

EVALUATION OF NAMED ENTITY RECOGNITION SOFTWARE PACKAGES FOR DATA MINING

I. Sokolov^{1,*}, *E. Antonov*^{1,2}, *A. Artamonov*¹

¹ National Research Nuclear University MEPHI
(Moscow Engineering Physics Institute), Moscow

² Plekhanov Russian University of Economics, Moscow

An important area of data analysis is natural language processing technology, including named entity recognition. A study was conducted to select the best software packages for this task in Russian news text. A corpus of 70 articles from various resources was compiled, and packages “Natasha”, “SpaCy”, “Stanza” and “DeepPavlov” were compared. Experiments involved both manual and programmatic extraction of named entities, and metric calculations were performed. The research findings indicated that a combination of “Natasha” and “Stanza” can achieve comprehensive entity extraction in Russian news text.

Важной областью анализа данных является технология обработки естественного языка, включая распознавание именованных сущностей. Проведено исследование для выбора наилучших программных пакетов для этой задачи в русском новостном тексте. Составлен корпус из 70 статей из разных источников, сравнивались пакеты «Natasha», «SpaCy», «Stanza» и «DeepPavlov». Эксперименты включали как ручное, так и программное извлечение именованных сущностей, а также вычисление метрик. Результаты исследования показали, что комбинация «Natasha» и «Stanza» позволяет достичь полного извлечения сущностей в русском новостном тексте.

PACS: 07.05.kf

* E-mail: idsokolov@kaf65.ru