Text Analysis with NLTK

Week 2

9 November 2020

CDCS Python Course Series

with Lucy Havens

Last Week

Text analysis

Analyzing unstructured data

Introduction to tools in Natural Language Toolkit (NLTK)

A Python library for text analysis

This Week

Structuring code around a research question

Using Python and NLTK

Using GitHub

A version control system

For Participants

Slides and assignment are uploaded to our team!

See the *Files* tab in the *Week 2 Assignment* channel

Research with NLTK

The building blocks:

- Word tokenization
- Sentence tokenization
- Paragraph segmentation
- Part-of-speech tags

Research with NLTK

Going further:

- Syntactic structure
  - For example: sentence structures represented as trees
- Shallow semantics
  - For example: named entity recognition, coreference resolution
- Dialogue and discourse
  - For example: dialog act tags, rhetorical structure

For more detail: https://www.nltk.org/book/ch06.html

Research with NLTK: Example

To identify **places** named in a text:

1. Tokenize words
2. Tokenize sentences
3. Tag parts of speech
4. Chunk or label to annotate named entities
5. Compare named places with places in a gazetteer

For more:
https://programminghistorian.org/en/lessons/geoparsing-text-with-edinburgh
https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da

DEMO

https://data.nls.uk/tools/jupyter-notebooks/exploring-edinburgh-ladies-debating-society/

Git and GitHub

Version control system

Code on your own computer

Local

Merge code with others in an online repository (“repo”)

Upstream

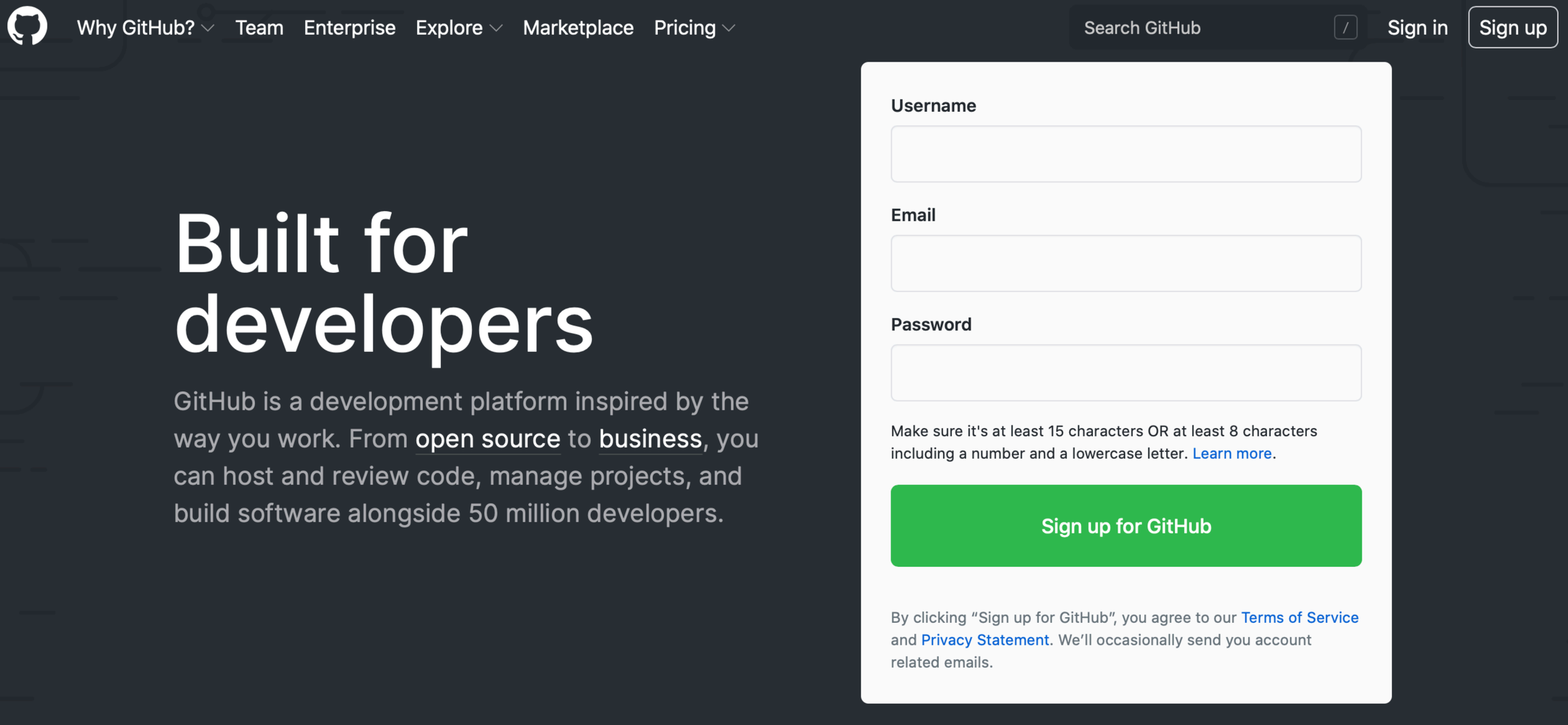
Reference: https://guides.github.com/introduction/git-handbook/

Git and GitHub

1. Create or clone into a repo
2. Create a branch off of the repo's `main` branch
3. Code locally
4. Add your code
5. Commit your code
6. [If coding with others] Open a pull request and review code
   [If coding on your own] Push your code.
7. [If coding with others] Merge your code

Reference: https://guides.github.com/introduction/git-handbook/

github.com


Why GitHub?
Team
Enterprise
Explore
Marketplace
Pricing
Search GitHub
/
Sign in
Sign up
Built for developers
GitHub is a development platform inspired by the way you work. From open source to business, you can host and review code, manage projects, and build software alongside 50 million developers.
Username
Email
Password
Make sure it's at least 15 characters OR at least 8 characters including a number and a lowercase letter. Learn more.
Sign up for GitHub
By clicking “Sign up for GitHub”, you agree to our Terms of Service and Privacy Statement. We’ll occasionally send you account related emails.

DEMO

Assignment

Helpful Resources - **for reference**, not required reading

Pick your own text to analyze! (ideas provided in case you're stuck)

Go Further - **optional** tutorials

Course2Week2 PDF file uploaded to the "Week 2 Assignment" channel of Teams

Thanks everyone!

Next course meeting: Friday, 10:00-11:00 AM BST

Office hours available on Wednesday (30 minutes)

To schedule, please message me on Teams!