Автоматическое составление краткого содержания текстового документа

Беляков Юрий

Декабрь 2020

Введение

Краткое содержание - текст, который содержит важнейшую информацию из исходного текстового документа, но при этом имеет значительно меньший объем. Задача автоматического составления краткого содержания (суммаризация) давно известна в области обработки естественного языка. Возможность автоматически составить краткое содержание позволяет, без необходимости читать полный объем текста, ознакомиться с его содержанием и оценить его релевантность к предмету поиска. Методы автоматической суммаризации разделяют на абстрактивные и экстрактивные. Абстрактивные методы выполняют задачу составления нового текста меньшего объема на основании исходного документа. Экстрактивные направлены на выбор подмножества множества предложений документа, которые содержат наибольшее количество информации. В данной работе будет применен экстрактивный графовый метод TextRank[1].

Постановка задачи

Цель данной работы - реализовать метод автоматической суммаризации TextRank[1] в варианте, предложенном в оригинальной статье, и продемонстрировать его работу на небольшом примере. С этой целью, каждому предложению текста будет сопоставлена оценка его важности. Краткое содержание будет составленно как набор из «самых важных» предложений текста.

Ход работы

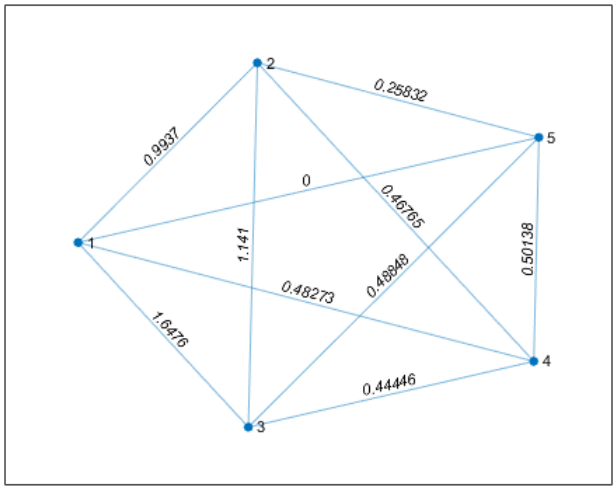
Рассмотрим небольшой пример текста из 5 предложений.

```
str = [
    "The fox jumped over the dog."
    "The lazy dog saw a fox jumping."
    "The quick brown fox jumped over the lazy dog."
    "There seem to be animals jumping over other animals."
    "There are quick animals and lazy animals"];
```

Для начала каждое предложение текста разбивается на список слов. Это необходимо для подсчета взаимосвязей между предложениями. В качестве такой связи рассматривается сходство двух предложений. Сходство ω_{ij} между двумя предложениями S_i и S_j определяем как нормализованное пересечение множеств слов w этих предложений. Формально:

$$\omega_{ij} = \frac{|\{w_k | w_k \in S_i \ \& \ w_k \in S_j|\}}{log(|S_i|) + log(|S_j|)}$$

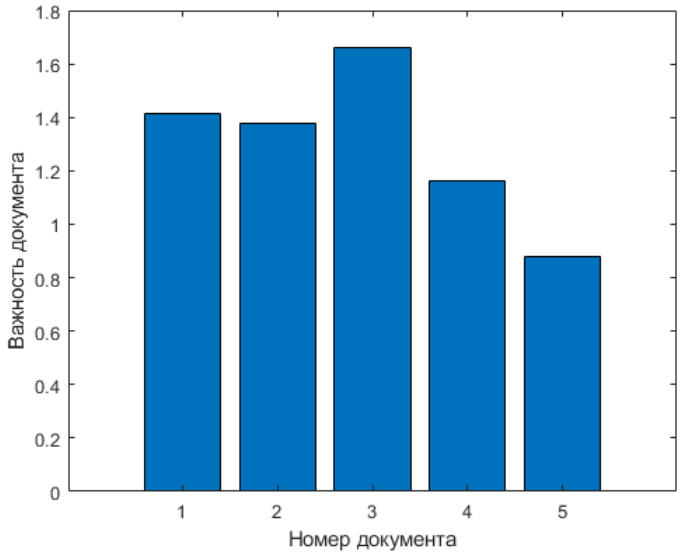
Затем составляется полный граф $G = (V, E)$, где V - множество предложений, а E - ребра, весами которых являются сходства между предложениями.



Построение краткого содержания производим путем выбора «важнейших» предложений. Важность $W(S_i)$ предложения S_i определяется формулой:

$$W(S_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{\omega_{ji}}{\sum_{V_k \in In(V_j)} \omega_{jk}} W(S_j)$$

где $In(V_i)$ - множество вершин, связаных с вершиной V_i, а d - параметр, выполняющий роль вероятности связи двух предложений и имеющий значение от 0 до 1. В данной работе используется значение 0.85. Значения вектора важности предложений итеративно пересчитываются пока не сработает критерий останова $||W_n - W_{n-1}|| < \epsilon$, где $\epsilon = 10^{-3}$, а W_n - вектор значений важности всех предложений на итерации n. В результате выполнения алгоритма получаем значения важности всех предложений.



Обычно, количество предложений для краткого выбирается в процентном соотношении от объема изначального документа. Для нашего примера потребуем, чтобы краткое содержание занимало 20% объема исходного текста. Таким образом, исходный текст из 5 предложений сократится до одного самого важного - предложения номер 3.

Литература

[1] Mihalcea, R. Tarau, P. (2004). TextRank: Bringing Order into Texts. Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing, July, .