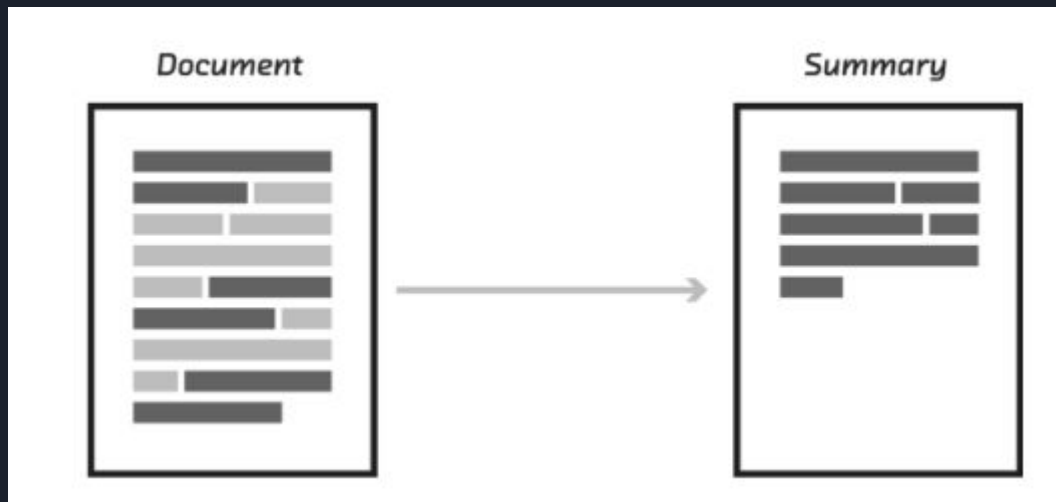


Суммаризация текста

Презентацию подготовил
Аристакесян Тигран

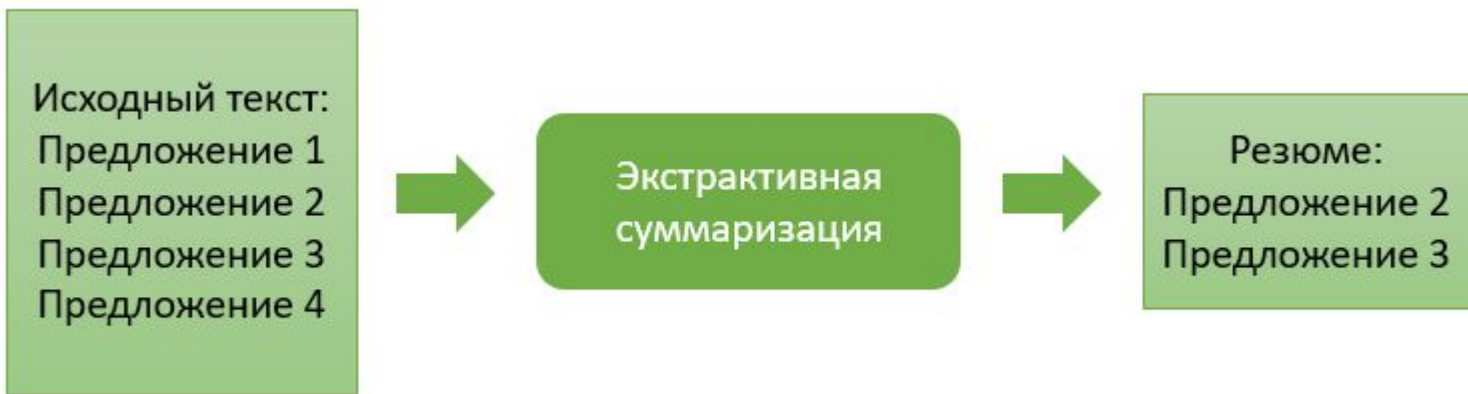
Суммаризация

Суммаризация текста — это автоматическое создание алгоритмом краткого варианта исходного текста с сохранением первоначального смысла.



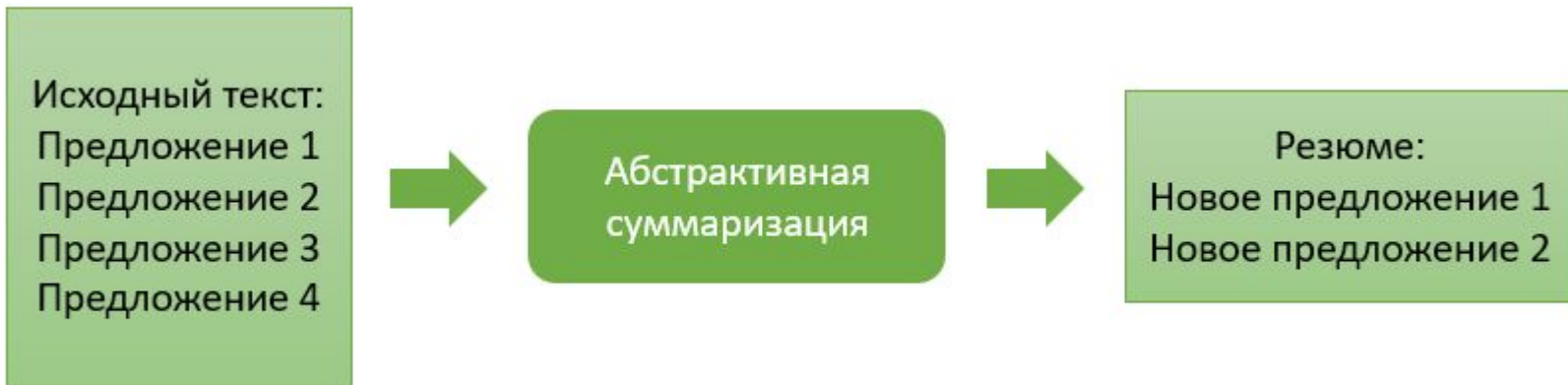
Подходы решения задачи

1) Экстрактивный



Подходы решения задачи


2) Абстрактивный





Экстрактивная суммаризация

- 1) Экстрактивная суммаризация на основе вхождения общих слов
- 2) Экстрактивная суммаризация на основе обученных векторных представлений




Экстрактивная суммаризация на основе вхождения общих слов

- 1) Разбиваем на предложения
- 2) Лемматизируем
- 3) Вычисляем схожесть каждой пары предложений

По формуле: отношение числа общих слов, встречающихся в обоих предложениях, к их суммарной длине

- 4) Строим граф взаимосвязи, вершины предложения, рёбра - наличие общих слов
- 5) Ранжируем по значимости
- 6) Выводим несколько самых значимых



Экстрактивная суммаризация на основе обученных векторных представлений

- 1) Разбиваем на предложения
- 2) Лемматизируем
- 3) Поиск векторного представления для каждого предложения
- 4) Вычисляем схожесть каждой пары предложений

По формуле косинусного расстояния

- 5) Строим граф взаимосвязи, вершины предложения, рёбра - наличие общих слов
- 6) Ранжируем по значимости
- 7) Выводим несколько самых значимых



Абстрактивная суммаризация

- 1) Токенизация текста
- 2) Определение метрики: ROUGE
- 3) Для обучающей выборки берём lead-3 строки
- 4) Обучение модели mT5 или mBart
- 5) Оценка



1. Экстрактивный подход

Преимущества:

1. Интуитивно понятна суть алгоритма
2. Относительная простота реализации

Недостатки:

- Качество содержания во многих случаях может быть хуже, чем написанное вручную человеком



2. Абстрактный подход:

Преимущества:

- Качественно реализованный алгоритм способен выдать результат наиболее близкий к ручному составлению резюме

Недостатки:

- Сложности при восприятии основных теоретических идей алгоритма
- Большие трудозатраты при реализации алгоритма