Документация приложения Аббревиатор

В данном документе аббревиатуры, которые могут попасть в итоговую таблицу списка сокращений (если пользователь не отказался от них) выделены в тексте зеленым, а те, что не должны – желтым.

# Список сокращений

Это имитация исходного списка сокращений, составленного вручную при написании документа.

|  |  |
| --- | --- |
| **Аббревиатура** | **Расшифровка** |
| TNF-α | Фактор некроза опухоли альфа (исходная расшифровка, отличающаяся от расшифровки в словаре) |
| FDA | Food and drug administration (исходная расшифровка, отличающаяся от расшифровки в словаре) |
| Cintro | Какой-то новый параметр или новое написание аббревиатуры, которого еще нет в словаре |
| АД | Артериальное давление (нет в тексте) |
| МИО | Не должно быть добавлено в итоговую таблицу сокращений, так как не является аббревиатурой (в тексте в кавычках) |
| XIX | Не должно быть добавлено в итоговую таблицу сокращений, так как не является аббревиатурой (римские цифры) |

# Общие принципы работы алгоритма

Скрипт автоматически находит аббревиатуры в тексте, подтягивает расшифровку из словаря или использует введенную пользователем и формирует список сокращений. Результат во многом зависит от взаимодействия с пользователем. Новые расшифровки сохраняются вместе с аббревиатурой и контекстом для последующей модерации.

**Основные этапы работы алгоритма**:

1. Поиск релевантного текста:
   * Разделы, указанные в списке пропускаемых, исключаются (например, «Список литературы»)
   * Алгоритм определяет начало и конец разделов по названию заголовка и его стилю (Heading) или жирному шрифту.
   * Поиск аббревиатур в таблицах и на рисунках не имплементирован.
2. Сканирование текста для поиска аббревиатур:
   * Аббревиатуры должны содержать минимум две заглавные буквы (латиница: A-Z, кириллица: А-Я).
   * Поддерживаются комбинации с цифрами, дефисами, скобками и символами (например, ICAM-1, GlyT1, TNF-α).
3. Фильтрация
   * Исключаются римские цифры и их комбинации с буквами (например, II, III, XIX, IV, VII, IIB, IVb, VIа, где **а** написано кириллицей, и IIIa, где **а** – латинская буква).
   * Термины из стоп-листа игнорируются (например, ПРОТОКОЛ КЛИНИЧЕСКОГО ИССЛЕДОВАНИЯ – написано заглавными буквами, но не является аббревиатурой).
   * Слова длиннее 8 букв, состоящие только из букв, также исключаются.
   * Слова в кавычках не обрабатываются как аббревиатуры («МИО», «RD-112»).

# Смешанные кириллические/латинские аббревиатуры

Скрипт проверяет аббревиатуры на наличие смешанных кириллических и латинских символов, сравнивает их с вариантами написания в словаре и подтягивает соответствующую расшифровку, если она есть.

**Примеры:**

* ATХ (рускоязычная аббревиатура, но первые две буквы написаны латиницей, последняя – кириллицей): если в словаре есть корректная форма ATX, она будет предложена пользователю.
* СТСАЕ (англоязычная аббревиатура полностью написана кириллицей): если правильная форма присутствует в словаре, она подтянется.
* ВГВ (последняя буква – кириллица, хотя обозначение вирусных гепатитов принято латиницей, например, A, B, C): расшифровка для верной формы будет предложена, если найдена в словаре.

Если правильной формы в словаре нет (например, АTM, где А – кириллица, а остальные буквы – латиница), пользователю будет предложено ввести расшифровку вручную. При этом неверно написанная аббревиатура отправляется на модерацию как кандидат для добавления в словарь.

# Многозначные аббревиатуры

Аббревиатуры могут иметь несколько расшифровок (например, RR может означать Relative Risk, Risk Ratio или R-R Интервал). Программа покажет контекст и предложит выбрать подходящий вариант или добавить свой.

# Однобуквенные аббревиатуры

Аббревиатуры, состоящие из одной буквы, не могут быть найдены в тексте, так как не соответствуют правилу поиска «как минимум две заглавные буквы». Однако программа дополнительно ищет все сокращения из словаря в тексте.

# Сокращения в таблицах

|  |  |
| --- | --- |
| **Рост** | **ИМТ (кг/м2)** |
| 25 | 75,5 |
| 85 | 1,5 |
| **Примечание:**  ИМТ – индекс массы тела. | |

Поиск аббревиатур в таблицах и на рисунках на данный момент не имплементирован.

# Проверки

Если в документе уже есть вручную составленная таблица сокращений, программа проведёт сравнение и выявит различия между существующей таблицей и новой.

Результаты проверки включают:

* Аббревиатуры, которые присутствуют в исходной таблице, но не были найдены алгоритмом.
* Аббревиатуры, которые обнаружены программой, но отсутствуют в существующей таблице.
* Разные аббревиатуры с одинаковой расшифровкой в итоговой таблице.
* Разные расшифровки для одной и той же аббревиатуры (если скрипт покажет такие, обратитесь, пожалуйста, к разработчику – скорее всего в ошибка в скрипте или словаре).

# Ограничения

1. Алгоритм исключает ограниченный список терминов. Если в тексте окажутся другие слова, которые систематически пишутся заглавными буквами, программа продолжит распознавать их как сокращения и предлагать ввести расшифровку. Чтобы добавить такие слова в список исключений, необходимо уведомить разработчика.
2. Могут встречаться сложные (не удовлетворяющие условиям поиска) аббревиатуры, например, фармакокинетические параметры: Vd, Cmax, T1/2. Указанные ФК параметры уже есть в словаре и будут автоматически найдены программой. Однако новые сложные аббревиатуры, такие как Cinf или Tbc, не будут обнаружены.
3. Аббревиатуры из двух слов (например, часто пишут NCI-CTC AE) программа распознает как отдельные сокращения. Если пользователь введет расшифровку для каждого компонента, а в словаре уже есть правильные версии написания (в данном случае NCI и CTCAE), в таблице окажутся все варианты. При подготовке документа следует учитывать, что на данный момент единственный способ обработки аббревиатур, состоящих из нескольких слов – обращение к разработчику с просьбой внести их в словарь.
4. Если одна и та же аббревиатура встречается в словаре и тексте в разных вариантах написания, например, AUC0-24 и AUC(0-24), все варианты попадут в итоговую таблицу (auc <большими буквами> в данном случае тоже попадет в таблицу, так как имеется в словаре и будет найдена поиском всех аббревиатур словаря в тексте).
5. Если раздел со списком литературы оформлен в другом стиле или под другим названием, скрипт продолжит поиск и может добавить в таблицу сокращений аббревиатуры из литературных ссылок.
6. Если при вводе расшифровки пользователь допустит ошибку, скрипт не сможет её исправить. В этом случае потребуется ручное редактирование таблицы сокращений.

# Список литературы

Аббревиатуры, использованные в списке литературы не добавляются в таблицу сокращений.

1. World Health Organization. WHO Guidelines on ...

# TODO

# Добавление новых аббревиатур и расшифровок

Когда в документе встречается новая аббревиатура, отсутствующая в словаре, программа предложит пользователю ввести расшифровку (описание). Процесс:

1. Программа покажет контекст в котором аббревиатура найдена (все куски текста.
2. Пользователь вводит расшифровку вручную. Или наживает х - «убрать»
3. Программа проверяет, есть ли в словаре **похожие** описания.
   * Если найдены, программа спросит, действительно ли необходимо добавить новую аббревиатуру, так как ранее для той же цели использовалась иная.
   * Пользователь может выбрать одну из имеющихся пар или все же добавить новую аббревиатуру с введенной расшифровкой.

**Пример:**

Документ содержит новую аббревиатуру ТТг. Пользователь вводит расшифровку «Тиреотропный гормон (гипофиз)». Программа находит похожую пару (сопоставляя расшифровки):

1. ТТГ – Тиреотропный гормон

Пользователь может выбрать 1, чтобы добавить в таблицу сокращений пару ТТГ – Тиреотропный гормон, или нажать n, чтобы добавить новую пару ТТг – Тиреотропный гормон (гипофиз).

**Важно:** Если выбрана существующая аббревиатура (ТТГ), пользователю рекомендуется проверить текст документа и вручную заменить ТТг на ТТГ. Если пользователь добавит новую аббревиатуру (ТТг) и обновит словарь – обе записи останутся. Это может привести к необходимости ручной проверки для единообразия сокращений в будущем.