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