МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра алгоритмических языков

ОТЧЁТ
По индивидуальному проекту
Тема: «Тональность»

Выполнила: студентка 324 группы
Мамиева Мария Алановна

Москва, 2025

СОДЕРЖАНИЕ :

АНАЛИЗАТОР ТОНАЛЬНОСТИ ТЕКСТА

1. РЕШАЕМАЯ ЗАДАЧА

Данная программа предназначена для анализа эмоциональной тональности текстов на русском языке.
Она выполняет следующие задачи:

- Токенизация (разбивка на отдельные слова)
- Лемматизация слов с определением частей речи
- Фильтрация шумовых слов (служебные части речи)
- Объединение слов в значимые фразы с модификаторами
- Учет модификаторов тональности (усилители, ослабители, отрицания)
- Анализ эмоциональной окраски с использованием тонального словаря
- Оценка тональности всего текста
- Генерация детальных отчетов и визуализации

2. АЛГОРИТМ РАБОТЫ МОДИФИКАТОРОВ ТОНАЛЬНОСТИ

2.1. Принцип работы

Сначала происходит токенизация текста, т.е разделение текста на отдельные токены(слова), удаление знаков препинания, приведение текста к нижнему регистру. Затем следует лингвистический анализ, в ходе которого слова приводятся к начальной форме (лемматизация) и определяются их части речи. На следующем этапе программа фильтрует шумовые слова, удаляя служебные части речи такие, как предлоги, союзы и местоимения, но сохраняя важные для анализа тональности модификаторы "не" и "ни". Особенностью программы является объединение слов в смысловые фразы, где модификаторы соединяются с основными словами, например, "не" + "хороший" образует фразу "не хороший". Ключевой этап - анализ тональности с использованием словаря эмоциональной окраски слов "kartaslovsent", из словаря берется значение value (значение тональности). Программа не просто определяет тональность отдельных слов, но и учитывает влияние модификаторов:
- Усилители ("очень", "крайне" и др.): умножают значение следующего слова на 2
- Ослабители ("слегка", "довольно" и др.): умножают значение следующего слова на 0.5
- Отрицания ("не", "ни"): инвертируют знак следующего слова (умножают на -1)

Все слова и фразы классифицируются на положительные, отрицательные и нейтральные основываясь на пороговых значениях параметра value. Программа генерирует четыре типа выходных данных: детальный анализ всех этапов обработки текста, таблицу статистики с разделением на значимые и шумовые слова, статистику с количественными показателями и визуализацию в виде диаграмм. На выход помимо файлов подается итоговое значение тональности текстов

2.2. Обоснование коэффициентов

1) Для слов меняющих окраску следующего слова.

Коэффициенты подобраны эмпирически на основе лингвистических познаниях автора проекта:

- Усилители (×2): значительно усиливают эмоциональную окраску
- Ослабители (×0.5): ослабляют, но не полностью нейтрализуют эмоцию
- Отрицания (×(-1)): полностью инвертируют эмоциональную оценку

2) Для тональности леммы.

Коэффициенты подобраны деление отрезка значений параметра value на три приблизительно равные группы:

- Положительная: value > 0.33
- Нейтральная: -0.33 <= value <= 0.33
- Отрицательная: value < 0.33

3) Для общей тональности текста:

Коэффициенты подобраны на основе анализа тестов:

- Крайне Положительный: > 0.5
- Положительный: > 0.1
- Нейтральный: от -0.1 до 0.1
- Отрицательный: < -0.1
- Крайне отрицательный: < -0.50

3. ОСНОВНЫЕ КОМПОНЕНТЫ ПРОГРАММЫ

3.1 Используемые библиотеки

pymorphy2 - морфологический анализатор для русского языка
csv - работа с CSV-файлами (тональный словарь)
os - работа с файловой системой
collections - расширенные структуры данных
re - регулярные выражения для токенизации
matplotlib - визуализация результатов
numpy - математические операции
sys - управление выполнением программы

3.2 Основные функции программы

segment_and_tokenize(text)
Назначение: Сегментация текста на токены (слова)
Вход: строка текста
Выход: список токенов в нижнем регистре

lemmatize_tokens(tokens)
Назначение: Лемматизация токенов и определение частей речи
Вход: список токенов
Выход: список кортежей (лемма, часть_речи)

get_part_of_speech(tag)
Назначение: Преобразование тегов pymorphy2 в сокращенные русские обозначения
Вход: тег части речи от pymorphy2
Выход: сокращенное русское обозначение части речи

remove_noise_words(lemmas_with_pos)
Назначение: Фильтрация шумовых слов (служебных частей речи)
Вход: список лемм с частями речи
Выход: кортеж (отфильтрованные леммы, шумовые слова)

combine_words(lemmas_with_pos)
Назначение: Объединение модификаторов с основными словами в фразы
Вход: список лемм с частями речи

Выход: список слов и фраз

load_tone_dict(dict_path='kartaslovsent.csv')
Назначение: Загрузка словаря тональности
Вход: путь к CSV файлу словаря
Выход: словарь {слово: значение_тональности}

calculate_phrase_tones(phrases, tone_dict)
Назначение: Расчет тональности для фраз с учетом модификаторов
Вход: список фраз, словарь тональности
Выход: список кортежей (упорядоченная, но неизменяемая последовательность элементов: фраза, тональность)

get_tone_category(value)
Назначение: Классификация тональности по числовому значению для отдельных слов/фраз
Вход: числовое значение тональности
Выход: текстовый тег ("положительное", "отрицательное", "нейтральное")

tone_category_for_text(value)
Назначение: Классификация общей тональности текста по удельным значениям
Вход: числовое значение удельной тональности
Выход: текстовый тег ("крайне положительный", "положительный", "нейтральный", "отрицательный", "крайне отрицательный")

create_statistics_table(phrases_with_tones, noise_words, tone_dict)
Назначение: Создание статистической таблицы по анализу
Вход: фразы с тональностью, шумовые слова, словарь тональности
Выход: список строк таблицы

calculate_comprehensive_statistics(tokens, lemmas_with_pos, filtered_lemmas, phrases, tone_results, tone_dict, noise_words)
Назначение: Расчет комплексной статистики анализа
Вход: различные данные анализа текста
Выход: словарь со статистическими показателями

create_plots(stats, filename)
Назначение: Создание визуализаций анализа
Вход: статистика, имя файла для сохранения
Выход: PNG файл с графиками

4. ТРЕБОВАНИЯ К СИСТЕМЕ

НЕОБХОДИМЫЕ ФАЙЛЫ:

kartaslovsent.csv - тональный словарь в формате CSV с разделителем ';'
Должен содержать колонки: term, value. Должен находиться в одной директории с исполняемым файлом.
Текстовые файлы для анализа (кодировка UTF-8)

МИНИМАЛЬНЫЕ ТРЕБОВАНИЯ:

Python 3.6 или выше, рекомендуется Python 3.10.18
Библиотека pymorphy2 версии 2.0 или выше
Библиотеки: matplotlib, numpy

УСТАНОВКА:

pip install pymorphy2 matplotlib numpy

kartaslovsent.csv по ссылке:
https://github.com/dkulagin/kartaslov/tree/master/dataset/kartaslovsent

5. ФОРМАТ РАБОТЫ ПРОГРАММЫ:

Пользователь вводит имя текстового файла для анализа

Программа создает четыре выходных файла:

[имя]_analysis.txt - детальный процесс обработки текста
[имя]_table.txt - таблица статистики по фразам и словам
[имя]_statistics.txt - комплексная статистика анализа
[имя]_plots.png - визуализация результатов в виде графиков

Программа выводит информацию о созданных файлах и общую тональность текста.

6. ВЫХОДНЫЕ ДАННЫЕ И ИХ ОПИСАНИЕ

1) Файл [имя]_analysis.txt последовательно отображает каждый этап работы алгоритма. Он начинается токенизации, затем следует раздел лемматизации с определением частей речи, где каждое слово приводится к начальной форме и классифицируется. Далее демонстрируется процесс фильтрации шумовых слов - удаляются служебные части речи. Далее раздел с формированием значимых фраз и расчетом их тональности, где показываются итоговые словосочетания с учетом влияния модификаторов и их финальная эмоциональная оценка.

2) Файл [имя]_table.txt организует всю полученную информацию в структурированном виде. В таблице представлены шесть ключевых колонок: сами слова и фразы с указанием частей речи для шумовых элементов, статус значимости, частота употребления в тексте, исходное значение тональности из словаря, итоговое значение после применения модификаторов и финальная тональность.

3) Файл [имя]_statistics.txt демонстрирует статистику по тексту: основные, дополнительные, суммарные показатели, результат - итоговое значение тональности текста.

4) Файл [имя]_plots.png содержит графическое представление данных в виде четырех диаграмм. Две круговые диаграммы в верхней части отображают пропорциональное распределение слов по тональностным категориям - первая показывает общее распределение всех слов текста, вторая - исключительно для значимых слов. Две диаграммы в нижней части демонстрируют нормализованные значения тональности по различным метрикам: левая диаграмма отображает тональность по основным показателям (словоупотребления, словоформы, леммы), правая - по значимым элементам (слова, леммы, фразы).

5) Итоговое значение тональности текста.

7. ОГРАНИЧЕНИЯ ПРОГРАММЫ

Учет контекста, программа не учитывает:

- Иронию и сарказм
- Многозначность слов в разных контекстах
- Сложные синтаксические конструкции
- Культурные и ситуационные особенности

Лингвистические ограничения:

- Обрабатывает только последовательные модификаторы, не учитывая порядка слов
- Не анализирует сложные словосочетания и идиомы

Технические ограничения:

Зависимость от размера тонального словаря
Точность зависит от правильности лемматизации
Не обрабатывает опечатки и нестандартные написания
Ограниченная обработка сложных модификаторов

8. ОПИСАНИЕ ТЕСТОВ

- test1.txt - текст только с положительными и нейтральными словами
- test2.txt - текст только из отрицательных слов и нейтральных слов
- test3.txt - текст с равным количеством отрицательных и положительных слов (нейтральный)
- test4.txt - текст с сарказмом
- test5.txt - текст где нейтральный посыл, но из-за отсутствия контекста может восприниматься отрицательно
- test6.txt - текст с двойными модификаторами
- test7.txt - отрывок из Гарри Поттера
- test8.txt - положительный отзыв о ВМК

9.ПРИМЕР РАБОТЫ ПРОГРАММЫ

Пример работы программы на тексте:
«Ужасный кошмарный провал вызывает отвратительное разочарование и горькое отчаяние. Мучительная болезненная тоска усугубляет гнетущую подавленность и досадную безнадежность. Отчаянная неприятная ситуация порождает злобу и непримиримую ненависть.»

Полученный результат: Текст: крайне отрицательный

Содержимое файла test_statistics.txt:

СТАТИСТИКА АНАЛИЗА ТЕКСТА
==========

ОСНОВНЫЕ ПОКАЗАТЕЛИ:
Словоупотреблений (токенов): 26
Словоформ (различных токенов): 24
Лемм (уникальных после лемматизации): 24

Значимых слов (всех после удаления шумовых): 22
Значимых лемм (уникальных после удаления шумовых): 22
Значимых фраз (после создания фраз): 22

ДОПОЛНИТЕЛЬНЫЕ ПОКАЗАТЕЛИ:
Слов с value > 0: 0
Слов с value < 0: 20
Слов с value = 0: 2
Шумовых слов: 4
Положительных слов (value > 0.33): 0
Отрицательных слов (value < -0.33): 19
Нейтральных слов (-0.33 ≤ value ≤ 0.33): 3

СУММАРНЫЕ ПОКАЗАТЕЛИ ТОНАЛЬНОСТИ:
Положительность (тональность слов с value > 0): 0.000
Отрицательность (тональность слов с value < 0): -17.940
Тональность (положительность + отрицательность): -17.940
Тональный разброс (положительность - отрицательность): 17.940

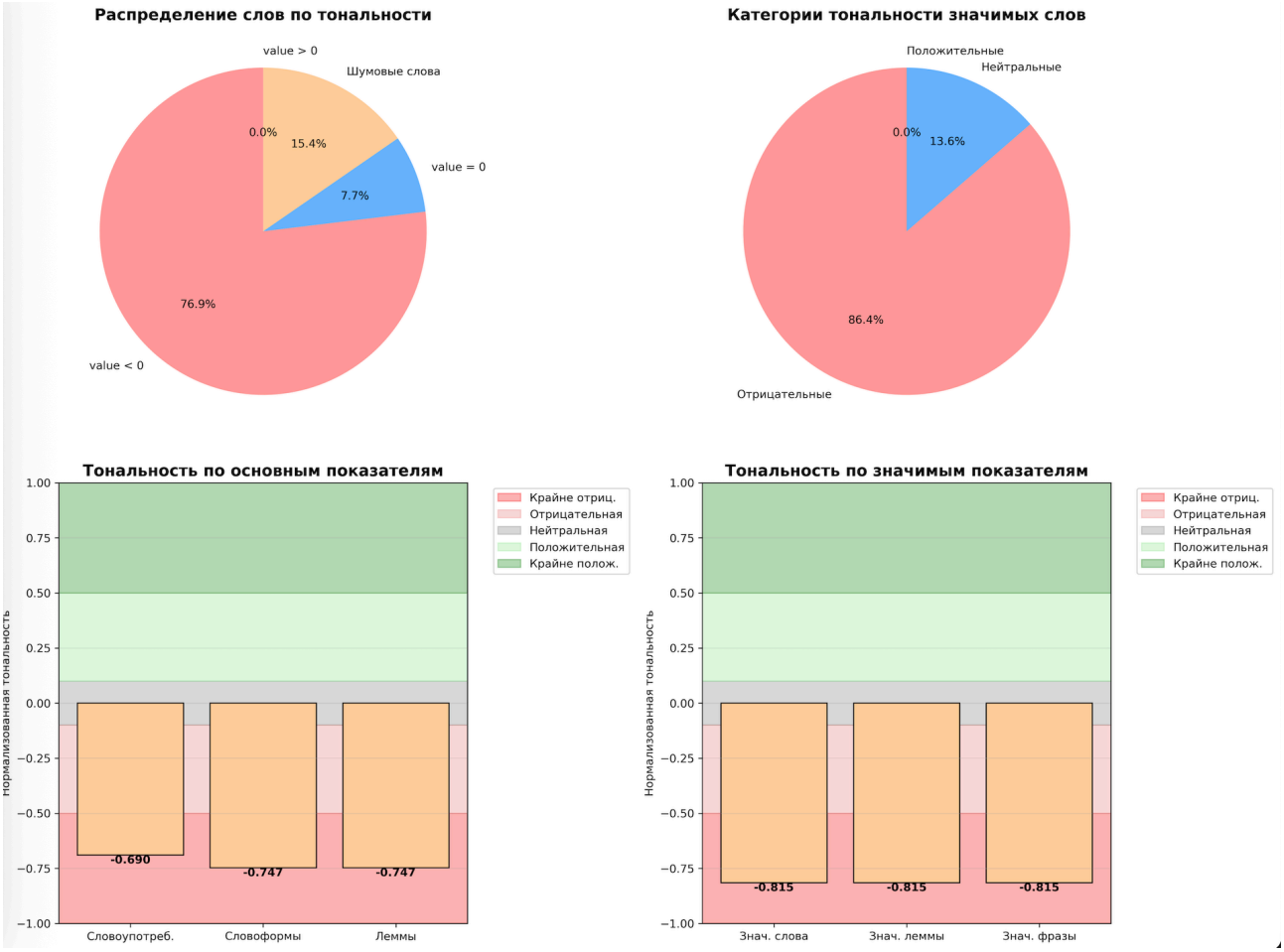
==========

РЕЗУЛЬТАТ:
Текст: крайне отрицательный

Содержимое файла test_table.txt:

Фраза/Слово	Статус	Количество	Базовая тональность	Суммарная тональность	Категория
вызывать	знач.	1	0.000	0.000	нейтральное
ситуация	знач.	1	0.000	0.000	нейтральное
и (СОЮЗ)	шум.	3	0.000	0.000	нейтральное
гнетущий (НЕИЗВ)	шум.	1	0.000	0.000	нейтральное
порождать	знач.	1	-0.120	-0.120	нейтральное
отчаянный	знач.	1	-0.550	-0.550	отрицательное
непримиримый	знач.	1	-0.580	-0.580	отрицательное
горький	знач.	1	-0.770	-0.770	отрицательное
досадный	знач.	1	-0.920	-0.920	отрицательное
ужасный	знач.	1	-1.000	-1.000	отрицательное
кошмарный	знач.	1	-1.000	-1.000	отрицательное
провал	знач.	1	-1.000	-1.000	отрицательное
отвратительный	знач.	1	-1.000	-1.000	отрицательное
разочарование	знач.	1	-1.000	-1.000	отрицательное
отчаяние	знач.	1	-1.000	-1.000	отрицательное
мучительный	знач.	1	-1.000	-1.000	отрицательное
болезненный	знач.	1	-1.000	-1.000	отрицательное
тоска	знач.	1	-1.000	-1.000	отрицательное
усугублять	знач.	1	-1.000	-1.000	отрицательное
подавленность	знач.	1	-1.000	-1.000	отрицательное
безнадёжность	знач.	1	-1.000	-1.000	отрицательное
неприятный	знач.	1	-1.000	-1.000	отрицательное
злоба	знач.	1	-1.000	-1.000	отрицательное
ненависть	знач.	1	-1.000	-1.000	отрицательное

Содержимое файла test_plots.png



10. КРАТКО О РЕЗУЛЬТАТАХ

Программа успешно решает задачу анализа тональности русскоязычных текстов, предоставляя:
- Полную цепочку обработки от токенизации до анализа тональности
- Учет лингвистических особенностей (модификаторы, части речи)
- Детальную статистику
- Визуализацию результатов

11. РЕШЕНИЕ ПРИКЛАДНОЙ ЗАДАЧИ

11.1 Описание прикладной задачи

Социологическое исследование.Анализ текстов по определенным темам:
- Мир
- Политика
- Религия
- Общество
- Наука
- Культура
- Общество

и поиск тональных особенностей каждой темы.

11.2 Алгоритм решения прикладной задачи

1.Подготовка данных

Был скачан датасет с русскими новостями за 2015-2018 года (https://www.kaggle.com/datasets/serart/nti-news-text-seriescsv?resource=download). Датасет - csv таблица с 4 столбцами (номер, текст статьи, достоверность, тема - #, text, is_train, class).

2.Выборка и категоризация

Были выбраны 7 тем(категорий, class): world, politics, religion, society, science, culture, economy. Был создан код (файл parser.py), который считывает датасет, создаёт папку text_categories_simple и в папке 7 файлов .txt с соответствующими названиями, в каждый файл сохраняет 100 новостей соответствующей категории.

3.Анализ тональности

Далее к каждому файлу была применена основная программа main.py и получены результаты тональности для сборников новостей. Для удобства был создан файл run_all.py который запускает последовательно все 8 тестов и все 7 категорий.

4.Сравнительный анализ

Был проведен анализ полученных результатов.

11.3 Пример полученных результатов

Тема - религия

Первые 10 строк таблицы:

Фраза/Слово	Статус	Количество	Базовая тональность	Суммарная тональность	Категория
спасти	знач.	6	1.000	6.000	положительное
безопасность	знач.	15	1.000	15.000	положительное
уважаемый	знач.	1	1.000	1.000	положительное
достижение	знач.	1	1.000	1.000	положительное
взаимопонимание	знач.	2	1.000	2.000	положительное
плодотворный	знач.	1	1.000	1.000	положительное
культурный	знач.	7	1.000	7.000	положительное
интеллектуальный	знач.	2	1.000	2.000	положительное
здравый	знач.	1	1.000	1.000	положительное
помощь	знач.	17	1.000	17.000	положительное

Графическое представление:

Стастистика:

СТАТИСТИКА АНАЛИЗА ТЕКСТА
==========

ОСНОВНЫЕ ПОКАЗАТЕЛИ:

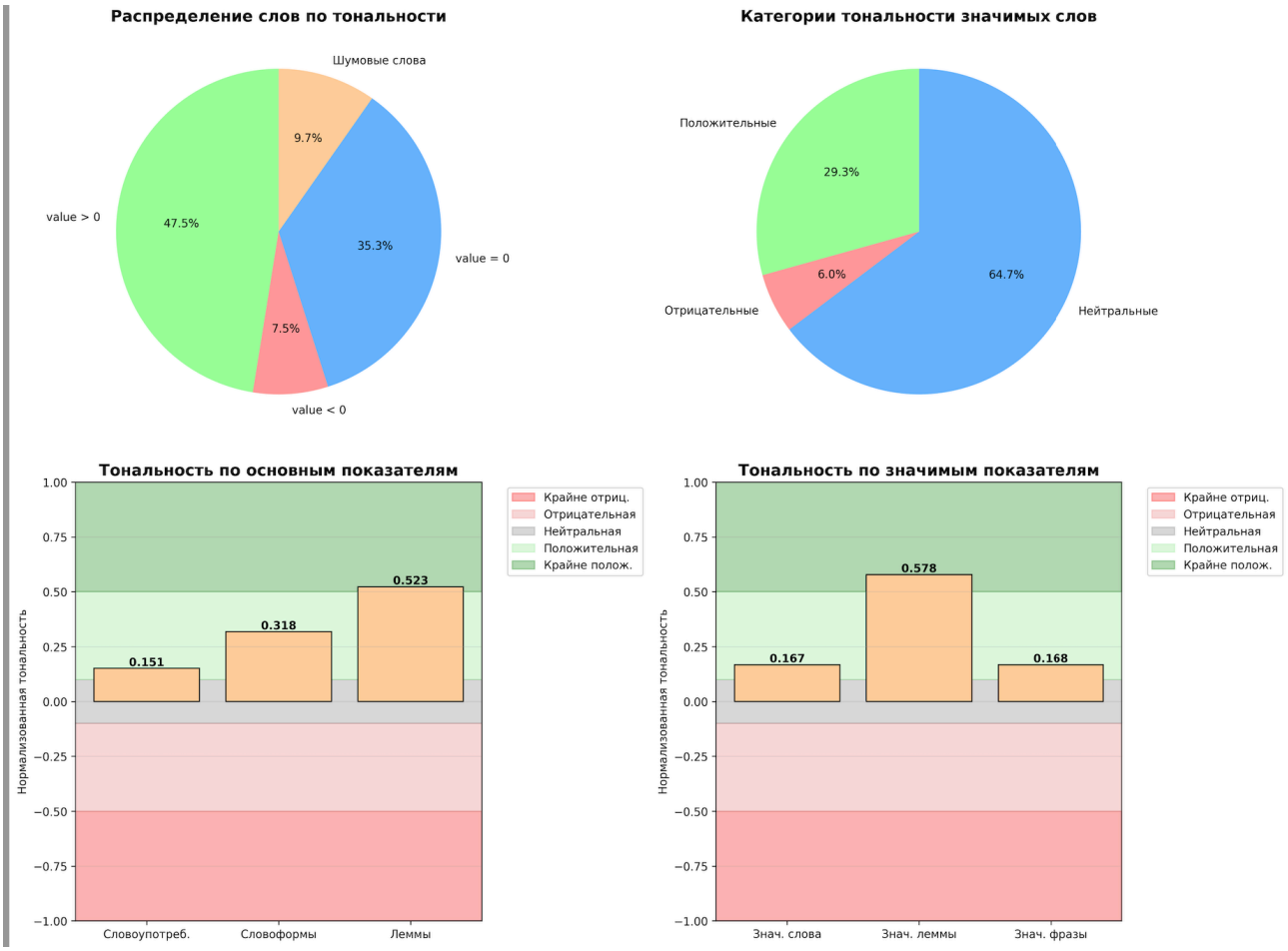
Словоупотреблений (токенов): 22734
Словоформ (различных токенов): 10809
Лемм (уникальных после лемматизации): 6570
Значимых слов (всех после удаления шумовых): 20520
Значимых лемм (уникальных после удаления шумовых): 5944
Значимых фраз (после создания фраз): 20501

ДОПОЛНИТЕЛЬНЫЕ ПОКАЗАТЕЛИ:
Слов с value > 0: 10780
Слов с value < 0: 1695
Слов с value = 0: 8026
Шумовых слов: 2214
Положительных слов (value > 0.33): 6011
Отрицательных слов (value < -0.33): 1222
Нейтральных слов (-0.33 ≤ value ≤ 0.33): 13268

СУММАРНЫЕ ПОКАЗАТЕЛИ ТОНАЛЬНОСТИ:
Положительность (тональность слов с value > 0): 4497.250
Отрицательность (тональность слов с value < 0): -1060.980
Тональность (положительность + отрицательность): 3436.270
Тональный разброс (положительность - отрицательность): 5558.230

==========

РЕЗУЛЬТАТ:
Текст: положительный



11.4 Результаты анализа для всех тем

Культура:
Общая тональность темы положительная, имеет самый высокий показатель итоговой тональности (нормализованная по количеству значимых фраз тональность) из всех тем. Также тональность, нормализованная по количеству значимых лемм, очень высокая (0.612), что может говорить о том, что используется много повторяющихся положительных фраз. Положительных слов в шесть раз больше, чем отрицательных. Шумовых слов (в процентах) меньше всего из всех тем.

Экономика:
Общая тональность темы положительная, имеет четвертый по высоте показатель итоговой тональности (нормализованная по количеству значимых фраз тональность) из всех тем. Тональность, нормализованная по количеству значимых лемм, очень высокая (0.618), что может говорить о том, что используется много повторяющихся положительных фраз, что характерно для официально-делового стиля речи. Положительных слов в четыре раз больше, чем отрицательных, а процент нейтральных слов очень высок (70%).

Политика:
Общая тональность темы нейтральная. Тональность, нормализованная по количеству значимых лемм, невысокая (0.3), что говорит о большом разнообразии окрашенных слов и их слабой окрашенность. Положительных слов в два с половиной раза больше, чем отрицательных, а процент нейтральных слов очень высок (70%). Также в этой теме довольно много шумовых слов.

Религия:
Общая тональность темы положительная, имеет второй по высоте показатель итоговой тональности (нормализованная по количеству значимых фраз тональность) из всех тем. Положительных слов в пять раз больше, чем отрицательных. Большое количество слов с тональностью больше нуля, т.е в текстах стараются использовать нейтральные слова имеющие скорее положительное влияние на читателя.Одни из самых часто встречающихся положительно окрашенные слов в тексте : семья, свобода, жизнь, папа, мама, церковь. Из отрицательных: наказание, запрет, проблема.

Наука:
Общая тональность темы положительная, третий по высоте показатель итоговой тональности (нормализованная по количеству значимых фраз тональность) из всех тем. Также тональность, нормализованная по количеству значимых лемм, очень высокая в сравнении с тональностью относительно значимых фраз, что может говорить о том, что используется много повторяющихся положительных фраз. Положительных слов в пять раз больше, чем отрицательных. Это тема имеет максимальный из всех процент шумовых слов.

Общество:
Общая тональность темы положительная. Тональность, нормализованная по количеству значимых лемм не сильно отличается от тональности, нормализованной по количеству значимых фраз. Это может говорить о том, что используются достаточно разнообразные окрашенные слова. Положительных слов в три раз больше, чем отрицательных, а процент нейтральных слов высок (70%).

Мир:
Общая тональность - нейтральная. Показатель тональности этой темы находиться ближе других рассмотренных к нулю. Положительных слов всего в два раза больше, чем отрицательных. Это может объяснить также низкое значение нормализованной по значимым леммам тональности (0.238). Также эта тема имеет высокое количество нейтральных слов. В теме практически отсутствуют отрицательные слова, имеющие высокое абсолютное значение тональности, т.е в текстах на эту темы отсутствуют "крайне отрицательные" слова.

РЕЙТИНГ ТЕМ ПО ТОНАЛЬНОСТИ:

1.Культура (0.192)
2.Религия (0.168)
3.Наука (0.162)
4.Экономика (0.145)
5.Общество (0.131)
6.Политика (0.099)
7.Мир (0.07)

ВЫВОДЫ:

Можно заметить, что во всех темах новостей стараются придерживаться общей положительности высказываний. Поэтому во всех темах общая тональность неотрицательна. Самую высокая тональность имеют духовные темы: религия и культура. Наименьшую темы, связанные с политикой. Социально-политическая тематика проявляет наибольшую нейтральность, что, вероятно, обусловлено необходимостью соблюдения баланса и объективности при освещении сложных общественных процессов.
Тема "Мир" демонстрирует практически нулевую тональность, что соответствует принципам нейтральности в освещении международных отношений.Научно-экономический блок демонстрирует умеренно-позитивную окраску, отражая оптимистические ожидания от технологического и экономического развития.

С файлами полученными в результате решения прикладной задачи можно ознакомиться(они находятся в папке text_categories_simple).

Литература :

- датасет с русскими новостями за 2015-2018 года (https://www.kaggle.com/datasets/serart/nti-news-text-seriescsv?resource=download)

- тональный словарь kartaslovsent.csv
https://github.com/dkulagin/kartaslov/tree/master/dataset/kartaslovsent

- описание построения графиков на Python https://skillbox.ru/media/code/biblioteka-matplotlib-dlya-postroeniya-grafikov/

- описание базовых функций библиотеки NumPY https://skillbox.ru/media/code/biblioteka-numpy-vsye-chto-nuzhno-znat-novichku/