KU LEUVEN

Automated URL retrieval for analyses of companies using Web Scraping and Machine Learning

by Koen Vanlommel and Nathan Vandeven
In collaboration with Statistic Flanders
Promotor: Prof. Dr. Bart Baesens
Daily supervisor: Manon Reusens

Research Questions

Scope & Relevance of our study

Context

- In 2023, 86.6% of European companies (more than 10 employees) had a website (Eurostat, 2023)
- Information available on websites can be very interesting for statistical research
  - Examples: price levels, tax numbers, addresses, company names,…
- By use of webscraping, these statistics can be transferred into key insights
- First up: finding the correct websites based on company names

Research Question 1

- Which search engine approach yields the highest performance?
- Options

Google Custom Search Engine

Google's Custom Search Engine

DuckDuckGo API

$$\text{Score} = \frac{1}{\text{rank}_{\text{Google}} \times \text{rank}_{\text{DuckDuckGo}}}$$

Multi-Search Engine Approach

Research Question 2

- What are the main challenges and limitations of web scraping and machine learning techniques?
- Building and validating the data collection pipeline
- Finding and overcoming challenges:
  - Rate limiting
  - Feature engineering
  - Data correctness
  - Missing data
- Selecting the best performing ML-models

Methodology & Results

Data Gathering

- Web Scraping Open Data KBO (Most Recent)
- Merge Useful Features for Analysis and Search Query
- Search Query Creation

EntityNumber	OfficialName	Abbreviation	ZipCode	Municipality	Street	HouseNumber	URL	NaceCode	SearchQuery
0206.653.946	Rijksinstituut voor Ziekte- en Invaliditeitsverzekering	RIZIV	1210	Sint-Joost-ten-Node	Galileelaan	5	www.inami.fgov.be	84301.0	Rijksinstituut voor Ziekte- en Invaliditeitsverzekering 1210 Sint-Joost-ten-Node

RQ1: Methodology

Google & DuckDuckGo Approach

- Skip Domain List
- API's Google and DDG
- Scrape First 5 Results

Multi-Search Approach

- Skip Domain List
- Combined API Usage
- Product of Rank (Top 5 is Scraped):

$$\text{Score} = \frac{1}{\text{rank}_{\text{Google}} \times \text{rank}_{\text{DuckDuckGo}}}$$

RQ1: Results

Missing Data

- Multi > DDG > Google CSE

	DDG_2	Google_2	Multi
Total Instances	17239	17239	17239
0 URLs	3	1642	0
1 URL	2	905	0
2 URLs	7	684	0
3 URLs	7	551	0
4 URLs	7	546	1
5 URLs	17213	12911	17238

Correct URL found

- Multi > DDG > Google CSE

Label	DDG	DDG_2	Google	Google_2	Multi
No URL	7983	7778	8784	9841	7019
URL1 match	8315	8560	7015	6643	7698
URL2 match	3493	4104	3319	2966	5368
URL3 match	2012	2739	2470	2304	3353
URL4 match	1300	2084	1401	1223	2737
URL5 match	1099	2022	949	939	2189

RQ2: Methodology

- Data Preprocessing
- Feature Transformation:
  - Length Metrics, Boolean Checks, Jaccard Similarity, Sequence Matching, Cosine Similarity, Levenshtein Distance, Character 3-Gram Overlap and Hamming Distance
  - Convert word similarity into numbers so the model understands
- Model Training & Evaluation
  - Models: LR, DT, RF, GB, SVM, NN and KNN
  - Performance metrics: F1 Weighted Average Score, Hamming Loss and Jaccard Score

RQ2: Results

DuckDuckGo

- Best Model: GB

Model	F1-Score (weighted avg)	Hamming Loss	Jaccard Score
LR	0.80	0.11	0.77
DT	0.71	0.15	0.63
RF	0.81	0.10	0.78
GB	0.82	0.10	0.79
SVM	0.81	0.10	0.78
NN	0.77	0.12	0.72
K-NN	0.78	0.12	0.75

Google

- Best Model: RF & GB

Model	F1-Score (weighted avg)	Hamming Loss	Jaccard Score
LR	0.86	0.07	0.85
DT	0.78	0.10	0.74
RF	0.87	0.06	0.86
GB	0.87	0.06	0.86
SVM	0.86	0.06	0.86
NN	0.83	0.08	0.81
K-NN	0.84	0.08	0.83

Multi

- Best Model: GB

Model	F1-Score (weighted avg)	Hamming Loss	Jaccard Score
LR	0.82	0.10	0.77
DT	0.73	0.15	0.62
RF	0.83	0.09	0.78
GB	0.83	0.09	0.79
SVM	0.83	0.10	0.78
NN	0.78	0.12	0.72
K-NN	0.80	0.11	0.75

Feature Importance

- Logistic Regression & Random Forest:
  - Levenshtein distance comparing company names to URLs boosts prediction power.
  - Domain names often match official company names.
- Decision Tree & Gradient Boosting:
  - Character n-grams are crucial for string similarity.
  - This metric significantly impacts model performance.
- Overall Findings:
  - URL features and similarity metrics enhance model prediction power.

Limitations & Future Research

Limitations

- Initial Dataset
- Search Query
- Rate Limiting
- Missing Data:
  - Special characters: # or !
  - Indexing differences
  - Keyword and content mismatching

Future Research

- Search Query selection
- Websites in multiple languages
- URLBERT: embeddings did not yet yield better results, could use extra research to better implement
- Hyperparameter tuning: Bayesian Optimization

KU LEUVEN

Thank you!
Questions?