

Inżynieria języka naturalnego

Tworzenie mapy dokumentów w języku polskim w oparciu o metody inżynierii języka naturalnego

Jacek Miiller, 179265 Roman Sorokowski, 179121
Bartosz Michałek, 179276

30 października 2013

Przegląd literatury oraz zapoznanie się z niezbędnymi podstawami teoretycznymi

W dokumencie [4] wprowadzono rozmytą miarę podobieństwa dokumentów, na bazie ujednolicania znaczeń słów. Jako wejście algorytmu wyznaczania podobieństwa posłużyły wyniki działania WordNetu. Następnie zbudowano architekturę Self Organizing Map, w której wyznaczanie zwycięskiego neuronu oparte było na algorytmie fuzzy C-means.

W dokumencie [2] opisano szczegółowo jakie cechy były ekstrahowane z dokumentów oraz sposób ich ekstrakcji. Cechy te posłużyły do nauki zmodyfikowanej przez autorów struktury Self Organizing Map.

W dokumencie [4] zastosowano połączenie naiwnego klasyfikatora Bayesa z siecią Self Organizing Map. Słowa z klasyfikowanych dokumentów stanowią wektor wejściowy do Bayesa, który wyznacza prawdopodobieństwa przynależności danego dokumentu do klasy. Prawdopodobieństwa te stanowią wejścia do SOM.

Dokument opisujący narzędzie Fextor [1], które będzie pomocne. Wyniki działania tego programu zostaną wykorzystane jako ciągi uczące Self Organizing Map.

Literatura

- [1] Bartosz Broda, Paweł Kędzia, Michał Marcińczuk, Adam Radziszewski, Radosław Ramocki, and Adam Wardyński. Fextor: A feature extraction framework for natural language processing: A case study in word sense disambiguation, relation recognition and anaphora resolution. In Adam Przepiórkowski, Maciej Piasecki, Krzysztof Jassem, and Piotr Fuglewicz, editors, *Computational Linguistics*, volume 458 of *Studies in Computational Intelligence*, pages 41–62. Springer Berlin Heidelberg, 2013.
- [2] Tommy W. S. Chow and M. K. M. Rahman. Multilayer som with tree-structured data for efficient document retrieval and plagiarism detection. *IEEE Transactions on Neural Networks*, 20(9):1385–1402, 2009.
- [3] Dino Isa, V. P. Kallimani, and Lam Hong Lee. Using the self organizing map for clustering of text documents. *Expert Syst. Appl.*, 36(5):9584–9591, 2009.
- [4] Francisco P. Romero, Arturo Peralta, Andrés Soto, José Angel Olivas, and Jesús Serrano-Guerrero. Fuzzy optimized self-organizing maps and their application to document clustering. *Soft Comput.*, 14(8):857–867, 2010.