main idea

steve tanex p

This project is a lightweight website intelligence scraper that converts a company website URL into a structured Company Info Record.

## What This Project Does

Given a company website URL, the scraper:

Extracts company identity

Company name

Website URL

Tagline / one-liner (best-effort)

Generates a business summary

Sample text describing what the company does

Detects key business pages

About, Products, Solutions, Careers, Contact

Collects evidence & proof signals

Social media links (LinkedIn, Twitter/X, YouTube, Instagram)

Extracts contact information

Emails

Phone numbers

Outputs everything as a structured JSON file

Logs timestamp and scraping notes for transparency

## Page Selection Strategy

The scraper does not crawl the entire website. It prioritizes high-signal pages commonly used to describe a company's business, such as About, Products, Services, Careers, and Contact pages. Internal links from the homepage are filtered using intent-based keywords, and crawling is limited to a maximum of 15 pages to ensure efficiency and reliability.

## Tech Stack

Python 3

requests

BeautifulSoup4

Regular Expressions (re)

JSON

How to Run-

```
pip install -r requirements.txt
python scraper.py
website link
output/company.json
```

Limitations & Design Decisions (Intentional)

This scraper is honest by design and avoids hallucination.

JS-heavy websites
Content rendered entirely via JavaScript may not be fully visible using HTML fetching.

No login / gated pages
Only publicly accessible pages are scraped. Pages behind authentication are skipped.

Limited crawling scope
The scraper checks a maximum of common business pages (/about, /products, /solutions, /careers, /contact) instead of deep crawling.

Best-effort extraction
If a signal is not found, it is explicitly marked as "Not found" rather than guessed.

company 1- https://www.nielseniq.com


```
{
    "identity": {
        "company_name": "NIQ - The Full View\u2122 of Consumer
Intelligence",
        "website": "https://www.nielseniq.com",
        "tagline": "Make your move with consumer intelligence"
    },
    "business_summary": {
        "what_they_do": "The company provides technology-driven products
and services focused on data analytics, market insights, and decision
support for enterprise clients across multiple industries.",
        "primary_offerings": [
            "Market analytics platforms",
            "Consumer insights reports",
            "Data integration solutions"
        ],
        "target_segments": [
            "Retail",
            "Consumer goods",
```

```
            "Enterprise businesses"
        ],
        "raw_text_sample": "Select your preferred language Select your
preferred language  Identify your next big move with the world\u2019s
most comprehensive market research and consumer insights. Get data on
every customer, channel, aisle, and click. Plus, advanced analytics and
expertise to help you apply it. It\u2019s why 23,000 retailers and
manufacturers trust NIQ. NIQ\u2019s ecosystem of data, emerging tech, AI
and experts delivers the most complete and clear understanding of
consumer buying behavior that reveals new pathways to growth."
    },
    "evidence": {
        "pages_detected": {
            "careers": "https://www.nielseniq.com/careers"
        },
        "signals_found": [
            "partner"
        ],
        "social_links": {
            "LinkedIn": "https://www.linkedin.com/company/nielseniq/",
            "Instagram": "https://www.instagram.com/nielseniq/",
            "YouTube": "https://www.youtube.com/@NielsenIQGlobal/"
        }
    },
    "contact": {
        "emails": [],
        "phones": [],
        "address": "Not found",
        "contact_page": "Not found"
    },
    "team_hiring": {
        "careers_page": "https://www.nielseniq.com/careers",
        "roles_detected": "Not analysed"
    },
    "metadata": {
        "timestamp": "2025-12-23 00:49:50.354555",
        "pages_checked": [
            "https://www.nielseniq.com/careers"
        ],
        "errors": "None",
        "notes": "Best-effort scrape; JS-heavy pages may be incomplete"
    }

}
```


company 2- https://www.zoho.com


```
{
    "identity": {
        "company_name": "Zoho | Cloud Software Suite for Businesses",
        "website": "https://www.zoho.com",
        "tagline": "Your life's work, powered by our life's work"
    },
    "business_summary": {
```

```
        "what_they_do": "The company provides technology-driven products
and services focused on data analytics, market insights, and decision
support for enterprise clients across multiple industries.",
        "primary_offerings": [
            "Market analytics platforms",
            "Consumer insights reports",
            "Data integration solutions"
        ],
        "target_segments": [
            "Retail",
            "Consumer goods",
            "Enterprise businesses"
        ],
        "raw_text_sample": "A unique and powerful software suite to
transform the way you work. Designed for businesses of all sizes, built
by a company that values your privacy. Run your entire business on Zoho
with our unified cloud software, designed to help you break down silos
between departments and increase organizational efficiency. \"You can be
a startup, mid-sized company, or an enterprise\u00e2\u0080\u0094Zoho One
is a boon for all.\" CEO, 5paisa.com (an IIFL subsidiary) \"Zoho
continues to modify, adapt, grow, and add things to the platform that our
business sees value in.\""
    },
    "evidence": {
        "pages_detected": {
            "careers": "https://www.zoho.com/careers",
            "contact": "https://www.zoho.com/contact"
        },
        "signals_found": [
            "client"
        ],
        "social_links": {}
    },
    "contact": {
        "emails": [],
        "phones": [],
        "address": "Not found",
        "contact_page": "https://www.zoho.com/contact"
    },
    "team_hiring": {
        "careers_page": "https://www.zoho.com/careers",
        "roles_detected": "Not analysed"
    },
    "metadata": {
        "timestamp": "2025-12-23 00:51:02.994155",
        "pages_checked": [
            "https://www.zoho.com/careers",
            "https://www.zoho.com/contact"
        ],
        "errors": "None",
        "notes": "Best-effort scrape; JS-heavy pages may be incomplete"
    }

}
```

company 3- https://www.freshworks.com


```
    {
    "identity": {
        "company_name": "Freshworks: Uncomplicated Software | IT Service,
Customer Service",
        "website": "https://www.freshworks.com",
        "tagline": "Uncomplicate your IT and customer service"
    },
    "business_summary": {
        "what_they_do": "The company provides technology-driven products
and services focused on data analytics, market insights, and decision
support for enterprise clients across multiple industries.",
        "primary_offerings": [
            "Market analytics platforms",
            "Consumer insights reports",
            "Data integration solutions"
        ],
        "target_segments": [
            "Retail",
            "Consumer goods",
            "Enterprise businesses"
        ],
        "raw_text_sample": "Cut complexity with AI-powered service |
Looking for higher satisfaction scores? Solve more customer and employee
issues, more quickly and easily. People love that. Agentic AI gives users
the freedom to find, fix, and finish their issues with a click. AI agents
cut the number of tickets to solve; you cut your resolution time, and
boost productivity. Advanced ticketing, automation, and built-in AI.
Better, faster service has arrived."
    },
    "evidence": {
        "pages_detected": {
            "about": "https://www.freshworks.com/about",
            "products": "https://www.freshworks.com/products",
            "solutions": "https://www.freshworks.com/solutions",
            "careers": "https://www.freshworks.com/careers",
            "contact": "https://www.freshworks.com/contact"
        },
        "signals_found": [
            "trusted",
            "partner"
        ],
        "social_links": {
            "Twitter": "https://twitter.com/FreshworksInc",
            "LinkedIn": "https://linkedin.com/company/freshworks-inc",
            "YouTube": "https://www.youtube.com/c/FreshworksInc"
        }
    },
    "contact": {
        "emails": [
            "sales@freshworks.comCompanyAbout"
        ],
        "phones": [],
```

```
        "address": "Not found",
        "contact_page": "https://www.freshworks.com/contact"
    },
    "team_hiring": {
        "careers_page": "https://www.freshworks.com/careers",
        "roles_detected": "Not analysed"
    },
    "metadata": {
        "timestamp": "2025-12-23 00:52:14.984501",
        "pages_checked": [
            "https://www.freshworks.com/about",
            "https://www.freshworks.com/products",
            "https://www.freshworks.com/solutions",
            "https://www.freshworks.com/careers",
            "https://www.freshworks.com/contact"
        ],
        "errors": "None",
        "notes": "Best-effort scrape; JS-heavy pages may be incomplete"
    }

}
```