LITHUANIAN COMPUTER SOCIETY
VILNIUS UNIVERSITY
INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES
LITHUANIAN ACADEMY OF SCIENCES

11th International Workshop on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Druskininkai, Lithuania, Hotel "Europa Royale"
http://www.mii.lt/DAMSS

November 28–30, 2019

VILNIUS UNIVERSITY PRESS
Vilnius, 2019

Co-Chairmen:

Dr. Saulius Maskeliūnas (Lithuanian Computer Society)
Prof. Gintautas Dzemyda (Vilnius University, Lithuanian Academy of Sciences)

Programme Committee:

Prof. Juris Borzovs (Latvia)
Prof. Albertas Čaplinskas (Lithuania)
Prof. Robertas Damaševičius (Lithuania)
Prof. Janis Grundspenkis (Latvia)
Prof. Janusz Kacprzyk (Poland)
Prof. Ignacy Kaliszewski (Poland)
Prof. Yuriy Kharin (Belarus)
Prof. Tomas Krilavičius (Lithuania)
Prof. Julius Žilinskas (Lithuania)

Organizing Committee:

Dr. Jolita Bernatavičienė
Prof. Olga Kurasova
Dr. Viktor Medvedev
Laima Paliulionienė
Dr. Martynas Sabaliauskas

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Prof. Olga Kurasova
olga.kurasova@mif.vu.lt
Tel. +370 5 2109 315

https://doi.org/10.15388/Proceedings.2019.8

ISBN 978-609-07-0325-0 (digital PDF)

The N-Grams Based Text Similarity Detection Approach Using Self-Organizing Maps and Similarity Measures

Rokas Štrimaitis[1], Olga Kurasova[2], Pavel Stefanovič[1]

[1] Vilnius Gediminas Technical University
[2] Institute of Data Science and Digital Technologies
Vilnius University
rokas.strimaitis@vgtu.lt

The word-level n-grams based approach is proposed to find similarity between texts. The approach is a combination of two separate and independent techniques: self-organizing map (SOM) and text similarity measures. SOM's uniqueness is that the obtained results of data clustering, as well as dimensionality reduction, are presented in a visual form. The four measures have been evaluated: cosine, dice, extended Jaccard's, and overlap. First of all, texts have to be converted to numerical expression. For that purpose, the text has been split into the word-level n-grams and after that, the bag of n-grams has been created. The n-grams' frequencies are calculated and the frequency matrix of dataset is formed. Various filters are used to create a bag of n-grams: stemming algorithms, number and punctuation removers, stop words, etc. All experimental investigation has been made using a corpus of plagiarized short answers dataset.