СЕРВІС ДЛЯ СКОРОЧЕННЯ ТЕКСТУ ЗА ДОПОМОГОЮ АЛГОРИТМУ ОБРОБКИ ТЕКСТУ TEXTRANK

Демченко А. Є.
Науковий керівник – к.т.н., доц. кафедри ШІ Золотухін О. В.
Харківський національний університет радіоелектроніки
(61166, Харків, просп. Науки, 14, каф. ШІ, тел. (057) 702-13-37)
e-mail: andrii.demchenko@nure.ua, факс (057) 702-13-37

The goal of this research paper is to create high quality algorithm, that can scrape text of article from third-party sites and summarize it and save sense of the text, but delete all useless with TextRank algorithm.

Мы живем в эру активного развития интернет технологий, когда получение и обработка информации является одной из составляющих жизни современного человека. Ежедневно, наш мозг получает и обрабатывает гигабайты информации, как визуальной, так и вербальной. Вся эта информация исходит из источников внешнего мира. Если брать во внимание интернет, которому человек может уделять от получаса до пяти часов в день, то можно сделать вывод, что человеческий мозг получает намного больше информации, чем было представлено выше.

Именно поэтому мы должны научится отсеивать информацию на правильную и неправильную, на полезную и бесполезную. Даже в хороших статьях, которые мы читаем может присутствовать достаточно много бесполезных вещей. Это связано с богатством и насыщенностью нашего языка. Речевые обороты, идиомы, цитаты и другие средства выразительности делают нашу речь красивой, придают ей эмоциональный и смысловой окрас, но в определенной мере отдаляют нас от понимания основной сути читаемого текста.

Целью данной исследовательской работы является разработка алгоритма для сокращения текста, с возможностью сохранения основной сути предложений и текста в целом, путем удаления наименее значимых частей предложения для получения краткого содержания статьи. Помимо сокращения текста до минимума данная разработка помогает автоматизировать процесс составления краткого описания для статей и сокращает процесс принятия решения о прочтении полной версии статьи до минимума.

В рамках данного исследования был проведен анализ данных, а в частности лексические и семантические связи между словами и предложениями в тексте для выделения четких правил сокращения текста.

Связь предложений и слов в тексте делает его осмысленным. Это достигается с помощью семантических особенностей языка, таких как местоимения (“мы”, “он”, “она”), указательные местоимения (“этот”, “тот”, “эти”), лексические отношения в тексте и словосочетания, ссылающиеся на один и тот же объект. Так, согласно Вольфану Дресслеру, связь в тексте

можно определить как предложения, продолжающие или взаимно дополняющие друг друга. Таким образом текст наполнен смыслом в результате объединения смысла всех предложений в общем на протяжении всего времени его прочтения.

Связи предложений в тексте видны, непосредственно, при прочтении текста, как единого целого, в то время как каждое взятое отдельно предложение может терять часть смысла.

Также был использован ACE Copus для изучения лексических шаблонов для связей между предложениями. При проведении анализа был выбран набор ключевых слов, появление которых в предложении однозначно утверждало бы его лексическую связь с предыдущим. Наличие таких слов как “однако”, “кроме того”, “соответственно”, “следовательно”, “кстати”, “таким образом”, “но”, “тем не менее”, “однажды”, “более того” в начале текущего предложения однозначно говорит о его связи с предыдущим. Также наличие этих слов в середине текущего предложения также может говорить о связи с предыдущим предложением.

В данном подходе было проанализировано что наличие местоимения в текущем предложении может говорить о наличии его связи с предыдущим предложением. Если в предложении фигурируют такие местоимения как “он”, “она”, “оно”, “они”, “его”, “её”, “их” это может говорить о наличии существительного к которому они относятся как в текущем, так и в предыдущем предложении. Однако не все местоимения могут однозначно говорить о связи между предложениями. Такие местоимения как “сам”, “сама” зачастую не создают связи.

Один из способов оценки работы алгоритма заключается в сравнении с суждениями человека, прочитавшего текст, или же авторов текста, которые могут однозначно утверждать, насколько точно отработал алгоритм и, какие важные части он упустил.

Сравнивать можно не только с вручную сделанным сокращением текста, как это делается на сайтах, но и с другими алгоритмами сокращения текста.

Литература

1. Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization.
2. Mihalcea and P. Tarau. 2004. TextRank - bringing order into texts.
3. Lin and E.H. Hovy. 2003a. Automatic evaluation of summaries using n-gram co-occurrence statistics.
4. Mihalcea, P. Tarau, and E. Figa. 2004. PageRank on semantic networks, with application to word sense disambiguation.
5. Radev, H. Y. Jing, M. Stys and D. Tam. Centroid-based summarization of multiple documents.