

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Математико-Механический факультет
Кафедра информационно аналитических систем

Суммаризация групп в социальных сетях

Дипломная работа студента 645 группы
Чурикова Никиты Сергеевича

Научный руководитель:

к.ф. - м.н., доцент ГРАФЕЕВА Н. Г.

Рецензент:

Руководитель департамента вычислительной биологии

ЯКОВЛЕВ П. А.

Заведующий кафедрой:

к.ф. - м.н., доцент МИХАЙЛОВА Е. Г.

Санкт-Петербург
2019 г.

Содержание

1	Аннотация	1
2	Введение	1
2.1	Постановка задачи	2
2.2	Обзор литературы	2
2.3	Полученные результаты	2
3	Алгоритмы, использованные в работе	2
3.1	Текст	2
3.2	Изображения	3
3.3	Оценки качества	3
4	Эксперименты	3
5	Заключение	3

1 Аннотация

Одной из задач обработки естественного языка является задача суммаризации текста. Ее целью является уменьшение размера исходного текста без потери ключевой информации. В данной работе мы решаем схожую проблему, но для информационных ресурсов в социальных сетях. В частности, необходимо рассмотреть задачу суммаризации текстов и картинок, поскольку это два основных источника информации. В тексте мы приводим численное обоснование выбранных методов, а также приводим оценку нашей суммаризации людьми.

2 Введение

В современном мире создается все больше и больше информации, которую мы можем потреблять. Новости, статьи, юмор постоянно меняются и создаются людьми. При таком потоке информации появляется потребность в инструментах, способных давать как можно больше информации с минимальными потерями.

При чтении новостей люди, как правило, не идут дальше новостных заголовков [1], для популярных технических статей создают краткие описания описывающие их достижения и основные моменты [3, 2], а визуальный контент нередко подчиняется единому шаблону.

В данной работе мы показываем, как используя современные достижения в области анализа данных можно извлекать полезную информацию из новостных ресурсов в социальной сети вконтакте [4], приводим

оценки людей нашей системы и приводим сравнение с наивными решениями.

2.1 Постановка задачи

В данной работе мы решили остановиться на двух основных современных видах медиа: тексте и изображениях. В данной работе мы не рассматриваем обработку видео, но есть предположения, что предложенные идеи насчет изображений можно было бы распространить на видеоинформацию.

Для текстовых ресурсов задача суммаризации была разбита на две подзадачи: 1) извлечение ключевых слов, присущих данному источнику информации и 2) автоматическое создание заголовков.

Для изображений – это сбор похожих изображений в кластера и показ некоторых одних изображений, иллюстрирующих каждую группу.

Через извлечение данной информации мы хотим добиться эффекта "чтения по диагонали".

2.2 Обзор литературы

Рассказать про литературу, которая рассматривает задачи выше.

2.3 Полученные результаты

Что является результатом работы (будет веб сервис, куда можно закинуть ссылку на группу), как оценивали качество (продолжить результаты работы алгоритмов толкерам), а также оценка качества по автоматизированным метрикам, и как они коррелируют с оценками людей. Сравниться с бэйзлайном.

3 Алгоритмы, использованные в работе

Нами были использованы как классические подходы, так и новые, основанные на нейронных сетях. В следующих секциях мы опишем их основные принципы, а также приведем ссылки на их реализации.

3.1 Текст

Для суммаризации текста мы воспользовались алгоритмом экстрактивной суммаризации основанном на TextRank [1], и моделью трансформера [2], обученной на датасете РИА новостей [3]. Для предобработки данных модели трансформера мы использовали byte pair encoding [4]. А также мы извлекали первое предложение из новости. Для TextRank и извлечения первого предложения не требуется обучающая выборка, что делает их

очень удобными в использовании. При этом, исследования показывают, что в задаче генерации заголовков, первое предложение в новости – это очень сильный бэйзлайн [], который трудно побить как экстрактивной, так и абстрактивной суммаризацией.

3.2 Изображения

Для суммаризации изображений мы реализовали алгоритм суммаризации изображений, описанный в статье []. Основная идея состоит в том, что из изображений извлекаются признаки, инвариантные к поворотам, эти признаки кластеризуют и индексы кластеров используют как признаки для латентного размещения Дирихле [].

3.3 Оценки качества

Для оценки качества текстовых моделей мы использовали метрику ROUGE-L F1 [], при этом мы считали ее на датасете РИА новостей.

Помимо этого, как для текстовых данных, так и для изображений, мы использовали Яндекс.Толоку [], чтобы привлечь людей к оценке качества наших результатов.

4 Эксперименты

Для обучения моделей были использованы 8 Tesla K80.

5 Заключение

На февраль 2019:

В данной работе мы ожидаем показать, что предложенные нами решения не хуже, а даже лучше предложенных бэйзлайнов как по автоматическим оценкам, так и по оценкам людей. Мы также представим код и ссылку на сервис, куда можно отправить ссылку на интересующую группу и оценить получившийся результат. Мы также планируем показать результаты в "одноклассниках" и рассказать об их мнении насчет полученного решения, поскольку год назад предлагалось совместное сотрудничество над данной проблемой.

Список литературы

- [1] DeMers, J. 59 percent of you will share this article without even reading it. — 2016. — aug. — <https://www.forbes.com/sites/jaysondemers/2016/08/08/>

59-percent-of-you-will-share-this-article-without-even-reading-it/
#71991fa2a648.

- [2] essence, A. Article essence. — 2019. — feb. — <https://opendatascience.slack.com/messages/C5VQ222UX>.
- [3] tldr arxiv. tldr arxiv. — 2019. — feb. — https://t.me/tldr_arxiv.
- [4] vk. vk. — 2019. — feb. — <https://vk.com/feed>.