Рекомендация онлайн-курсов

Тема: NLP

Avosya

команда

Александр Никулин:
2-й курс НИУ ВШЭ СПб GitHub: @Howuhh

Аня Батаева:
3-й курс НИУ ВШЭ СПб GitHub: @fyzbt

Вадим Володин:
2-й курс СПБГУ GitHub: @PolyProgrammist

Оля Силютина:
3-й курс НИУ ВШЭ СПб GitHub: @olgasilyutina

Ярослав Соколов:
2-й курс СПБГУ GitHub: @SokolovYaroslav

идея

Extension в Хроме для пользователей Stackoverflow

бизнес-задача

Адаптация новых пользователей Stackoverflow с целью их удержания

планы по реализации

- Content-based рекомендательная система
- НЛП-алгоритмы (topic modelling)
- Текстовые данные Stackoverflow
- Данные Coursera
- ~~Chrome Extension~~ (Telegram bot)

данные

Questions: 289,122
Id
Body
Title

Courses: 533
Id
Описание курса
Описание недель
Рейтинг

Предобработка текстовых данных: стемминг, удаление стопслов, regexp и все дела

текущая стадия

Бот в Телеграме (@EduStack)

На вход
ссылка на вопрос со Stackoverflow

Алгоритмы внутри
.1 knn по тегам и ключевым словам из постов на Stackoverflow
.2 knn на основе tf-idf по текстам вопросов
.3 усреднение результатов моделей

На выход
Максимально релевантный курс

метрика качества моделей: пример

Облако слов по данным вопроса

https://stackoverflow.com/questions/7851077/how-to-return-index-of-a-sorted-list

Облако слов по данным курса

https://www.coursera.org/learn/python-representation

что пытались

- Word2Vec – для affinity propagation (уменьшить размер матрицы)
- Affinity Propagation – кластеризация вопросов и курсов
- LDA – разбивка курсов по латентным топикам
- Label Propagation – кластеризация вопросов и курсов
- SVD – для уменьшить размерность

планы на будущее

Получить данные о пользователях и построить user-based рекомендательную систему, чтобы сравнить её с текущей.