Структурные и статистические методы анализа эмоциональной окраски текста

Лукашкина Юлия

МГУ имени М. В. Ломоносова, факультет ВМК,

кафедра ММП

Научный руководитель:
к.ф-м.н., доцент
Чехович Юрий Викторович

28 мая 2015

Содержательная постановка задачи

- Анализ эмоциональной окраски (тональности) текста — это определение мнения автора по отношению к объектам, которые описываются в тексте.
- Пример:
  - “I did’n like the movie at all”, негативная тональность,
  - “I loved the movie”, положительная тональность отзыва.
- Цель работы: построить алгоритм автоматического определения тональности текста.

Формальная постановка задачи

Дано:

- Коллекция текстовых документов D
  - обучающая выборка документов с известными тональностями.
  - выборка документов с неизвестными тональностями
- Множество употребляемых слов W
- $\forall d \in D$ — последовательность слов $W_d = (w_1, \ldots, w_{n_d})$, где n_d — длина документа d.

Найти:

- Тональность документов $t \in T = \{0, 1\}$ (0 — негативный класс, 1 — позитивный).

Подход «мешок слов»

Любой документ может быть представлен моделью «мешка слов»[1]

$$d = (w_1, \ldots, w_{|W|}),$$

- $w_i = tf_i$ (частотный подход),
- $w_i = I[tf_i > 0]$ (бинарный подход),
- $w_i = tf_i \times idf_i = tf_i \times \log \frac{N}{df}$ (TF-IDF),

tf_i — частота встречаемости i-го слова в документе,
idf_i — обратная частота документа,
N — количество документов,
df — количество документов, содержащих данное i-ое слово.

[1] Pang Lee Bo Pang, Lillian Lee. Thumbs up?

N-граммы

Определение

N-грамма (N-gram) — подпоследовательность из N элементов некоторой последовательности.

- Пример: «There are some very good actors in the film.»
  - Униграммы: «There», «are», «some», «very», «good», «actors», «in», «the», «film»
  - Биграммы: «There are», «are some», «some very», «very good», «good actors», «actors in», «in the», «the film»
- Униграмы, биграммы, 3-граммы, N-граммы.
- Униграммы и биграммы, 1-N граммы.

Структурные методы

Алгоритм с использованием RST:[2]

- для каждого слова рассчитать оценку за положительный и негативный класс с помощью корпуса SentiWordNet;
- тональность всего документа — усреднение всех оценок;
- учитывать некоторые отношения: отрицание и усиление.

[2] Maite Taboada. Extracting sentiment as a function of discourse structure.

Другие методы

- PMI с помощью Google:
  - расчет тональности отдельного слова вычислением поточечной взаимной информации (PMI) с эталонными словами «excellent» и «poor»,
  - расчет тональности всего документа усреднением значений тональностей его слов;
- автоматическая разметка текста на части речи (part-of-speech tagger);
- введение различных весов для слов из разных частей текста (введения, заключения и т.д.).

Модификация TF-IDF

Была проведена модификация TF-IDF метода.
Предлагается:

- использовать биграммную модель представления документа,
- произвести отбор признаков с помощью ограничений на их частоту встречаемости,
- веса признаков вычислять с помощью TF-IDF,
- нормировать веса.

Композиция методов

- Голосование по большинству.
- Композиция наивного байесовского классификатора и SVM.

Обзор данных

Для проведения экспериментов были выбраны два набора данных:

- Movie Review Data — коллекция отзывов на кинофильмы
- Multi-Domain Sentiment Dataset — коллекция отзывов покупателей на различные продукты (книги, электроника и т.д.).

Все наборы данных сбалансированы по классам.

Предобработка данных

- приведение текста к нижнему регистру;
- стемминг;
- удаление редко и часто встречающихся слов.

Результаты экспериментов

	dvd	эл-ка	книги	к/ф
(1,2)	82.75, NN	79, NN	81.75, NB	86.5, SVM
(n,t)	82.5, NN	78.75, SVM	81.5, NN	86.75, SVM
(l, Δ(k))b	83.5, NB	79.25, SVM	82, NN	87, SVM
SVM vote	82.25	81.75[3]	82.25	87.5[4]
NN vote	84.5[5]	81	81.5	85.75
NB vote	83.75	80.25	85	85.5
NB + SVM	83.25	81.25	85.25[6]	86.5

[3] (n,t)+(l,k)+(l,k)b
[4] (1,2)+(1,3)+(n,t)+(l,Δ(k))+(l,Δ(t))b
[5] (1,2)+(l,Δ(*k*))+(l,Δ(*k*))b
[6] (1,3)+(1,4)+(n,t)

Выводы

- Наилучшие результаты удалось получить с помощью применения композиции алгоритмов.
- Экспериментально было установлено, что нормирование признаков дает небольшой прирост качества (1–2 %).
- Точность классификации всех выборок выше у TF-IDF модели, чем у простой бинарной (частотной) n-граммной модели.
- Частотное представление избыточно, и при бинарном представлении результаты получаются лучше.
- Среди n-граммных моделей наилучшие результаты получаются при комбинации униграмм и биграмм.

Положения выносимые на защиту

- Программная реализация различных методов, используемых для анализа эмоциональной окраски текстов;
- Реализация программного стенда для проведения экспериментов;
- Проведение экспериментов и сравнение различных методов на наборах реальных данных;
- Реализация алгоритмической композиции методов.