Datasheet: qs-ocrized-text

Motivation

For what purpose was the dataset created?

At Quicksign, we thrive to deliver smart and powerful document analysis tools for digital onboarding. To do so, our R&D team applies state-of-the-art research from computer vision, deep learning and natural language processing. We work at lot with optical character recognition (OCR) and although we deploy lots of efforts to make it as accurate as possible, we sometimes have to deal with noisy text due to recognition errors. To our surprise, very few public datasets for text classification address this problem. From IMDB and Amazon reviews to Toxic Tweets classification, existing datasets deal with user-generated content which can be considered "clean".

Leveraging a previous dataset of more than 400,000 annotated document images, we applied Tesseract OCR to generate two new text datasets. We reuse the existing classification labels. By combining the generated text files and the existing labels, this repository constitutes a new text classification dataset. We hope this help the field go further into automated document image analysis.

Who created this dataset and on behalf of which entity?

This dataset was built by Nicolas Audebert, Catherine Herold and Kuider Slimani while employed in the Quicksign Research and Development team on behalf of Quicksign. The original document images dataset from which the texts have been extracted were created respectively by Adam Harley et al. at Ryerson University (RVL-CDIP) and Jayant Kumar et al. at University of Maryland (Tobacco3482).

Composition

What do the instances that comprise the dataset represent?

Each instance is a pair between a text document labeled with its type. Documents have been extracted from the Truth Tobacco Industry Documents archive which houses corporate documents that have been made public during litigation between the US governement and several major tobacco companies.

How many instances are there in total ?

The dataset consists in 399,999 + 3,482 = 403,481 text files and as many labels.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset does not contain all possible instances. It is a subset from the still-updated (as of April 2019) Truth Tobacco Industry Documents which contains more than 14 million unique documents and more than 90 million pages.

What data does each instance consist of?

Each instance is a text file encoded in UTF-8. The text was extracted from digitized documents using optical character recognition.

Is there a label or target associated with each instance?

Each text file is accompanied by a label indicating the document category it belongs to (e.g. "email" or "scientific report").

Is any information missing from individual instances?

One text file is missing from the QS-OCR-Large dataset since the corresponding image in the RVL-CDIP dataset was corrupted (`2500126531_2500126536.tif`). Some text files might be empty due to failures of the OCR: absence of detected text in the corresponding image. Otherwise, everything is included in the dataset.

Are relationships between individual instances made explicit (e.g.,users' movie ratings, social network links) ?

Due to the corporate natures of the documents, especially emails, some people and entities might be named and appear in several documents. No relationships between instances are explicited in the dataset and we are not aware of stronger relationships than just appearing the same corpus (i.e. the Truth Tobacco Industry Documents public archive).

Are there recommended data splits ?

The QS-OCR-Large comes with a predefined training/validation/testing split according to the one used by Harley et al. in their ICDAR'15 paper for the RVL-CDIP.

The QS-OCR-Small does not come with such a split and we recommend evaluating models using k-fold cross-validation.

Are there any errors, sources of noise, or redundancies in the dataset?

The dataset contains a significant part of noise due to the OCR processing. Spelling errors, missing words and spurious words are common. Some text files can be identical or near-identical due to the images containing originally the same text.

Is the dataset self-contained, or does it link to or otherwise rely onexternal resources (e.g., websites, tweets, other datasets)?

The text dataset is self-contained. However, its generation relies on the availability of the Tobacco3482 and RVL-CDIP datasets. For multimodal learning, e.g. document classification based on both text and image, these datasets are also required. The Tobacco3482 dataset has been archived by the Internet Archive. The RVL-CDIP is only available through Google Drive as far we know.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?

This dataset only contains data that has been ruled publicly accessible and is already available in the Truth Tobacco Industry Documents archive.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

Not that we know of.

Does the dataset contain data that might be considered sensitive in any way?

The dataset contains data regarding the internal organization of tobacco companies, although these are already public on the Truth Tobacco Industry Documents archive.

Collection process

How was the data associated with each instance acquired?

The text was extracted using OCR and the labels were reused from the Tobacco3482 and RVL-CDIP datasets.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

The text was extracted from document images using the Tesseract OCR engine.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)

We do not know how Harley et al. and Kumar et al. sampled the images from the larger TTID archive.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g.,how much were crowdworkers paid)?

The dataset was built by employees of the R&D team at Quicksign.

Over what timeframe was the data collected?

Document images cover several years and therefore so do the texts. The dataset was built in 2019 over several weeks.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

Does the dataset relate to people?

This dataset relates to people in that the texts have been authored by people and might refer to others.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was obtained through the RVL-CDIP and Tobacco3482, which have been built using documents from the TTID archive.

Were the individuals in question notified about the data collection?

Unknown. Since documents have been made public during legal procedures (e.g. litigations) of which the involved institutions are aware.

Did the individuals in question consent to the collection and use oftheir data?

No, see previous question.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT-feature extraction, removal of instances, processing of missing values)?

No specific preprocessing was used. Tesseract was directly applied to the original TIFF images.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g.,to support unanticipated future uses)?

The text dataset is the raw data computed by Tesseract.

Is the software used to preprocess/clean/label the instances available?

Yes, on Github.

Uses

Has the dataset been used for any tasks already?

At the time of first release, the dataset has only been used internally at Quicksign.

Is there a repository that links to any or all papers or systems that use the dataset?

No.

What (other) tasks could the dataset be used for?

The dataset could be used for anything related to modeling or understanding OCRized documents. This includes self-supervised/unsupervised modeling of documents with plausible OCR errors.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Not that we are aware of.

Are there tasks for which the dataset should not be used?

The dataset should not be used to model generic noisy language. Human errors, especially when talking or writing text, do not follow the same distribution as OCR errors. OCR might confuse similarly looking characters such as “l” and “|” which is not something that a human might do when typing on a keyboard. Therefore, this dataset is only reprentative of OCRized text, not general natural language.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes, the dataset is publicly available on the Internet.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed (including code) on Github: https://github.com/Quicksign/ocrized-text-dataset. The dataset does not have a DOI.

Will the dataset be distributed under a copyright or other intellectualproperty (IP) license, and/or under applicable terms of use (ToU)?

The crawled data copyright from the TTID archive belongs to the authors of the documents. There is no license although this work depends on the previous publications:

- A. W. Harley, A. Ufkes, K. G. Derpanis, “Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval,” in ICDAR, 2015.
- J. Kumar, P. Ye and D. Doermann, “Structural Similarity for Document Image Classification and Retrieval”, in Pattern Recognition Letters, November 2013.

It is expected that these are cited when using this dataset to acknowledge their work in agregating and labeling the original document images.

Have any third parties imposed IP-based or other restrictions on thedata associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to thedataset or to individual instances?

Unknown.

Maintenance

Who is supporting/hosting/maintaining the dataset?

Nicolas Audebert is supporting and maintaining the dataset. The dataset is hosted on Quicksign’s public Github repository.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The recommended contact point is the Github repository issues.

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, bywhom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset might be updated depending on how OCR performance improves in the near future. News will be posted on the Github repository if this is the case.

Will older versions of the dataset continue to be supported/hosted/maintained?

Older versions stay available on the releases section of the Github repository. Obsolete version will be tagged as such.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Will these contributions be validated/verified? If not, why not? Is there a process for communicating/distributing these contributions to other users?

Others may do so and should contact the original authors about incorporating fixes/extensions. Pull requests are welcomed on the repository to include new information and contributions will be curated and merged by the authors.