ВЫДЕЛЕНИЕ КОЛЛОКАЦИЙ

ПЛАН

МФТИ

- Что такое коллокации
- Зачем выделять коллокации
- Взаимная информация
- Выделение коллокаций по PMI
- Комбинированный подход
- Другие статистические методы
- Простая эвристика

ЧТО ТАКОЕ КОЛЛОКАЦИИ

МФТИ

- Коллокация — устойчивое словосочетание
- Мы для простоты будем рассматривать биграммы, но на N-граммы все обобщается
- Примеры:
  - ставить условия
  - назначать встречу
  - крейсер «Аврора»

ЗАЧЕМ ВЫДЕЛЯТЬ КОЛЛОКАЦИИ

МФТИ

- Идея 1 — более качественные признаки
- Идея 2 — визуализация текстовых данных:
  - Представленные в тексте темы
  - Тематическое моделирование
  - Кластеризация
  - Понижение размерности и визуализация

ВЗАИМНАЯ ИНФОРМАЦИЯ

МФТИ

- PMI – Pointwise Mutual Information

$$PMI(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

- Совместное вхождение более вероятно, чем бы для независимых событий
- Вместо вероятностей используются частотные оценки

ВЫДЕЛЕНИЕ КОЛЛОКАЦИЙ ПО PMI

МФТИ

- $PMI > t$
- Порог t подбираем для конкретного датасета

КОМБИНИРОВАННЫЙ ПОДХОД

МФТИ

- Вариант 1:
  - По PMI делаем отсечение по порогу
  - И берем N самых частых биграмм
- Вариант 2:
  - Пересекаем топN по PMI и топM по частотам

ДРУГИЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

МФТИ

- По матожиданию и дисперсии разности позиций слов
- t-тест
- χ^2-квадрат тест
- Отношение правдоподобий

ПРОСТАЯ ЭВРИСТИКА

МФТИ

- Вариант 1:
  Взять N самых частотных биграмм
- Вариант 2:
  Взять N биграмм с самой большой документной частотой

ИТОГ

МФТИ

- Что такое коллокации
- Зачем выделять коллокации
- Взаимная информация
- Выделение коллокаций по PMI
- Комбинированный подход
- Другие статистические методы
- Простая эвристика

ЯЗЫКОВЫЕ МОДЕЛИ

ПЛАН

МФТИ

- Общая постановка задачи
- Частный пример постановки
- N-граммные языковые модели
- Возможные применения

ОБЩАЯ ПОСТАНОВКА ЗАДАЧИ

МФТИ

- Оценка распределения вероятностей последовательностей слов

$$P(\omega_1, \ldots, \omega_m)$$

ЧАСТНЫЙ ПРИМЕР

МФТИ

- Вероятность при условии предыдущих слов

$$P(\omega_i|\omega_1, \ldots, \omega_{i-1})$$

N-ГРАММНЫЕ ЯЗЫКОВЫЕ МОДЕЛИ

МФТИ

$$P(\omega_1, \ldots, \omega_m) = \prod_{i=1}^{m} P(\omega_i | \omega_1, \ldots, \omega_{i-1}) \approx$$

N-ГРАММНЫЕ ЯЗЫКОВЫЕ МОДЕЛИ

МФТИ

$$P(\omega_1, \ldots, \omega_m) = \prod_{i=1}^{m} P(\omega_i | \omega_1, \ldots, \omega_{i-1}) \approx$$

$$\approx \prod_{i=1}^{m} P(\omega_i | \omega_{i-(n-1)}, \ldots, \omega_{i-1})$$

N-ГРАММНЫЕ ЯЗЫКОВЫЕ МОДЕЛИ

МФТИ

$$P(\omega_1, \ldots, \omega_m) = \prod_{i=1}^{m} P(\omega_i | \omega_1, \ldots, \omega_{i-1}) \approx$$

$$\approx \prod_{i=1}^{m} P(\omega_i | \omega_{i-(n-1)}, \ldots, \omega_{i-1})$$

$$P(\omega_i | \omega_{i-(n-1)}, \ldots, \omega_{i-1}) =$$

$$= \frac{\text{count}(\omega_{i-(n-1)}, \ldots, \omega_{i-1}, \omega_i)}{\text{count}(\omega_{i-(n-1)}, \ldots, \omega_{i-1})}$$

N-ГРАММНЫЕ ЯЗЫКОВЫЕ МОДЕЛИ

МФТИ

- Обычно:
  - $N = 2$ (биграммы)
  - $N = 3$ (триграммы)

N-ГРАММНЫЕ ЯЗЫКОВЫЕ МОДЕЛИ

МФТИ

- Проблемы
  - Для больших N мало статистики
  - Малые N недостаточно хорошо моделируют осмысленный текст
  - Плохие оценки вероятностей, если мало статистики

ПРИМЕНЕНИЯ ЯЗЫКОВЫХ МОДЕЛЕЙ

МФТИ

- Оценивать вероятность появления текста (например, можно использовать её в классификаторах)
- Генерировать тексты

Вся Солнечная система заполнена микроскопическими пылинками и астероидами, и в советское время такие опыты? Воды выступала дальняя половина пояса астероидов, в этой фазе человек испытывает так называемую сонную амнезию, в том числе предназначенные для управления маркетинговыми кампаниями в социальных сетях, могут предотвращать скапливание тромбоцитов, что в одной из самых высокоэффективных вариантов горючего для космических ракет?

РЕЗЮМЕ

МФТИ

- Общая постановка задачи
- Частный пример постановки
- N-граммные языковые модели
- Возможные применения

АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТА

ПЛАН

МФТИ

- Примеры, применения и трудности
- Простая постановка задачи и простое решение
- Наблюдения из практики
- Возможные постановки задачи
- Данные

ПРИМЕРЫ, ПРИМЕНЕНИЯ И ТРУДНОСТИ

ПРИМЕР НА АНАЛИЗ ТОНАЛЬНОСТИ

МФТИ

- Я купил этот телефон две недели назад.
  Всё изначально было хорошо.
  Отличный звук, батарея жила долго.
  Но вчера он перестал работать.
- Объективные и субъективные предложения
- Характеристика текста в целом и отдельных предложений
- Характеристики: общее впечатление, звук, батарея

ПРИМЕНЕНИЯ SENTIMENT ANALYSIS

МФТИ

- Для потребителя: анализ отзывов на товары
- Для организаций: замена опросов и фокус-групп
- Политика: результаты выборов и мнение избирателей
- Биржевые торги: анализ оценок экспертов и предсказание курсов

SA: ОСНОВНЫЕ ТРУДНОСТИ

МФТИ

- Тексты от пользователей отличаются от текстов, прошедших редактуру
- Люди используют различные наборы слов в зависимости от пола, возраста, страны проживания...
- Слова меняют эмоциональную окраску в зависимости от предмета описания
- Сарказм
- Каждый сайт с отзывами навязывает некоторую модель написания текста

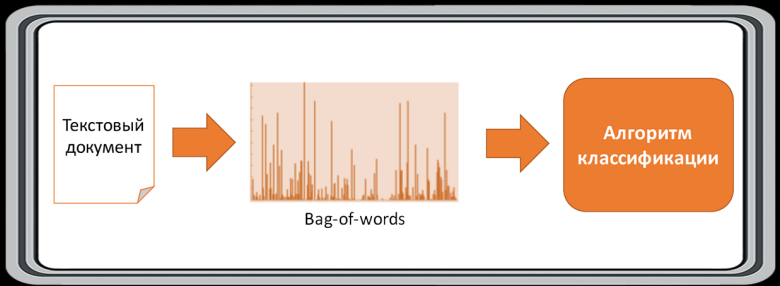
ПРОСТАЯ ПОСТАНОВКА ЗАДАЧИ И ПРОСТОЕ РЕШЕНИЕ

ПРОСТАЯ ПОСТАНОВКА ЗАДАЧИ

МФТИ

- 2 класса — позитивный и негативный
- Признаки — обычный мешок слов

ПРОСТОЕ РЕШЕНИЕ ЗАДАЧИ



НАБЛЮДЕНИЯ ИЗ ПРАКТИКИ

ЛЕММАТИЗАЦИЯ И СТЕММИРОВАНИЕ

МФТИ

- Лемматизация — приведение в нормальную форму:
  - летел → лететь
  - самолетами → самолет
  - идешь → идти
  - шел → идти
- Стеммирование — выделение основы слова:
  - летел → лет
  - самолетами → самолет
  - идешь → ид
  - шел → шел
- Часто не улучшает качество в сентимент-анализе

ОТРИЦАНИЯ

МФТИ

- Отзывы разной тональности с одинаковым представлением в модели мешка слов:
  - Это лучшая модель, экран не отвратительный, как в прошлой
  - Это не лучшая модель, экран отвратительный, как в прошлой

- Простейший способ – объединять с частицей «не» в один токен:
  - не отвратительный $\rightarrow$ не_отвратительный
  - не лучшая $\rightarrow$ не_лучшая

РАЗГОВОРНЫЙ ЯЗЫК, ОШИБКИ

МФТИ

- Частоты буквенных n-грамм вместо частот слов позволят похожим образом учитывать в тексте разные варианты написания одного слова

- Пример с триграммами:
  - ужасно $\rightarrow$ (ужа, жас, асн, сно)
  - ужаааасно $\rightarrow$ (ужа, жаа, ааа, аас, асн, сно)

- В текстах с этими словами будет хотя бы три общих токена

ПУНКТУАЦИЯ

МФТИ

- Сижу в кино на «Вспомнить все» :)
- Вчера купил новый айфон, сложно описать эмоции словами :(
- Очень рекомендую эту модель!

ВОЗМОЖНЫЕ ПОСТАНОВКИ ЗАДАЧИ

МУЛЬТИКЛАССОВАЯ КЛАССИФИКАЦИЯ

МФТИ

- Классы отзывов:
  - Положительные
  - Негативные
  - Нейтральные

- Проблема: отнести негативный к нейтральному — не так плохо, как в позитивному

ПРОГНОЗИРОВАНИЕ ОЦЕНКИ

МФТИ

- Обучать алгоритм предсказывать не класс, а оценку
- В этом случае, конечно, надо решать задачу регрессии
- Плюс: алгоритм начинает чувствовать разную цену ошибок
- Минус: плохо интерпретируемый функционал качества

- Два класса — позитивные и негативные отзывы
- Когда не уверены — говорим, что отзыв «без яркой эмоциональной окраски»
- При доработке:
  - Повышаем качество вне серой зоны
  - Уменьшаем её размер

УРОВНИ SA

МФТИ

- Документ
  - Положительное или отрицательное мнение или отношение выражает данный документ?
- Предложение
  - Предположение: «маленький документ», содержащий только одно мнение
  - Фактически — промежуточный этап

УРОВНИ SA

МФТИ

- Предложение
  - Предположение: «маленький документ», содержащий только одно мнение
  - Фактически — промежуточный этап

- Аспект

  Примеры:
  - отличный звук
  - батарея живет долго
  - дисплей яркий

МФТИ

ДАННЫЕ

ВАРИАНТ 1: ВЗЯТЬ ГОТОВЫЙ ДАТАСЕТ

МФТИ

- https://www.cs.cornell.edu/people/pabo/movie-review-data/
- http://www.sananalytics.com/lab/twitter-sentiment/
- http://inclass.kaggle.com/c/si650winter11/data
- http://nlp.stanford.edu/sentiment/treebank.html

ВАРИАНТ 2: ПАРСИТЬ САЙТ С ОТЗЫВАМИ

МФТИ

- Примеры:
  - сайты с отзывами на фильмы
  - сайты интернет-магазинов
  - сайты с отзывами на работу организаций и компаний

ИТОГ

МФТИ

- Примеры, применения и трудности
- Простая постановка задачи и простое решение
- Наблюдения из практики
- Возможные постановки задачи
- Данные

АННОТИРОВАНИЕ

ПЛАН

МФТИ

- Постановки задачи
- Какие бывают методы
- Простые методы без учителя
- Трудности

ПОСТАНОВКИ ЗАДАЧИ

МФТИ

- Сократить текст
- Выделять в тексте ключевые предложения
- Получить краткую аннотацию для коллекции текстов

ПРИМЕР

Промышленная группа «Базовый Элемент» объединяет около 100 российских и международных предприятий, работающих в энергетической, горнодобывающей, металлургической, машиностроительной, авиационной, финансовой, сельскохозяйственной и других отраслях.

Обширный и диверсифицированный портфель активов «Базового Элемента» представляет собой единую бизнес-структуру, предприятия которой эффективно взаимодействуют между собой, реализуя партнерские программы, позволяющие усилить конкурентные преимущества отдельных компаний группы и холдинга в целом.

В состав «Базового Элемента» входят лидеры крупнейших промышленных отраслей. Среди них – ведущий мировой производитель алюминия РУСАЛ, крупнейший в России частный производитель электроэнергии «ЕвроСибЭнерго» (входят в группу En+), автомобильный холдинг номер один в России «Группа ГАЗ» (входит в холдинг «Русские машины»), а также компания «Главстрой», лидер строительного рынка Москвы и Санкт-Петербурга.

Масштабная и активная деятельность предприятий «Базового Элемента» вносит существенный вклад в развитие российской промышленности, экономики и социальной инфраструктуры. Бизнес-группа обеспечивает около 1% ВВП России и инвестирует значительные средства в развитие регионов страны. «Базовый Элемент» – один из крупнейших работодателей в России. В течение последних пяти лет компания создала более 15 тыс. новых рабочих мест и планирует создать еще несколько десятков тысяч рабочих мест к 2025 году.

Стратегия роста «Базового Элемента» направлена на укрепление лидерских позиций предприятий группы за счет повышения эффективности производства, расширения промышленной базы и реализации крупных инновационных проектов, а также на содействие социально-экономическому развитию территорий своего присутствия и обеспечение экологической безопасности производства.

МЕТОДЫ РЕШЕНИЯ

МФТИ

- Обучение с учителем (supervised learning)
  - Нужна разметка: какие предложения оставить (решаем задачу классификации)
  - Признаки предложений: длина, количество слов с заглавной буквы/терминов, встречаемость слов из предложения во всем тексте и др.
  - Любой классификатор

МЕТОДЫ РЕШЕНИЯ

МФТИ

- Обучение без учителя (unsupervised learning)
  - Подсчет «значимости» предложений на основе их содержания
  - Выделение групп предложений, относящихся к одной общей идее

ПРОСТОЙ БЕЙЗЛАЙН

МФТИ

- Рассматривается корпус, состоящий из отдельных предложений документа и самого документа
- Признаки — частоты слов
- Cosine similarity между документом и предложениями — ранг предложений
- Аннотация — предложения с рангом выше заданного порога (или предложения с самым высоким рангом)

КЛАСТЕРИЗАЦИЯ ПРЕДЛОЖЕНИЙ

МФТИ

- Признаки предложений — частоты слов
- Кластеризуем предложения k-Means
- Оставляем ближайшие к центрам кластеров предложения

ПОНИЖЕНИЕ РАЗМЕРНОСТИ

МФТИ

- Во всех предыдущих примерах мы получали представление предложений в модели мешка слов
- Можно пробовать преобразовывать пространство признаков

БОЛЕЕ СЛОЖНЫЕ МЕТОДЫ

МФТИ

- TextRank
- Аннотирование с помощью нейросетей

РЕЗЮМЕ

МФТИ

- Постановки задачи
- Какие бывают методы
- Простые методы без учителя
- Трудности