Автоматическая разметка решений через Gemini API

Последовательность выполнения

Этап 1: Поиск необработанных текстов

Система сканирует директорию TXT_DIR, в которой хранятся все тексты, собранные по результатам скрапинга, и для каждого файла с расширением .txt выполняет проверки:

- Извлекает doc_id из имени файла через `filename.replace(`
- Формирует путь к потенциальному JSON-файлу как `os.path.join(JSON_DIR, f"{doc_id}.json")`
- Проверяет существование JSON через `os.path.exists(json_path)`
- Проверяет размер TXT-файла через `os.path.getsize(txt_path) > 1024` (более 1 КБ)
- Читает содержимое и проверяет отсутствие строки "Text not found"

Только файлы, прошедшие все проверки, добавляются в список `unprocessed` для обработки. Это помогает обработать файлы, в отношении которых со стороны Google были ошибки или проблемы со связью.

Этап 2: Обработка через Gemini с прогрессивными таймаутами

Функция `process_with_gemini()` использует трехуровневую систему таймаутов:

1. Установка обработчика сигнала через `signal.signal(signal.SIGALRM, timeout_handler)`
2. Активация таймера через `signal.alarm(timeout)` с последовательными значениями [180, 300, 480] секунд
3. Выполнение запроса к API через `model.generate_content(prompt, safety_settings=safety_settings)`
4. При успехе — отключение таймера через `signal.alarm(0)`
5. При TimeoutError — экспоненциальная задержка через `time.sleep(2 ** attempt)` и переход к следующему таймауту

Этап 3: Извлечение JSON из ответа модели и сохранение с метаданными сессии

Система использует регулярное выражение для извлечения JSON-объекта из текстового ответа, потому что Gemini может вернуть JSON с дополнительным текстом до или после.

Функция `process_with_gemini_retry_with_session()` добавляет к результату метаданные:

```
result['session_metadata'] = {
```

```
    'processed_at': datetime.now().isoformat(),
    'session_id': CURRENT_SESSION['session_id'],
    'period': CURRENT_SESSION['period']
}
```

Это важно для дальнейшей сборки csv-таблицы для ручной обработки результатов разметки.

Этап 4: Логирование завершенной сессии

При наличии глобальной переменной `CURRENT_SESSION` система:

- Устанавливает `end_time` и `status = 'completed'`
- Подсчитывает `total_processed`, `success_count`, `error_count`
- Загружает существующий лог через `json.load()` или создает новый
- Добавляет текущую сессию в список `session_log['sessions']`
- Сохраняет обновленный лог в `OUTPUT_DIR/session_log.json`

Ключевые технические особенности

Кеширование на уровне файловой системы

Проверка `if os.path.exists(json_path): return None` выполняется в самом начале функции, до любых API-запросов. Это обеспечивает:

- Нулевую стоимость повторной обработки уже размеченных документов
- Возможность прерывания и продолжения обработки в любой момент
- Защиту от дублирования результатов при параллельном запуске

Отключение фильтров безопасности

Настройки `safety_settings` устанавливают `HarmBlockThreshold.BLOCK_NONE` для всех категорий. Это необходимо, потому что решения ФАС могут содержать описания товаров или рекламы, которые ложно срабатывают на фильтры.

Трассируемость через session_id

Формат session_id `{timestamp}_{start_date}_{end_date}` позволяет:

- Уникально идентифицировать каждый запуск пайплайна
- Определить период обработки без обращения к базе данных
- Фильтровать результаты по времени создания или по периоду документов

Отрывки из промпта на разметку

Not valuable (for RAG)

The case is exclusively about the method of distribution, technical/formal aspects, or specific administrative requirements. Examples of such non-valuable topics include:

- Lack of prior consent to receive advertising (ст. 18).

- Improper use of physical advertising structures, vehicles, road signs (ч. 10 ст. 5, ч. 10.3 ст. 5, ст. 19, 20).

- Missing audio/visual warnings about advertising on TV/radio (ст. 14, 15).

- Missing "реклама" marker in print media (ст. 16).

- Requirement for the advertiser to have a license or special permission (advocates, children camps, touristic services, micro-financial organisations) or accreditation or to be included in certain registries.

- Violations of the following articles of the Federal Law "On Advertising": "ч. 14 ст. 28", "ч. 16 ст. 18.1", "п. 7 ст. 7", "ч. 1 ст. 18", "п. 5 ч. 2 ст. 21", "ч. 2.1 ст. 21", "ч. 2 ст. 27", "п. 8 ч. 2 ст. 21", "ч. 13 ст. 28", "ч. 2 ст. 20", "ч. 2 ст. 18", "ч. 9 ст. 19", "ч. 6 ст. 7", "п. 6 ч. 2 ст. 21", "ч. 1 ст. 15", "ч. 5 ст. 19", "ч. 1 ст. 14", "п. 3 ч. 2 ст. 21", "ч. 10 ст. 19".

- Other technical errors in real-world or Internet ad placement.

- Cases where the advertices terms of purchase of products and services (prices, promotions, discounts) turned out to be inconcistent with actual terms.

- Procedural descisions to commence the proceedings.

Scoring Criteria

(for valuable cases, scale 1-10):

- High Value (Score 8-10): Precedent-Setting and Complex Content Cases (e.g., "gray areas" like superlatives, incorrect comparisons, unfair competition, new trends like influencer ads; complex industry-specific violations, cases involving deep linguistic and imaging analysis of advertiser's intentions).

- Medium Value (Score 4-7): Important but More Common Content Violations (e.g., misleading guarantees; use of medical expert images; violations of requirements to languale; concerns medicinal products and services or food supplement; combination of violations).

- Low Value (Score 1-3): Standard and Obvious Content Violations (e.g., single common violations like failure to include a mandatory warning; other plain violations where no specific or deep consideration was given by FAS uninteresting cases where no violation was found).

Примеры JSON-ов

Пример ответа нейросети по кейсу, не релевантному ФЗ «О рекламе». В приведенном тексте ключ — «is_relevant_to_ad_law», а значение — «false».

41c25dfd-ceee-4220-8d1a-ccd985e9ba17.json


```
{
  "is_relevant_to_ad_law": false,
  "reason": "Дело не относится к ФЗ 'О рекламе', а касается
законодательства о защите конкуренции (ст. 15 ФЗ 'О защите
конкуренции') в части бездействия органа местного самоуправления
при контроле за установкой рекламных конструкций."
}
```


Пример ответа нейросети по кейсу, релевантному ФЗ «О рекламе», но не ценному для RAG. В приведенном тексте много ключей помимо «is_relevant_to_ad_law», например, value_reason, fas_division и т.п.

c498fe0d-32d5-40b9-8d3c-4820d3db8213.json


```
{
  "is_relevant_to_ad_law": true,
  "value_score": 0,
  "value_reason": "ПОНЯТИЕ_РЕКЛАМА: Дело касается способа распространения рекламы без согласия
абонента (ст. 18) и квалификации звонка в качестве рекламы, а не ее содержания.",
  "case_markup": {
    "violation_found": "да",
    "fas_division": "Татарстанское УФАС России",
    "defendant_info": {
      "name": "Российский Фонд Образовательных Программ «Экономика и Управление»",
      "industry": "Образовательные услуги"
    },
    "ad_description": {
      "content": "Телефонный звонок с предложением принять участие в образовательном курсе по
договорному праву.",
      "content_cited": "Алло. – <…>, здравствуйте. Меня зовут <…>, Российский фонд «Экономика и
Управление». Вы ранее участвовали в наших мероприятиях. Сейчас мы собираем заявки на курс «неделя
договорного права» в Москве с 2 по 6 июня. Также проходит в онлайн формате. Могу отправить письма на
почту, посмотрите?",
      "platform": "Телефонный звонок (подвижная радиотелефонная связь)"
    },
    "violation_summary": "Телефонный звонок с предложением образовательного курса «неделя договорного
права» был осуществлен на номер абонента без его предварительного согласия на получение рекламы. Это
является нарушением установленного законом порядка распространения рекламы по сетям электросвязи.",
    "fas_arguments": "Ключевой тезис: ФАС признала звонок ненадлежащей рекламой, распространенной без
предварительного согласия абонента. Юридическое и фактическое обоснование: Закон о рекламе
устанавливает презумпцию отсутствия согласия на получение рекламы по сетям электросвязи, и бремя
доказывания наличия такого согласия лежит на рекламораспространителе. Также ФАС квалифицировала звонок
как рекламу, поскольку он направлен на привлечение внимания к образовательной услуге и адресован
неопределенному кругу лиц. Контрдовод: Довод Фонда о том, что звонок не являлся рекламой, а
представлял собой предложение о возобновлении сотрудничества, был отклонен. Аргумент о наличии
согласия на получение рекламы, включенного в согласие на обработку персональных данных, также был
отвергнут, так как такая форма согласия носит безальтернативный характер и не является добровольным и
недвусмысленным волеизъявлением потребителя. Цитата: 'Согласие на получение рекламных материалов
закрепленное в согласии на обработку персональных данных носит безальтернативный характер согласия на
получение рекламы, а также не соответствует фактическому волеизъявлению потребителя.'",
    "legal_provisions": [
      "ч. 1 ст. 18"
    ]
  }
}
```