**Software Requirements Specification**

Система автоматизованого формування текстових корпусів (САФТК)

1.0

11.06.2025

Горєлов Д.О.

1 Вступ

1.1 Мета документа

Цей документ визначає програмні вимоги до системи автоматизованого формування текстових корпусів (САФТК) з покращеними методами нормалізації та токенізації українських текстів.

1.2 Область застосування

Система призначена для автоматизованого створення структурованих текстових корпусів з різноманітних джерел для дослідження природної мови, розробки мовних технологій та тренування моделей.

1.3 Цільові користувачі

Цільовими користувачами системи чи корпусу є:

– дослідники в галузі обробки природної мови та корпусної лінгвістики;

– розробники мовних технологій та NLP-систем.

1.4 Контекст використання

САФТК функціонує як посередник між сирими текстовими джерелами та готовими до використання лінгвістичними корпусами, забезпечуючи стандартизацію, нормалізацію та структурування текстових даних згідно з сучасними вимогами корпусної лінгвістики.

2 Загальний опис системи

2.1 Призначення продукту

САФТК є модульною платформою для автоматизації повного життєвого циклу текстових корпусів: від отримання сирих текстових даних до створення анотованих, нормалізованих та токенізованих корпусів, готових для лінгвістичного аналізу та машинного навчання.

2.2 Основні функції системи

Система реалізує конвеєрну архітектуру, де кожен етап обробки є незалежним модулем з чітко визначеними входами та виходами:

– інтеграція джерел. Уніфікований доступ до різноманітних текстових джерел;

– структурний парсинг. Вилучення та збереження логічної структури документів;

– нормалізація за стандартами. Приведення текстів до національних та міжнародних стандартів;

– лінгвістична оцінка та виправлення. Автоматизований аналіз якості та коректності текстів з можливим подальшим автоматичним виправленням;

– токенізація. Розбиття тексту на логічні одиниці;

– анотація. Збереження кожного етапу обробки в анотованому документів;

– збереження. Архівування документів з кожного етапу обробки.

2.3 Обмеження поточної версії

Поточна реалізація має наступні обмеження:

– реалізовано компоненти лише для роботи з джерелом «Суспільне»;

– компонент виправлення тексту не реалізовано, використовується заглушка;

– немає підтримки відслідкування прогресу;

– компонент перевірки тексту Language Tool суттєво уповільнює обробку.

3 функціональні вимоги

3.1 Отримання та обробка даних

FR-001: Інтеграція з джерелами даних

– Система ПОВИННА отримувати дані з різноманітних джерел (вебсайти, електронні документи, медіаресурси) за визначеним часовим інтервалом;

– Система ПОВИННА зберігати метадані для кожного запису: джерело, автор, заголовок, дата публікації, посилання;

– Система ПОВИННА забезпечувати механізм доступу до оригінального контенту;

– Система ПОВИННА обробляти помилки доступу до джерел без зупинки загального процесу.

Примітка: Поточна реалізація працює з новинним порталом "Суспільне Новини"

FR-002: Структурний парсинг контенту

– Система ПОВИННА витягувати з вихідних даних основний текстовий контент;

– Система ПОВИННА зберігати структуру документу: заголовки, абзаци, цитати, списки;

– Система ПОВИННА відфільтровувати сторонній вміст (реклама, навігаційні елементи, коментарі);

– Результатом цього етапу ПОВИНЕН бути документ з базовою структурою тексту.

Примітка: Поточна реалізація включає парсинг HTML-сторінок та декодування захищених CloudFlare електронних адрес

3.2 Нормалізація українських текстів

Система ПОВИННА розширити стандартні методи нормалізації специфічними покращеннями для обробки новинних текстів українською мовою з метою забезпечення однорідності представлення телефонних номерів, апострофів та лапок.

FR-003: Нормалізація телефонних номерів

– Система ПОВИННА розпізнавати українські телефонні номери у різних форматах;

– Система ПОВИННА приводити номери до стандартних форматів згідно з дизайн-системою державних сайтів України [1]: +380 (XX) XXX-XX-XX для двозначних кодів операторів та +380 (XXX) XX-XX-XX для тризначних спеціальних номерів;

– Система ПОВИННА обмежувати нормалізацію українськими номерами (код країни +380 або початок з 0);

– Система ПОВИННА перевіряти відповідність кодів операторів списку Vodafone [2];

– Система ПОВИННА генерувати попередження для неоднозначних випадків.

FR-004: Нормалізація лапок

– Система ПОВИННА використовувати «лапки-ялинки» (U+00AB та U+00BB) для зовнішніх лапок та “лапки-лапки” (U+201C та U+201D) для внутрішніх лапок згідно з чинним правописом [3, c. 246];

– Система ПОВИННА забезпечувати правильне чергування внутрішніх і зовнішніх лапок;

– Система ПОВИННА розпізнавати контекст для заміни универсального символу U+0022;

– Система ПОВИННА генерувати помилки для незбалансованих лапок.

FR-005: Нормалізація апострофів

– Система ПОВИННА заміняти всі варіанти символів апострофа на єдиний стандарт U+02BC;

– Система ПОВИННА розрізняти апострофи та одинарні лапки за контекстом;

– Система ПОВИННА обробляти подвійні послідовності апострофів;

– Система ПОВИННА генерувати попередження для символів у нестандартних позиціях.

FR-006: Загальна нормалізація тексту

– Система ПОВИННА нормалізувати пробільні символи;

– Система ПОВИННА генерувати унікальні ідентифікатори для кожної виявленої проблеми;

– Система ПОВИННА класифікувати проблеми як попередження або помилки;

– Результатом ПОВИНЕН бути нормалізований документ зі списком наявних помилок чи попереджень.

3.3 Оцінка якості та покращення тексту

FR-007: Лінгвістичний аналіз

– Система ПОВИННА інтегруватися з інструментами перевірки української мови;

– Система ПОВИННА виявляти граматичні, орфографічні та стилістичні помилки;

– Система ПОВИННА присвоювати унікальні ідентифікатори кожній виявленій проблемі;

– Система ПОВИННА класифікувати проблеми за типами та рівнем критичності.

Примітка: Поточна реалізація використовує LanguageTool для аналізу

FR-008: Автоматичне виправлення

– Система ПОВИННА підтримувати ітеративний процес виправлення виявлених помилок;

– Система ПОВИННА обмежувати кількість ітерацій виправлення (максимум 3);

– Система ПОВИННА зберігати історію всіх виправлень для аудиту;

– Результатом ПОВИНЕН бути оцінений документ з виправленнями та списком залишкових проблем.

Примітка: У поточній версії компонент виправлення реалізований як заглушка

3.4 Семантична токенізація

FR-009: Збереження семантичної цілісності

Система ПОВИННА зберігати як єдині токени наступні категорії конструкцій:

– слова з дефісом, наприклад, складні слова «будь-який», «бозна-що», а також словоформи на кшталт «2022-му», «37-річна»;

– числові діапазони, зокрема конструкції «15-21», «№1-4», що позначають певний діапазон;

– дати та час, наприклад, формати дат «18.05.2021» і часу «22:37»;

– порядкові номери, зокрема «№11», «№3,4»;

– електронні адреси, наприклад «rename@city.kharkiv.ua»;

– телефонні номери у різних форматах, наприклад «+38 (095) 568 38 77»;

– одиниці виміру, наприклад «м²», «км²/місяць», «м/с»;

– температурні показники, зокрема записи на кшталт «-21°С»;

– грошові суми, наприклад «$1,461»;

– спеціальні скорочення, що відповідають чинному ДСТУ 3582:2013 [4], наприклад, «м.» (місто), «вул.» (вулиця), «пл.» (площа) та «ред.» (редакція);

– слова з апострофом, наприклад «сімʼя»;

– числа з розділювачами, зокрема десяткові дроби «234,5» та числа з пробілами між розрядами «1 000 000»;

– приблизні значення, наприклад, конструкції з тильдою «~100».

FR-010: Поділ на речення та токени

– Система ПОВИННА коректно розбивати текст на речення з урахуванням скорочень

– Система ПОВИННА підтримувати як поділ на слова, так і поділ на речення

– Результатом ПОВИНЕН бути токенізований документ з ієрархією речень та токенів

– Система ПОВИННА зберігати всі попередні ідентифікатори проблем

3.5 Збереження та архівування результатів

FR-011: Багаторівневе збереження

– Система ПОВИННА зберігати документи після кожного етапу обробки:

– base/ - результат парсингу з метаданими;

– normalized/ - нормалізовані тексти з інформацією про проблеми;

– evaluated/ - тексти з результатами лінгвістичного аналізу;

– tokenized/ - токенізовані документи з повною ієрархією.

– Система ПОВИННА використовувати структурований XML-формат;

– Система ПОВИННА забезпечувати цілісність аналітичної інформації на всіх етапах.

FR-012: Управління файлами

– Система ПОВИННА генерувати унікальні імена файлів з транслітерацією;

– Система ПОВИННА створювати необхідну структуру каталогів автоматично.

4 АРХІТЕКТУРНІ ВИМОГИ

4.1 Принципи архітектури

AR-001: Модульність

– Система ПОВИННА складатися з незалежних компонентів;

– Кожен компонент ПОВИНЕН мати чітко визначену відповідальність;

– Заміна компонентів ПОВИННА бути легкою.

AR-002: Конвеєрна обробка

– Компоненти ПОВИННІ бути об'єднані в єдиний конвеєр;

– Вихідні дані одного етапу це вхідні дані наступного;

– Система ПОВИННА забезпечувати послідовну трансформацію.

AR-003: Розширюваність

– Додавання нових компонентів ПОВИННО вимагати лише відповідної реалізації абстрактного класу з подальшою підминою старої реалізації.

4.2 Компоненти системи

– AbstractPipeline: координація конвеєра;

– AbstractSource: отримання записів;

– AbstractRecord: отримання змісту;

– AbstractParser: парсинг змісту;

– AbstractNormalizer: нормалізація тексту;

– AbstractEvaluator: оцінка якості;

– AbstractCorrector: виправлення помилок;

– AbstractTokenizer: токенізація;

– AbstractAnnotator: управління форматом документів;

– AbstractSaver: збереження результатів.

5 Нефункціональні вимоги

5.1 Продуктивність

NFR-001: Швидкість обробки

– Час обробки одного документа НЕ ПОВИНЕН перевищувати 30 секунд в нормальних умовах.

NFR-002: Використання ресурсів

– Споживання оперативної пам'яті НЕ ПОВИННО перевищувати 8GB при обробці типових документів;

– Система ПОВИННА ефективно використовувати дисковий простір з можливістю очищення тимчасових файлів;

– Система НЕ ПОВИННА блокувати CPU на тривалий час без можливості переривання.

5.2 Надійність та відновлюваність

NFR-003: Обробка помилок

– Система ПОВИННА продовжувати роботу при помилках обробки окремих документів.

NFR-004: Відновлення після збоїв

– Система ПОВИННА зберігати проміжні результати для можливості відновлення.

Перелік джерел посилання

1. Орфографія та синтаксис. Дизайн система державних сайтів України. URL: https://design.gov.ua/ua/teksty-i-kontent/orfografiya-ta-sintaksis (дата звернення: 03.04.2025).
2. Як дізнатися, який оператор використовує код? | Vodafone Україна. Vodafone Україна - Мобільний зв'язок, Інтернет, Тарифи та Послуги. URL: https://www.vodafone.ua/support/faq/jak-diznatysj-kod-operatora (дата звернення: 03.04.2025).
3. Український правопис / ред.: Є. І. Мазніченко та ін. Наук. думка, 2019. 390 с.
4. ДСТУ 3582:2013. Інформація та документація. Бібліографічний опис. Скорочення слів і словосполучень українською мовою. Загальні вимоги та правила (ISO 4:1984, NEQ; ISO 832:1994, NEQ). На заміну ДСТУ 3582-97 ; чинний від 2014-01-01. Вид. офіц. Київ : Мінекономрозвитку України, 2014. 15 с.