Table of Contents

Sprint 1

Group Basic Information

- GitHub repository: Link
- Project Board: Link

Name	GitHub Id	Group Name
Ton That Tu Nguyen	tonthattuit	Headline Grabber

Individual Reflection

1. What you planned to do?

- Add new option "--target-dir" or "-td" for specifying destination folder of html report.
- Scrap headlines from Reuters website.
- Build unit test for scrapping functionality, specially for New York Times news verification.
- Export unit test report and code coverage report.

2. What you did not do?

- Initialize the desktop UI for this project.

3. What problems you encountered?

- The different IDEs need different configuration for import statement.
- Documentation for local setup which is not provided.

4. Issues completed

- HG12 - Implement logic to specify the output destination through a command line parameter
- HG20 – Unit test for implementation of parsing HTML data from New York Times
- HG24 - Add configuration for parsing of HTML data from REU (Reuters)

5. Files you worked on

- HG24 - Add configuration for parsing of HTML data from REU (Reuters)
  - .python-version
  - requirements.txt
  - src/headline_grabber.py
  - src/headline_grabber/**main**.py
  - src/headline_grabber/configurations/enums/scraper_engine.py
  - src/headline_grabber/configurations/sites/**init**.py
  - src/headline_grabber/configurations/sites/rtrs.yaml
  - src/headline_grabber/models/display_document.py
  - src/headline_grabber/models/pipeline_context.py
  - src/headline_grabber/pipeline_steps/classify_subject.py
  - src/headline_grabber/pipeline_steps/display_report.py
  - src/headline_grabber/pipeline_steps/filter_sites.py
  - src/headline_grabber/pipeline_steps/group_by_similarity.py
  - src/headline_grabber/pipeline_steps/pipeline_step.py
  - src/headline_grabber/pipeline_steps/prepare_for_display.py
  - src/headline_grabber/pipeline_steps/score_sentiment.py
  - src/headline_grabber/pipeline_steps/scrape_text.py
  - src/headline_grabber/pipeline_steps/text_similarity.py
  - src/headline_grabber/pipelines/**init**.py
  - src/headline_grabber/pipelines/pipeline.py
  - src/headline_grabber/validators/click/validate_site_name.py

---

- HG20 – Unit test for implementation of parsing HTML data from New York Times
  - .gitignore
  - README.md
  - pyproject.toml
  - requirements.txt
  - src/headline_grabber/configurations/sites/**init**.py
  - src/headline_grabber/pipeline_steps/**init**.py
  - src/headline_grabber/pipeline_steps/scrape_text.py
  - tests/headline_grabber/**init**.py
  - tests/headline_grabber/pipeline_steps/**init**.py
  - tests/headline_grabber/pipeline_steps/scrape_text_test.py
  - tests/headline_grabber/pipeline_steps/test_data/**init**.py
  - tests/headline_grabber/pipeline_steps/test_data/nyt.html
  - tests/headline_grabber/pipeline_steps/test_data/scrape_text_data.py

---

- HG12 – Implement logic to specify the output destination through a command line parameter
  - src/headline_grabber/**main**.py

- src/headline_grabber/models/user_preferences.py
- src/headline_grabber/pipeline_steps/display_report.py
- src/headline_grabber/pipelines/**init**.py
- src/headline_grabber/validators/click/option_validator.py
- src/headline_grabber/validators/click/validate_site_name.py

6. What you accomplished?

HG24 - Headline Grabber for Reuters

- Remove "src." from import statement to have project works with VS Code.
- Add rtrs.yaml for defining the scrapping items.
- Add requirements.txt for pulling dependencies with pip
- Add headless & disable gpu option for Selenium - Firefox Driver

HG20 - Unit test for parsing HTML data from NYT

- Add new libs for unit testing: pytest, pytest-html, pytest-cov.
- Introduce exception class ScrapeTextException.
- Refactor code for scraping text functionality.
- Update README.md for unit test information.
- Add unit test for parsing HTML of New York Times.
- Introduce libraries to export HTML unit test result report.
- Introduce libraries to scan and provide code coverage.

HG12 - Add new option for specifying target directory of report

- Add new option --target-dir or -td.
- Introduce a new class for Option validation which contains all of validation methods.
- Add a new validation for target path.
- Refactor the code for display_report.py.
- Remove the validate_site_name.py.