Medium.com Web Crawler with Django

SDLC Model – Waterfall Model

Timeline-

Phase	Date	Work
Requirement Analysis	03/06	Understanding requirements of the project
	04/06	Understanding prerequisites – technologies and framework
Design	04/06	Identifying core and additional features Determining the most suitable language and framework Determining which technology stack will be efficient and easiest to implement a web crawler Potential technology stacks: Spring Boot - more experience with java thus Easier Learning Curve Django - Limited experience and steeper learning curve Database used - SQLITE3
Implementation	05/06	Decided upon python with Django and started with the config Learning beautiful soup and locating elements using beautifulsoup Encountering Infinite scrolling website problem.
	05/06	Encountering Infinite scrolling website problem.
	06/06	Understanding use of pagination, Ajax, Jquery in websites for autoscrolling. Finding a solution with selenium. Discarded due to it being inelegant and unsuitable
	07/06	Deciding to Implement the solution using Scrapy library. Learning Scrapy.
	08/06	Locating medium(dot)com archives and implementing scraping on archives thus solving the infinite scrolling issue
	08/06	Learning Django and Scrapy's integration in Django Storing Scraped Data in sqlite3.

	09/06	Implementing makeshift Webpage to display scraped data. Extracting Data from sqlite3 DB and displaying it on a webpage. Serving Static Html pages on webpage
	10/06	Fine tuning Code and Frontend to make it more user friendly. Added dynamic elements to web page.
Testing	10/06	Unit Testing Integration Testing Fixing Bugs – Search Feature failing for 2nd search when scraping using tags
	11/06	Regression Testing and minor bug fixes