ТЕХНИЧЕСКОЕ ЗАДАНИЕ ПО ПРОЕКТУ “ВЫШКА.СЛОВАРИ”

1. **ОПИСАНИЕ:**  
   Финальный продукт -- платформа двуязычных переводных электронных словарей с удобной навигацией и разными типами поиска (возрождение идей “Яндекс.Словари”). Базой данных служат электронные .xml-словари, преобразованные согласно нашему TEI-шаблону. Архитектура состоит из собственно базы данных и web-оболочки. Пользователь сайта вводит input, система обращается к базе данных и выводит пользователю словарную статью. Типы предоставляемой в статье информации указаны в TEI-шаблоне.
2. **ЭТАПЫ:**

|  |  |
| --- | --- |
| осень 2017 | посмотреть на структуру словарей, TEI-шаблон, макет и пользовательские сценарии; доступные словарные ресурсы |
| январь-февраль 2018 | распарсить и проиндексировать словари по TEI-шаблону |
| март-апрель 2018 | решение задач лексикографического поиска, подготовка вспомогательного контента |
| май 2018 | тестирование |
| осень 2018 | создать базу данных для каждого из словарей и вспомогательного контента (бек-энд) |
| осень 2018 | создать web-оболочку (франт-энд) |

\* время условное и показывает ударный момент каждого этапа

1. **ЗАДАЧИ:** 
   1. Парсинг словарей
      1. изучить исходные словари
      2. внести поправки в TEI-шаблон
      3. типология конвертации для конкретных словарей
   2. Индексация словарей
      1. Типы индексов
      2. Проиндексировать элементы в различных полях
      3. Индексация элементов в свободных полях (толкования, примеры)
      4. Единый индекс для всех словарей
   3. Подтягивание семантических векторов
      1. взять существующие модели, добавить для других языков
      2. ..
      3. ..
   4. (Здесь должны быть задачи для вспомогательного контента -- откуда берется произношение, биграммы, гиперонимы, как решается search by regular expressions, что происходит, если пользователь неправильно набирает слово и т.п.)
   5. Скрипты для валидации, оценка качества словарей

##### 3.А. Словари в работе

Стас -- немецко-французский (в обе стороны)

Маша -- англо-французский (в обе стороны)

Оля -- немецко-испанский (в обе стороны)

Наташа -- англо-немецкий (в обе стороны)

все -- русско-английский

Оля -- русско-немецкий

Стас -- русско-украинский

Маша -- русско-белорусский

Наташа -- ???

## **ТЕСТИРОВАНИЕ:**

* 1. Валидность TEI-разметки (скобки, атрибуты, индексы и т.д.)
  2. Валидность содержания (пустые поля, сопоставление со списками разрешенных элементов, валидность по шаблонам/регулярам)
  3. Качество унифицированных индексов (по всем словарям) -- контроль повторов (орфография!), значимых/случайных отсутствий и т.п.
  4. Плотность, связность и прочие графовые свойства словарей (может быть проектом по курсу соцсетей)
  5. Тестирование онлайн-функций

