УДК 004.41

И.В. Капелько, магистрант; Д.В. Шиман доц.
(БГТУ, г. Минск)

ТРЕХУРОВНЕВАЯ ТОКЕНИЗАЦИЯ ДЛЯ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТА

Автоматическое реферирование – это составление коротких изложений материалов, дайджестов, т.е. извлечение наиболее важных сведений из одного или нескольких документов и генерация на их основе лаконичных отчетов.

В основе алгоритма лежит метод TF-IDF (от англ. TF — term frequency, IDF — inverse document frequency) – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.

Перед применением каких-либо методов реферирования необходимо использовать предварительную обработку текста: удаление стоп-слов, исправление грамматических ошибок.

Под трехуровневой токенизацией подразумевается, что токеном будет являться не только какое-то отдельное слово, но и предложение и абзац. Токеном первого уровня будет абзац текста, который включает в себя коллекцию токенов второго уровня – отдельные предложения этого абзаца. Токеном третьего уровня является слово. Для каждого слова будет выделена его лемма – неизмененная, исходная форма слова, а для нее рассчитана TF-IDF. Где TF – отношение числа вхождений некоторой леммы к общему числу слов документа. Таким образом, оценивается важность леммы в пределах отдельного документа. IDF – это обратная частотность документов. Она измеряет непосредственно важность термина. В моем алгоритме ее можно выразить формулой:

$$IDF=\log(\max(TF)/f),$$

где f – частота вхождения леммы в корпус текста.

После расчёта TF-IDF у каждого слова есть свое числовое значение, выражающее его вес в тексте. Для расчёта веса предложений необходимо сложить значения токенов (слов), которые входят в состав этого предложения и разделить на их количество. Далее ту же операцию необходимо применить и для абзацев: сложить значение весов предложений и разделить на количество этих предложений в абзаце.

Таким образом можно выделить наиболее значимые слова, а на их основе выделить наиболее важные части текста.