From company websites to business research: Beyond words

Josep Domenech

Universitat Politecnica de Valencia

Company websites are rich sources of information

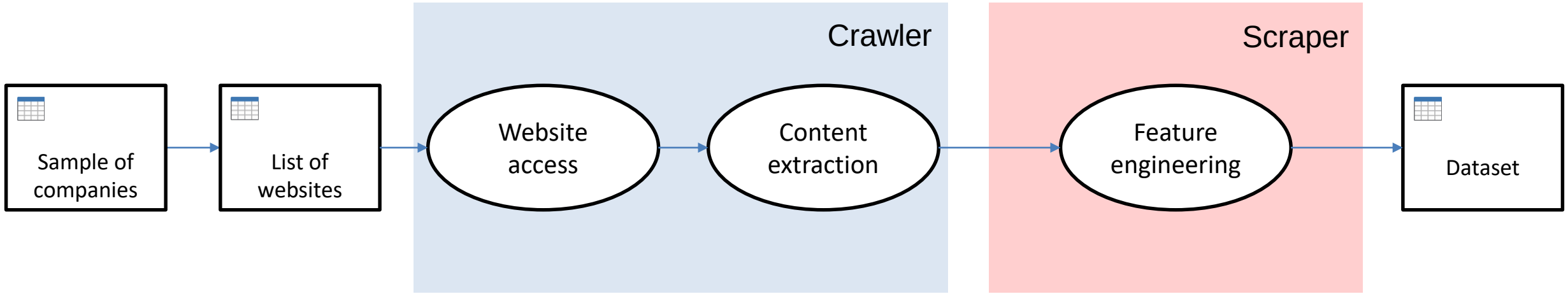
Main advantages

- Fresh information
- Very high granularity
- Wide coverage
- Non-intrusive
- Scalable
- Inexpensive

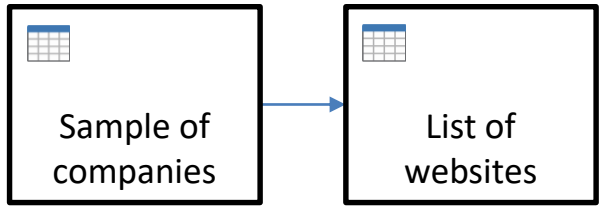
Limitations

- Sampling bias
  - Incomplete coverage
- Reporting bias
  - Most information is self-reported
  - Selective disclosure
  - Not conveninent for some topics
- URL changes
- URL ambiguity
- Language ambiguity

The process [1,9]

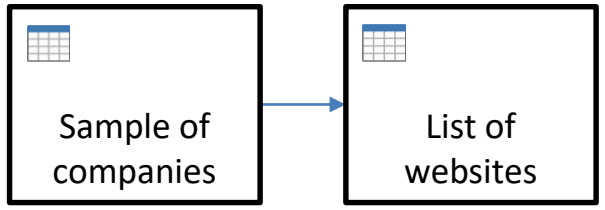


From the sample of companies to the list of websites



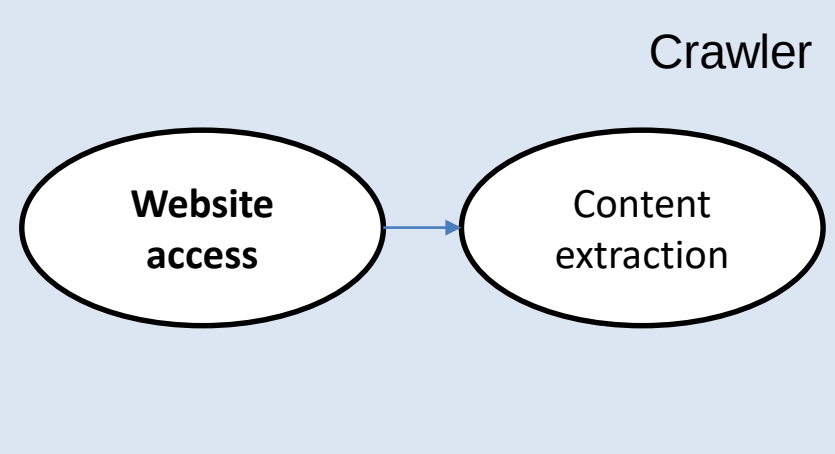
- Challenges
  1. What's the company's homepage URL?
     - Directories are not completely reliable
       - **Missing websites**: How to find them? → Search engines
       - **Old URLs**. HTTP redirects may help with this
       - **Inaccurate URLs**. Different approaches to check accuracy:
         - » Barcaroli et al. (2016)
         - » Bottai et al. (2022) [2]

From the sample of companies to the list of websites



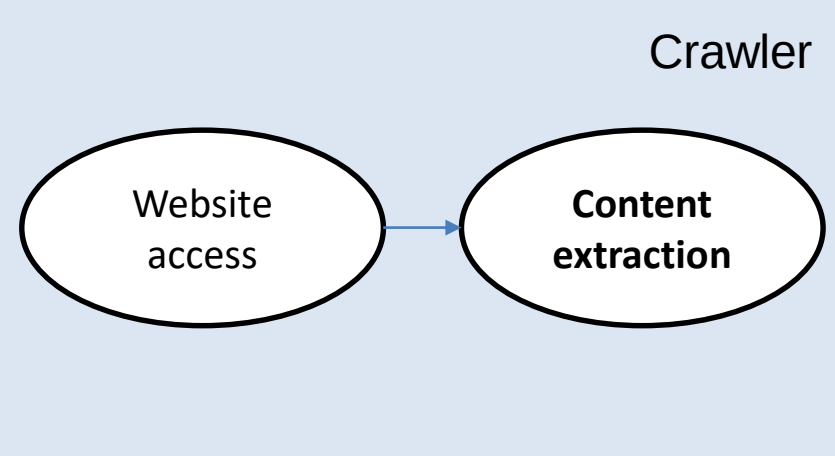
- Challenges
  2. Companies and websites are not one-to-one
     - One company with many websites
       - (e.g., one per brand)
     - One website for many companies
       - (e.g., franchisees)
     - Companies without website
       - No online activity or using third-party platforms

The Crawler



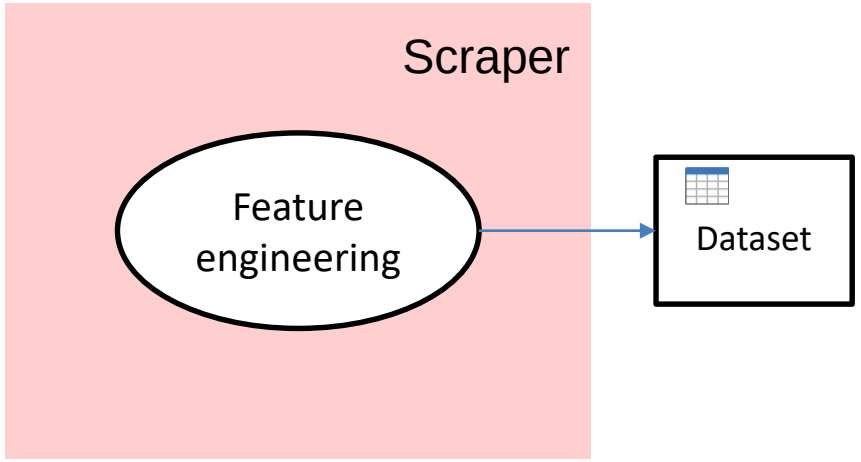
- Website access
  - Definition of website
    - Which domains and subdomains are included?
    - Which objects should be accessed? And processed?
      - Crawling depth
      - Content types of interest (generally HTML and PDF)
  - Technical challenges
    - Expired domains or unavailable servers
    - Browser technology (JavaScript or Flash)
    - Loops and content deduplication
    - CAPTCHAs (generally used by CDNs)
  - Ethical considerations
    - Netiquette
    - Avoiding server overload (and bans)

The Crawler



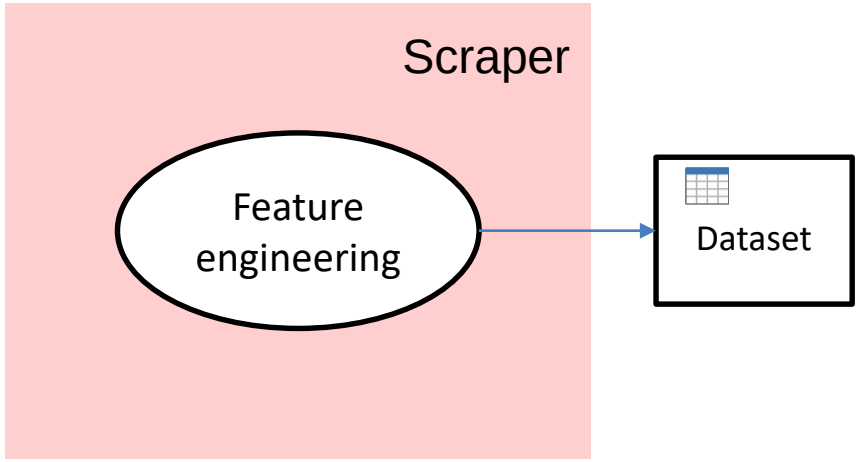
- Content extraction
  - Relevant contents:
    - DNS / Whois [3-7]
    - Server headers [1]
    - Content types
      - HTML
        - » **Text**[1-6,8-13]
        - » Links [1,9]
        - » Code [1-2,9-12]
      - PDF
        - » Text [8, 13]
        - » Metadata [8]
      - Other

The Scraper



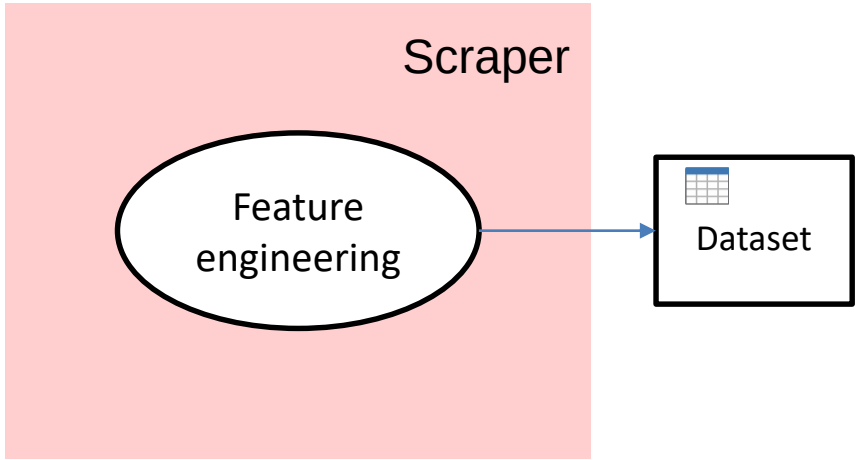
- Objective:
  - Create minable view from unstructured data
- Techniques for text (and code):
  1. Bag-of-Words[1-6,8-13]
     - Simple method: it just computes frequencies
     - Definition of a dictionary of words of interest
     - It may include a prior stemming process

The Scraper



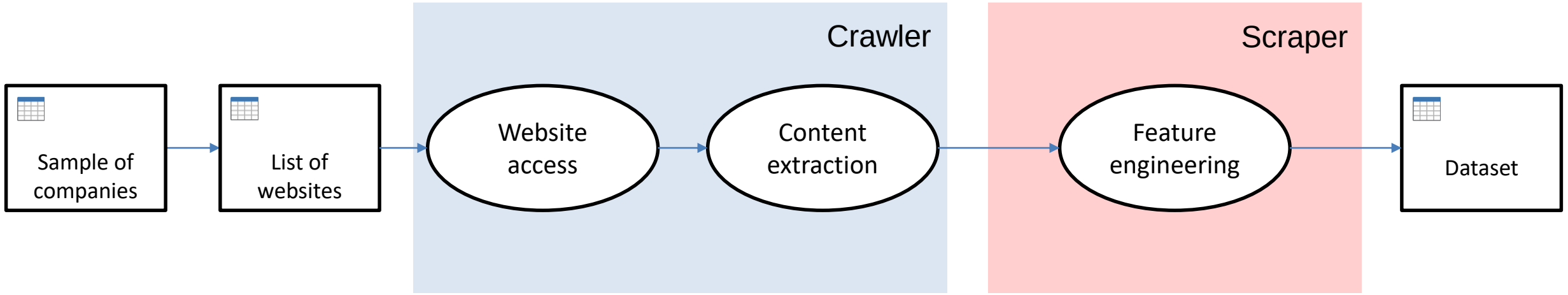
- Objective:
  - Create minable view from unstructured data
- Techniques for text (and code):
  2. Embeddings [*]
     - They may work at different level:
       - Word
       - Sentence
       - Document…
     - It transforms input into numeric vectors
     - Difficult interpretation
     - Good for clustering

The Scraper



- Objective:
  - Create minable view from unstructured data
- Techniques for text:
  3. Topic modeling (e.g., LDA)
  4. Transformers (e.g., BERT)
     - Applied to text classification

The process



How to consider a longitudinal approach?

Dynamics of companies through their websites

- Challenges regarding dynamic studies:
  - Access to website contents:
    - Past? → Wayback machine [2, 7, 21-23]
      - Works well for popular sites
      - Limited coverage both in terms of websites and time frequency
    - From now on
      - Which contents to crawl? All of them?
  - Storage of website contents:
    - Which contents to store?
    - Change detection?
  - Feature engineering of differences in contents: f(ΔW)
    - Or differences in features of contents: Δf(W) [10, 12]

Summary

- Company websites are a rich source of information about companies
- Sample Construction:
  - Challenges in mapping companies to websites
- Crawling Process:
  - Technical (e.g., expired domains, JavaScript-heavy sites) and ethical considerations
- Analysis Techniques:
  - Textual Content: NLP techniques from BoW to encoders.
  - Non-Textual Content reveals technological choices and .
- Dynamic vs. Static Analysis

References

1. Domenech, J., de la Ossa, B., Pont, A., Gil, J. A., Martinez, M., & Rubio, A. (2012). An intelligent system for retrieving economic information from corporate websites. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 573-578). IEEE.
2. Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2022, September). Unconventional data for policy: Using Big Data for detecting Italian innovative SMEs. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good* (pp. 338-344).
3. Domenech, J., Martinez-Gomez, V., & Mas-Verdú, F. (2014). Location and adoption of ICT innovations in the agri-food industry. *Applied Economics Letters*, 21(6), 421-424.
4. Blazquez, D., & Domenech, J. (2014). Inferring export orientation from corporate websites. *Applied Economics Letters*, 21(7), 509-512.
5. Rizov, M., Vecchi, M., & Domenech, J. (2022). Going online: Forecasting the impact of websites on productivity and market structure. *Technological Forecasting and Social Change*, 184, 121959.
6. Blazquez, D., & Domenech, J. (2018). Web data mining for monitoring business export orientation. *Technological and economic development of economy*, 24(2), 406-428.
7. Blazquez, D., Domenech, J., & Debón, A. (2018). Do corporate websites' changes reflect firms' survival?. *Online Information Review,* 42(6), 956-970.
8. Blazquez, D., Domenech, J., & Garcia-Alvarez-Coque, J. M. (2018). Assessing technology platforms for sustainability with web data mining techniques. *Sustainability*, 10(12).
9. Blazquez, D., Domenech, J., Gil, J. A., & Pont, A. (2019). Monitoring e-commerce adoption from online data. *Knowledge and Information Systems*, 60(1), 227-245.
10. Crosato, L., Domenech, J., & Liberati, C. (2021). Predicting SME's default: Are their websites informative?. *Economics Letters*, 204, 109888.
11. Crosato, L., Domenech, J., & Liberati, C. (2022). Non-conventional data and default prediction: the challenge of companies' websites. In *4th **International Conference on Advanced Research Methods and Analytics (CARMA 2022)***.
12. Crosato, L., Domenech, J., & Liberati, C. (2023). Websites' data: a new asset for enhancing credit risk modeling. *Annals of Operations Research*, 1-16.
13. Domenech, J., Garcia-Bernabeu, A., & Diaz-Garcia, P. (2023). Productivity, digital footprint and sustainability in the textile and clothing industry. In *5th **International Conference on Advanced Research Methods and Analytics (CARMA 2023)***.


CARMA 2024
26-28 June 2024
Internet and Big Data in Economics and Social Sciences
Valencia, 26-28 June
https://carmaconf.org/