**Real Estate Listing Scraper for Zillow**

**Project Overview:**

The goal of this project is to develop a web scraping tool using Scrapy to extract real estate listing data from Zillow, including property details, pricing, location, and images. The scraped data will be stored in a structured format for further analysis and use.

**Project Scope:**

**Target Website:**

[Zillow](https://www.zillow.com/)

**Data to be Scraped:**

1. Property Title

2. Price

3. Location (City, State, ZIP Code)

4. Number of Bedrooms

5. Number of Bathrooms

6. Property Description

7. Property Images

**Features:**

**1. Crawling Logic:**

- The spider should start from the main page and follow links to individual real estate listing pages.

- Implement logic to handle pagination and navigate through multiple pages of listings.

**2. Data Extraction:**

- Utilize XPath or CSS selectors to extract relevant information from the HTML structure of the pages.

- Handle cases where information may be missing or structured differently.

**3. Data Storage:**

- Store the scraped data in a structured format such as CSV, JSON, or a database.

- Include fields for each data point extracted.

**4. Error Handling:**

- Implement robust error handling to manage issues that may arise during scraping.

- Log errors and exceptions for debugging purposes.

**5. Respect Zillow's Policies:**

- Ensure compliance with Zillow's terms of service and scraping policies.

- Check and respect the `robots.txt` file for any crawling restrictions.

**6. User-Agent Rotation:**

- Implement user-agent rotation to mimic human-like behavior and avoid being blocked.

**7. Image Downloading:**

- Include functionality to download property images and store them locally.

**8. Data Cleaning and Validation:**

- Perform data cleaning and validation to handle inconsistencies in the scraped data.

- Ensure that the stored data is accurate and reliable.

**9. Scheduling:**

- Explore scheduling options to run the spider at regular intervals for periodic updates.

**Deliverables:**

1. Scrapy project directory containing the spider(s), settings, and other necessary files.

2. Source code for the spider(s) with inline comments for better understanding.

3. A README file explaining how to set up and run the spider, along with any additional instructions.

**Timeline:**

The project is expected to be completed within 2 weeks from the start date. Milestones and progress will be tracked weekly.

**Constraints:**

1. The web scraping tool must be developed in Python using the Scrapy framework.

2. Compliance with Zillow's terms of service and scraping policies is mandatory.

3. The tool should operate responsibly, avoiding any adverse impact on Zillow's servers.

**Risks and Mitigations:**

**- Risk:** Zillow updates its website structure.

**- Mitigation:** Regularly check and update XPath or CSS selectors as needed.

**- Risk:** Zillow blocks the scraper due to excessive requests.

**- Mitigation:** Implement user-agent rotation and use Scrapy's built-in mechanisms to handle delays.

**Approval:**

This requirement document is to be reviewed and approved by all stakeholders before development begins.

***Ahmad S.***