# Автоматическое извлечение оценочных слов для конкретной предметной области

Четверкин Илья
ВМК МГУ
Лукашевич Наталья
НИВЦ МГУ

# Определение задачи

- Огромное количество отзывов о товарах и услугах в интернете
  - *Неплохой фильм, главное не выключить его в начале, где он напоминает просто ужасную пародию на Адреналин. Ну а в целом в фильме есть, как и положительные (адреналиновые, захватывающие и интересные сцены) так и отрицательные (неоднозначный финал, не везде удачная режиссура) качества.*

- Поиск и классификация отзывов – актуальная и востребованная задача

# Решение задачи

- Сбор дополнительных знаний
  - Объекты обзоров
  - Атрибуты объекта
  - ***Оценочные слова и выражения***

- ***Неожиданная*** *развязка и новые герои делают этот фильм* ***непохожим*** *на предшественника.*

- **Задача**: выделение оценочных слов для конкретной предметной области (кинематограф)

# Оценочные слова характерные для фильмов

*затянутый*

*унылый*

*предсказуемый*

*асоциальный*

*атомный*

*безликий*

*вдумчивый*

*никакущий*

*бойня*

*ахинея*

*плеваться*

*зацепить*

# План

- Формирование текстовых коллекций
- Вычисление характеристик слов
- Комбинирование характеристик методами машинного обучения

# Данные

- Для решения задачи выделения оценочных слов было подготовлено 4 корпуса
  - *Корпус мнений* (30 тысяч отзывов с пользовательскими оценками от 1 до 10)
  - *Корпус описаний* (20 тысяч описаний объектов)
  - *Новостной корпус* (1 млн. документов)
  - *Малый корпус* (составлен из частей корпуса мнений)
- Предварительная морфологическая обработка всех данных
- Слова разделяются на прилагательные и неприлагательные

# Малый корпус

- Составные части
  - Предложения, заканчивающиеся на «!»
  - Предложения, заканчивающиеся на «…»
  - Короткие предложения не более, чем из 7 слов
  - Предложения, содержащие слово «фильм», без других существительных
  - Короткие отзывы, состоящие из одного предложения

- Размер малого корпуса примерно в 2.5 раза меньше, чем у корпуса мнений

# Характеристики

- Для каждого слова вычисляется 18 характеристик
  - Частотные (6 характеристик)
    - Частота слова во всем корпусе
    - Количество документов (отзывов), в которых встречается слово
    - Частота слов с большой буквы
  - По парам корпусов (10 характеристик)
    - TFIDF
    - «Странность»
  - Отклонение от средней оценки
  - Существительные, связанные с «оценочными» прилагательными

# Частотность

| | | | |
|---|---|---|---|
| ФИЛЬМ | 45251 | МОЖНО | 4469 |
| ОЧЕНЬ | 11838 | ПОНРАВИТЬСЯ | 4384 |
| БЫТЬ | 9190 | ЛЮДИ | 3822 |
| СМОТРЕТЬ | 8511 | ГЛАВНЫЙ | 3529 |
| МОЧЬ | 6645 | ИГРА | 3204 |
| ЖИЗНЬ | 5643 | ПРОСМОТР | 3131 |
| ХОРОШИЙ | 5477 | КОНЕЦ | 3127 |
| ГЕРОЙ | 5394 | ИСТОРИЯ | 3041 |
| АКТЕР | 5229 | ЧЕЛОВЕК | 3036 |
| ПОСМОТРЕТЬ | 4835 | ЛЮБОВЬ | 3009 |
| ВРЕМЯ | 4810 | РОЛЬ | 2899 |
| СЮЖЕТ | 4809 | КОМЕДИЯ | 2823 |
| КИНО | 4633 | СКАЗАТЬ | 2748 |

# TFIDF

**TFIDF(l) = β + (1- β)*tf(l)*idf(l)**

tf(l) – частота леммы l, в исследуемом корпусе

Idf(l) = log((|c|+0.5)/df(l))/log(|c|+1)

df(l) – количество документов, в которых встречалась лемма l в контрастной коллекции

β = 0.4

|c| - количество документов в коллекции.

# Странность

**Weirdness = (FRL/FRC)/(FRLC/FRCC)**

FRL - частотность леммы, в исследуемой коллекции
FRC - число словоупотреблений, во всей исследуемой коллекции
FRLC - частотность леммы в контрастной коллекции
FRCC - число словоупотреблений в контрастной коллекции.

- Вместо частотности можно использовать количество документов, в которых встретилась лемма

# Отклонение от средней оценки

$$dev(l) = \left| \frac{\sum_{i=1}^{n} m_i k_i}{k} - \frac{\sum_{i=1}^{n} m_i}{n} \right|$$

$$\sum_{i=1}^{n} k_i = k$$

l – рассматриваемая лемма

n – общее количество отзывов

mi – оценка i-го отзыва

ki – число словоупотреблений леммы в i-ом отзыве (если не употребляется, тогда 0)

# Существительные связанные с прилагательными

- *Идея:* с оценочными прилагательными согласуются неоценочные существительные

- Из выдачи классификатора были взяты первые 200 прилагательных
  - Точность 90%
- Извлечены существительные, следующие сразу за этими прилагательными, и подсчитана их частотность
  - Пример: Я сегодня посмотрел *отличный* **фильм**!

# Оценка для отдельных характеристик

- Лучшие показатели по количеству оценочных слов в первой тысяче, по группам
  - Прилагательные
    - Частотные: 58.7%
    - По двум корпусам: **64**%
    - Отклонение от средней оценки: 56.3%
  - Неприлагательные
    - Частотные: 21.4%
    - По двум корпусам: **41.7**%
    - Отклонение от средней оценки: 30.6%

# Классификация

- Необходимо классифицировать слова на оценочные и неоценочные
- Для работы оставлено 10 тысяч слов из корпуса мнений с самой высокой частотой
- Отдельная работа с прилагательными и остальными частями речи
- Разметку производили два эксперта

# Машинное обучение

- Алгоритмы машинного обучения
  - Метод k ближайших соседей
  - «Наивный» Байесовский алгоритм
  - Нейронные сети (1,2,3 слоя)
  - Логистическая регрессия
  - Метод опорных векторов (скалярное и радиальное ядра)

- Оценка работы алгоритмов
  - F-мера
  - Количество оценочных слов, попавших в первую 1000 слов, упорядоченных по байесовской «вероятности»

# Прилагательные

- Лучший показатель по F-мере 71.08% для двухслойной нейронной сети
- Количество оценочных слов в первой тысяче 69.1% для логистической регрессии
- Начало лучшего списка

| | |
|---|---|
| ДОБРЫЙ | КРАСИВЫЙ |
| ЗАМЕЧАТЕЛЬНЫЙ | СМЕШНОЙ |
| ВЕЛИКОЛЕПНЫЙ | ЛЮБИМЫЙ |
| ПОТРЯСАЮЩИЙ | ОТЛИЧНЫЙ |
| | ТРОГАТЕЛЬНЫЙ |

# Неприлагательные

- Несбалансированные данные
  - Уменьшение неоценочных слов в обучающей выборке
  - Понижение порога классификации

- Результирующая F-мера 50.59%
- Количество оценочных слов в первой тысяче 50.9%

- Необходимы дополнительные характеристики

# Отбор характеристик

- Отдельный интерес представляет насколько существенной является каждая характеристика➔*Отбор характеристик*
- Отбор проводился с помощью генетического алгоритма
- Оценка качества набора признаков производилась с помощью Correlation-based Feature Selection (CFS)

# Основные характеристики

- Прилагательные:
  - Количество обзоров, в которых встречается слово
  - Частота слов с большой буквы
  - Отклонение от средней оценки
  - TFIDF *малый-новости*
  - Странность по документам *мнения новости*
  - Странность по документам *мнения-описания*
- Не прилагательные:
  - Частотность в корпусе описаний
  - Существительные, связанные с «оценочными» прилагательными
  - Странность по документам *мнения новости*
  - Странность по частотности *малый-описания*
- Вывод: все корпуса оказались важными для решения задачи

# Выводы

- Предложен новый подход к извлечению оценочных слов на основе 4-х корпусов
  - Для каждого слова разработано 18 характеристик
  - Получены неплохие результаты классификации
  - Произведен отбор характеристик

# Планы

- Предполагается исследование более сложных типов оценочных конструкций (многословных, композиционных).
- Проверка устойчивости метода извлечения слов для другой предметной области
  - Формирование качественного списка общезначимых оценочных слов, используя списки по нескольким предметным областям


# Вопросы?