Информационная система анализа тональности текстов

Студент: Никаноров Георгий Максимович

Научный руководитель: к.т.н., доц. Филиппович Андрей Юрьевич

МГТУ им. Н.Э. Баумана, кафедра ИУ5

Цель работы:

Разработка системы для автоматизации анализа тональности текста и его фрагментов (слово, предложение)

- Под тональностью *слова* понимается позитивная или негативная эмоциональная окраска слова.
- Под тональностью *текста* понимается позитивное или негативное отношение его автора

Позитивные слова	Негативные слова
РАДОСТЬ **ПРОГРЕССИВНЫЙ** **ПОБЕДА** **УДОВОЛЬСТВИЕ** **ХОРОШО** **УДАЧА** **ПРОЧНОСТЬ** **ПРЕСТИЖ** **ЗДРАВЫЙ**	**ПРЕСТУПНОСТЬ** **КАТАСТРОФА** **ПОГИБНУТЬ** **ПРЕСТУПЛЕНИЕ** **КРУШЕНИЕ** **УГОЛОВНЫЙ** **УБИЙСТВО** **ВЗРЫВ** **ПОСТРАДАВШИЙ**

Примеры

Отзывы о ноутбуке

«Очень ***красивый*** дизайн, это наверное первое что бросается в глаза (но это не главное). Экран маленький и очень ***четкий***.»

«Процессор ***далёк от совершенства***. Оперативной памяти скорее всего придётся добавить, но это зависит от пользователя. И до превращения в замену КПК и E-book читалки ***не хватает*** пары кнопок на дисплее.»

Статья из газеты

«В результате ***ДТП*** в Домодедовском районе Подмосковья ***пострадали*** шесть человек. ***Инцидент*** произошел в субботу на восьмом километре Каширо-Симфелопольского шоссе.»

Актуальность работы

- Пользователи сети Интернет могут оставлять свои отзывы о товаре или услуге, высказывать мнение о людях и событиях.
- В связи с активным развитием и распространением Интернета задача компьютерного анализа тональности текста оказалась востребованной на рынке.
- Исследование субъективного образа объекта, естественно возникающего или намеренно формируемого в информационном поле, является составляющей
  - обеспечения эффективной политики и бизнеса,
  - оценки эффективности PR и рекламных компаний,
  - выбора целевой аудитории для маркетинга товаров и услуг. [1]

Возможные области применения анализа тональности

- Сбор отзывов пользователей сети Интернет

  Такая система может служить альтернативой порталам похожим на Epinions.com, которые созданы для сбора отзывов и ограничены в плане тематики, и собирать информацию о людях, политических событиях и т.д.

- Рекомендательные системы

  Например система не будет рекомендовать продукт, если он получил много отрицательных отзывов.

- Системы размещения баннеров

  Реклама будет размещена на странице с текстом, имеющим позитивную тональность, и не будет размещена на странице с негативным текстом

Типы оценочных шкал

- Бинарная шкала

  (позитивная/негативная тональность)

- Рейтинг, степени позитивности

  определить оценку относительно дифференцированной шкалы (например, одна - пять "звезд")

- Сравнительная оценка

  сравнительное предложение “оптика Canon лучше, чем Sony и Nikon” выражают сравнительное отношение:

  (лучше, {оптика}, {Canon}, {Sony, Nikon}).

Методы формализации текста

- Лемматизация
- Частота встречаемости слова, фразы
- Позиция слова или фразы в документе

- Часть речи (ЧР) слова и другая грамматич. информация
  - Основная информация о тональности – из прилагательных.
  - Отсеивание ненужных слов по ЧР, отсеивание названий и имён

- Анализ слов инвертирующих тональность

  Отрицательные частицы, наречия, прилагательные и некоторые глаголы.

  Пример: «Якобы хороший ноутбук» = «плохой ноутбук»

Методы формализации текста

- Синтаксический анализ
  - Фильтрация, снятие омонимии
  - Определение объекта (персона, организация, товар, событие)
  - Группировка слов, определение тональности группы
- Тема текста может задавать тональность всего документа

Подходы к решению задачи анализа тональности

- Использование словарей, составленных вручную
- Автоматизированное составление лексикона
- Лингвистическая эвристика

  Правила, основанные на наблюдениях
- Машинное обучение
  - Исходные данные для обучения созданы вручную
  - Исходные данные получены с использованием автоматизированной обработки

Применение правил для анализа тональности словосочетаний

Тональность сущ.	**Тональность прил.**	**Тональность группы**	**Пример**
Нейтрально	Любая	Тональность прилагательного	Плохой руководитель, Хороший руководитель
Любая	Нейтрально	Тональность существительного	Новый олигарх, новый победитель
Позитив	Негатив	Негатив	Плохой защитник демократии
Позитив	Позитив	Позитив	Отличный защитник демократии
Негатив	Любой	Негатив	Слабый диктатор, Сильный диктатор

При наличии хотя бы одного слова с негативной тональностью общая тональность группы негативна; в противном случае общая тональность позитивна, если присутствует хотя бы одно слово с позитивной тональностью.[1]

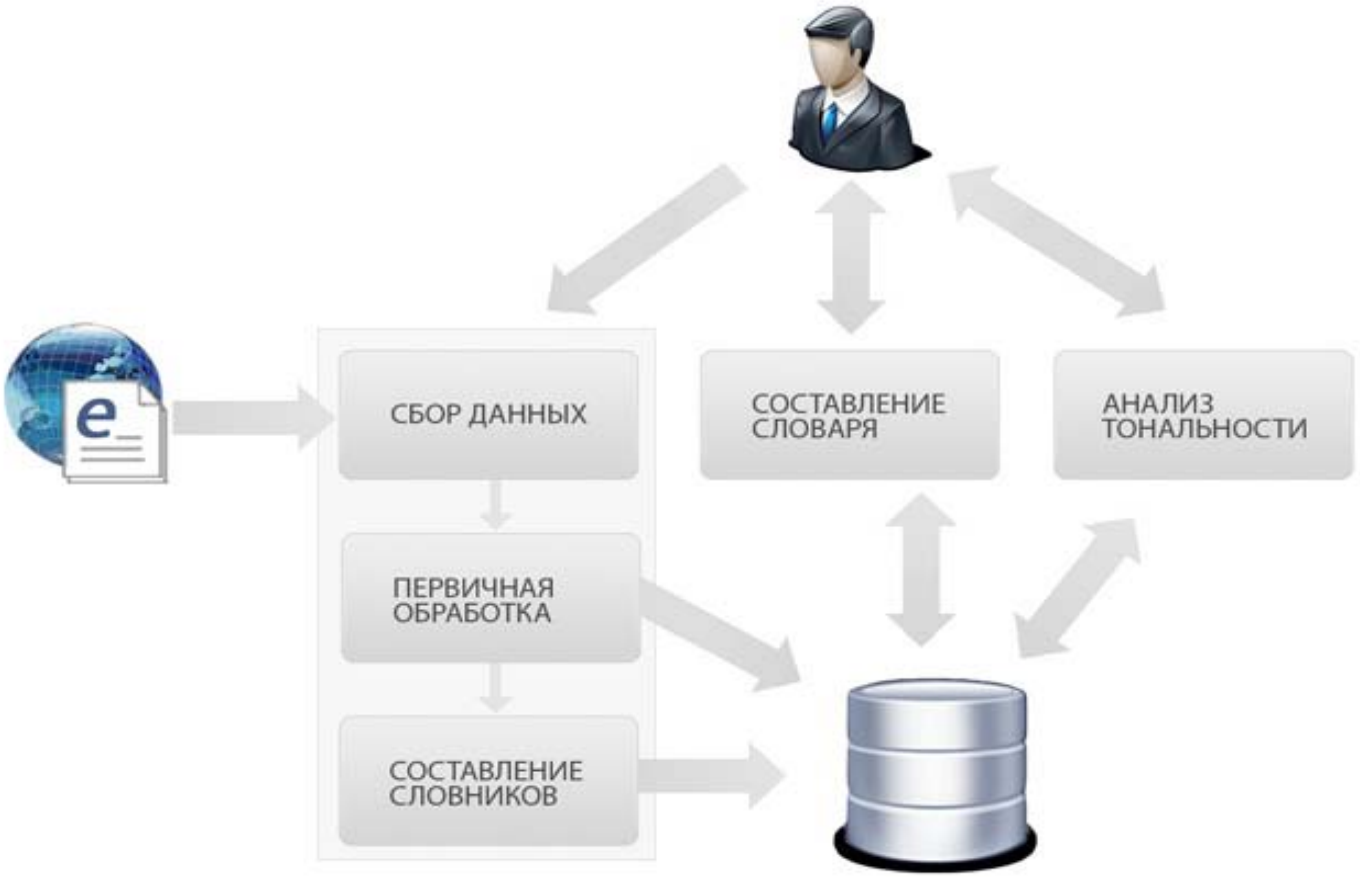
Проблемы при решении задачи анализа тональности

- Использование метафор.

  Пример: «Этот безумный фильм меня впечатлил». В этом предложении присутствует слово «безумный», которое негативно окрашено.
- Зависимость тональности от контекста.

  Например, комментарий «прочитай книгу» может быть расценен как позитивный на сайте отзывов о книгах и как негативный на сайте отзывов о фильмах.
- Обработка тональности зависит от предметной области текста.

Структура системы

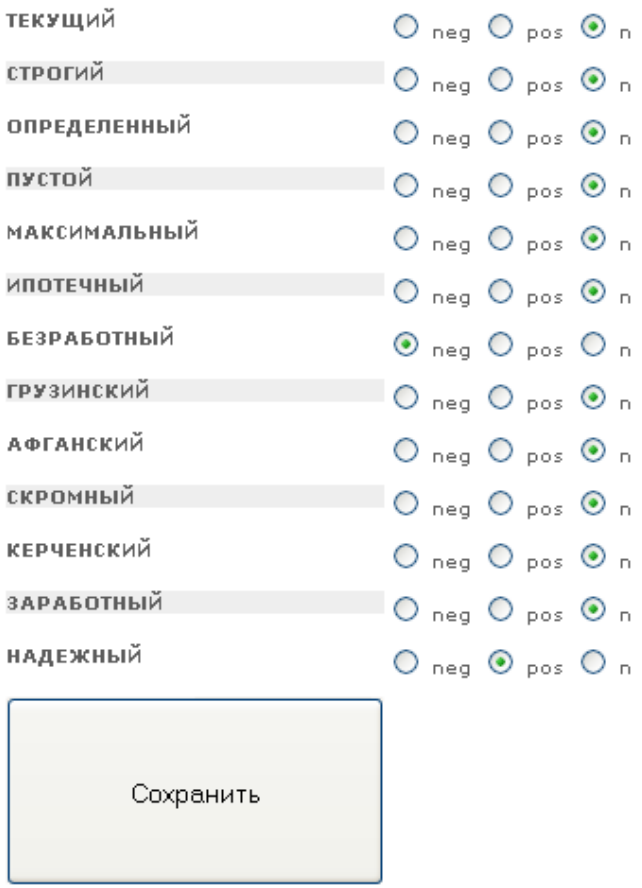


Модуль сбора данных и первичной обработки

- Сохранение текстов из html-документов
- Удаление стоп-символов
- Разделение текста на слова
- Определение начальной формы слова
- Определение части речи слова
- Подсчет частоты появления слова

Модуль составления словаря

- Выборка слов по части речи
- Сортировка по частоте появления
- Вывод слов, для которых определена тональность



Модуль анализа тональности

Исходные данные:

- 600 текстов с сайтов gazeta.ru и lenta.ru
- 223 слов с позитивной тональностью
- 397 слов с негативной тональностью

Модуль анализа тональности

Матрица связанности слов:

- Для каждой пары слов вычисляется частота совместной встречаемости (в предложении)
- Более точные частоты могут быть получены из поисковых систем (в ПС «Яндекс» оператор “&”)

	Слово 1	Слово j	Слово n
Слово 1 …	F 11	F 1j	F 1n
Слово k	F k1	F kj	F kn
… Слово n	F n1	F nj	F nn

Методы оценки тональности

$$L(Wi, POSj) = \log_2 \left(\frac{\frac{Freq(W_i, POS_j, ADJ_p) + t}{Freq(W_{all}, POS_j, ADJ_p)}}{\frac{Freq(W_i, POS_j, ADJ_n) + t}{Freq(W_{all}, POS_j, ADJ_n)}} \right)$$

W_i - слово

ADJ_p, ADJ_n – множество исходных слов и словосочетаний с заданной тональностью

POS_j = {Прилагательное, наречие, существительное, глагол}

$t = 0.5$ – константа сглаживания

Необходимо определить пороговые значения для получения тональности слова и предложения

$L > T_p$ – позитивное $T_n < L < T_p$ – нейтральное

$L < T_n$ – негативное [2]

Распределение хи-квадрат

- Критерий проверки значимости связи между двумя переменными.
- Ожидаемые частоты можно вычислить по таблице (между переменными нет зависимости)
- ожидаемые частоты не должны быть очень малы

$$\chi^2(w) = \sum_{x \in (w, \neg w)} \sum_{y \in (pos, neg)} \frac{\{f(x,y) - \hat{f}(x,y)\}^2}{\hat{f}(x,y)}$$

	pos	*neg*
w	f	f
$\neg w$	f	f

$$L_{\chi^2}(w) = \begin{cases} \chi^2(w), & \text{если } P(w|neg) < P(w|pos) \\ -\chi^2(w) \end{cases}$$

Результаты работы

Примеры полученных слов и значения критериев

слово	Хи-квадрат	Логарифм	Логарифм (только прилагательные)
БОЛЬШОЙ	19.16	1.73	1.73
УГОЛОВНЫЙ	-14.69	-1.56	-1.04
ПОСВЯЩЕННЫЙ	6.99	2.11	2.26

Графическое представление тональности текста

Во время перестрелки в ходе который был убит сотрудник милиции он получил ранения, однако смог скрыться. Розыск других подозреваемых продолжается. Перестрелка между преступниками и сотрудниками милиции произошла после того, как трое злоумышленников в масках поздно вечером 25 декабря напали на бензоколонку на 81 м километре МКАД.

Рекомендация о тщательных проверках рейсов направляющихся в США получена из Вашингтона. Чрезвычайные меры безопасности уже приняты на Тайване и в Сиднее, а аэропорт Мельбурна по указанию австралийского правительства последует их примеру с воскресенья. Вызвавший такое беспокойство инцидент произошел 25 декабря на борту лайнера А330 следовавшего из Амстердама в Детройт с 278 пассажирами на борту. По версии следствия 23 летний нигериец Абдул Мудаллад попытался взорвать в салоне самолета самодельное взрывное устройство

Оценка результатов работы системы

- Результаты оценивались двумя людьми, была произведена выборка полученных после обработки слов (только прилагательные)

- Точность для позитивных слов – 20%
- Точность для негативных слов – 40%

Дальнейшая работа

- Анализ эффективности различных подходов
- Применение системы для составления корпуса документов (дальнейшее использование при машинном обучении)
- Создание интернет-сервиса на основе системы