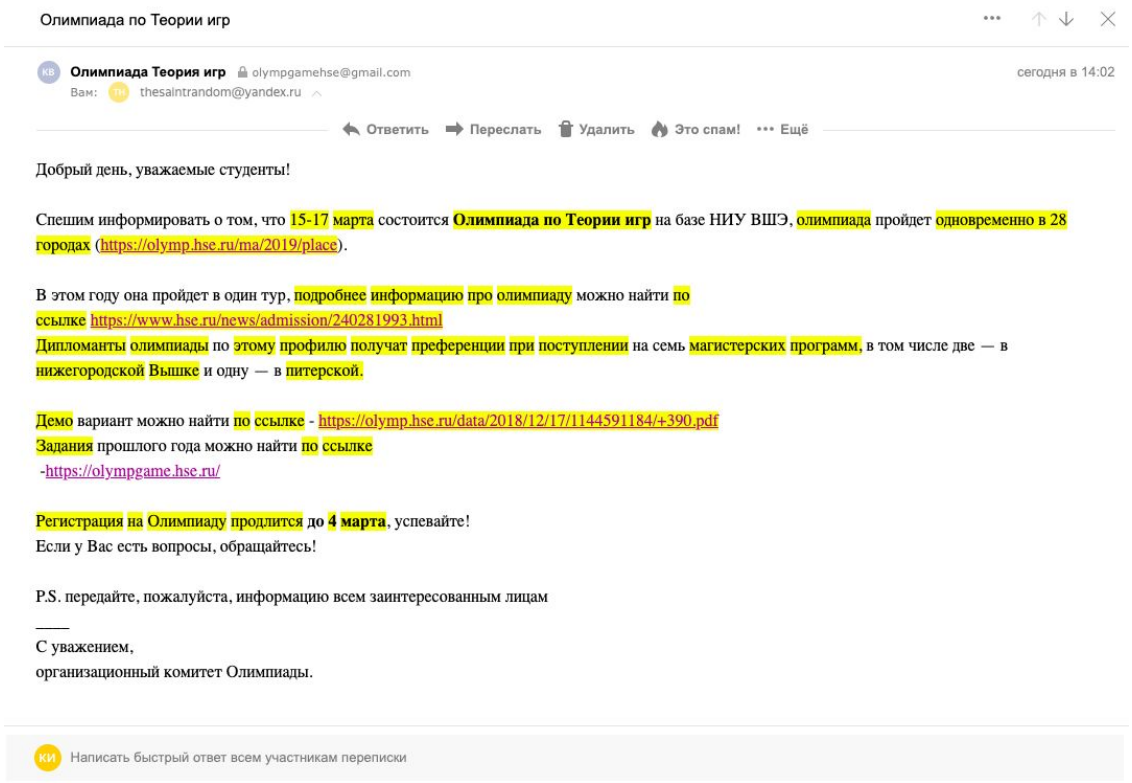
Подсветка важного в текстах электронной почты

Автор: Кирилл Булгаков, 471 гр.
Научный руководитель: к. т. н., доц. Т.А. Брыксин
Рецензент: ведущий разработчик ООО "Яндекс Технологии" А.С. Селиванов
Консультант: аналитик ООО “Интеллиджей Лабс” Н.И. Поваров

Санкт-Петербургский Государственный Университет
Кафедра системного программирования, 2019

Проблема



Олимпиада по Теории игр

Олимпиада Теория игр olympgamehse@gmail.com сегодня в 14:02
Вам: thesaintrandom@yandex.ru

Ответить Переслать Удалить Это спам! Ещё

Добрый день, уважаемые студенты!

Спешим информировать о том, что 15-17 марта состоится **Олимпиада по Теории игр** на базе НИУ ВШЭ, олимпиада пройдет одновременно в 28 городах (https://olymp.hse.ru/ma/2019/place).

В этом году она пройдет в один тур, подробнее информацию про олимпиаду можно найти по ссылке https://www.hse.ru/news/admission/240281993.html
Дипломанты олимпиады по этому профилю получат преференции при поступлении на семь магистерских программ, в том числе две — в нижегородской Вышке и одну — в питерской.

Демо вариант можно найти по ссылке - https://olymp.hse.ru/data/2018/12/17/1144591184/+390.pdf
Задания прошлого года можно найти по ссылке
-https://olympgame.hse.ru/

Регистрация на Олимпиаду продлится до **4 марта**, успевайте!
Если у Вас есть вопросы, обращайтесь!

P.S. передайте, пожалуйста, информацию всем заинтересованным лицам

—
С уважением,
организационный комитет Олимпиады.

Написать быстрый ответ всем участникам переписки

Постановка задачи

Цель

- Реализовать решение для подсветки важного в текстах электронной почты на платформе браузерных расширений

Задачи

- Выбрать алгоритмы и определить метрики качества
- Сравнительный анализ алгоритмов
- Спроектировать архитектуру и реализовать решение на платформе браузерных расширений Chromium

Алгоритмы выделения важного

По словам	По предложениям

Алгоритмы выделения важного

По словам	По предложениям
TF-IDF ● Статистическая мера ● Частота в документе, редкость в корпусе документов	

Алгоритмы выделения важного

По словам	По предложениям
TF-IDF ● Статистическая мера ● Частота в документе, редкость в корпусе документов	
Word2Vec ● Векторное разложение ● Норма вектора как степень важности	

Алгоритмы выделения важного

По словам	По предложениям
TF-IDF ● Статистическая мера ● Частота в документе, редкость в корпусе документов	**TextRank** ● Текст в виде графа ● Вес вебра - степень важности ● Ищем путь с самым большим весом ● Готовая реализация: nltk
Word2Vec ● Векторное разложение ● Норма вектора как степень важности	

Алгоритмы выделения важного

По словам	По предложениям
TF-IDF ● Статистическая мера ● Частота в документе, редкость в корпусе документов	**TextRank** ● Текст в виде графа ● Вес вебра - степень важности ● Ищем путь с самым большим весом ● Готовая реализация: nltk
Word2Vec ● Векторное разложение ● Норма вектора как степень важности	**LSA** ● Текст в виде матрицы ● Используем разложение матриц для уменьшения размерности ● Готовая реализация: sumy

Оценка качества

- Нет возможности использовать обычные метрики, такие как precision/recall/F-measure и пр.
- Было решено использовать экспертные оценки с использованием сервиса Яндекс.Толока

Платформа Яндекс.Толока

[Машинное обучение, часть 2] Появилось новое домашнее задание
Когда: 25 фев. в 13:50
Кому: mymail@yandex.ru
От кого: noreply@compscicenter.ru

На сайте опубликовано новое задание «Вступительный тест 2 семестр (Машинное обучение, часть 2, весна 2019)», его необходимо сдать до 20:20 11 февраля. Описание домашнего задания ниже, в письме может отображаться некорректно.

Тест по первому семестру

Это письмо отправлено автоматически и не требует ответа.

[Машинное обучение, часть 2] Появилось новое домашнее задание
Когда: 25 фев. в 13:50
Кому: mymail@yandex.ru
От кого: noreply@compscicenter.ru

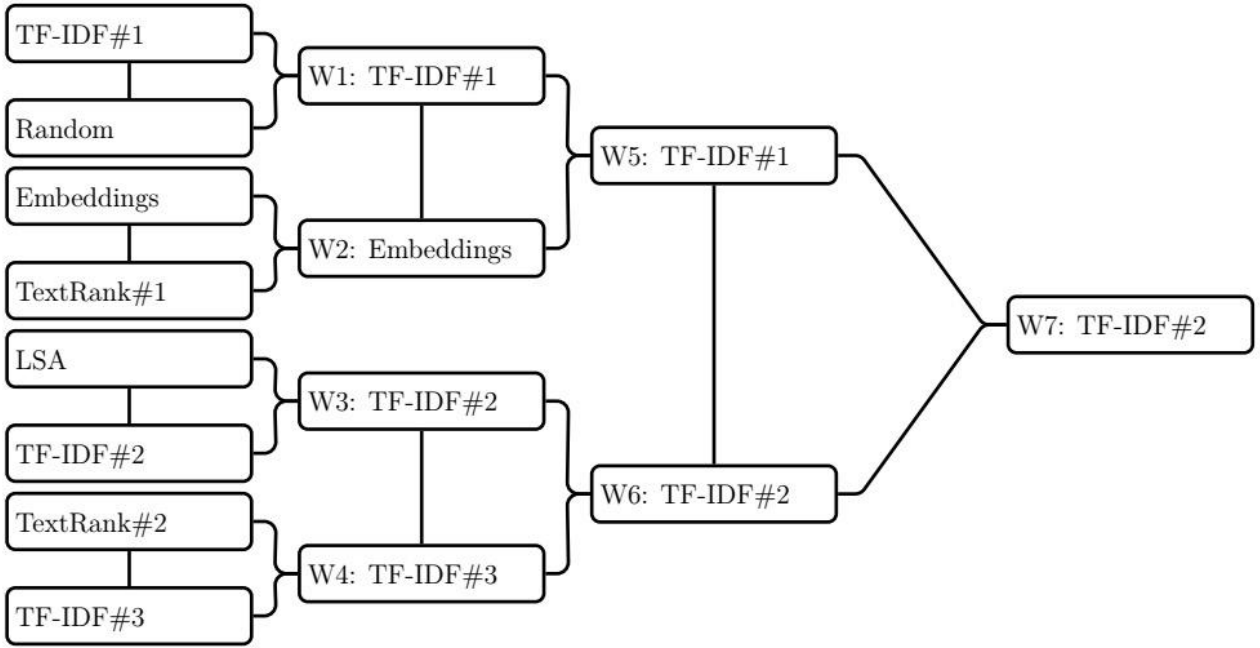
На сайте опубликовано новое задание «Вступительный тест 2 семестр (Машинное обучение, часть 2, весна 2019)», его необходимо сдать до 20:20 11 февраля. Описание домашнего задания ниже, в письме может отображаться некорректно.

Тест по первому семестру

Это письмо отправлено автоматически и не требует ответа.

1 ● Левый лучше 2 ○ Правый лучше

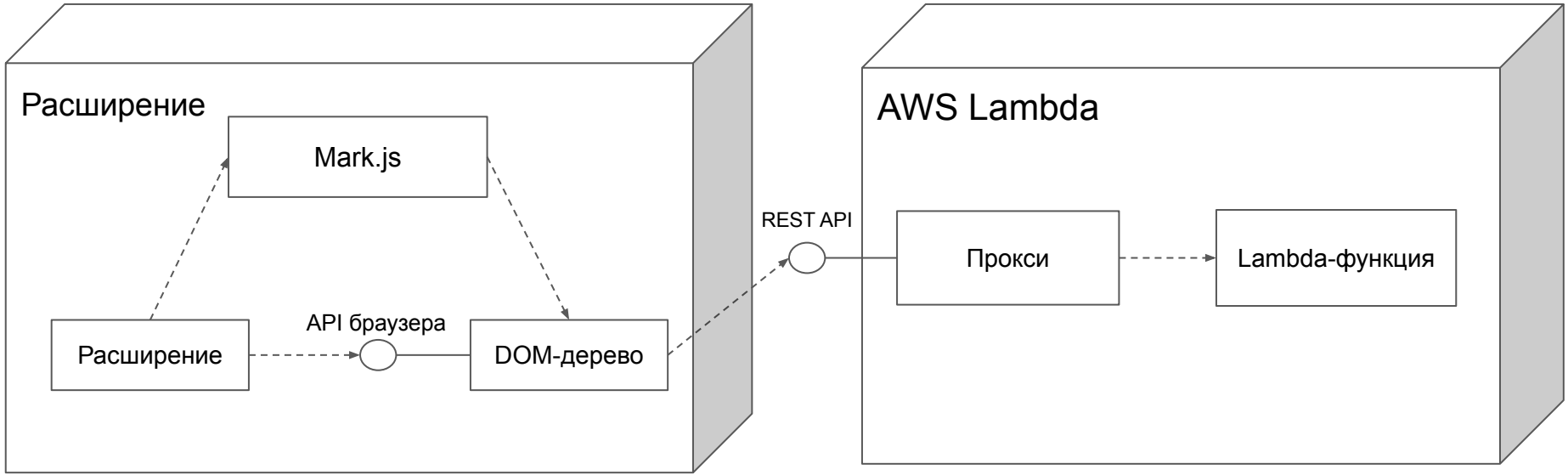
Турнирная сетка
для сравнительного анализа алгоритмов



Утилиты для автоматизации сравнительного анализа алгоритмов

- Автоматическое создание скриншотов при помощи Selenium
- Генерация заданий и сбор результатов при помощи API Яндекс. Толока
- Анализ результатов

Архитектура решения



Реализация алгоритмов

- Python
- Упаковка приложения в Docker-контейнер
- Автоматическое развертывание при помощи API AWS Lambda

Результаты

- Выбраны и реализованы следующие алгоритмы для подсветки важного в тексте: TF-IDF в трех вариациях, TextRank в двух вариациях, алгоритм с использованием норм векторов, алгоритм с использованием LSA. Для оценки качества использованы экспертные оценки
- Спроектирован и выполнен на базе платформы Яндекс.Толока эксперимент с целью выявления лучшего алгоритма. Лучшим алгоритмом оказался TF-IDF на национальном корпусе русского языка
- Спроектирована и реализована клиент-серверная система на базе расширений Chromium (использованы технологии AWS Lambda, Docker и языки JavaScript и Python)