Инструкция для разметки датасета SciMDIX

1. Сущности

Задача: выделить сущности в текстах научных статей.
Рассматриваются два типа сущностей: TERM и VALUE.

Начало или окончание сущности не может быть посередине слова, т.е. слова нужно захватывать целиком вместе с суффиксами, окончаниями, аффиксами. Если сущность написана с орфографической ошибкой или опечаткой, то ее все равно следует выделять.

VALUE (Значение)

Числовые значения в сочетании с дополнительной информацией (контекстом или единицей измерения), количественные или качественные показатели, используемые для описания конкретных данных, которые можно измерить или оценить.
К типу VALUE относятся:

1. абсолютные числа (***11, 0,74, 0.89***);
2. проценты (***37,7%, 81,3%, 11.5%***);
3. даты (***2019-2021 гг***);
4. единицы измерения, в том числе, стандартные и нестандартные (***174 гена, 58 человек, 41 вопрос, 75% детей, 18 до 25 лет***).

TERM (Термин)

Слова или фразы, используемые в определенной предметной области для точного обозначения конкретных понятий, явлений или объектов.
К типу TERM относятся:

- термины, состоящие из одного слова, в том числе, являющиеся аббревиатурой (***ПО, БД, бустинг, интерфейс, УЗИ, ВСС***);
- понятия, написанные через дефис, содержащие латинские символы или цифры в составе слова (***СУИQT, SPARQL-запрос, n-грамма, web-сервис, f-мера, c.G3785A***);

Более конкретно по областям знаний можно выделить некоторые группы терминов, но не обязательно ими ограничиваться.
<u>Информационные технологии:</u>

- названия языков программирования (***Python, Kotlin, Java, C++***);
- названия библиотек (***Pytorch, Keras, pymorphy2***);

- названия методов, архитектур, алгоритмов, задач, моделей и др., в том числе аббревиатуры (***zero-shot learning***, ***long short-term memory, LSTM, text-to-speech, задачи NLP, модель конечности “кинетических скульптур” Тео Янсена***) и др.

Лингвистика:

- названия терминов из морфологии, синтаксиса, семантики, прагматики, фонологии (***казахско-русский билингвизм, детская речь, грамматика казахского языка, категории функциональной грамматики, каузальные понятия, синтагматическое соотношение, семантическая категория, парадигма языковых единиц, теория функционально-семантического поля***) и др.

Медицина:

- заболевания, симптомы, диагнозы, патологии (***синдром удлиненного интервала QT, обморок, расстройства аутистического спектра***);
- названия генов, препаратов (***c.G3785A, I2splice, Δ8bp***);
- процедуры, манипуляции (***электрокардиография, компьютерная томография, генетический скрининг***) и др.

Психология:

- способности, эмоции, ощущения (***уровень финансовой независимости, управленческие навыки***);
- процессы, состояния, заболевания (***когнитивные процессы, поведенческое состояние, эмоциональное состояние, депрессия, психическое расстройство***);
- статистические показатели (***коэффициент внутренней согласованности альфа Кронбаха***) и др.

1.1. Одно и то же слово в разных контекстах может являться или не являться термином. Чтобы понять это, необходимо вникать в суть, в общий смысл текста, исходить из того, что термин описывает специфику конкретного исследования, и в заданном контексте не должен быть просто абстрактным понятием.

1.2. Разметка должна быть согласована со смыслом высказываний. Например, в статье по лингвистике, предметом исследования которой являются понятия [***термин “Covid”***] и [***дискурс*** “***Пандемия Covid-19***”], а не сама болезнь, самостоятельным термином не может выступать просто слово “Covid”. Нужно избегать двусмысленности, но при этом стремиться, чтобы разметка была компактная. Как правило, термин встречается в конкретном тексте несколько раз как самостоятельная смысловая единица. Такой признак может служить вспомогательной проверкой.

1.3. Особую сложность представляет выделение многословных терминов. Многословным термином считается цепочка слов максимальной длины, которая при отбрасывании некоторых элементов преобразуется в более общий термин. Как правило, это названия программных продуктов, методов, алгоритмов, задач, подходов ([***Quantum GIS***], [***метод k ближайших соседей***], [***метод SPH***], [***метод опорных векторов***], [***задача оптимального управления***]).

1.4. Кавычки и скобки не рекомендуется включать в состав сущности. Если в тексте встретилось название системы, записанное в кавычках, то сущностью считается только само название без кавычек. Например: [***система поддержки принятия решений***] *(*[***СППР***]*) "*[***DSS Invect 2020***]*"*. Если в состав сущности входит название в кавычках, то открывающие и закрывающие кавычки включаются ([***методика "Словесно-цветовой интерференции"***], [***Оксфордская методика "Счастье"***], [***системы "Домашняя смарт теплица"***]).
Если в тексте встретилось общее понятие, после которого без кавычек сразу идет конкретное имя собственное, называющее это понятие (обычно широко известное), то возможны два варианта: выделено только конкретное имя собственное (***операционная система*** [***Android***]***, язык программирования*** [***Java***], ***ген*** [***CYP21A2***]), либо дополнительно выделено общее понятие как отдельный термин ([***операционная система***] [***Android***]***,*** [***язык программирования***] [***Java***]). Второй вариант, как правило, избыточный (так как без общего понятия смысл высказывания не искажается), поэтому менее предпочтительный.

1.5. Терминами НЕ являются общеупотребимые слова, такие как "решение задачи", "данные", "запись", "безопасность человека", "культурный код", "лекарства", "стационар", "больница", "здравоохранение", "болезнь" и т.д. Такие слова как "система", "алгоритм", "метод", "процесс" могут быть включены в состав термина, конкретного понятия ([***системы "Домашняя смарт теплица"***], [***метод SVM***]). Если в тексте упоминается конкретный метод / алгоритм / подход, ранее обозначенный полным названием, то слова "метод", "алгоритм", "подход" могут считаться терминами, поскольку ссылаются на полную версию понятия. Однако просто абстрактно используемые слова "метод", "алгоритм", "подход" не желательно отмечать как термины, так как сами по себе они не являются достаточно конкретными.
Важно избегать двусмысленности. Рассмотрим пример многословного термина: [***модель структурной организации единого информационного пространства***]. В данном контексте одно слово "модель" выделить недостаточно, так как оно не несет полной информации об особенностях понятия. Поэтому слово "модель" может быть включено в состав многословного термина для уточнения. Например: [***задачи NLP***] и [***методы NLP***]; [***архитектура LSTM***] и [***модель LSTM***]. В данных примерах не следует выделять как самостоятельные термины отдельно слова "задачи", "методы", "архитектура", "модель", "NLP" или "LSTM", так как возникнет неоднозначность, что создаст дополнительные трудности при дальнейшей обработке.
Другие примеры многословных терминов: [***транзакционная модель***]***,*** [***математическая модель упругопластических сред***]***,*** [***теоретико-модельный подход***]***,*** [***задача геонавигации***]***,*** [***информационно-аналитической системы***].

1.6. Если не возникает двусмысленности, то предпочтение отдается конкретным названиям, а не общим понятиям ([***мутации***] в гене [***CYP21A2***]). В данном примере название гена однозначное, но относящееся к нему слово "гене" может быть опущено из-за избыточности, поэтому его не требуется включать. Однако если название гена не

конкретизировано, то словосочетание [***мутаций в генах***] является термином, который описывает изменения.

1.7. Слова “дети”, “пациенты”, “женщины”, “мужчины”, “подростки”, “медсестры”, “врачи” и пр. могут входить в состав сущности типа VALUE ([***88% пациентов***] с [***21-OHD***], [***43 руководителя***]), если сообщается об их количестве, либо в состав сущности типа TERM ([***аутичных детей***], [***детей-аутистов***], [***учителя-мужчины***], [***студентов педагогических специальностей***], [***студентов-мусульман***]), если к ним относятся слова определенной специфики. В остальных случаях они не являются терминами из-за слишком большой общности (*детей с* [***РАС***], *детей с* [***аутизмом***], *подростков с* [***суицидальными мыслями***], *жизнестойкость* [***подростков***]).

1.8. Поскольку информационные технологии применяются для решения большого круга задач в разных областях, то в текстах из области информационных технологий в качестве сущностей могут быть выделены разновидности данных ([***ЭЭГ-данные***], [***гамма-каротаж***] и пр.), а также могут считаться терминами понятия из других предметных областей, если они непосредственно связаны с постановкой задачи или ее решением, и в целом, предметом исследования конкретной статьи ([***спектр гамма-излучения***], [***метроритмическая характеристика***], [***генетическая последовательность***], [***глаголов***] [***казахского языка***]).

1.9. Терминами НЕ являются имена людей, названия стран, городов, регионов, университетов, так как они больше выступают в роли вспомогательной, а не основной информацией для рассматриваемых областей знаний. Однако имена собственные включаются в состав термина, если таким образом обозначены их названия ([***коэффициент внутренней согласованности альфа Кронбаха***], [***модель конечности “кинетических скульптур” Тео Янсена***], [“**дискретный опросник эмоций**” ***Хармона-Джонса***]). Если метод не имеет названия и характеризуется только именем собственным, в таком случае имя собственное включается в состав термина ([***метода Ньютона***], [***метод Якоби***], [***уравнение Пуассона***]).

1.10. В состав термина не могут быть включены глаголы, наречия, местоимения. В сущность не могут включаться такие слова как “различных”, “некоторых”, “их”, “предложенный”, “таких”, “разработанный” и т.д. В этих случаях надо разделять цепочку слов на несколько более мелких.

1.11. Термин не может начинаться с предлога, но может содержать его внутри. Предлог “для” обычно не включается в состав сущности. Часто этот предлог употребляется в значении “используется для” или “применяется для”, а значит, между сущностями, которые он соединяет, есть отношение использования “used_for” (см. раздел Отношения, следующий этап). Пример: *для* [***автоматизированной системы управления***] *разработаны* [***модель двухфакторной аутентификации***] *и последовательный* [***алгоритм генерации временного пароля***]; [***интеллектуальной информационно-аналитической системы***] *для* [***оценки состояния здоровья***

студентов]. В казахском языке послелоги включаются в состав термина, если их не получается отделить.

1.12. Прилагательные могут быть включены в состав термина, но термин не может состоять ТОЛЬКО из прилагательных. Например, слова “казахского”, “турецким”, “тюркских” не являются терминами без относящихся к ним существительных.

1.13. При перечислении каждое понятие выделяется самостоятельно, если оно является существительным или содержит в составе существительное, согласованное в роде, числе и падеже ([***методы опроса***], [***интервью***], [***наблюдения***]). Отдельно стоящее прилагательное не считается термином, и должно быть выделено только вместе с относящимся к нему существительным: *казахского, английского и* [***турецкого языка***]; *фонетического, грамматического, орфоэпического,* [***синтаксического анализа***] *слова; когнитивные, языковые,* культурологические и [***прагматические характеристики***]. Чтобы избежать противоречий при связывании сущностей на следующем этапе, решено не рассматривать в качестве терминов такие цепочки слов как *казахского и русского языков*, *лингвокогнитивного и лингвориторического подходов*, *прозаических и драматических текстов* . В нашей постановке мы не рассматриваем “разорванные сущности”.

1.14. Названия областей исследования или применения результатов являются терминами ([***социолингвистика***], [***антропоцентрической лингвистики***], [***юриспруденции***]). Однако профессии специалистов, работающих в соответствующей области, такие как “экономисты”, “социолингвисты”, “юристы” и пр. не являются терминами, если статья не про людей, а про явления из научной области знаний.

1.15. Названия электронных научных ресурсов и баз данных являются терминами ([PubMed], [Cochran], [Scopus]).

2. Отношения

Задача: обозначить отношения между двумя уже выделенными сущностями в тексте.
Одна сущность может участвовать в нескольких отношениях одновременно. Если в предложении перечислены несколько однородных сущностей, которые семантически связаны с другой сущностью, то в каждой такой паре следует указывать отношение. В одном предложении могут встретиться нескольких отношений. Классы отношений были выбираны на основе следующих критериев.

1. Отношение должно связывать между собой сущности в научных текстах. Например, актантами семантического отношения Communication-Topic (an act of communication is about a topic) выступают не научные термины, поэтому такое отношение не подходит.
2. Отношение должно толковаться однозначно. Например, не рассматривается семантическое отношение Entity-Destination, так как оно имеет также косвенное значение (перемещение, воздействие, назначение).
3. Поскольку на следующем этапе планируется связывание сущностей с Wikidata, то названия отношений взяты из Wikidata.

Отношение	**Пояснение**	**Примеры**	**Wikidata**
HAS_CHARACTERISTIC	Указывает, что сущность обладает определенным свойством, характеристикой или качеством, которое не является числом. X **ассоциирован с** Y; X **с** Y; X **обладает свойством** Y; X **has_characteristic** Y	[трех пациентов] с [СУИQT] у [одного пациента] была обнаружена [мутация c.G3785A] [мутации] в гене [CYP21A2]	https://www.wikidata.org/wiki/Property:P1552
HAS_PART	Указывает на структуру или состав объекта, показывая, из каких компонентов он состоит. X **состоит из** Y X **has_part** Y	[кардиопанель] состоит из [174 генов] [опросник] состоит из [трех блоков]	https://www.wikidata.org/wiki/Property:P527
HAS_USE	Отношение использования: X **используется для/в качестве** Y; X **применяется для/в качестве** Y. В состав Y могут входить следующие слова: - создания, - анализа, - изучения,	для [генетического скрининга] использовалась панель – [Illumina TruSight Cardio panel] [метод статистической обработки]	https://www.wikidata.org/wiki/Property:P366

	- исследования, - сравнения, - решения, - выполнения, - вычисления. X **позволяет** - **создавать** Y, - **анализировать** Y, - **изучать** Y, - **исследовать** Y, - **сравнивать** Y, - **решать** Y, - **выполнять** Y, - **вычислить** Y, - **автоматизировать** Y, - **управлять** Y. Направление будет противоположное, если аргументы поменяны местами: Y **выполняется с помощью** X. X **has_use** Y	использовался для [анализа текстов] [диагностического инструментария] по [оценке предпринимательской активности]	
HAS_VALUE	Указывает на числовое значение, связанное с объектом, например, количество участников, элементов и др. X **имеет значение** Y; **значение** X **находится в диапазоне** Y; **значение** X **варьируется от** Y1 **до** Y2. X **quantity** Y	[дефицит 21-гидроксилазы] составляет более [90-95% случаев] [мутации I172N] ([37,7%]) и [I2splice] ([26,2%]), за ними следовали [повреждения Δ8bp] ([11,5%]) и [Q318X] ([9,8%])	отсутствует
SUBCLASS_ OF	Таксономическое отношение между объектом и множеством, показывающее иерархию, определяет, что одна сущность является подтипом или конкретным примером другой сущности: X **подкласс от** Y; X **– это** Y. X **subclass_of** Y	[архитектуры]: [многослойный перцептрон] и [сверточные нейронные сети] [сердечно-сосудистым и нарушениями], включая [СУИQT]	https://www.wikidata.org/wiki/Property:P279 apple is a <subclass of> fruit

		[расстройства аутистического спектра] (PAC) – это [неврологические расстройства] [криптографического метода шифрования] [Base64] разработанная под [операционную систему] [Windows]	
SYNONYM	Отношение синонимии, показывающее, что два термина используются взаимозаменяемо; часто это отношение связывает полное название с аббревиатурой или переводом, который приводится рядом в скобках: X (Y) X **synonym** Y	[графический процессор] ([GPU]) [сверточная нейронная сеть] ([CNN]) [внезапная сердечная смерть] ([BCC]) [Дефицит 21-гидроксилазы|T17|TERM] ([21-OHD|T18|TERM]) [расстройства аутистического спектра] ([PAC])	https://www.wikidata.org/wiki/Property:P5973 human <has characteristic> gender

ПРИМЕЧАНИЕ: Глагол, используемый для выражения отношения (например, “применяется”) и существительное, которое вступает в отношение, но при этом характеризует процесс (например, “применение чего-то”) имеют схожую семантику. Как правило, глагол соответствует отношению, а отглагольное существительное является сущностью или ее частью.