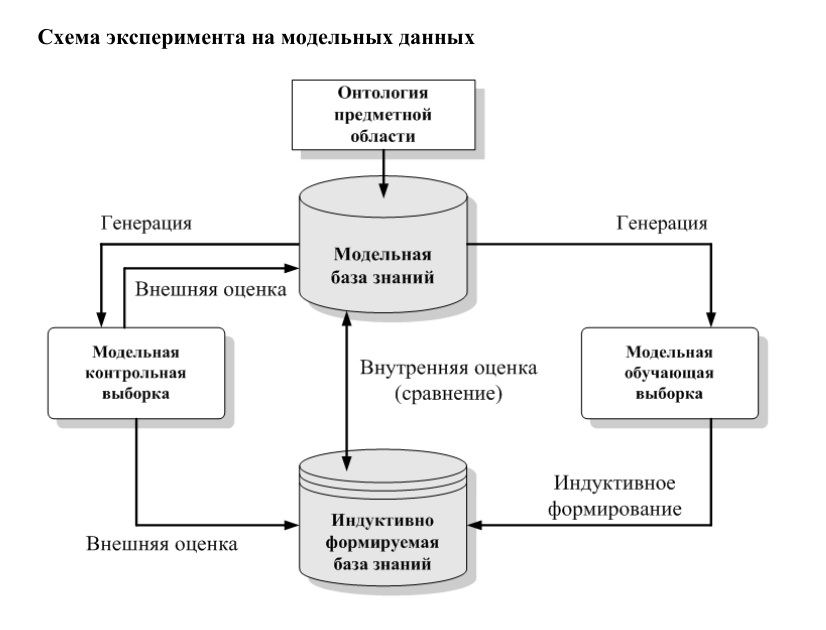
1. Описание функционала

Реализована программа генерации модельной базы знаний и программа генерации обучающей выборки на основе базы знаний.

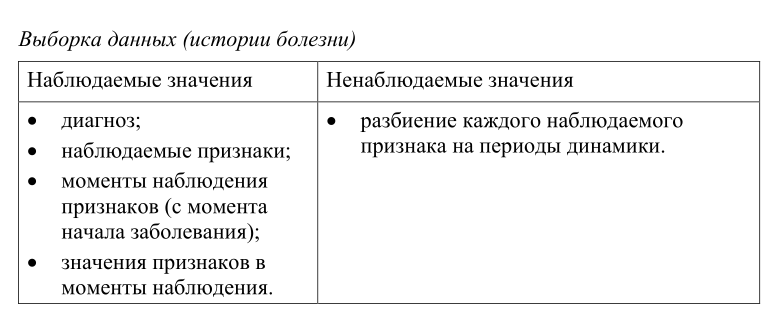
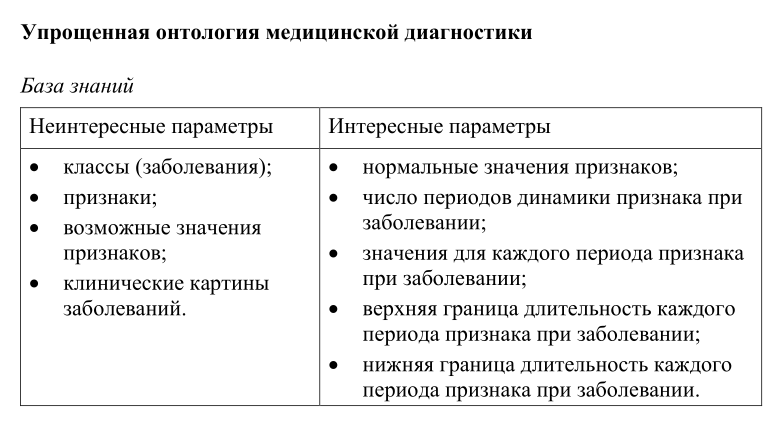
Данные программы необходимы для тестирования алгоритмов, извлекающих базы знаний из **Датасетов**.

Принципиальная схема работы такая:



Реализованные программы осуществляют генерацию «Модельной базы знаний» и «Модельной обучающей выборки». «Модельная обучающая выборка» фактически является **Датасетом**.

Для генерации **Модельной базы знаний** применялась **Упрощённая онтология медицинской диагностики**.



Было написано 2 программы: KnowledgeBaseGenerator.py и LearningSetGenerator.py для генерации **модельной Базы Знаний** и генерации **модельной выборки (Датасета)** соответственно. Для настройки параметров генерации существует файл Config.py.

Программа для генерации Модельной базы знаний KnowledgeBaseGenerator.py имеет флаг -o для обозначения выходного файла.

Пример:

python KnowledgeBaseGenerator.py -o ./data/model\_knowledgebase.json

Программа для генерации Модельной обучающей выборки (Датасета) LearningSetGenerator.py имеет имеет флаги -o для обозначения выходного файла и -i для обозначения входного файла.

Пример:

python LearningSetGenerator.py -i ./data/model\_knowledgebase.json -o ./data/learning\_set.json

Имена файлов могут быть произвольными. Далее в качестве примера будут использоваться те же имена, что и выше.

Цепочка работы программ следующая:

[ KnowledgeBaseGenerator.py ] -> model\_knowledgebase.json -> [ LearningSetGenerator.py ] -> learning\_set.json

**Формат знаний, использованных в программах для данной онтологии, следующий:**

На самом верхнем уровне файл содержит 2 поля:

"features" - массив описаний признаков (или симптомов)

"classes" - массив описаний классов (или заболеваний)

Каждый признак в "features" содержит имя признака "name", массив всех возможных значений признака "values", и массив нормальных значений признака "normal\_values".

Значения в "values" имеют формат "f{индекс признака}value{индекс значения}". Префикс f{индекс признака} нужен скорее в отладочных целях, чтобы ошибки вроде попадания значений одного признака в значения других признаков были очевидны.

Значения "normal\_values" генерируются просто как первые N значений "values".

Каждое описание класса (заболевания) в "classes" содержит поля:

"name" - название заболевания

"symptoms" - Клиническая Картина; массив объектов, описывающих Периоды Динамики для каждого признака заболевания.

Объекты в "symptoms" содержит поля:

"feature" - название признака

"periods" - массив с описанием периодов динамики; массив упорядоченный, поэтому в описании периодов динамики отсутствует поле с индексом.

Объекты в "periods" содержит поля:

"duration\_lower" - нижняя граница длительности периода

"duration\_upper" - верхняя граница длительности периода

"values" - массив значений для периода