Описание к большой программе

Требования для запуска программы

Для работы программы необходимо поместить файл *kartaslovsent.csv* в одну папку с программой, так как это словарь тональностей, на котором основывается анализ текстов, иначе программа не будет работать.

Чтобы запустить тестовые файлы необходимо указывать путь папки, например, input/input1.txt .

Информация по содержанию архива:

mood.py - файл с программой

kartaslovsent.csv - словарь тональностей

input{1-22}.txt - файлы с тестовыми текстами

mumu.txt - рассказ Муму

gamlet_tragedy.txt - Гамлет

texts.txt - тексты, на которых проверялась программа, тексты размечены по количеству тональных слов и score

README.txt - информация о работе с программой и о необходимых библиотеках

Report.pdf - этот файл

Постановка задачи

Составить программу, выполняющую для русскоязычного текста:

1. Морфологический анализ слов текста
2. Определение тональной окрашенности текста, т.е. степени использования в нем оценочных слов, соотношение слов с различной тональной окраской.
3. Вычисление статистических характеристик текста:
    - общее число словоупотреблений
    - число различных слов

- число различных тональных слов
- подсчет наиболее частотных слов текста

4. вывод подсчитанных характеристик и оценок в читабельной форме.

В качестве прикладной задачи была выбрана задача определения стиля текста (научно-деловой, публицистический, художественный)

Метод решения:

При решении задачи пользуемся библиотеками nltk, re, pandas, pymorphy, collections и os.path. Библиотека nltk используется для токенизации текста, re - для очистки текста от знаков препинания, pymorphy3 - для лемматизации, pandas - для работы с csv-таблицей, в которой находится словарь тональностей.

Основные этапы решения задачи:

1. Предобработка текста: понижаем регистр текста для удобства обработки и последующего анализа
2. Получение тонального словаря из заранее загруженной базы данных
3. Токенизация текста.

   Реализовано два варианта работы программы. Первый вариант не подразумевает удаление стоп-слов из текста. Критерии для прикладной задачи выводились именно для этого варианта. Второй вариант подразумевает удаление стоп-слов.
   Пользователю дается выбор, какой вариант запускать.
4. Повторная предобработка текста: удаляем токены с пунктуацией
5. Лемматизация
6. Тональная оценка текста: получение средней тональности слов текста, получение количества слов с положительной, отрицательной и нейтральной оценкой
7. Анализ полученных результатов: вывод о тональности текста на основе нескольких критериев, решение прикладной задачи

Предобработка нужна для того, чтобы избавиться от “лишних” токенов, которые не нужно обрабатывать при дальнейшем анализе, а именно, от знаков препинания.

Морфологический анализ

Воспользуемся библиотекой pymorphy3. Используя инструмент MorphAnalyzer(), выделим первую форму слова как наиболее частотную и сохраним ее нормальную форму, проведем лемматизацию токенов. В качестве дополнительного функционала пользователю предлагается провести морфологический разбор слов анализируемого текста.

Вычисление статистических характеристик

На данном этапе мы уже работаем с “чистым” текстом, поэтому для подсчета количества словоупотреблений считаем длину списка токенов. Для подсчета количества уникальных слов достаточно выделить уникальные элементы из списка токенов и рассмотреть длину нового списка. Подсчет частотных слов и тональных слов производятся внутри функции analyze_sentiment. Статистические характеристики выводим пользователю на экран. Также предлагаем пользователю получить подробный разбор от pymorphy для каждого слова из текста. Такой разбор будет записан в файл, уведомление о формировании файла пользователь получит на экране.

Подсчет процента тональности

Тональность слова определяем с помощью данных из тонального словаря русского языка kartaslovsent. В данном словаре представлено порядка 46 тысяч элементов, каждый из которых имеет тег ‘PSTV’, ‘NGTV’, ‘NEUT’, а также value, принимающее значение в диапазоне от -1 до 1.

Сформируем pandas.DataFrame из csv-таблицы для удобства работы с данными.

Проходим по каждому элементу списка лексем. Отдельно обрабатываем случай с лексеммой “не”, так как эта частица в сочетании с положительной лексемой должна давать отрицательную тональность. Проверяем, есть ли такая лемма в нашем тональном словаре, если да, то увеличиваем счетчик sentimental_score на значение value этой леммы. В этот же момент происходит увеличение счетчика тональных слов, в зависимости от tag (может принимать значения PSTV”, “NGTV” или “NEUT”) рассматриваемой леммы.

На экран пользователя будут выведены следующие данные:

- Число положительных слов
- Число отрицательных слов

- Число нейтральных слов
- Общее число словоупотреблений
- Число различных (уникальных) слов

Решение прикладной задачи

Прикладная задача: определение стиля текста (научно-деловой, публицистический, художественный)

В данной задаче мы не будем выделять разговорный стиль, так как не имеем хорошего размеченного стилистического словаря, а статистические предположения слабо отражают особенности разговорного стиля. Разговорный стиль также содержит много различных сленговых слов, которые не включены в наш словарь и поэтому не добавят никакой тональности. При определении публицистического и художественного стилей возможны накладки из-за того, что могут происходить наслоения критериев для этих стилей. Чтобы избежать такого, вероятно, стоит использовать “слова-маркеры” для разных стилей текста.

Попробуем решить нашу задачу исходя из статистики. За неимением четких критериев проведем небольшое исследование: рассмотрим N текстов различных стилей (заранее знаем, в каком стиле написан проверяемый текст) и попробуем найти закономерности в результатах.

Начнем исследование с научно-делового стиля, который должен быть нейтральным, т.е. в нем не должно быть превосходства положительных/ отрицательных слов. Убедимся в этом рассмотрев несколько текстов написанных в научно-деловом стиле:

Анализ научно-делового стиля

> Текст из файла “input8.txt”
> Результат работы программы:
>
> Тональная окраска текста (средняя тональность слов): 0.0198
> Тональная окраска текста (на основе количества тональных слов): Нейтральный

> Тональная окраска текста (как комбинация двух мер): Скорее позитивный
>
> Текст написан в научно-деловом стиле
>
> Число положительных слов: 1
> Число отрицательных слов: 3
> Число нейтральных слов: 30
> Общее число словоупотреблений: 56

> Текст из файла “input21.txt”
> Результат работы программы:
> Тональная окраска текста (средняя тональность слов): 0.1050
> Тональная окраска текста (на основе количества тональных слов): Нейтральный
> Тональная окраска текста (как комбинация двух мер): Скорее позитивный
> Текст написан в научно-деловом стиле
> Число положительных слов: 85
> Число отрицательных слов: 33
> Число нейтральных слов: 804
> Общее число словоупотреблений: 1437

> Текст из файла “input22.txt”
> Результат работы программы:
>
> Тональная окраска текста (средняя тональность слов): 0.0796
> Тональная окраска текста (на основе количества тональных слов): Нейтральный
> Тональная окраска текста (как комбинация двух мер): Скорее позитивный

> Текст написан в научно-деловом стиле
> Число положительных слов: 50
> Число отрицательных слов: 27
> Число нейтральных слов: 473
> Общее число словоупотреблений: 945

Анализ художественного стиля

Приведем несколько примеров текстов, которые использовались для выявления критериев для классификации художственных текстов.

> Текст из файла “mumu.txt”
> Результат работы программы:
> Тональная окраска текста (средняя тональность слов): 0.0631
> Тональная окраска текста (на основе количества тональных слов): Нейтральный
> Тональная окраска текста (как комбинация двух мер): Скорее позитивный
>
> Текст написан в художественном стиле
> Число положительных слов: 615
> Число отрицательных слов: 545
> Число нейтральных слов: 2921
> Общее число словоупотреблений: 8433

> Текст из файла “gamlet_tragedy.txt”
> Результат работы программы:
>
> Тональная окраска текста (средняя тональность слов): 0.0798
> Тональная окраска текста (на основе количества тональных слов): Нейтральный

> Тональная окраска текста (как комбинация двух мер): Скорее позитивный
>
> Текст написан в художественном стиле
> Число положительных слов: 2877
> Число отрицательных слов: 2251
> Число нейтральных слов: 7714
> Общее число словоупотреблений: 26254

> Текст из файла “input13.txt”
> Результат работы программы:
>
> Тональная окраска текста (средняя тональность слов): 0.0707
> Тональная окраска текста (на основе количества тональных слов): Нейтральный
> Тональная окраска текста (как комбинация двух мер): Скорее позитивный
>
> Текст написан в художественном стиле
> Число положительных слов: 4
> Число отрицательных слов: 6
> Число нейтральных слов: 24
> Общее число словоупотреблений: 68

> Текст из файла “input20.txt”
> Результат работы программы:
>
> Тональная окраска текста (средняя тональность слов): 0.0607
> Тональная окраска текста (на основе количества тональных слов): Нейтральный
> Тональная окраска текста (как комбинация двух мер): Скорее позитивный

> Текст написан в художественном стиле
> Число положительных слов: 13
> Число отрицательных слов: 15
> Число нейтральных слов: 34
> Общее число словоупотреблений: 114

Анализ публицистического стиля

> Текст из файла “input18.txt”.
> Результат работы программы:
>
> Тональная окраска текста (средняя тональность слов): 0.1638
> Тональная окраска текста (на основе количества тональных слов): Нейтральный
> Тональная окраска текста (как комбинация двух мер): Скорее позитивный
>
> Текст написан в научно-деловом стиле
>
> Число положительных слов: 174
> Число отрицательных слов: 19
> Число нейтральных слов: 643
>
> Общее число словоупотреблений: 1523

> Текст из файла “input17.txt”
> Результат работы программы:
>
> Тональная окраска текста (средняя тональность слов): 0.2170
> Тональная окраска текста (на основе количества тональных слов): Нейтральный
> Тональная окраска текста (как комбинация двух мер): Скорее позитивный

> Текст написан в публицистическом стиле
> Число положительных слов: 57
> Число отрицательных слов: 10
> Число нейтральных слов: 133
> Общее число словоупотреблений: 314

> Текст из файла “input9.txt”
> Результат работы программы:
> Тональная окраска текста (средняя тональность слов): 0.1546
> Тональная окраска текста (на основе количества тональных слов): Нейтральный
> Тональная окраска текста (как комбинация двух мер): Скорее позитивный
> Текст написан в публицистическом стиле
> Число положительных слов: 3
> Число отрицательных слов: 1
> Число нейтральных слов: 5
> Общее число словоупотреблений: 13

Результаты экспериментов

	Тональная окраска текста				Количество слов			
Текст	средняя тональность	кол-во тональных слов	комбинация двух мер	Стиль	положительных	отрицательных	нейтральных	всего
input8.txt	0.0198	Нейтральный	Скорее позитивный	Научно-деловой	1	3	30	56
input21.txt	0.105	Нейтральный	Скорее позитивный	Научно-деловой	85	33	804	1437
input22.txt	0.0796	Нейтральный	Скорее позитивный	Научно-деловой	50	27	473	945

Результаты экспериментов для научно-делового стиля

	Тональная окраска текста				Количество слов			
Текст	средняя тональность	кол-во тональных слов	комбинация двух мер	Стиль	положительных	отрицательных	нейтральных	всего
mumu.txt	0.0631	Нейтральный	Скорее позитивный	Художественный	615	545	2921	8433
gamlet_tragedy.txt	0.0798	Нейтральный	Скорее позитивный	Художественный	2877	2251	7714	26254
input13.txt	0.0707	Нейтральный	Скорее позитивный	Художественный	4	6	24	68
input20.txt	0.0607	Нейтральный	Скорее позитивный	Художественный	13	15	34	114

Результаты экспериментов для художественного стиля

	Тональная окраска текста				Количество слов			
Текст	средняя тональность	кол-во тональных слов	комбинация двух мер	Стиль	положительных	отрицательных	нейтральных	всего
input18.txt	0.1638	Нейтральный	Скорее позитивный	Научно-деловой	174	19	643	1523
input17.txt	0.217	Нейтральный	Скорее позитивный	Публицистический	57	10	133	314
input9.txt	0.1546	Нейтральный	Скорее позитивный	Публицистический	3	1	5	13

Результаты экспериментов для публицистического стиля

В результате экспериментов были установлены следующие критерии:

1. В научно-деловом стиле преобладают нейтральные слова, поэтому такой стиль будем определять исходя из соотношения нейтральных слов ко всем тональным словам в тексте. В результате экспериментов выведен “оптимальный” критерий для научно-делового стиля, в таком тексте нейтральные слова должны составлять не менее 50% всех слов в тексте.
2. Видно, что в художественном тексте встречается много различных тональных слов, поэтому зависимость от тональных слов установить не получается, значит в качестве критерия будет выступать sentimental_score (средняя тональность слов). Эмперически был выведен диапазон для средней тональности `abs(sentiment_score) > 0.3 or abs(sentiment_score) < 0.11`
3. В публицистическом стиле можно выделить превосходство положительных и отрицательных слов над нейтральными, поэтому будем учитывать, что нейтральных слов должно быть меньше половины всех слов в тексте, а также будем учитывать некоторые тональные оценки в диапазонах, которые не покрываются правилом для художственных текстов.

   `sentiment_counter['Нейтральный'] / total_word_count < 0.5 or (abs(sentiment_score) < 0.29)`

Выводы

Задача выполнена для русскоязычных текстов. Опечатки, некорректные слова, ирония, сарказм и другие языковые аномалии не будут иметь влияния на анализ текста. В данной реализации, если слово не встретилось в словаре, то оно не добавляет никакой тональной окраски в статистику. Такой подход приводит к очевидной проблеме: мы не можем никак оценить тексты, состоящие из слов, которые не представлены в нашем словаре. Эта проблема может быть решена комбинированием нескольких словарей.

Таким образом, видно, что чем “богаче” используемый словарь, тем лучше результат получается.

При решении прикладной задачи были выявлены следующие трудности:

1. Не все функциональные стили можно классифицировать, используя имеющиеся данные.
2. Могут происходить “наложения” при классификации публцистических и художественных текстов.

3. Отсутствие формализованных критериев для каждого из функциональных стилей.

Возможные решения:

1. Использование стилистических словарей.
2. Использование “маркеров” или наборов “маркеров”, характерных для того или иного стиля.
3. Использование более сложных инструментов, позволяющих проводить более серьезный анализ контекста.

Таким образом, нельзя говорить, что классификация текстов, выполняемая программой, корректна на 100%, так как она основана исключительно на статистике.