
⁹² EU Merger Decisions. *Stanford Computational Antitrust*. https://law.stanford.edu/
⁹³ wp-content/uploads/2024/12/Bernhardt-Dewenter.pdf

⁹⁴ European Commission, D.-G. for C. (2025). *European Commission Competition Case Search*.
⁹⁵ https://competition-cases.ec.europa.eu/

⁹⁶ Patakyová1, M. T. (2020). Competition Law in Digital Era – How to Define the Relevant
⁹⁷ Market? *4th International Scientific Conference on Economics & Management*. https:
⁹⁸ //doi.org/10.31410/EMAN.2020.171

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

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

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