GOV_UK_Webpages

The workflow scrapes gov.uk webpages content and api information. Evaluates webpages content and performance.

Workflow runs once a week and produces four files for every scrape. Two first files contain information about most of gov.uk webpages, two next about 75 most important services. Each group contains file refered as predictions containing information about accessing service and webpage performance, and second file refered as cpm containing information about content performance.

Output

Workflow uses 'GOV.UK Content API' to get links to most of webpages on gov.uk domain. These links are used to create output inside 'all' folder. Links used to create output inside '75_services' folder are taken from 'links_75_services.csv' file. In each subfolder is created csv file with results after every scrape.

Example folder structure:

- all
    - predictions
        - 2023-04-17.csv
        - 2023-04-24.csv
    - cpm
        - ...
- 75_services
    - predictions
    - cpm

Variables of predictions files:

- **link** - relative link of webpage i.e. without prefix 'https://www.gov.uk/'
- **date** - date of webpage scrape in 'yyyy-mm-dd' format e.g. '2023-04-17'
- **primary_department** - primary department webpage belongs to, value taken from API results
- **Service Name** - name of service associated with webpage
- **format** - for 'all' folder content type of website, one of 'simple_smart_answer','form','licence','answer','guide','transaction', value taken from API; for '75_services' it is '75_services'
- **online_route** - This is output of trained model to determine more accurately whether the service can be used online, e.g. it enables services to be labelled as online (Y), even if they don't have a 'Start Now' button, but can be used online. Also, indicates a 'N' if they have start button but then redirect to a paper form.
- **phone_route** - This is output of trained model to determine more accurately whether the service has a phone channel. The service has a phone channel, either a helpline or some other phone number; suggests presence of customer support team or a telephone route to completing a transaction
- **post_route** - This is output of trained model to determine more accurately whether the service has a post channel. The service has an address provided, either for post or for face to face completion of the transaction

- **uses_gov_gateway** - Whether HMRC's identity verification system 'Government Gateway' is used by the service
- **number_of_paper_documents** - Number of documents attached to webpage, 'attached' means docs are provided in specific html form; for '75_services' documents being on webpages associated with service webpage (i.e. link to webpage is on service webpage) are added.
- **front_end_on_gov_uk** - If the page uses front end version of GOV.UK and version of it
- **number_of_email_addresses** - If webpage has 'mailto' html element. It is not working well.
- **service_requires_any_document_upload** - This is output of trained model to determine more accurately whether the service requires document upload. It is my subjective opinion that webpage suggests requirement for document upload.
- **service_requires_fill_document_and_upload_it** - The service requires the download, filling in and upload of a PDF form. Page has to contain words (download or pdf) and (fill or documents) and (scan or upload)
- **service_requires_document_upload_without_filling_it** - The service requires the scan or upload other than the download, filling in and upload of a PDF form. Page has to contain words (scan or upload) and cannot (download or pdf) and (fill or documents)
- **has_progress_button** - Whether the service page has an online progression button e.g. 'Start Now', button matching is done basing on html attribute
- **name_of_progress_button** - Name of an online progression button e.g. 'Start Now', 'Sign in', 'Apply Now', 'Buy Now'
- **service_link_under_progress_button** - webpage link redirected from progress button button
- **Start button takes to doc** - Whether progression button takes user to a document/ PDF that has to be printed
- **is_service_website_using_govuk_design_system** - Whether the service page (after progression button) uses GOV.UK 'govuk-template' (or not) for content design and features.
- **is_service_website_on_servicegovuk_domain** - If the service page (after progression button) is on service.gov.uk domain
- **front_end_service** - If the service page (after progression button) uses front end version of GOV.UK and version of it
- **is_link_to_service_broken** - Whether the service page (after progression button) is broken
- **PSI_metric_indicates_service_domain_needs_improvement** - PageSpeed insights: performance metrics detected some problems with user experience of website.
- **PSI_metric_indicates_service_domain_has_serious_issues** - PageSpeed insights performance metrics detected significant problems with user experience of website

Variables of cpm files:

- **link** - relative link of webpage i.e. without prefix 'https://www.gov.uk/'
- **date** - date of webpage scrape in 'yyyy-mm-dd' format e.g. '2023-04-17'
- **primary_department** - primary department webpage belongs to, value taken from API results
- **Service Name** - name of service associated with webpage
- **format** - for 'all' folder content type of website, one of 'simple_smart_answer','form','licence','answer','guide','transaction', value taken from API; for '75_services' it is '75_services'
- **Sentences over 25 words** - number of sentences with over 25 words
- **Sentences over 30 words** - number of sentences with over 30 words
- **Sentences over 35 words** - number of sentences with over 35 words

- **Paragraphs over 5 sentences** - number of paragraphs with over 5 sentences
- **Headers over 65 characters** - number of headers with over 65 characters (including the title)
- **Link availability issues** - number of statuses of all links with errors
- **Link accessibility issues** - number of same URLs with different anchor text; same anchor text with different URLs; one word links
- **PDF count** - number of PDF attachments
- **Words to avoid** - number of words to avoid ("Machine readable style guide.xlsx")
- **Flesch reading ease** - number of flesh reading ease
- **Bold text** - number of detect bold text (excluding table headers)
- **Subjectivity** - subjectivity (using sentiment analysis)
- **Time sensitive references number** - number of time sensitive references ("This year" etc)
- **Anchor text ending with full stop** - number of links which anchor text ending with full stop
- **Walls of text** - number of walls of text (>250 words without bullets or other formatting)
- **Basic errors in opening paras** - number of basic errors in opening paragraphs (errors in links, sentences over 25 words, paragraph more than 5 sentences)
- **Read more type link section** - webpage ending with a 'Read more' type section of links

Steps during workflow execution

1. Using 'GOV.UK Content API' links to webpages with some basic information are downloaded. File 'links_75_services.csv' is added to previous files. Files are saved to 'General_Scrape' folder.
2. For all links from previous step usinng 'GOV.UK Content API' detailed information about links is scraped and html content of webpages. Files are saved to 'Files_To_Download' folder.
3. With data from previous step features and CPM metrics of particular webpage are extracted. Files are saved to 'Data_Scrape' folder.
4. Most useful features from previous step are selected and are used to generate ML models predictions and rule based results. Files are saved to 'Transform_Files' folder.
5. Predictions results and CPM metrics are formated and saved to bucket with final results.

ML models and rule based results information

ML models were created with manually labelled dataset(over 500 labelled webpages) to train and test performance of rules and models predicting output metrics. Metric I wanted to maximise was Cohen's kappa score due to high imbalance in targets distributions. Kappa score can range from 0 and 1. Anything what gives kappa score lower than 0.5 might be considered as worse than random choice and should not be used.

1. Start button – Kappa=0.99. Best results were achieved used rule-based approach. Mistakes are due to mislabelling caused by updated webpages over time.
2. Start button takes to doc – Kappa=1. Best results were achieved used rule-based approach.
3. GOV Gateway/Verify – Kappa=1. Best results were achieved used rule-based approach.
4. Online – Kappa=0.75. Best results were achieved used ML models.
5. Use Phone – Kappa=0.82. Best results were achieved used ML models.
6. Use Post – Kappa=0.74. Best results were achieved used ML models.
7. N documents – Kappa=0.96. Best results were achieved used rule-based approach. Mistakes are due to mislabelling caused by overlapping documents having different formats.
8. Upload doc – Kappa=0.92. Best results were achieved used ML models.

Pickled models:

- online.pkl
- phone.pkl
- post.pkl
- upload.pkl

Useful information about sources/method used

1. Front end of webpages is detected using fe-detection repository, only bin/fe-test.sh is needed. Front end version is updated with time, thus to enable workflow to detect the most recent version of front end one needs to update fe-test.sh file in this repository.
2. PSI metrics are extracted from PageSpeed Insights.
3. Original CPM documentation
4. Original CPM colab notebook
5. Original Machine readable style guide used to extract information in CPM task.
6. Notebook '10. Overall metrics.ipynb' contain code used to generate overall metrics for digital assessment of service and CPM. Not used currently, left for reference.