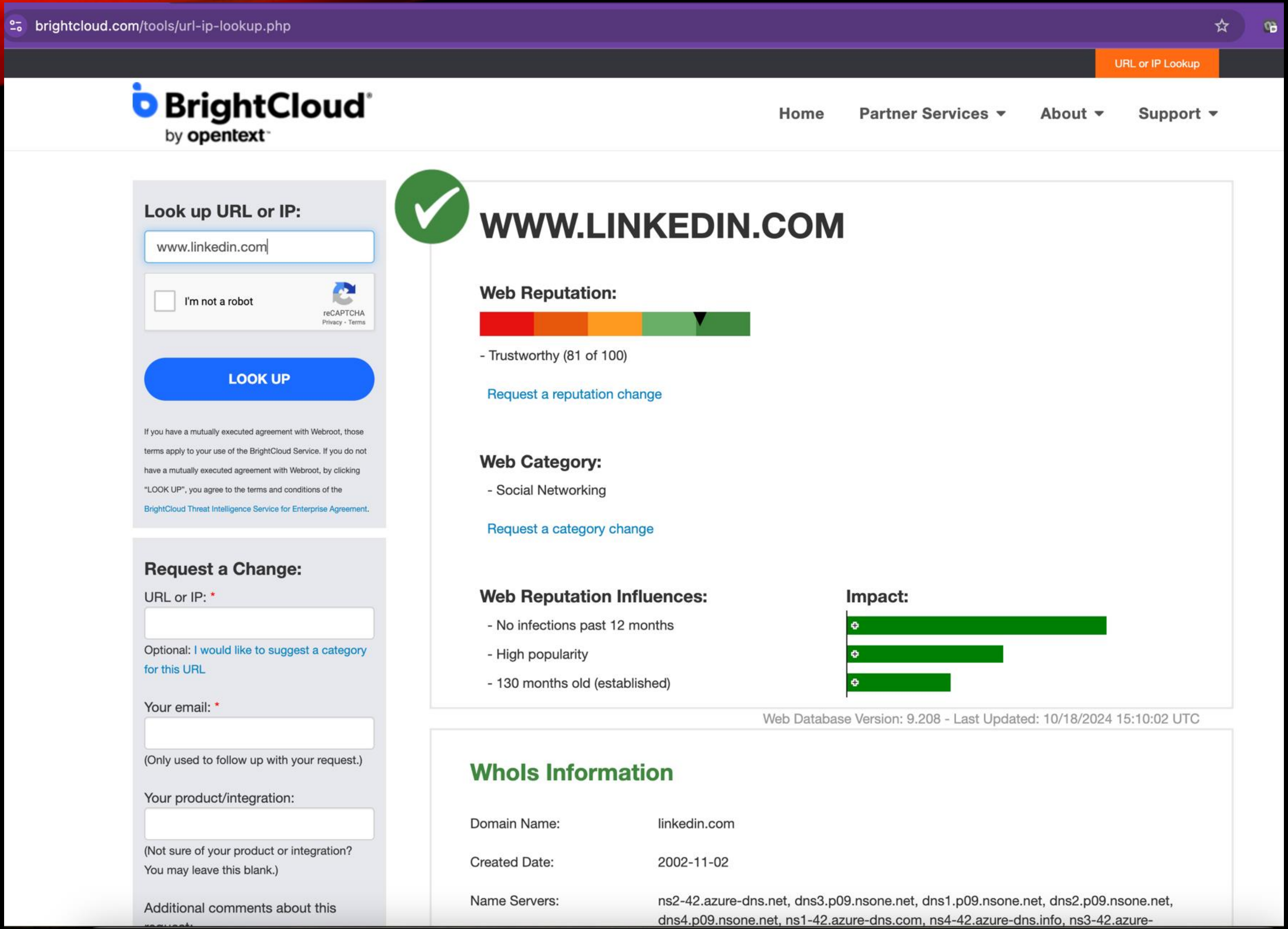
URL CLASSIFICATION

- Suman Kanukollu
- https://www.linkedin.com/in/suman-kanukollu/
- https://www.youtube.com/@sumankanukollu

INTRODUCTION

- Categorize & filter
Millions of WEB sites in real time?



Look up URL or IP:

www.linkedin.com

If you have a mutually executed agreement with Webroot, those terms apply to your use of the BrightCloud Service. If you do not have a mutually executed agreement with Webroot, by clicking "LOOK UP", you agree to the terms and conditions of the BrightCloud Threat Intelligence Service for Enterprise Agreement.

Request a Change:

URL or IP: *

Optional: I would like to suggest a category for this URL

Your email: *

(Only used to follow up with your request.)

Your product/integration:

(Not sure of your product or integration? You may leave this blank.)

Additional comments about this request:

WWW.LINKEDIN.COM

Web Reputation:

- Trustworthy (81 of 100)

Request a reputation change

Web Category:

- Social Networking

Request a category change

Web Reputation Influences:	Impact:
- No infections past 12 months	+
- High popularity	+
- 130 months old (established)	+

Web Database Version: 9.208 - Last Updated: 10/18/2024 15:10:02 UTC

WhoIs Information

Domain Name:	linkedin.com
Created Date:	2002-11-02
Name Servers:	ns2-42.azure-dns.net, dns3.p09.nsone.net, dns1.p09.nsone.net, dns2.p09.nsone.net, dns4.p09.nsone.net, ns1-42.azure-dns.com, ns4-42.azure-dns.info, ns3-42.azure-

SOLUTION OVERVIEW

- Collect data using web scraping (beautiful soup)
- Apply NLTK techniques
- Spacy (large model) extract 18-named Entities
- Sentence-transformers/all-MiniLM-L6-v2 (384-dim embedding vector)
- 786 Vector = 384-Title + 384-Text + 18-NE's
- Encoder only Transformer (with 6-multi-head attention blocks)
- Given a URL, this model processes the content and classifies it into categories like 'Business,' 'Shopping,' or 'News', 'Job Search' ….etc

DEMO (WITH CHROME EXTENSION)

REAL WORLD APPLICATIONS

1. WEB Security
2. Add targeting
3. SEO and spam detection
4. Parental control
5. Cyber Security

CONCLUSION

This project showcases the potential of transformers for efficient URL classification, with real world applications in security, marketing, content filtering and much more.