Агрегация и персонализация новостного текстового контента

Дарина Дементьева
Научный руководитель: д.ф.-м.н. К.В. Воронцов

6 декабря, 2018

Agenda

Жизненная проблема

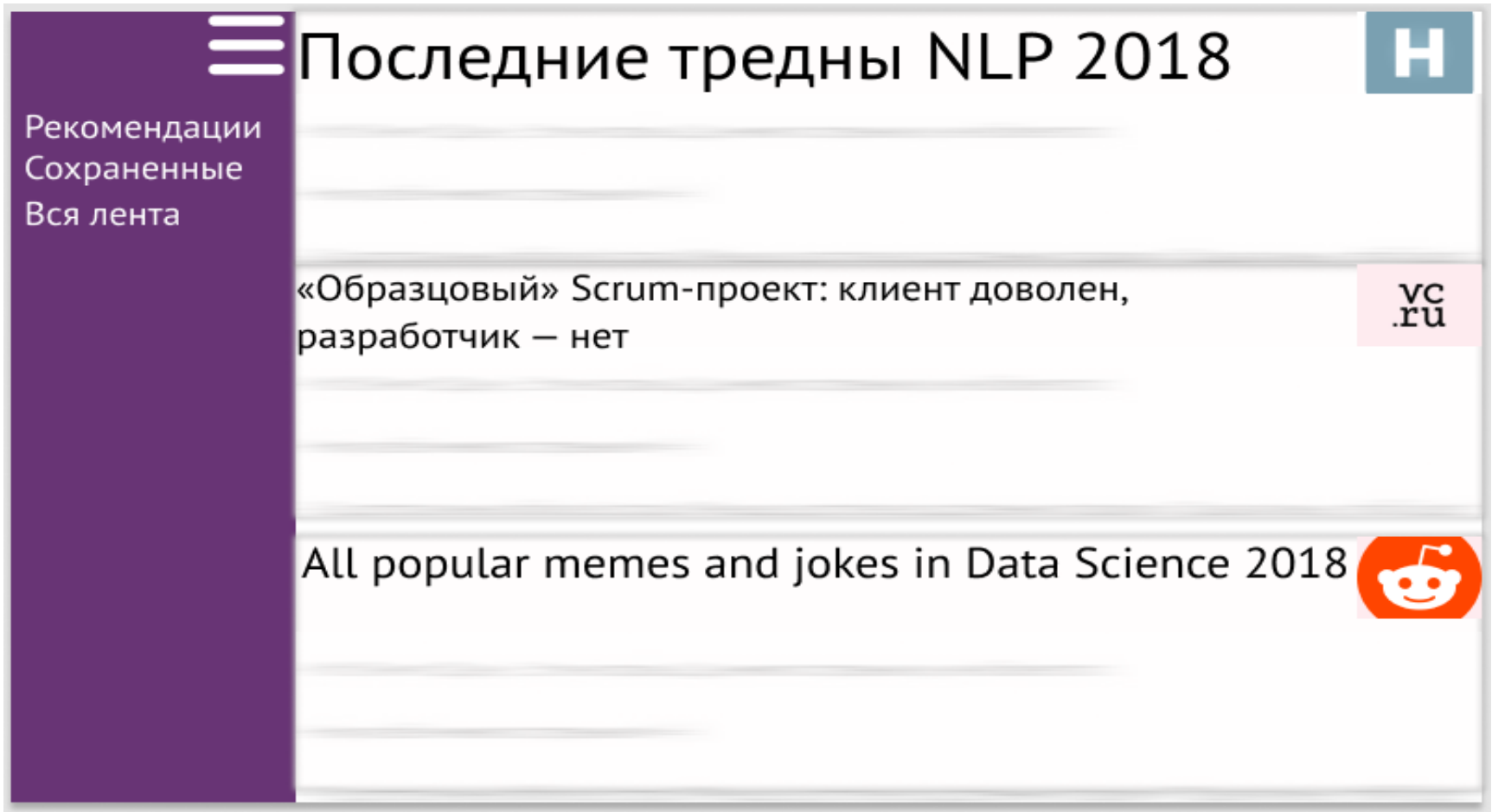
Средняя ежедневная лента: $\approx$ 35 постов
Действительно интересные и полезны: $\approx$ 5-7 постов
Всего затрачивается времени: $\approx$ 3 часа (1 час утром, 2-2.5 часа вечером)

Идеальное чтение

Как хотелось бы читать статьи:

- Как можно меньше времени тратить на пролистывание всей ленты
- Сохранять полезные статьи по темам с кратким содержанием
- Получать рекомендации на основе сохраненных мною статей

Создание нового сервиса



Создание нового сервиса

- Агрегация постов и статей с указанных источников
- Возможность сохранения/добаление в любимые с кратким конспектом
- Автоматическое подстраивание ежедневных рекомендаций на основе предпочтений
- Подсказка новых тематик на основе трендов

Создание точной рекомендательной системы

Дано: Папка с сохраненными полезными текстами и поступившая в ленту новая статья

Определить: Выносить новую статью в ежедневную рекомендательную ленту (т.е. считать, что статья будет интересна и полезна пользователю)

Критерий качества: Пользователь в течения дня отмечает из рекомендованных статей как полезные $>80\%$, а из основной ленты $<20\%$

Решаемые задачи

- Определение удобного векторного представления документов:
  - doc2vec, sen2vec
  - Вектор тем
- Определение способа измерения расстояния для сравнения вектора запроса и новой статьи

Покрытие терминологии и тематики документа

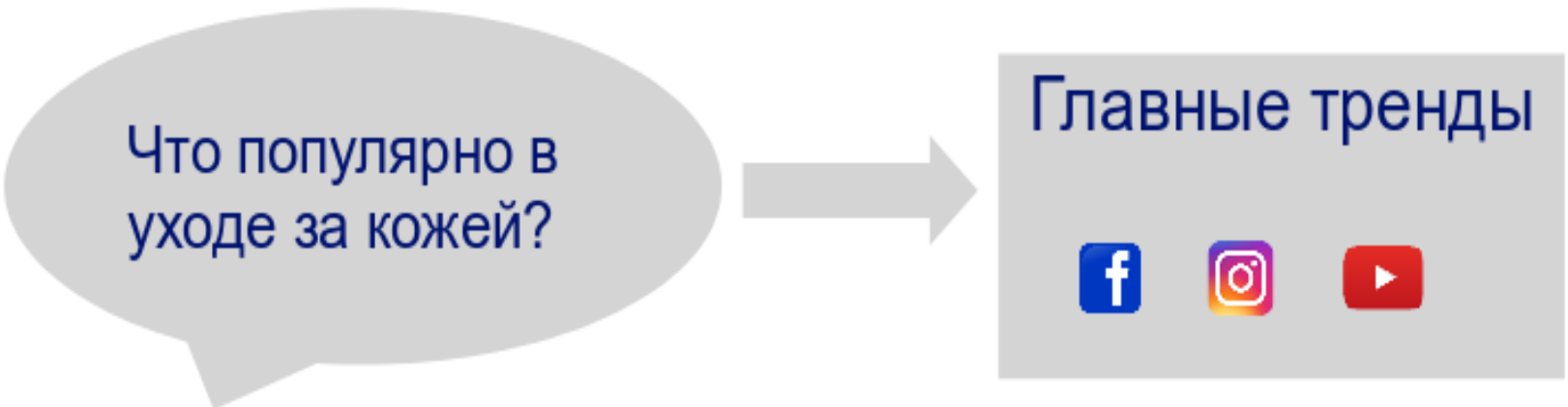
Дано: S_d — множество предложений документа d
Найти: Суммаризацию данного документа a $\subset S_d$
Критерий качества: Суммаризация должна покрывать основные факты, описанные в документе

Фоновое отслеживание трендов по выбранным тематикам

Дано: Набор документов, интересных польхователю, и лента всех новостей с агрегированных ресурсов
Найти: Трендовые темы, связанные с интересами пользваотеля, которые ему стоит порекомендовать
Критерий качества: Тема действительно является трендом, и пользователь отмечает ее как полезную

Проект для компании Beiersdorf (Nivea)

Цель: мониторить тренды в сфере косметики по соц сетям

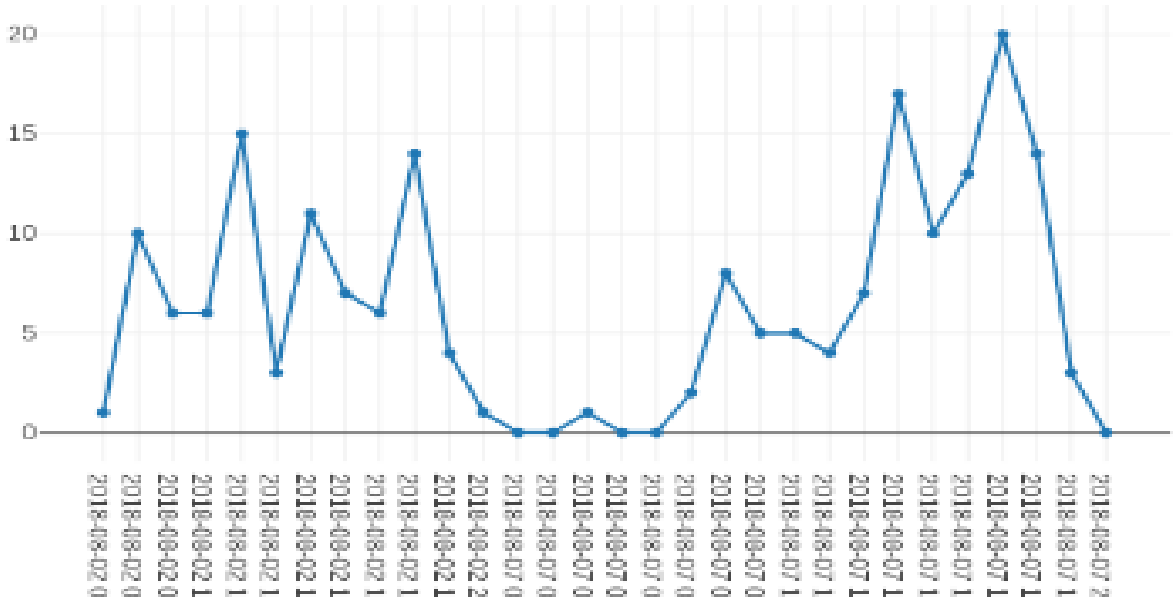


Основные шаги алгоритма и используемые ресурсы

- Сбор данных:
  - Извлечение ключевых слов из запроса: rake
  - Исследование вохможностей получения данных из соц.сетей
  - Написание своего инструмента для сбора постов по ключевым словам из Instagram и Medium
- Предобработка текста:
  - Определение языка: Apache OpenNLP
  - Удаление стоп слов, хэштегов, эмоджи, лемматизация
- Мониторинг трендовых тем
  - Веделение тем: BigARTM
  - Выделение новых трендовых тематик: Twitter Trend Detection

Визулизация тематических кластеров

Визуализации упоминания тем во времени



Mentions of hashtag **#treatment** per hour on 1-2 weeks of July

Следующие шаги

- Создание прототипа описанного сервиса:
  - Усовершенствование инструмента собирания контента из блогов и журналов
  - Источники контента: habr и vc.ru
- Первоначальный сбор данных:
  - Использование подписок на habr и vc.ru
  - Уже размеченные статьи как полезные из данных источников
- Построение более чувствительной рекомендательной системы:
  - Определение способа измерения расстояния для сравнения вектора запроса и новой статьи
  - Проверка на новостной ленте из указанных источников за последние 4 месяца: процентное соотношение корректно отнесенных к полезным и добавленный шум
- Тестирование и оценка качества на нескольких пользователях

Спасибо!
@dementyeva_ds