Dateparser Better Language Detection

GSOC 2021, Zyte

ABOUT ME

Personal Information

Name	Gavish Poddar
University	G.D Goenka University, Gurgaon
Major	Bachelor in Computer Application
GitHub	gavishpoddar
Portfolio Website	Link
Location	Assam, India
Time Zone	IST(UTC + 5:30)

BACKGROUND

I am an undergraduate student in my second year at GD Goenka University in Gurgaon, India, studying Computers. I first got familiar with computers and aspects of software development in high school, and have been programming for the last **5 years**. Seeing the advancements and ever-increasing utility of the field continued to fuel my desire to learn more. I was comfortable in Python and was involved in game development projects during my freshman year and wanted to learn something new.

Hence, in the pursuit of exploring other fields of computer science, I began to dive into the realm of **Web Spider** right from libraries like **Scrapy** to **Selenium, Ultimate Sitemap Parser,** and

Newspaper3k.

In my current year, I worked on some [React based portfolio builder], Data Science, and Visualization, Websockets. And a lot of scraping related libraries

PROJECT DESCRIPTION

As described on the idea page “Currently language detection is rudimentary and often causes incorrect interpretation of dates.” **Implementing an optional language detection library to improve language detection.**

In the last four weeks (6th to 9th week) if the above task is completed, fixing as much as possible the issues with the `search_dates` function. Currently, there are over 27 open issues.

PROJECT GOALS

1. Discussing and Testing different open-source language detection libraries.
2. **Integrating the language detection library** as an optional tool.
3. Fixing as many issues as possible `search_dates`

MENTORS

1. Marc Hernández
2. Adrián Chaves

COMMITMENTS

I have semester examinations between 24th April - 5th May (however, this is subject to change given the current COVID 19 pandemic).

My summer vacations start on May 07, 2021, as of now. I am keeping a close eye on the COVID19 pandemic situation as it may force the institute to change the dates for the same. I can easily devote about 50 hours per week till August.

Afterward (from August 1st week), my institute reopens and I will have to spare some time for regular academics, hence I could then devote about 35-40 hours a week, but according to my timeline, the primary and secondary goals would be almost complete by that time, so everything

is manageable throughout the timeline. I might spend a little bit less time during weekdays but I plan to make it up during weekends. Overall, I'll still be able to give 40 hours a week.

PRE-GSOC

- I have gone through the following material at least twice and am doing a revisit of the same currently.
    - Codebase
        1. dateparser/tests
        2. dateparser/search
        3. dateparser/languages
        4. Dateparser - the core itself
- I have also discussed issues and created PR
- I have even started interacting with and building a community of developers with people I have never met before.

GSOC

PROPOSED TIMELINE

TIMESPAN	PLAN
May 17th - May 28th (**Community Bonding Period**)	- Getting familiarized with the scrapinghub Team and Mentors - Setup GSoC Blog. - Make notes and workflow diagrams on - Contact & Collaborate with other GSoC students in scrapinghub.
May 28th - June 7th (**Community Bonding Period**)	- Discussing all the possible solutions and testing the integrations suggested on #612 and Stack Overflow solution
Coding Officially Begins	
June 8th - June 23rd **(Week 1 & Week 2)**	- Integrating the library as an optional tool for language detection in the code-base
June 24th - July 9st **(Week 3 & Week 4)**	- Creating settings for the optional library - Rectifying other dependent functions to support the changes - Complete the documentation of the

	entire project till now and prepare a submission for **Phase 1 Evaluations**
July 10th - July 17th **(Week 5)**	- Writing test's and documenting the optional language detect library and updating README - Creating PR
July 18th - July 25th **(Week 6)**	- Fixing Issue #856 - By adding code similar to sentence splitter in tests it solved the issue
July 26th - August 2nd **(Week 7)**	- Issue #846 & #833 are similar requesting custom date formats. - Implementing & Documenting custom date format and adding custom settings.
August 3rd - August 10th **(Week 8)**	- Fixing Issue #843 Adding support for month abbreviations in search_dates
August 11th - August 16th **(Week 9)**	- Fixing micro issues #706 - Writing tests for after the above fixes.
Phase 1 evaluations (July 12th - July 16th)	
July 14th - July 16th	- Submit PR for review - Documenting current developments
Phase 2 evaluations (August 16th - August 23rd)	
August 16th - August 23th	- Work on the initial reviews received on the PR - Finalise documentation - Refactor code if necessary - Buffer time - Wrap up - All the merged PRs/ changelog would be submitted to Google for evaluation
Final Submission and Evaluation - August 23rd	

*(**Note**: The timeline is to provide me an idea and keep my work up to pace. I shall strictly adhere to it and plan on submitting individual PRs so that it's easier to review. Tests are an important component of this project. I'll also write tests and document the code simultaneously.)*

Post GSoC & Stretch Goals

- If I finish sooner than expected, I plan to contribute to the **dateparser** and **Scrapy** project, helping out the selected student.
- I would continue fixing issues in this project's repository (since by that time I'll be well acquainted with the codebase and if possible implement some new features into upcoming releases).
  I would primarily start working on fixing the 429s Error with a middleware if that's still by then #4424

Experience with the proposed tech stack

For more than a year now, I have been working in the field of Software Development. Through these years, I have been primarily using Python.

It's been a really exciting journey of self-taught coding and I have been fortunate to work on several projects right from simple programs to full-blown libraries. I have played with all kinds of relevant libraries like scrapy, ultimate-sitemap-parser, Selenium, etc.

Why Me?

I'm an enthusiastic developer, with rich experience in working on team projects. Also, I've known the importance of the various phases of software engineering. While developing anything I do stress the design phase and testing phase. Also, I am a quick learner. It takes me very little time to go through particular documentation and use its API.

I document code while writing it. It really helps the future development process.

I am known to meet hard deadlines. Some of the projects that I have worked on spanned around 4-5 months.

I promise to adhere strictly to the deadlines mentioned in the above table. I know my capabilities perfectly and have formulated the timeline accordingly. I would communicate with my mentor every week and will discuss my progress, objectives, or any issue that I am facing.

This project will be a very good learning opportunity for me. It will be my first contribution to such a big organization and a noble initiative is impacting the healthcare sector.

I feel with the amount of time I've invested in Application Development and the experiences gained through the work I've done make me a strong candidate and a perfect fit for this project.

I'm confident with the set of skills I possess, I'll be able to manage and successfully complete the project within the proposed timeline.