УДК 004.912

ПОСТРОЕНИЕ СЕМАНТИЧЕСКИХ АНАЛОГИЙ ПО ДАННЫМ ТЕМАТИЧЕСКИХ ТЕКСТОВЫХ КОРПУСОВ

Бабиков И.А. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.т.н., Ковальчук С.В.
(Университет ИТМО, г. Санкт-Петербург)

Аннотация
Формализация текстовых знаний тематических корпусов. Построение семантических аналогий с помощью векторных представлений слов. Сравнение обученных на тематических корпусах векторных представлений с эталонными.

Введение. Растущий объем неструктурированных текстовых данных позволяет искать пути для обеспечения информационный поддержки, в частности для системы поддержки принятия врачебных решений. Формирование семантических аналогий с помощью векторных представлений слов может помочь в формализации знаний данной предметной области.
В общем доступе существуют эталонные векторные представления слов, обученные на большом количестве текстов. Цель работы состояла в том, чтобы построить эмбеддинги, характеристики которых будут максимально похожи на эталонные.

Основная часть. С целью извлечения качественных текстовых эмбеддингов по данной предметной области принято решение собрать тексты статей с портала PubMed для медицинской тематики, а также для сравнения тексты архивов статей ICCS по теме “Computer Science”.
Перед загрузкой текстовых данных необходимо провести шаг предобработки: приведение слов к нормальной форме, удаление стоп-слов (шумовых слов), применение различных эвристик по ограничению длины слов.
Для построения данных эмбеддингов применялась традиционная модель в области обработки естественного языка, Word2Vec с настройкой различных значений гиперпараметров.
Для оценки качества полученных эмбеддингов тематической области для каждого набора гиперпараметров модели строилось распределения косинусной меры для каждого слова из словаря, на котором обучалась модель. Качественным распределением считается такое распределение, которое наиболее похоже на такое же распределение косинусной меры для эталонных эмбеддингов.

Выводы. В ходе выполнения данной работы были собраны различные текстовые тематические корпуса, проведена их предварительная предобработка (удаление стоп-слов, лемматизация) и построена модель модель Word2Vec с различными гиперпараметрами для извлечения векторных представлений (эмбеддингов) терминов данной области знаний. Полученные векторные представления сравнивались с эталонными эмбеддингами.