Национальный исследовательский университет
Высшая школа экономики
Факультет компьютерных наук

Отчет по итоговому домашнему заданию

Реферирование/аннотирование текста на русском языке на основе статистики и эвристических правил

студент 1 курса магистратуры
Зуев Кирилл Александрович

Москва, 2019

Содержание

1 Постановка задачи

Реализовать один или несколько методов автоматического аннотирования (построения реферата) текста. Провести экспериментальное исследование построения аннотаций нескольких выбранных текстов с использованием различных параметров для реализованных методов.

2 Уточнение постановки задачи

Для начала, когда на вход программе подается «сырой» текст, нужно его разбить на предложения для того, чтобы из них оставить наиболее значимые. Это я и выполнил в первой части программы. Сделал я это достаточно примитивно, по соответсвующим знакам препинания, которые обычно обозначают конец предложения. А именно: точка (**.**), восклицательный (**!**) и вопросительный (**?**) знаки.

Я реализовал несколько методов аннотирования текстов:

- **SumBasic**;
- **TF-IDF**;
- на основе **графа предложений**.

Были рассмотрены разные степени сжатия текста:

- 20%;
- 30%;
- 40%.

Для метода на основе графа предложений я использовал разные пороги степени сходства предложений:

- 0.2;
- 0.3;
- 0.4.

Для проведения экспериментов я выбрал следующие тексты:

- Рецензия на фильм «Мстители: Финал»;
- Рецензия на 3 эпизод нового сезона «Игры престолов»;
- Комментарий Noize MC о произошедшей стычке 1 мая в Лужниках;

3 Описание реализованных методов

Стоит отметить, что в каждом из методов «слова» представляют собой леммы (нормальные формы) лексем, встречающихся в тексте.

3.1 SumBasic

Идея метода состоит в том, что наиболее частотные слова текста с большой вероятностью должны оказаться в аннотации.

Для выбора очередного предложения подсчитывается вес всех предложений:

$$weight(S_j) = \sum_{w_i \in S_j} \frac{p(w_i)}{|\{w_i \mid w_i \in S_j\}|},$$

где S_j — j-е предложение текста, w_i — i-е слово в предложении, а $p(w_i)$ — вероятность появления слова w_i в тексте:

$$p(w_i) = \frac{n}{N},$$

где n — число вхождений слова w_i в текст, а N — общее число слов в тексте.

В аннотацию на каждом шаге выбирается предложение с наибольшим весом. После выбора предложения происходит пересчёт весов для предложений, не вошедших в аннотацию. И выбор уже происходит из оставшихся предложений.

3.2 TF-IDF

Данный метод отличается от метода **SumBasic** вычислением «ценности» слова. Теперь она определяется не вероятностью его появления в тексте, а таким образом, что чем реже оно встречается в различных предложениях, тем оно более информативно. Вес слова $p(w_i)$ определяется по формуле:

$$p(w_i) = tf(w_i, S_j) * idf(w_i, T),$$

где T — текст, S_j — j-е предложение текста, w_i — i-е слово в предложении.

Значения функций tf и idf вычисляются следующим образом:

$$tf(w_i, S_j) = \frac{n_{w_i}}{\sum_k n_{w_k}},$$

$$idf(w_i, T) = log\frac{|T|}{|\{S_j \in T \mid w_i \in S_j\}|},$$

где n_{w_k} — число вхождений слова w_k в предложение S_j, $\sum_k n_{w_k}$ — общее число слов в предложении S_j, $|T|$ — число предложений в тексте, а $|\{S_j \in T \mid w_i \in S_j\}|$ — число предложений из текста, в которых встречается слово w_i.

В аннотацию на каждом шаге выбирается предложение с наибольшим весом. После каждого выбора происходит пересчёт весов для предложений, не вошедших в аннотацию. И выбор уже происходит из оставшихся предложений.

3.3 На основе графа предложений

В данном методе для каждого предложения ставится в соответствие вектор, значения в каждой координате которого определяются количеством вхождений определенного слова в это предложение.

Далее строится граф, вершинами которого являются сами предложения, а наличие дуг между парой вершин означает, что соответствующие предложения имеют схожесть не ниже заданного порога θ:

$$similarity(v_i, v_j) = cos(v_i, v_j) = \frac{(v_i, v_j)}{||v_i||_2 \ ||v_j||_2} \geq \theta,$$

где v_i, v_j — векторы соответственно i-го и j-го предложений.

Для каждой вершины графа вычисляется значение центральности по ее степени (числу ребер, инцидентных с этой вершиной). Предложения с наибольшими значениями центральности выбираются в аннотацию.

4 Код программы

Реализация программы находится в приложенных файлах **Big_HW.html** и **Big_HW.ipynb**.

5 Полученные результаты

Результаты эксперимента можно увидеть в конце кода программы, где представлены аннотации, полученные для всех рассматриваемых текстов, для всех моделей и параметров.

6 Выводы по исследованию

По результатам проведенного эксперимента можно заметить, что метод **SumBasic** выдает предложения, в которых встречаются наиболее частотные слова, но чаще всего они не отражают всей сути исходного текста. Метод **TF-IDF** в большинстве случаев выдает короткие предложения, что тоже редко передает смысл текста. Метод на основе **графа предложений**, по моему мнению, дает наиболее удачные аннотации, учитывающие схожесть предложений, а не использующие исключительно частотность слов. Из рассмотренных порогов образования дуг в графе я считаю оптимальным значением — 0.2, поскольку оно допускает больше связей, чем 0.3 или 0.4 (но с меньшей степенью схожести). По объему аннотации я считаю, что оптимально выбрать 30 или 40% предложений от исходного текста, поскольку в 20% труднее передать всю суть, а больше уже будет слишком много для короткой аннотации.