**II. МЕТОДЫ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ**

**12. АНАЛИЗ ОТЗЫВОВ**

**12.1. Постановка задачи**

Развитие сетевых технологий и, не в последнюю очередь появление социальных сетей, дало возможность пользователям явно выражать свое мнение относительно качества услуг, товаров, фильмов, книг, результатах деятельности компаний или политических деятелей.

Наличие подобной информации привело к необходимости ее анализа с точки зрения установления обратной связи между поставщиками услуг и их потребителями. При этом, количество отзывов может достигать десятков тысяч, и обработка отзывов вручную оказывается невозможной. В связи с этим широкое распространение получили автоматизированные подходы к анализу тональности текстов (sentiment analysis).

Однако, непосредственно перед анализом необходимо проводить процедуру отсеивания имеющихся отзывов с точки зрения выявления их дубликатов.

Особенно этим «грешат» форумы на сайтах компаний, являющихся поставщиками таких услуг. Не является секретом то, что часть отзывов (особенно положительных) пишут менеджеры данных компаний. Но, еще чаще для их написания привлекаются специальные люди – копирайтеры. В интернет имеется множество публикаций на тему как правильно писать подобные отзывы.

Задачей настоящей работы является освоение технологии отсеивания «заказных» отзывов.

**12.2. Описание алгоритма**

Для решения выше указанной задачи мало подходят стандартные методы классификации текстов (типа ***bag of words***) прежде всего из-за небольших объемов сравниваемых текстов – максимум 1-2 стр. формата А4. Получающиеся при этом частотные словари являются статистически неустойчивыми и, как следствие, результаты классификации являются не надежными.

**Понятие семантического спектра**

Для характеристики смысла текста вводится понятие семантического спектра. Данный спектр представляет собой своеобразную хэш-функцию заданной последовательности символов.

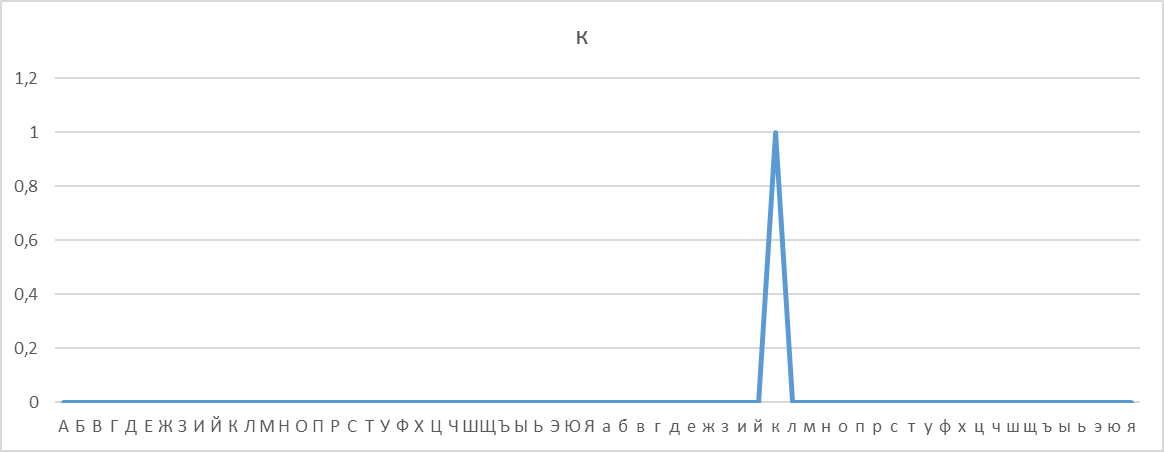
Указанная функция формируется следующим образом:

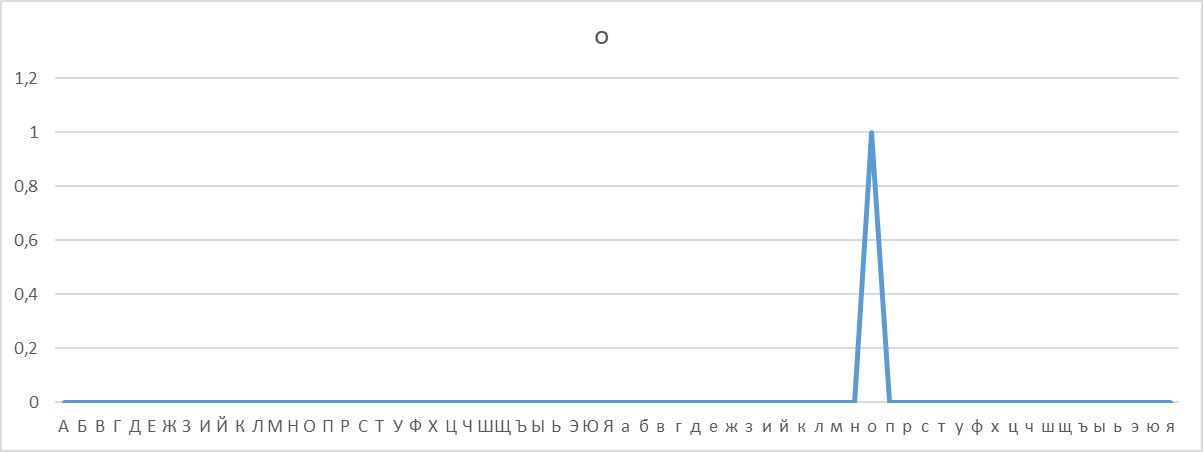
- задается набор символов, которые ожидаются в обрабатываемом тексте. Если это русскоязычные тексты, то в этот набор можно включить все строчные и прописные русские символы, цифровые символы и символы-разделители (точка, запятая, двоеточие и т.д.);

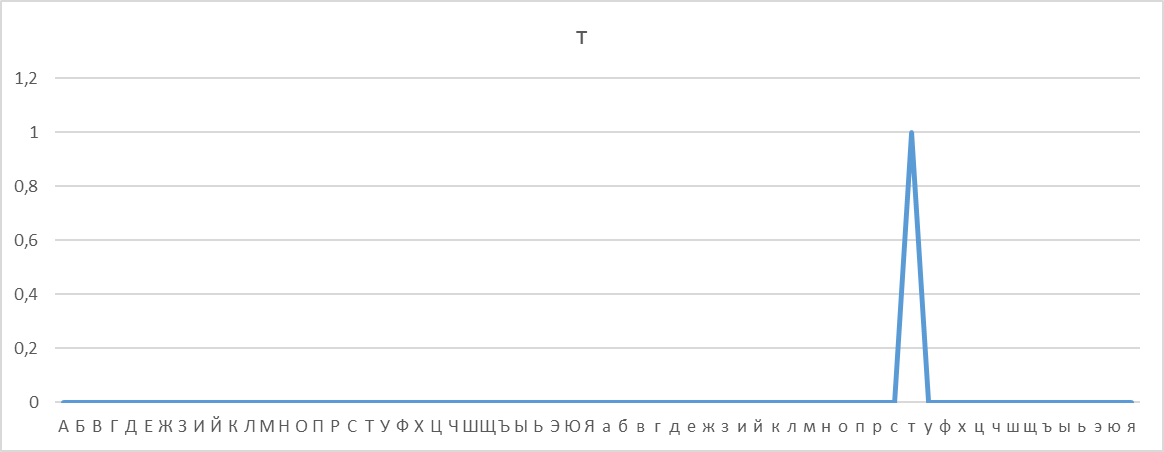
- выбранный набор символов упорядочивается. Упорядочивание наиболее естественно произвести в порядке возрастания ASCII-кодов символов;

- для каждого символа формируется стартовое значение семантического спектра. Формирование спектра производится по правилу: если номер элемента спектра равен номеру данного символа, то этот элемент спектра приравнивается 1. В противном случае элемент спектра равен 0.

Например. На рис. 5-7 показаны спектры букв «к», «о» и «т».







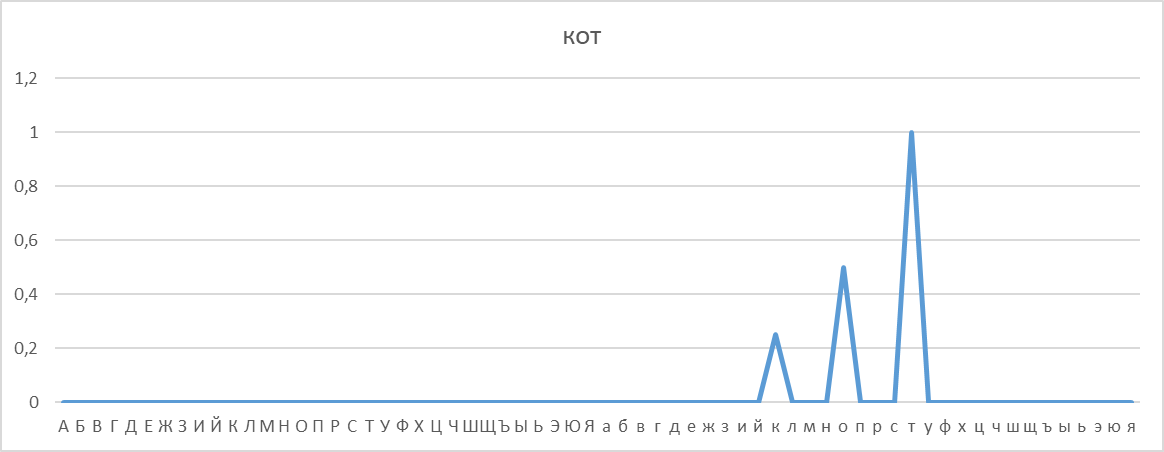
Для получения спектра последовательности символов (текста) выполняются следующие операции:

- задается начальный спектр текста, все элементы которого равны 0;

- поочередно берутся символы текста и их спектры поэлементно совмещаются с соответствующими элементами спектра текста по формуле:

(1)

На рис. 8 показан полученный спектр для последовательности символов «кот».



Особенности формулы (1) приводят к следующим эффектам:

1. Получаемый спектр позволяет учитывать не только вхождение символа в последовательность, но и его порядок в этой последовательности.
2. Чем раньше символ появляется в последовательности, тем меньше его значение в спектре. Это приводит к тому, что для очень длинных последовательностей символов «сигнал» от начальных символов станет практически незаметным.

Наличие второго эффекта, в частности, означает, что нельзя механически применять уравнение (3) для очень длинных последовательностей символов.

Чтобы не потерять информацию о начальных символах длина обрабатываемой последовательности ограничивается длиной очередного слова в тексте. Типичная длина русских слов (5-6 символов) позволяет сохранять информацию о начальных символах слова в получающемся спектре. Получившийся спектр слова можно затем по той же формуле (3) объединять с начальным (нулевым) семантическим спектром текста.

Однако, нельзя и весь текст сводить к набору слов, поскольку для большого текста вклад спектров начальных слов в общую семантику текста будет постепенно нивелироваться. Чтобы этого не произошло, производится формирование промежуточных семантических спектров предложений. Для русского языка средняя длина простых предложений составляет 5-6 слов. Такой длины еще вполне достаточно, чтобы не потерять информацию о семантике начальных слов.

Если продолжить аналогичные рассуждения дольше, то просматривается следующая иерархическая последовательность получения семантики текста с помощью формулы (3):

***стартовая семантика символов > семантика слов > семантика простых предложений > семантика составных предложений > семантика абзацев > семантика параграфов > семантика глав (разделов) > семантика текста.***

**Алгоритм обработки**

1. Обрабатываемый текст просматривается и все входящие в него слова заменяются базовыми словоформами.

Например:

Исходная фраза: «часто проводятся акции на разные товары»

преобразуется к виду: «часто проводить акция на разный товар»

1. Базовые словоформы заменяются условным базовым синонимом.

Например:

вышеприведенная фраза: «часто проводить акция на разный товар»

преобразуется к виду: «частый проводить поступок на отличный товар»

1. Формируется семантический спектр текста
2. С помощью шагов 1 – 3 формируются спектры других сравниваемых текстов.
3. Полученные спектры попарно сравниваются по некоторому критерию. В данной работе в качестве меры сходства семантики текстов используется косинусное сходство:

где n – размер спектра;

xi- i-ый элемент спектра в первом тексте;

yi- i-ый элемент спектра во втором тексте.

Значение R интерпретируется как некоторый угол между векторами X и Y в многомерном пространстве. Если вектора совпадают, угол между ними равен 0 и значение R будет равно 1. Если вектора не совпадают, то максимальный угол между ними будет равен 90 градусов и значение R будет равно 0.

**12.3 Пример**

***Задание***

***Требования к используемому ПО:***

- версия ОС: Windows 10;

- MS Office 2016

- файл «Анализ отзывов.xlsm».

- файлы «Отзыв1.txt», «Отзыв2.txt», «Отзыв3.txt», «Отзыв4.txt», «Отзыв5.txt», «Отзыв6.txt».

***Подготовка файлов***

Указанные выше файлы скачаны с сайта торговой сети «Лента» в формате \*.txt.

Однако даже такой простой формат может быть представлен в различной кодировке.

Поэтому:

1. Необходимо перевести их в кодировку Windows (или убедиться в том, что файлы уже находятся в этой кодировке).

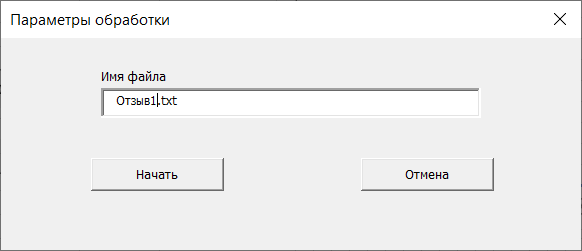
Для этого выбранный файл загружается в Word и пересохраняется в указанной кодировке.

1. Заменить в файлах букву «ё» на букву «е».

Для этого загрузить файл в «Блокнот» выделить весь текст (Ctrl/A) и с помощью пункта меню «Заменить» произвести указанную замену.

***Проведение анализа***

1. Открыть файл «Анализ отзывов.xlsm».
2. Нажать кнопку «Обработать текст»
3. В появившейся форме задать имя обрабатываемого файла.



1. Нажать кнопку «Начать».

В результате в строке 5 в ячейках F18:IZ18 сформирован семантический спектр обрабатываемого текста.

1. Скопировать содержимое данных ячеек в ячейки F19:IZ19.
2. Повторить шаги 2 – 4 для другого текста и скопировать полученный спектр в строку 20.
3. Аналогично получить спектры остальных файлов, каждый раз копируя получаемый спектры в новые строки (21, 22, и т.д.).
4. Произвести расчет меры сходства полученных спектров. Для расчета используется функция CompareText (находится в категории «Определенные пользователем»).

Для более компактного размещения результатов расчетов рекомендуется следущее размещение данных:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |
|  | text1 | text2 | text3 | text4 | text5 | text6 |  |
| text1 | 1 |  |  |  |  |  |  |
| text2 | 0,444665 | 1 |  |  |  |  |  |
| text3 | 0,602006 | 0,7841 | 1 |  |  |  |  |
| text4 | 0,484063 | 0,6223 | 0,642163 | 1 |  |  |  |
| text5 | 0,288512 | 0,385 | 0,36231 | 0,51119 | 1 |  |  |
| text6 | 0,631552 | 0,7467 | 0,737299 | 0,67563 | 0,47 | 1 |  |
|  |  |  |  |  |  |  |  |

Непосредственно для расчетов необходимо установить курсор в выбранную ячейку и в строке формул вызвать функцию CompareText.

Пример вызова функции для расчета, сходства между первым и вторым текстами:

=CompareText(F$19:IZ$19;F20:IZ20)

1. Строго обоснованных критических значений мер сходства не существует – все оставляется на усмотрение аналитика. В данном случае принято, что если мера сходства более 0,6, то тексты идентичны по смыслу.
2. Из данных таблицы следует, что тексты 1, 3, 4 и 6 одинаковы по смыслу и являются дубликатами. Оригинальными являются тексты 2 и 5.

**12.4. Задания**

1. Скачать из интернет отзывы о товарах или услугах выбранных компаний.
2. Произвести процедуру сравнения полученных отзывов.

Лишний раз напоминаем:

- скачанные тексты сохранить в формате \*.txt в кодировке Windows;

- заменить в текстах букву «ё» на букву «е».

**Примерный список компаний:**

1. Магнит (Чебоксары)
2. Магнит косметик (Чебоксары)
3. Магазин «Победа» (Чебоксары)
4. Пятерочка (Чебоксары)
5. МВидео
6. Сеть обувных магазинов «Лапоток»
7. МакДональдс
8. Ситилинк
9. METRO
10. Читай-город
11. Светофор
12. Эльдорадо
13. DNS
14. Перекресток (Чебоксары)
15. Виртуальный мир
16. Стройград
17. Автосервис (Чебоксары)