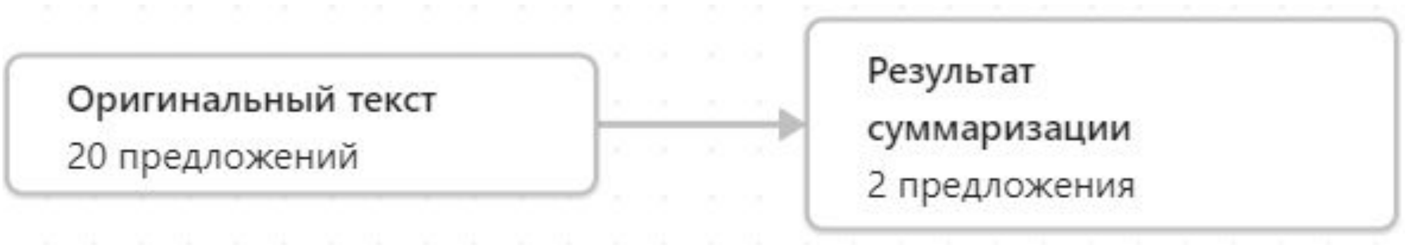
Суммаризация текста

Ефремов Иван 21.Б04-ПУ

Введение

Суммаризация - процесс автоматической обработки естественного языка, направленный на создание краткого и информативного изложения исходного текста, сохраняя при этом его семантическую информацию.



Оригинальный текст
20 предложений

Результат суммаризации
2 предложения

Задачи суммаризации

- **Сжатие информации** - сокращение объёма данных без значимых смысловых потерь

- **Улучшение читаемости** - выделение главной информации из текста

- **Автоматизация генераций заголовка или аннотации**

Метрика ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) — это набор метрик для оценки качества автоматической суммаризации текста, сравнивая сгенерированное резюме с эталонным (референтным) резюме, написанным человеком.

Чем больше перекрытие, тем выше оценка ROUGE, указывающая на лучшее качество сгенерированного резюме.

- ROUGE-N (N-gram precision): Измеряет перекрытие N-грамм (последовательности из N слов) между сгенерированным и референтным резюме.
- ROUGE-L (Longest Common Subsequence): Измеряет длину наибольшей общей подпоследовательности между сгенерированным и референтным резюме. Более устойчива к перестановке слов, чем ROUGE-N.
- ROUGE-S (Skip-bigram precision): Измеряет перекрытие скип-биграмм, то есть биграмм с пропусками. Например, для предложения “The cat sat on the mat” скип-биграммами будут “The cat”, “The sat”, “The on”, и так далее.

Виды суммаризации

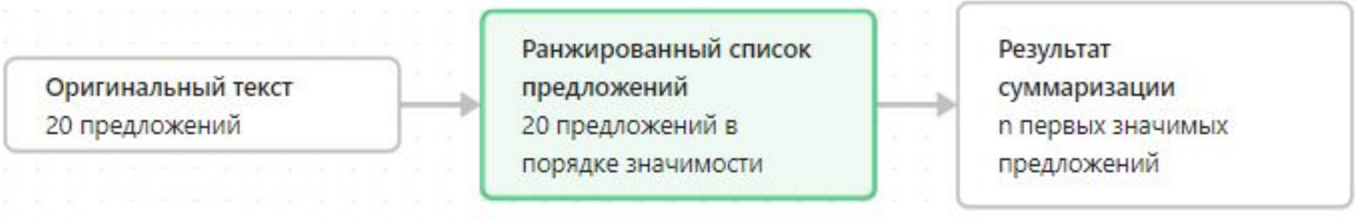
Экстрактивная суммаризация - выбор существующих предложений или отдельных фраз из исходного текста. Резюме содержит только информацию исходного текста.

Абстрактивная суммаризация - генерация новых фраз и предложений на основе исходного текста. Резюме может содержать как информацию напрямую из текста, так и не присутствующую в тексте в явном виде

Экстрактивная суммаризация

Задача экстрактивной суммаризации сводится к ранжированию предложений предложений по значимости и выборе определённого количества из самых значимых из них.

Количество может разнится в зависимости от целей суммаризации.



Оригинальный текст
20 предложений

Ранжированный список предложений
20 предложений в порядке значимости

Результат суммаризации
n первых значимых предложений

Эвристический алгоритм Луна (Luhn)

Алгоритм Luhn, разработанный в 1950-х годах, является одним из самых ранних и простых алгоритмов экстрактивной суммаризации.

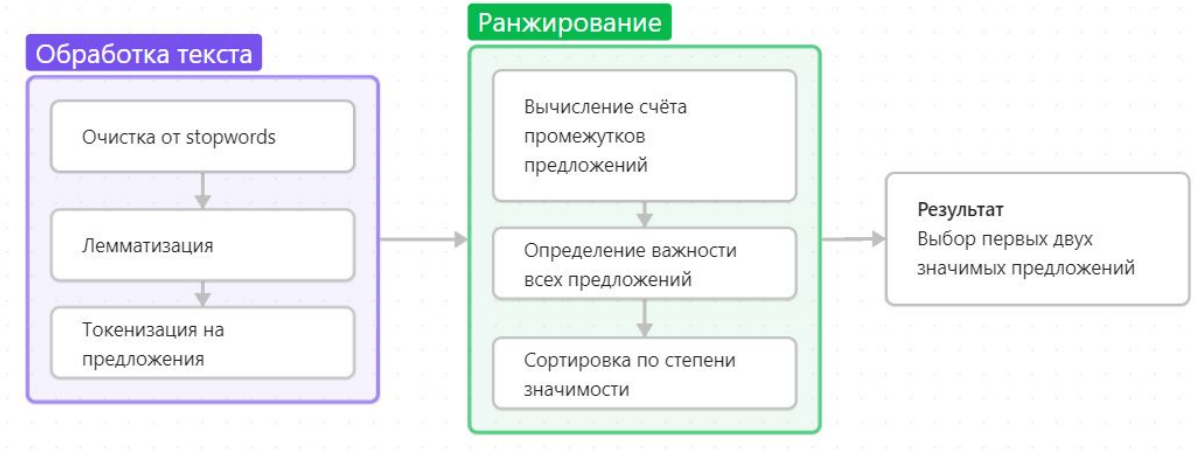
Он основан на предположении, что самыми важными в тексте будут предложения, содержащие наибольшее количество “важных” слов во всём тексте, учитывая их сгруппированность внутри предложения.

Метрика важности предложения по Лунну:

1. Предложения разбиваются на интервалы, начинающиеся и заканчивающиеся с важных слов и содержащих не более 4 не важных
2. Каждому интервалу присваивается счёт:

$$Score = \frac{\text{Количество важных слов}^2}{\text{Длинна интервала}}$$

3. Важность предложения = максимальному счёту его промежутков



Обработка текста

Очистка от stopwords

Лемматизация

Токенизация на предложения

Ранжирование

Вычисление счёта промежутков предложений

Определение важности всех предложений

Сортировка по степени значимости

Результат
Выбор первых двух значимых предложений

Изначальный текст

Never mind cats having nine lives. A stray pooch in Washington State has used up at least three of her own after being hit by a car, apparently whacked on the head with a hammer in a misguided mercy killing and then buried in a field -- only to survive. That's according to Washington State University, where the dog -- a friendly white-and-black bully breed mix now named Theia -- has been receiving care at the Veterinary Teaching Hospital. Four days after her apparent death, the dog managed to stagger to a nearby farm, dirt-covered and emaciated, where she was found by a worker who took her to a vet for help. She was taken in by Moses Lake, Washington, resident Sara Mellado. "Considering everything that she's been through, she's incredibly gentle and loving," Mellado said, according to WSU News. "She's a true miracle dog and she deserves a good life." Theia is only one year old but the dog's brush with death did not leave her unscathed. She suffered a dislocated jaw, leg injuries and a caved-in sinus cavity -- and still requires surgery to help her breathe. The veterinary hospital's Good Samaritan Fund committee awarded some money to help pay for the dog's treatment, but Mellado has set up a fundraising page to help meet the remaining cost of the dog's care. She's also created a Facebook page to keep supporters updated. Donors have already surpassed the $10,000 target, inspired by Theia's tale of survival against the odds. On the fundraising page, Mellado writes, "She is in desperate need of extensive medical procedures to fix her nasal damage and reset her jaw. I agreed to foster her until she finally found a loving home." She is dedicated to making sure Theia gets the medical attention she needs, Mellado adds, and wants to "make sure she gets placed in a family where this will never happen to her again!" Any additional funds raised will be "paid forward" to help other animals. Theia is not the only animal to apparently rise from the grave in recent weeks. A cat in Tampa, Florida, found seemingly dead after he was hit by a car in January, showed up alive in a neighbor's yard five days after he was buried by his owner. The cat was in bad shape, with maggots covering open wounds on his body and a ruined left eye, but remarkably survived with the help of treatment from the Humane Society.

Эталон

Theia, a bully breed mix, was apparently hit by a car, whacked with a hammer and buried in a field . "She's a true miracle dog and she deserves a good life," says Sara Mellado, who is looking for a home for Theia .

Результат работы алгоритма

Never mind cats having nine lives. A stray pooch in Washington State has used up at least three of her own after being hit by a car, apparently whacked on the head with a hammer in a misguided mercy killing and then buried in a field -- only to survive.

ROUGE:

```
rouge-1 F1: 0.3037974634513
rouge-2 F1: 0.17977527592980694
rouge-l F1 : 0.2784810077551675
```

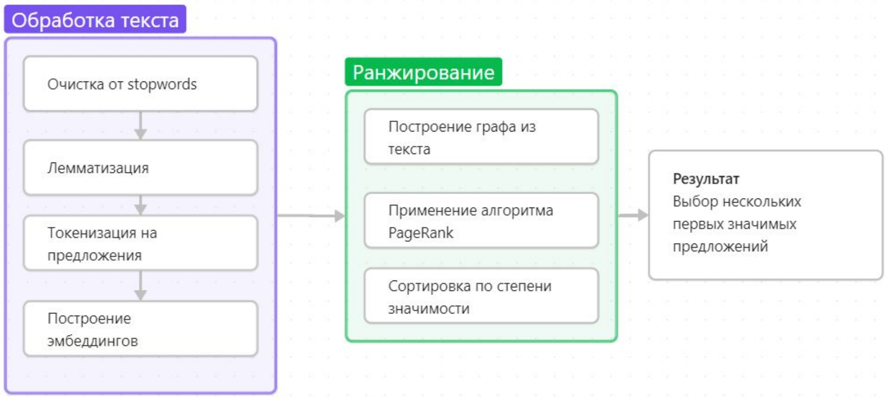
Text Rank

TextRank — это алгоритм экстрактивной суммаризации, основанный на методе PageRank, используемом в поисковых системах для ранжирования веб-страниц. Вместо веб-страниц, TextRank ранжирует предложения в тексте, определяя их важность на основе их взаимосвязей.

Алгоритм предполагает, что важные предложения тесно связаны с другими важными предложениями.

Text Rank

1. Строятся эмбеддинги предложений
2. Между полученными векторами вычисляется косинусовое сходство
3. Строится граф предложений, где узлы - предложения, рёбра - значения косинусового сходства
4. К полученному графу применяется алгоритм PageRank


Обработка текста
Очистка от stopwords
Лемматизация
Токенизация на предложения
Построение эмбеддингов
Ранжирование
Построение графа из текста
Применение алгоритма PageRank
Сортировка по степени значимости
Результат
Выбор нескольких первых значимых предложений

PageRank

Алгоритм PageRank — итеративный алгоритм, определяющий важность страницы, основываясь на входящих ссылках и важности страниц, ссылающихся на нее. Предполагается, что ссылки от важных страниц наделяют авторитетностью и страницу, на которую они ссылаются.

1. Каждому предложению присваивается начальный рейтинг в виде 1/кол-во предложений
2. В каждой итерации предложению i устанавливается рейтинг по формуле:

$$PageRank(i) = \frac{(1-d)}{N} + d\sum \frac{weight(i,j) * PageRank(j)}{L(j)}$$

d - коэф. затухания = 0.85
N - общее число страниц
j - остальные предложения
L(j) - кол-во предложений = N-1

Изначальный текст

Never mind cats having nine lives. A stray pooch in Washington State has used up at least three of her own after being hit by a car, apparently whacked on the head with a hammer in a misguided mercy killing and then buried in a field -- only to survive. That's according to Washington State University, where the dog -- a friendly white-and-black bully breed mix now named Theia -- has been receiving care at the Veterinary Teaching Hospital. Four days after her apparent death, the dog managed to stagger to a nearby farm, dirt-covered and emaciated, where she was found by a worker who took her to a vet for help. She was taken in by Moses Lake, Washington, resident Sara Mellado. "Considering everything that she's been through, she's incredibly gentle and loving," Mellado said, according to WSU News. "She's a true miracle dog and she deserves a good life." Theia is only one year old but the dog's brush with death did not leave her unscathed. She suffered a dislocated jaw, leg injuries and a caved-in sinus cavity -- and still requires surgery to help her breathe. The veterinary hospital's Good Samaritan Fund committee awarded some money to help pay for the dog's treatment, but Mellado has set up a fundraising page to help meet the remaining cost of the dog's care. She's also created a Facebook page to keep supporters updated. Donors have already surpassed the $10,000 target, inspired by Theia's tale of survival against the odds. On the fundraising page, Mellado writes, "She is in desperate need of extensive medical procedures to fix her nasal damage and reset her jaw. I agreed to foster her until she finally found a loving home." She is dedicated to making sure Theia gets the medical attention she needs, Mellado adds, and wants to "make sure she gets placed in a family where this will never happen to her again!" Any additional funds raised will be "paid forward" to help other animals. Theia is not the only animal to apparently rise from the grave in recent weeks. A cat in Tampa, Florida, found seemingly dead after he was hit by a car in January, showed up alive in a neighbor's yard five days after he was buried by his owner. The cat was in bad shape, with maggots covering open wounds on his body and a ruined left eye, but remarkably survived with the help of treatment from the Humane Society.

Эталон

Theia, a bully breed mix, was apparently hit by a car, whacked with a hammer and buried in a field . "She's a true miracle dog and she deserves a good life," says Sara Mellado, who is looking for a home for Theia .

Результат работы алгоритма

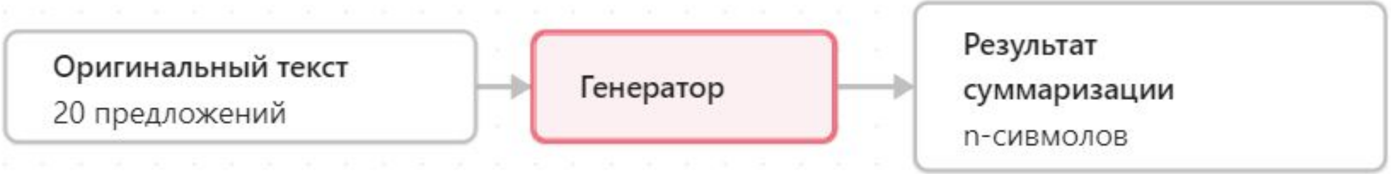
The veterinary hospital's Good Samaritan Fund committee awarded some money to help pay for the dog's treatment, but Mellado has set up a fundraising page to help meet the remaining cost of the dog's care. That's according to Washington State University, where the dog -- a friendly white-and-black bully breed mix now named Theia -- has been receiving care at the Veterinary Teaching Hospital."

ROUGE:

```
rouge-1 F1: 0.13953487893996774
rouge-2 F1: 0.01960783832948984
rouge-l F1: 0.13953487893996774
```

Абстрактивная суммаризация

Задача абстрактивной суммаризации сводится к генерации новых текстовых данных на основе изначального текста



Оригинальный текст
20 предложений

Генератор

Результат суммаризации
n-сивмолов

На основе трансформеров

Наиболее эффективным способом работы с генерации семантически полных новых данных является генерация текста с помощью нейросетей-трансформеров.

Трансформеры глубоко понимают семантику текста, так как обрабатывают весь текст сразу и имеют “механизм внимания”, что позволяет модели уделять различную степень внимания разным частям входного текста при обработке каждого элемента, что делает модель гораздо более эффективной для понимания взаимосвязей между словами.

BART

BART (Bidirectional and Auto-Regressive Transformers) — это архитектура трансформера, разработанная Facebook AI Research

- Bidirectional кодирование: Позволяет модели понять контекст слова с обеих сторон, что приводит к более точному пониманию смысла.
- Autoregressive декодирование: Обеспечивает последовательную генерацию текста, что важно для задач, требующих создание хорошо сформулированных предложений.
- Обучение на шумных данных: Обучение на данных с добавлением шума делает модель более устойчивой к ошибкам и позволяет ей лучше генерировать текст. Это особенно важно для абстрактной суммаризации, где требуется генерировать новый текст, а не просто копировать фрагменты исходного текста.

Изначальный текст

Never mind cats having nine lives. A stray pooch in Washington State has used up at least three of her own after being hit by a car, apparently whacked on the head with a hammer in a misguided mercy killing and then buried in a field -- only to survive. That's according to Washington State University, where the dog -- a friendly white-and-black bully breed mix now named Theia -- has been receiving care at the Veterinary Teaching Hospital. Four days after her apparent death, the dog managed to stagger to a nearby farm, dirt-covered and emaciated, where she was found by a worker who took her to a vet for help. She was taken in by Moses Lake, Washington, resident Sara Mellado. "Considering everything that she's been through, she's incredibly gentle and loving," Mellado said, according to WSU News. "She's a true miracle dog and she deserves a good life." Theia is only one year old but the dog's brush with death did not leave her unscathed. She suffered a dislocated jaw, leg injuries and a caved-in sinus cavity -- and still requires surgery to help her breathe. The veterinary hospital's Good Samaritan Fund committee awarded some money to help pay for the dog's treatment, but Mellado has set up a fundraising page to help meet the remaining cost of the dog's care. She's also created a Facebook page to keep supporters updated. Donors have already surpassed the $10,000 target, inspired by Theia's tale of survival against the odds. On the fundraising page, Mellado writes, "She is in desperate need of extensive medical procedures to fix her nasal damage and reset her jaw. I agreed to foster her until she finally found a loving home." She is dedicated to making sure Theia gets the medical attention she needs, Mellado adds, and wants to "make sure she gets placed in a family where this will never happen to her again!" Any additional funds raised will be "paid forward" to help other animals. Theia is not the only animal to apparently rise from the grave in recent weeks. A cat in Tampa, Florida, found seemingly dead after he was hit by a car in January, showed up alive in a neighbor's yard five days after he was buried by his owner. The cat was in bad shape, with maggots covering open wounds on his body and a ruined left eye, but remarkably survived with the help of treatment from the Humane Society.

Эталон

Theia, a bully breed mix, was apparently hit by a car, whacked with a hammer and buried in a field . "She's a true miracle dog and she deserves a good life," says Sara Mellado, who is looking for a home for Theia .

Результат работы алгоритма

Theia, a one-year-old bully breed mix, was hit by a car and buried in a field. She managed to stagger to a nearby farm, dirt-covered and emaciated. She suffered a dislocated jaw, leg injuries and a caved-in sinus cavity. A fundraising page has raised more than $10,000 for her care.

ROUGE:

```
rouge-1 F1: 0.35294117167058825 rouge-2
F1: 0.13861385656308223
rouge-l F1: 0.3058823481411765
```