# Wymagający złamania wierszy tytuł pracy w języku polskim

(English title)

Grzegorz Ciesielski

Praca licencjacka

**Promotor:** dr Jan Chorowski

Uniwersytet Wrocławski
Wydział Matematyki i Informatyki
Instytut Informatyki

13 sierpnia 2019

## Streszczenie

Poniższa praca sprawdza hipotezę, że można nauczyc sieć neuronową rozpoznawać tekst bez korzystania z podpisanych próbek tekstu, a jedynie z rokładów prawdopodobieństw wystąpienia pewnych sekwencji liter (n-gramów) w danych uczących.

---

Below thesis validates hypothesis whether it is possible to teach a neural network text recognition without labeled data, but using sequential output statistics (in form of n-grams) only.

# Spis treści

# Rozdział 1.

# Introduction

Most machine learning models require two main factors to function: well designed model and (usually a lot of) data. Preparing a model usually does not raise problems, as there are a lot of them available online and their perfromance usually allows them to learn on semi-advanced, affordable machines. But to do so, we also need a second factor: data. Especially in text recognition it seems that we cannot build a decent model without any labeled samples, and gathering such data requires a lot of human effort. Hence in this work we will study possibility of training neural net to recognize written text without labeled data, but only using sequential statistics regarding ouptut, as in [1], except that we will use smaller, simpler and higher-variance dataset and more complex model of neural net.

## 1.1. Problem formulation

As mentioned before, we consider problem of learning text classifier using dataset consisting pure (not labeled) sequences of data and their output statistics. In language processing obtaining such dataset is considerbaly simpler that labeled data, as it is enough to get images of written text in specific language and for output statistics use easily obtainable n-grams of this language. Specifically: our classifier predicts sequences $(y_1, \dots y_n)$ (labels) from input sequences $(x_1, \dots x_n)$. The algorithm has access to dataset $D = \{(x_1^k, \dots x_n^k) : k = 1, \dots M\}$ of sequences, and to n-gram probabilities of such dataset - which is denoted as:

$$p(i_1, \dots i_n) = p(y_1^k = i_1, \dots y_n^k = i_n)$$

where $i_1, \dots i_n \in \{0, 1, \cdots C\}$, and $y_j^k$ is the label of letter $x_j^k$.

## Rozdział 2.

# Output Distribution Match

# Bibliografia

[1] Yu Liu, Jianshu Chen, Li Deng. “Unsupervised Sequence Classification using Sequential Output Statistics”