FRANCIS SHARON J

Final Project

XENON

A web scraper / summarization tool built with python

AGENDA

- PROBLEM STATEMENT
- PROJECT OVERVIEW
- END USERS
- SOLUTION AND ITS VALUE PROPOSITION
- THE ‘’WOW’’ FACTOR
- MODELING

PROBLEM STATEMENT

To analyze the web scraped DOM content and predict / categorize their intensity

PROJECT OVERVIEW

XENON is a web scraper and a sentiment detector built into a single unified system

1. XENON scrapes the DOM files (page content) from the specified URl
2. It summarizes and segments the content using an NLP engine
3. The segments are the sent to a sentiment classifier which categorizes the sentiments assign them a score based on their level of intensity

WHO ARE THE END USERS?

The target audience for XENON is **cyber security firms that specialize on reconnaissance** and **growing MSME firms**

YOUR SOLUTION AND ITS VALUE PROPOSITION

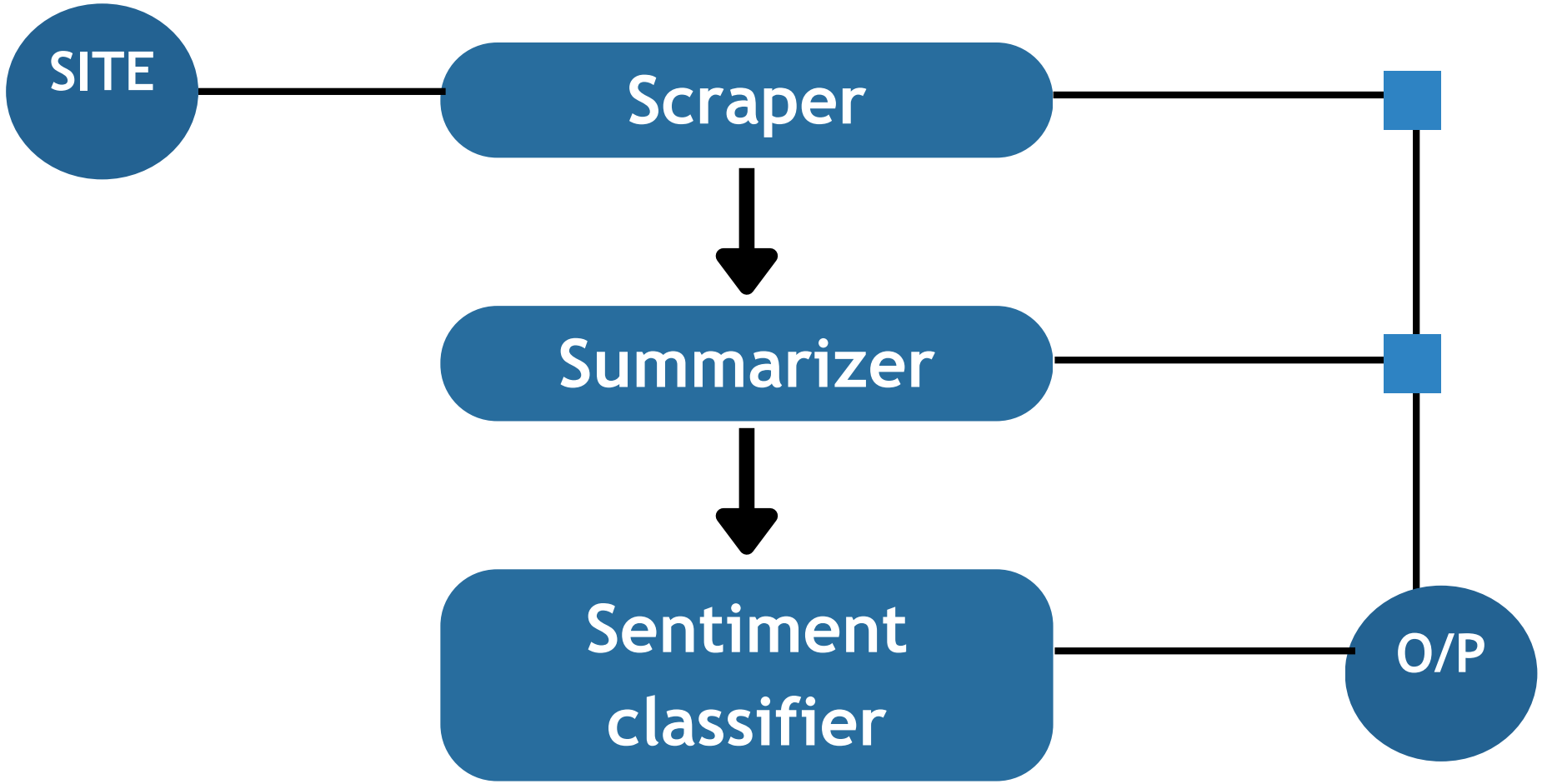
The XENON highlights are as follows,

- A really effecient scraping tool that takes an average of 2 seconds to completely scrape a 500 words long website
- An accurate summarizer that summarizes the contents and segments them based on the context
- An intuitive sentiment analysis tool that uses the context provided by the summarizer to effeciently predict the emotional intensity of the segmented content

THE WOW IN YOUR SOLUTION

Apart from scraping/summarizing sites available on the clearnet XENON can also scrape sites from the darknet. Thus by setting a threshold value, XENON can be used to effectively monitor a particular domain and define a notification if the content there gets out of hand

MODELING

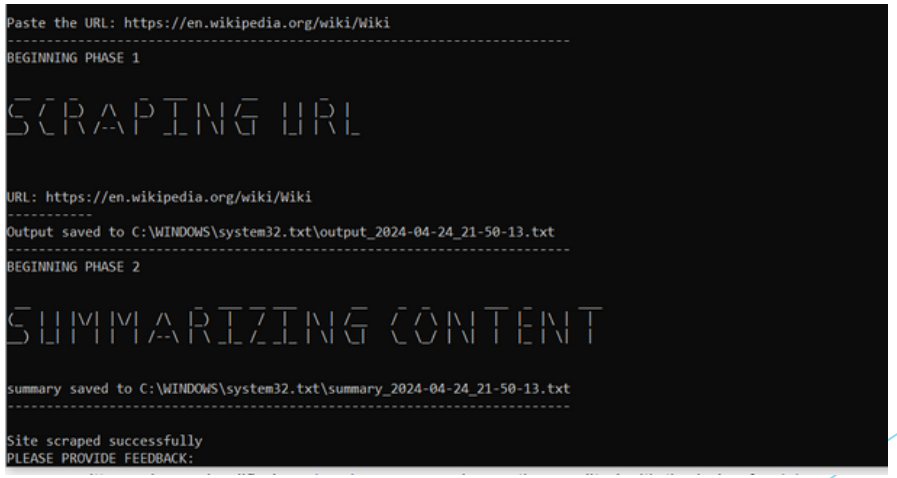


RESULTS

1


Administrator: Command Prompt - python D:\Desktop\XENON\xenon.py
[nltk_data] Package stopwords is already up-to-date!
THE SCRAPER
AND SUMMARIZATION FRAMEWORK
MODE-SELECTION
Press [0] for [[clearnet]] and [1] for [[darknet]] ->_

2


Paste the URL: https://en.wikipedia.org/wiki/Wiki
BEGINNING PHASE 1
SCRAPING URL
URL: https://en.wikipedia.org/wiki/Wiki
Output saved to C:\WINDOWS\system32.txt\output_2024-04-24_21-50-13.txt
BEGINNING PHASE 2
SUMMARIZING CONTENT
summary saved to C:\WINDOWS\system32.txt\summary_2024-04-24_21-50-13.txt
Site scraped successfully
PLEASE PROVIDE FEEDBACK: