МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

В начале поднимается вопрос, почему вообще задавать вопросы на естественном языке удобно и нужно.

Сама работа, по сути, состоит из следующих этапов:  
- анализируются все методы и классы

- по git анализируются все методы по принципу – до/после.

- для каждого метода строится текст на естественном языке

- эти тексты обрабатываются (убирается лишняя информация) и строятся их векторные представления

- вопрос, который мы задали на естественном языке – мы тоже строим в векторном представлении

- для векторного запроса и документов рассчитывается косинусное сходство

- N лучших документов поступает на обработку в BERT

Обзор существующих решений:  
- CODEnn (особо тяжелая DL, требующая больших ресурсов, правильный ответ может находиться далеко в ранжировании)

- NCS (обучение без учителя, лучше, чем стандартные методы, но в целом точность плохая, требует параллельный корпус)

- UNIF (подход проще чем CODEnn, лучше, чем NCS)

В целом важно обучать на основе настоящих вопросов пользователей. NCS и UNIF – код закрыт, проблематично ими воспользоваться.

Дальше поднимается проблема документаций, их актуализация и поддержка. И почему плохо формировать текст на естественном языке основываясь на них.

Рассмотрение Hipikat системы.

Опирается на документацию, чаты, код, отчеты. Проблематика – необходимо погружение и профессионализм пользователя, который сможет задавать конкретные узконаправленные запросы. Нет глубокого анализа. Ориентирование на использование узконаправленных сервисов.

BERT

Хорошее решение для получения ответа на вопрос на естественном языке + открытый код, хорошее базовое обучение на запросах вопрос-ответ из Википедии.

Далее описывается более подробно анализ исходного кода, извлечение данных из git. Извлечение релевантного документа по FastText – предсказывает окружающие слова на основе выбранного одного слова, лучше, чем IDF и BM25 (которые требуют, чтобы нормализованные токены в точности совпадали у запроса и документа).

В конце приводятся различные эксперименты – замена слов, односложный запрос и т.д.