НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Проект "Краулер для автоматической категоризации сайтов"

Мартынов Н.И.,Кубышкина Е.К.,Хайкова С.П.,Антонов А.А.
МИЭМ им. А.Н.Тихонова ОП "Прикладная математика"
Руководители проекта - Буров В.В., Зонтов Ю.В.

Видео-презентация -
https://www.youtube.com/watch?v=R1UhPWDIELY

НИУ «Высшая школа экономики»

12 ноября 2018 г.

Суть задачи

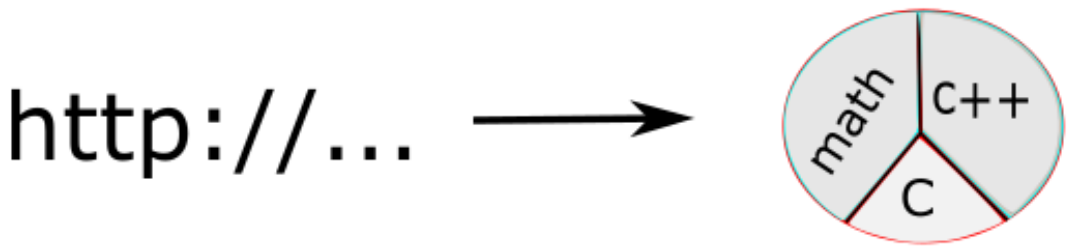
Ожидаемый результат

Задача:

- Разработка инструмента, способного распознавать содержимое сайтов

Хотим получить:

- Web-сервис, в который мы загружаем адрес сайта и на выходе получаем сферы, к которым с определённой вероятностью принадлежит контент сайта



Актуальность задачи

Зачем нужен наш web-сервис?

Где можно это применять:

1. Использование в разработке антивирусных программ
   - Оценка безопасного перехода на незнакомый сайт
2. Оценка качества работы поисковых систем
   - Сравнение содержимого предложенных сайтов с тем, что было в пользовательском запросе
3. Осуществление "родительского контроля"
   - Запретить доступ на определённые сайты в зависимости от их содержания

Как мы хотим достичь результата

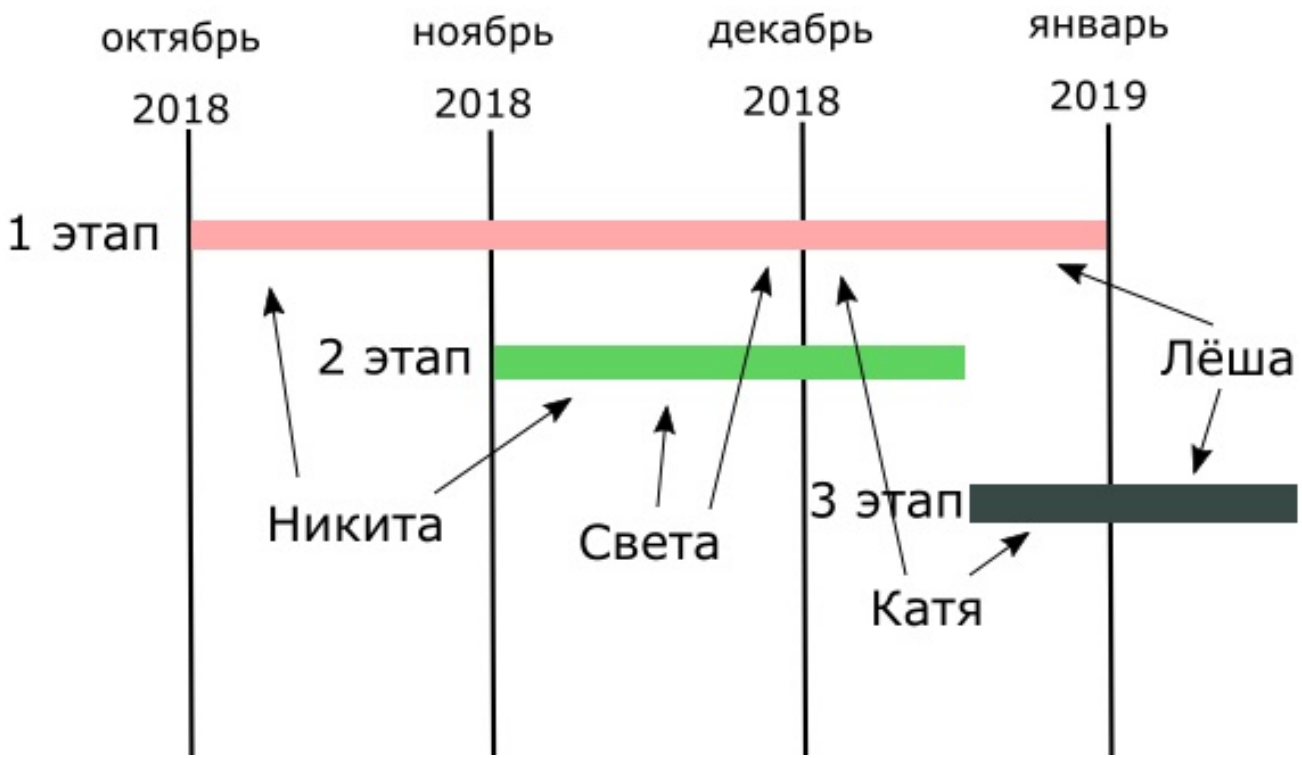
План работ

1. Составление "базы знаний":
   - октябрь 2018г - январь 2019г
   - наполнение базы данных, которая хранит характеристики уже откатегоризированных сайтов
2. Разработка алгоритма категоризации:
   - начало ноября 2018г - середина декабря 2018г
   - на основе введённой метрики сходства и объектов из "базы знаний"делать вывод о характере контента сайта
3. Разработка web-interface:
   - середина декабря 2018г - январь 2019г
   - интерфейс, через который пользователь будет говорить нашему алгоритму, про какой сайт он хочет всё знать

Как мы хотим достичь результата

Распределение ролей

Как мы это представляем:



Спасибо за внимание!

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ