# ТЕХНИЧЕСКИЕ НАУКИ

## СИНТАКСИЧЕСКИЙ РАЗБОР ПРЕДЛОЖЕНИЯ ДЛЯ ВЕКТОРИЗАЦИИ ТЕКСТА

**Иванов Н.Н.**

*Иванов Николай Николаевич - кандидат физико-математических наук, доцент, кафедра ЭВМ,*
*Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Республика Беларусь*

***Аннотация:*** *рассматривается подход к векторизации текста после предварительного синтаксического разбора предложений. Это позволяет более глубоко оценивать взаимосвязь слов, игнорируя пары слов с малой семантической значимостью. При классификации научных и формализованных документов учет частей речи позволяет более точно оценить уровень зависимости документов. При исследовании эмоциональной окраски художественного произведения синтаксический разбор не играет такой важной роли.*

***Ключевые слова:*** *векторизация текста, синтаксический разбор предложения, word2vec.*

После задач обработки числовых данных, анализа изображений и видеопотока внимание на себя обратила задача обработки текстов, точнее подзадача кластеризации текстов. Здесь основной текущей задачей является кластеризация текстов с целью помочь исследователю или потенциальному читателю ориентироваться в огромном количестве публикаций [1].

Простейшим примером кластеризации является разделение заданного множества публикаций на фиксированное количество классов, в элементарном случае в качестве исходной информации даны тезаурасы классов. Обычным алгоритмом решения задачи в такой постановке является классификатор k-средних или кластеризация нейронной сетью Кохонена.

Более сложные алгоритмы используют «мешок слов» – статистическую информацию о словах, составляющих документ. Следующим шагом в кластеризации текстов стало использование оцифровки текста, точнее, кодирование слов документа векторами, при этом все слова сравниваемых документов, кроме стоп-слов, заменяются векторами одинаковой размерности, равной, примерно 200.

В настоящее время для векторизации применяется приложение word2vec, разработанное компанией *Google* для ускорения поиска в глобальной сети. Модуль *word2vec* сканирует фразу и отмечает в ней слова, расположенные в предложении близко друг к другу. Модуль захватывает *n* рядом стоящих слов, этот факт обозначается как *n-gram*, кроме того, он может не учитывать вообще *k* некоторых слов, пропуская их, эта ситуация названа *k-skip*. Числовые векторы-коды строит искусственная нейронная сеть глубокого обучения. Близкие по значению слова приложение кодирует векторами, близкими в некоторой метрике векторного пространства. Получаемые при этом векторы на глубине одной-двух операций арифметики сложения и вычитания сохраняют смысловую связь слов-прообразов векторов. Однако, это приложение не предназначено для анализа и сравнения текстовых документов.

В работе [2] дан краткий, но исчерпывающий обзор методов векторизации текстов и документов.

В этой заметке предлагается при выполнении процедуры векторизации текста выполнять предобработку текста синтаксическим разбором предложении и учитывать слова в предложении как часть речи. Основным словом в предложении следует

считать подлежащее, следующим по значимости сказуемое, затем в зависимости от характера документа и поставленной задачи, обстоятельство, дополнение, определение.
В лингвистике отмечается, что семантическая сущность членов предложения неоднородна, это связано с различием в них уровня синтаксической абстракции. Кроме того, и в различных языках имеются отличия семантической составляющей одинаковых частей речи [3]. Предлагаемый алгоритм векторизации ориентирован на кластеризацию формальных текстов на русском языке. Он состоит из четырех этапов:
1) синтаксический разбор предложения;
2) предобработка текста, включающая, в частности, удаление стоп-слов;
3) процедура gram-skip, связывающая подлежащее со сказуемыми и другими членами предложения;
4) обработка полученных статистических данных сверточной и/или глубокой нейронной сетью.

Как показали предварительные эксперименты, такой подход приносит результаты для кластеризации формальных документов, как то: научных статей, описания патентов, сухих изложений новостей.
Для задачи сравнения эмоциональной окраски художественного произведения предлагаемый здесь метод не пригоден – в этой задаче важны тонкие нюансы речи, порожденные культурой и традициями народа и местности.
Список литературы
1. Aggarwal Charu C., Zhai Cheng Xiang. Mining text data. Springer Science & Business Media, 2012. P. 524.
2. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов // Труды ИСП РАН, 2017. № 2 ( 29). С. 161–200.
3. Fillmore C.J. Types of lexical information // Studies in syntax and semantics/ Ed. by F. Kiefer. Dordrecht, 1969. Pp. 109–137.