

# ОЧИЩЕННЯ ТЕКСТОВИХ ДОКУМЕНТІВ

Originally: [Denoising dirty documents](#) @ Kaggle

Ірина Максименко

09.05.2016

**Optical Character Recognition** (OCR) is the process of getting type or handwritten documents into a digitized format. If you've read a classic novel on a digital reading device or had your doctor pull up old healthcare records via the hospital computer system, you've probably benefited from OCR.

OCR makes previously static content editable, searchable, and much easier to share. But, a lot of documents eager for digitization are being held back. Coffee stains, faded sun spots, dog-eared pages, and lot of wrinkles are keeping some printed documents offline and in the past.

This competition challenges you to give these documents a machine learning makeover. Given a dataset of images of scanned text that has seen better days, you're challenged to remove the noise. Improving the ease of document enhancement will help us get that rare mathematics book on our e-reader before the next beach vacation.

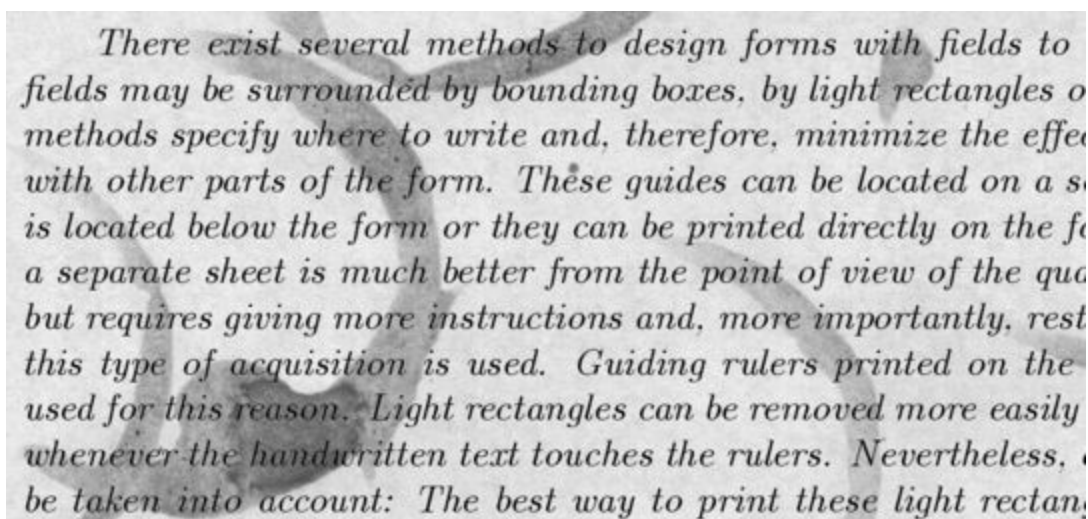
## ОПИС ПРОБЛЕМИ

Досить актуальною є ідея конвертації друкованих документів в електронний вигляд. Після сканування отримане зображення може містити різного роду пошкодженості: плями, згини та інший так званий “шум”. Звичайно, існують фільтри, які можуть бути використані для покращення якості зображення. Проте, метою даної роботи є дослідити можливості застосування алгоритмів машинного навчання для очищення таких сканованих документів і приведення їх до більш “читабельного” вигляду.

## DATASET

Дані для цього проекту, як і його ідея, були взяті з однойменного дослідження, розміщеного на сайті [www.kaggle.com](http://www.kaggle.com).

Ось кілька яскравих прикладів “забруднених” документів:



*There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles or methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a separate sheet is located below the form or they can be printed directly on the form. A separate sheet is much better from the point of view of the quality requires giving more instructions and, more importantly, restricts type of acquisition is used. Guiding rulers printed on the form are not a reason. Light rectangles can be removed more easily with filters than handwritten text touches the rulers. Nevertheless, other practical account: The best way to print these light rectangles is in a different*

There are several classic spatial filters for reducing frequency noise from images. The mean filter, the median filter, the opening filter are frequently used. The mean filter is a filter that replaces the pixel values with the neighborhood average, reducing the image noise but blurs the image edges. The median filter replaces the pixel value with the median of the pixel neighborhood for each pixel, thereby reducing salt and pepper noise. Finally, the opening closing filter is a mathematical morphology operation that combines the same number of erosion and dilation morphological operations to eliminate small objects from images.

The main goal was to train a neural network in a supervised manner to take a noisy image and output a clean image from a noisy one. In this particular case

Дані зображення містять англійський текст різного змісту, надрукований різними шрифтами і спотворений кількома видами “забруднень” - плями, згини, пом’ятість тощо.

Було використано два підходи до вирішення проблеми “очищення” таких зображень - кластеризацію і нейронну мережу.

## КЛАСТЕРИЗАЦІЯ

Ідея застосування кластеризації для даної задачі полягає в наступному:

- Завантажити зображення як матрицю, елементами якої є яскравості пікселів

- Для кожного пікселя визначити “фічі”, які будуть його характеризувати (його власна яскравість, середнє значення яскравості його сусідів, стандартне відхилення яскравості його сусідів, розкид яскравості його сусідів та ін)
- Розбити всі пікселі зображення на кластери відповідно до їх характеристик (“фіч”) використовуючи алгоритм k-means
- Знайти серед цих кластерів той, що відповідає за фон (він буде мати найвище середнє значення яскравості)
- Змінити значення яскравості пікселів таким чином, щоб висвітлити фон і дефекти, але залишити текст достатньо чітким для читання

Такий підхід має певні особливості. А саме, результат залежить від вибраних “фіч”, тому багато часу затрачається на підбір таких параметрів, щоб текст і фон потрапили в різні кластери. А також, потрібно правильно опрацьовувати пікселі різних кластерів, щоб текст залишився читабельним.

Переваги:

- Швидкість виконання
- Незалежність від типу “забруднення”
- Незалежність від розміру зображення
- Не потрібно попередньо навчати алгоритм
- Не потрібно збирати датасет

Недоліки / Можливі покращення:

- Велика залежність від вибору фіч
- Подекуди недостатня якість очищеного зображення



## РЕЗУЛЬТАТИ КЛАСТЕРИЗАЦІЇ

Початкове зображення:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the quality but requires giving more instructions and, more importantly, restricts this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles

Зображення, розбите на кластери:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the quality but requires giving more instructions and, more importantly, restricts this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles

Кінцеве, очищене зображення:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the quality but requires giving more instructions and, more importantly, restricts this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectangles

Початкове зображення:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the quality but requires giving more instructions and, more importantly, restricts this type of acquisition is used. Guiding rulers printed on the form are n reason. Light rectangles can be removed more easily with filters th handwritten text touches the rulers. Nevertheless, other practical account: The best way to print these light rectangles is in a differ

Зображення, розбите на кластери:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the quality but requires giving more instructions and, more importantly, restricts this type of acquisition is used. Guiding rulers printed on the form are n reason. Light rectangles can be removed more easily with filters th handwritten text touches the rulers. Nevertheless, other practical account: The best way to print these light rectangles is in a differ

Кінцеве, очищене зображення:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the quality but requires giving more instructions and, more importantly, restricts this type of acquisition is used. Guiding rulers printed on the form are n reason. Light rectangles can be removed more easily with filters th handwritten text touches the rulers. Nevertheless, other practical account: The best way to print these light rectangles is in a differ

Початкове зображення:

There are several classic spatial filters for reducing frequency noise from images. The mean filter, the median filter, and the opening filter are frequently used. The mean filter is a filter that replaces the pixel values with the neighborhood average, thereby blurring the image noise but blurs the image edges. The median filter replaces the pixel values with the median of the pixel neighborhood for each pixel, thereby reducing the image noise but blurs the image edges. Finally, the opening closing filter is a mathematical morphology operation that combines the same number of erosion and dilation morphological operations to eliminate small objects from images.

The main goal was to train a neural network in a supervised learning task to learn to map a noisy image to a clean image from a noisy one. In this particular case,

Зображення, розбите на кластери:

There are several classic spatial filters for reducing frequency noise from images. The mean filter, the median filter, and the opening filter are frequently used. The mean filter is a filter that replaces the pixel values with the neighborhood average, thereby blurring the image noise but blurs the image edges. The median filter replaces the pixel values with the median of the pixel neighborhood for each pixel, thereby reducing the image noise but blurs the image edges. Finally, the opening closing filter is a mathematical morphology operation that combines the same number of erosion and dilation morphological operations to eliminate small objects from images.

The main goal was to train a neural network in a supervised learning task to learn to map a noisy image to a clean image from a noisy one. In this particular case,

Кінцеве, очищене зображення:

There are several classic spatial filters for reducing frequency noise from images. The mean filter, the median filter, and the opening filter are frequently used. The mean filter is a filter that replaces the pixel values with the neighborhood average, thereby blurring the image noise but blurs the image edges. The median filter replaces the pixel values with the median of the pixel neighborhood for each pixel, thereby reducing the image noise but blurs the image edges. Finally, the opening closing filter is a mathematical morphology operation that combines the same number of erosion and dilation morphological operations to eliminate small objects from images.

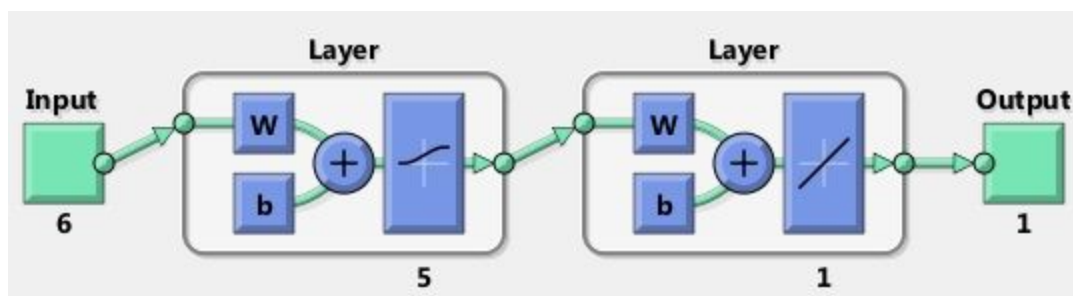
The main goal was to train a neural network in a supervised learning task to learn to map a noisy image to a clean image from a noisy one. In this particular case,

## НЕЙРОННА МЕРЕЖА

Ідея застосування нейронної мережі для даної задачі полягає в наступному:

- Завантажити пошкоджені зображення як вхідні дані для нейронної мережі і вже очищені - як очікуваний результат
- Для кожного пікселя визначити “фічі”, які будуть його характеризувати (його власна яскравість, середнє значення яскравості його сусідів, стандартне відхилення яскравості його сусідів, розкид яскравості його сусідів та ін)
- Побудувати нейронну мережу, на вхід якої буде іти кожен піксель (зі своїми характеристиками), а очікуваним результатом буде значення яскравості відповідного пікселя з уже очищеного зображення.





- Після навчання мережі, використати її для передбачення яскравості кожного окремого пікселя зображення.

Такий підхід має ряд недоліків:

- Оскільки кожен піксель опрацьовується окремо, навіть використання вже навченої мережі є досить повільним
- Необхідна наявність вже чистих зображень для навчання
- Результат очищення зображення залежить від даних, на яких вчилася мережа.

## РЕЗУЛЬТАТИ ВИКОРИСТАННЯ НЕЙРОННОЇ МЕРЕЖІ

Початкове зображення:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the qu but requires giving more instructions and, more importantly, rest this type of acquisition is used. Guiding rulers printed on the used for this reason. Light rectangles can be removed more easily whenever the handwritten text touches the rulers. Nevertheless, be taken into account: The best way to print these light rectan

Кінцеве, очищене зображення:

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the quality requires giving more instructions and, more importantly, restricts type of acquisition is used. Guiding rulers printed on the form are n reason. Light rectangles can be removed more easily with filters th handwritten text touches the rulers. Nevertheless, other practical account: The best way to print these light rectangles is in a differ

There exist several methods to design forms with fields to fields may be surrounded by bounding boxes, by light rectangles o methods specify where to write and, therefore, minimize the effect with other parts of the form. These guides can be located on a s is located below the form or they can be printed directly on the fo a separate sheet is much better from the point of view of the quality requires giving more instructions and, more importantly, restricts type of acquisition is used. Guiding rulers printed on the form are n reason. Light rectangles can be removed more easily with filters th handwritten text touches the rulers. Nevertheless, other practical account: The best way to print these light rectangles is in a differ

There are several classic spatial filters for reducing frequency noise from images. The mean filter, the median filter, and the opening filter are frequently used. The mean filter is a filter that replaces the pixel values with the neighborhood average, thereby reducing the image noise but blurs the image edges. The median filter replaces the pixel values with the median of the pixel neighborhood for each pixel, thereby reducing the image noise but blurs the image edges. Finally, the opening closing filter is a mathematical morphology operation that combines the same number of erosion and dilation morphological operations to eliminate small objects from images.

The main goal was to train a neural network in a supervised manner to learn to map a noisy image to a clean image from a noisy one. In this particular case,

There are several classic spatial filters for reducing frequency noise from images. The mean filter, the median filter, and the opening filter are frequently used. The mean filter is a filter that replaces the pixel values with the neighborhood average, thereby reducing the image noise but blurs the image edges. The median filter replaces the pixel values with the median of the pixel neighborhood for each pixel, thereby reducing the image noise but blurs the image edges. Finally, the opening closing filter is a mathematical morphology operation that combines the same number of erosion and dilation morphological operations to eliminate small objects from images.

The main goal was to train a neural network in a supervised manner to learn to map a noisy image to a clean image from a noisy one. In this particular case,

## НАСТУПНІ КРОКИ

Наступними кроками в цьому проєкті можуть бути:

- Підвищення точності кластеризації шляхом вибору кращих “фіч”
- Пришвидшення використання нейронної мережі (можливо використовувати “sliding window” замість того щоб опрацьовувати кожен піксель окремо)
- Реалізувати ідею Stacked Denoising Autoencoders, яку використовують для очищення будь яких (не тільки текстових) зображень (<https://papers.nips.cc/paper/4686-image-denoising-and-inpainting-with-deep-neural-networks.pdf>)