Отчет по выполненой работе на тему

«Поиск и автоматический анализ научных публикаций».

Выполнил:

учащийся группы

МАГ15 ПМИ

Соколов Артём

ВШЭ 2016г.

Изначально было запланировано что поиск и анализ публикаций будет основываться на тематическом моделировании. Это сравнительно новое направление в машинном обучении подразумевает что поиск будет выполняться на основе анализа документа вцелом, т.е. с учетом текста, ссылок, авторов. В нашей группе мы разделились так что 4 человека отвечали непосредственно за исследование возможностей по созданию тематической модели на основе научных работ. 2 человека искали и исследовали возможности языка R и два человека делали то же самое для языка python с целью дальнейшего сравнения.

Я принимал участие в работе второй группы. Мы нашли очень перспективную библиотеку для создания и работе с тематическими моделями BigARTM, написанной на языке c++, но имеющей, также, API для языка python (http://bigartm.org/). Коллегами была найдена и выложена большая выборка научных статей на медицинские темы. Мы обработали перевели эти документы в формат vowpal wabbit, поддерживаемый BigARTM и далее, разделили наши усилия.

Коллега занимался исследованием возможностей модели в направлении мультимодальности (т.е. учитывания не только текста, но и ссылок и авторов для нахождения релевантных), а я занимался построением иерархической тематической модели. Такая модель помимо матриц документы-темы и темы-слова содержит матрицу темы-подтемы для каждых соседних уровней иерархии. ([www.machinelearning.ru/](http://www.machinelearning.ru/)). Как я выяснил, BigARTM иммет интерфейс для построения таких моделей, но не в основном, стабильном бранче. Я написал код на питоне для того что бы построить трёхуровневую модель на основе текстов документов. Подбирались параметры и регуляризаторы для того что бы сделать матрицы разреженными (см. <http://bigartm.org/> для дополнительной информации о принципах работы BigARTM). Количество тем и подтем подбиралось экспериментально. Какой-то абсолютной числовой характеристики, по которой можно было бы судить о чистоте тем найдено не было. Я приведу здесь примеры для двух уровней, при этом я уверен что, полученные результаты еще возможно значительно улучшить.



Таблица тем Уровень 1



Таблица части тем Уровень 2

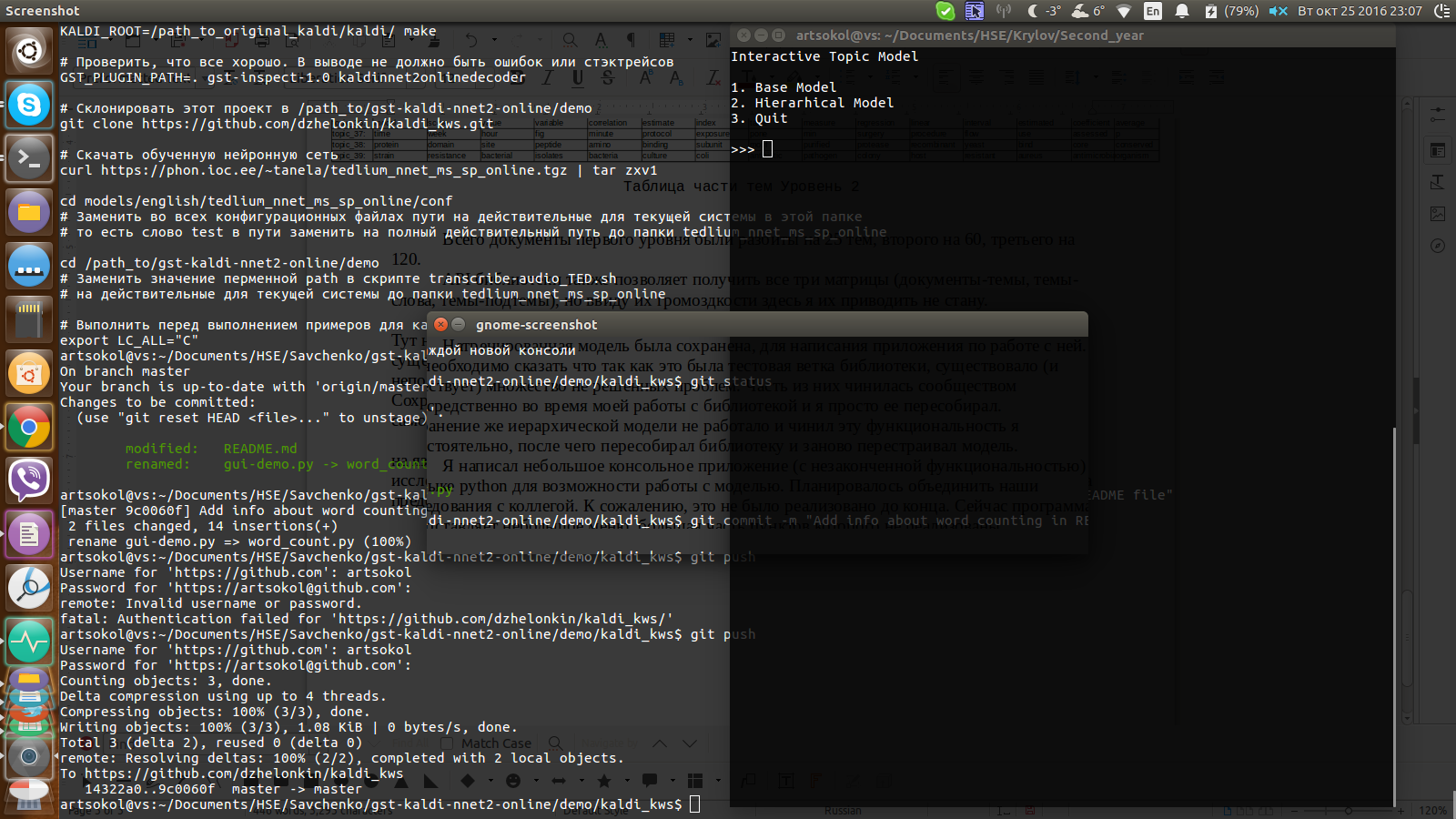
Всего документы первого уровня были разбиты на 25 тем, второго на 60, третьего на 120.

API библиотеки также позволяет получить все три матрицы (документы-темы, темы-слова, темы-подтемы), но ввиду их громоздкости здесь я их приводить не стану.

Натренированная модель была сохранена, для написания приложения по работе с ней.

Тут необходимо сказать что так как это была тестовая ветка библиотеки, существовало (и существует) множество не решенных проблем. Часть из них чинилась сообществом непосредственно во время моей работы с библиотекой и я просто ее пересобирал. Сохранение же иерархической модели не работало и чинил эту функциональность я самостоятельно, после чего пересобирал библиотеку и заново перестраивал модель.

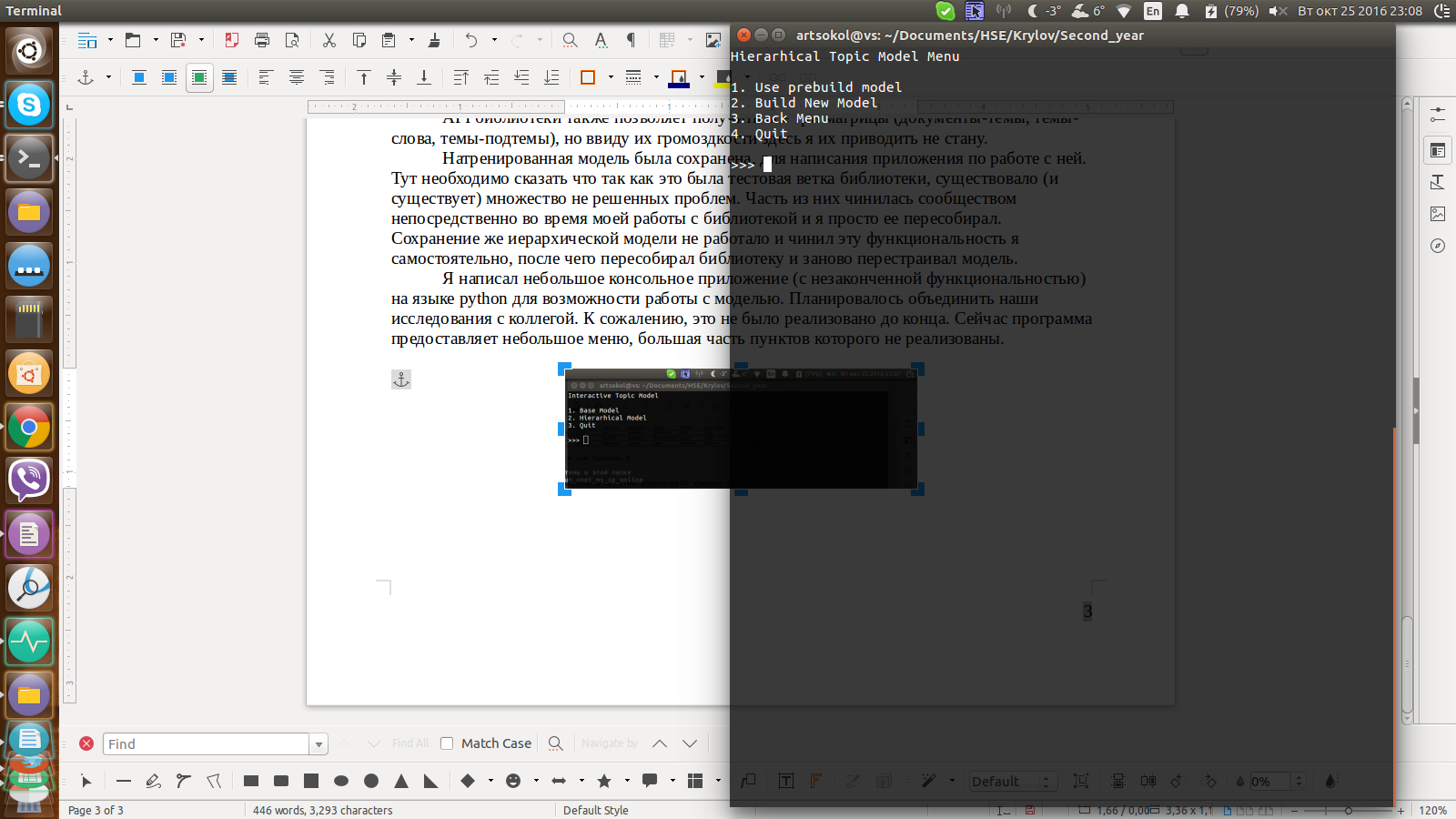
Я написал небольшое консольное приложение (с незаконченной функциональностью) на языке python и с BigARTM API идля возможности работы с моделью. Планировалось объединить наши исследования с коллегой. К сожалению, это не было реализовано до конца. Сейчас программа предоставляет небольшое меню, большая часть пунктов которого не реализованы.



Верхний уровень меню

**Base model** — уровни меню для работы с обычной, не иерархической моделью(не реализована).

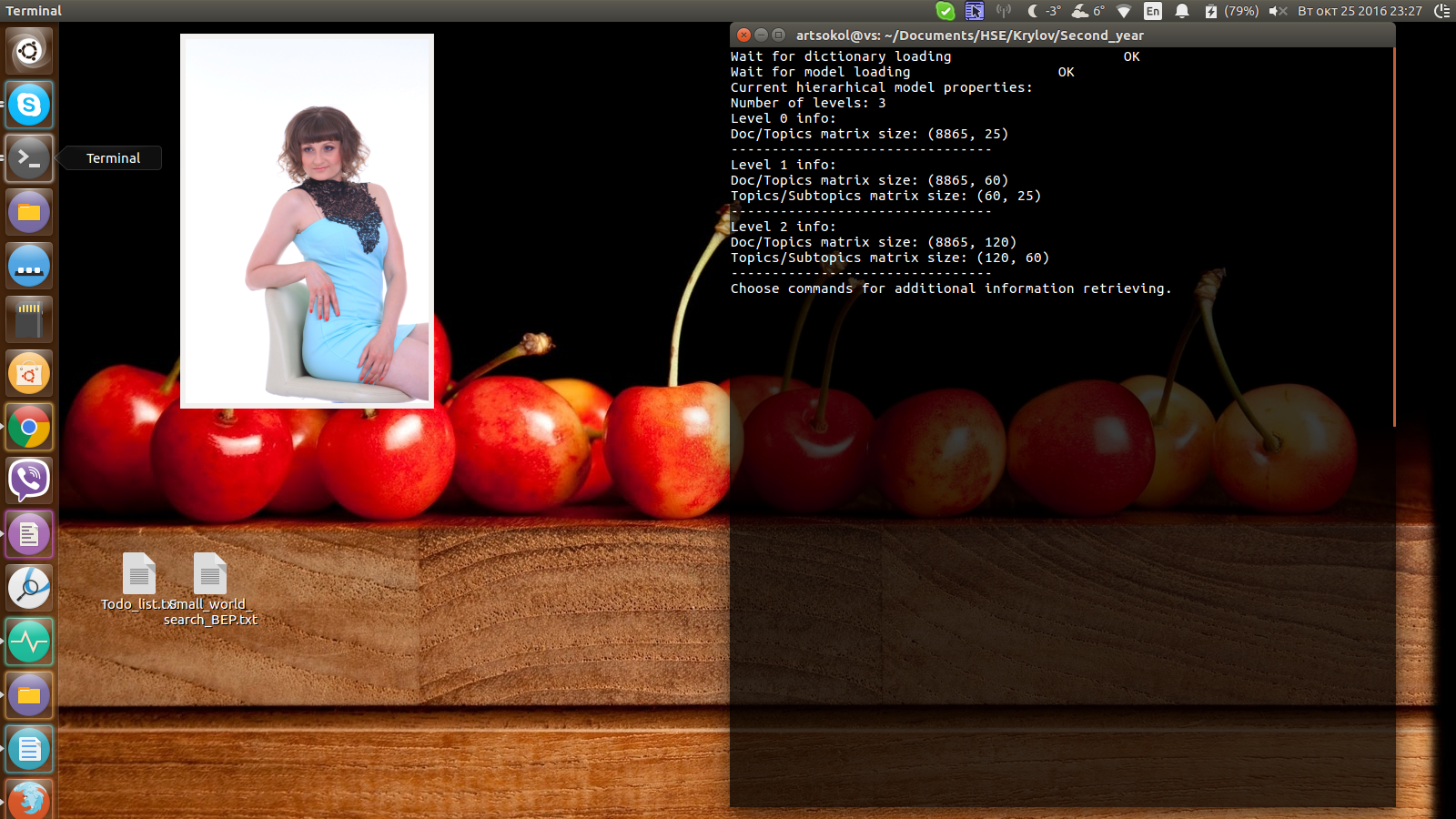
**Hierarhical Model** — работа с иерархической моделью.



Второй уровень меню

**Use prebuild model** — загрузка построенной модели с возможностью дальнейшего получения информации.

**Build New Model** — задание параметров для построения и сохранения новой модели.



Краткие данные о загруженной модели

Следующим этапом планировалось предоставить терминал пользователю для запросов информации, матриц, коэффициентов модели в целом и отдельных уровней.