# Описание работ по усовершенствованию программного пакета TopicMiner в 2016 году

## С. Кольцов

Адаптация пакета сентимент-анализа SentiStrength в программный комплекс TopicMiner, тестирование словаря на этно-окрашенных текстах. Применение сентимент-анализа для исследования уровня напряженности межэтнических отношений.

**Особенности работы SentStrength**

В рамках совместного сотрудничества между лабораторией интернет-исследований и разработчиком ПО SentStrength профессором Маком Телволом (Mike Thelwall), Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, был предоставлен исходный код программы SentStrength. Анализ предоставленного кода показал следующий алгоритм работы программы.

**Алгоритм работы SentStrength**

Текст разбивается на предложения. Рассчитывается ‘score’ каждого предложения по описанной ниже формуле (считается, что у каждого слова есть положительная и отрицательная оценки). Под словом понимается токен. Токен – может быть слово, а может быть смайлик. Оценки берутся из словаря; именно в словаре учитывается влияние сопутствующих слов усилителей, типа ‘very big’. Соответственно, у слова ‘big’ оценка будет меняться в зависимости от наличия/отсутствия слова ‘very’.

**Расчет тональности предложения.**

Расчет тональности предложений происходит по следующим формулам:

Позитивная оценка:

iPositive = Int((iTotalPos + 0.5) / iWordTotal.

Негативная оценка:

iNegative = Int((iTotalNeg + 0.5) / iWordTotal),

где, iWordTotal – число слов в предложении, iTotalPos – максимальная положительная оценка слова в предложении. iTotalNeg – максимальная отрицательная оценка слова в предложении.

Таким образом, тональной оценкой предложения является максимальная оценка, поделенная на общее число слов в предложении.

**Расчет тональности текста.**

При условии, что у каждого предложения в тексте рассчитаны тональные оценки (как положительные, так и отрицательные), оценка всего текста считается так:

Позитивная оценка текста:

iPositive = Int((iPosTotal + 0.5) / iSentencesTotal).

Негативная оценка текста:

iNegative = Int((iNegTotal + 0.5) / iSentencesTotal),

где iNegMax – максимальная негативная оценка среди всех оценок по предложениям, iPosMax – максимальная позитивная оценка среди всех оценок по предложениям, iPositive – позитивная оценка текста, iNegative – негативная оценка текста, iSentencesTotal – общее число предложений в тексте.

Данный подход хорошо работает для коротких текстов типа твитов, однако в случае больших текстов, например, постов из социальной сети ‘Живой Журнал’, такой подход не применим, в силу того, что значительная часть слов в больших текстах имеет нулевую тональную оценку. Поэтому для больших текстов тональность будет близка к нулю.

Исходя из этого, а также из алгоритма препроцессинга на основе программы ‘mystem’ (разработка компании Yandex), для мониторинговой информационной системы ‘TopicMiner’ был реализован модифицированный алгоритм расчета сентиментных оценок для текстов из социальных сетей. Алгоритм расчета тональности базируется на том, что в большом тексте могут встречаться как положительные, так и отрицательные слова, но большая часть слов все же имеет нулевую тональность. Поэтому алгоритм рассчитывает сумму всех положительных оценок и сумму всех отрицательных оценок в каждом тексте. Далее он подсчитывает количество слов с положительной и отрицательной оценками. Итоговые сентимент-оценки больших текстов рассчитываются по следующим формулам:

iPositive = Int((sum_pos) / iPos_words_Total)

iNegative = Int((sum_neg) / iNeg_words_Total),

где sum_pos – сумма всех позитивных оценок в тексте, iPos_words_Total – количество слов с позитивной оценкой в тексте, где sum_neg – сумма всех негативных оценок в тексте, iNeg_words_Total – количество слов с негативной оценкой в тексте.

Таким образом, сентимент-оценки в тексте являются средними значениями, и не зависят от размера текста. В случае коротких текстов, данный вариант практически совпадает с результатами работы первого варианта программы SentStrength.

**Практическая реализация нового варианта SentStrength в информационной системе.**

В качестве исходного словаря для сентмент-анализа были использованы результаты гранта РГНФ ‘Разработка общедоступной базы данных и краудсорсингового веб-ресурса для создания инструментов сентимент-анализа’, номер заявки: 14-04-12031, 2013г. А именно, словарь размера порядка 7 тысяч слов с сентимент-оценками по социально-политическим темам. Далее данный словарь был дополнен набором этнофолизмов с тональными оценками, которые отражали специфику этничности.

Применение полученного словаря в информационной системе проводится в два этапа. На первом этапе словарь с сентимент-оценками проходит через процедуру лемматизации. На выходе получается бинарный файл, который в дальнейшем подключается к результатам тематического моделирования. Второй этап состоит собственно в применении словаря. В силу того, что результатами ТМ является две матрицы: 1. Матрицы распределения слов по темам. 2. Матрица распределения документов по темам, сентимент-анализ реализован для двух указанных матриц. Пользователь путем нажатия на кнопку запускает процесс расчета сентмент-оценок для распределения слов и документов по темам. В результате расчета в информационной системе у пользователя появляются две таблицы: 1. Таблица, содержащая слова, отсортированные по вероятности в каждой теме, а также сентимент-оценки для каждого слова в каждой теме. Данная таблица может быть выгружена во внешний файл в формате csv. 2. Таблица, содержащая документы, которые отсортированы по вероятности в каждой теме, и также тональные оценки каждого текста. Данная таблица также может быть выгружена во внешний файл в формате csv. В силу того, что данная мониторинговая система может работать с миллионами документов, пользователю предоставляется возможность указания размера выгрузки слов и документов по темам. Например, пользователь может указать, что надо рассчитать и выгрузить лишь 100 (или 500) наиболее вероятностных слов, и документов вместе с тональными оценками, и другими метаданными, включая геотеги. Это позволяет оптимизировать работу эксперта-аналитика.

Разбиение работы со словарем на два этапа обусловлено тем, что словарь может как пополнятся, так и усекаться по желанию аналитика. Например, в данный словарь могут быть добавлены слова с сентмент-оценками, отражающими отношение к тем или иным политическим действиям, отношению к террористическим организациям, к религиозным сектам или группам смерти. Поэтому данная информация система за счет только модернизации словаря может быть использована как для мониторинга этнической напряженности, так и для отслеживания политической напряженности или мониторинга групп смерти или поиска тем, связанных с террористическими организациями.

Другим немаловажным дополнением версии мониторинговой системы 2016 года является автоматическая подсказка релевантных тем, при помощи цветового выделения содержимого тем в таблице распределения слов по темам. В силу того, что число тем в тематическом моделировании может меняться от десятка тем до нескольких сотен, поиск актуальных тем среди огромного множества сложен. Для этого реализована следующая опция. Пользователь может подключить внешний файл с набором слов, который характеризует то или иное событие (например, список этнофолизмов). Программа, используя данный список, подсвечивает разными цветами темы, в которых были встречены эти слова. Красный цвет сигнализирует, что в данной теме найдено максимальное число слов из заданного списка. Всего реализовано 4 цвета: Красный цвет (максимальное число слов). 2. Зеленый цвет. 3. Синий цвет. 4. Желтый цвет (минимальное число слов). Темы, в которых среди наиболее

вероятностных слов слова из списка не были найдены, подсвечены не будут. Таким образом, меняя наборы слов, можно достаточно быстро ориентироваться в результатах тематического моделирования.

Таким образом, результатом работы в 2016 году является новая версия офлайновой информационной системы мониторинга, которая включает в себя сентимент-анализ и систему подсказок.