ПРЕПОДАВАНИЕ ИТ В РОССИИ XVI
14 – 15 мая 2018 г. Москва

Проектно-ориентированный подход к преподаванию дисциплины «Анализ данных»

А.Л. Калабин, Тверской государственный технический университет

При участии
Туляков Андрея, Кузнецова Дмитрия и Сябро Николая


Тверь

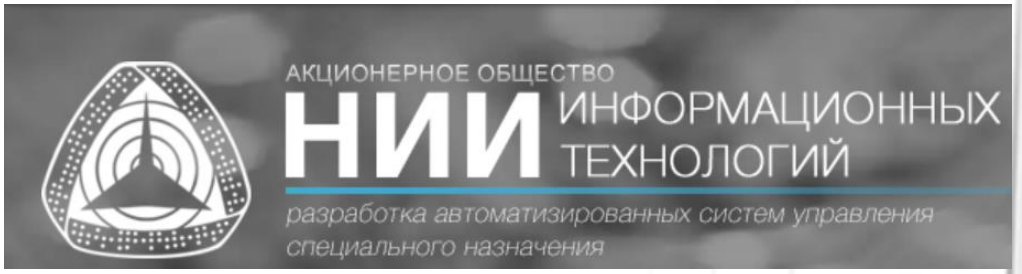

ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

КОГО ГОТОВИМ: ИТ студенты ФИТ ТвТГУ

Форма обучения	Направление	Всего
Очная	Прикладная информатика (по областям)	62
Очная	Информатика и вычислительная техника	115
Очная	Программная инженерия	54
Очная	Информационные системы и технологии	65
Всего		296
	Магистратура	78

4 КИТА ВЫСШЕЙ ШКОЛЫ

Актуальность работы



Актуальность работы

Главное – соответствие актуальным требованиям индустрии.

Это основной критерий качества образования.

При обсуждении с работодателями результатов начального опыта работы молодых специалистов, что студенты неплохо умеют решать отдельные этапы проекта и малые задачи, но весьма слабы в системе целого проекта.

Третий приход ИИ в ИТ.

Способствует трудоустройству выпускников (hh.ru).

Учебный проект

что

и

как

делают студенты.

Учебный проект

как

делают студенты

Группа магистров;
Архитектор – интегратор проекта;
Персональные задания и ответственность;
Программная реализация отдельных методов с поэтапным выводом промежуточных результатов;
Презентация теории;
Демонстрация тестирования;
Анализ реальных данных с извлечением знаний.

Этап 1

Общее ознакомление с предметом и освоение известных средств.
Освоение возможных вариантов средств проведения интеллектуального анализа данных средствами Microsoft SQL Server с использование надстроек для пакета Microsoft Office.
Набор данных для анализа, доступный для скачивания по адресу:
http://russiandmaddins.codeplex.com/

Другое средство - интерпретатор языка R и среда R-Studio.

Выполняются лабораторные работы для изучения методов, примерного объема предложенного в учебном пособие.

* Нестеров С. А. Базы данных. Интеллектуальный анализ данных: учеб.пособие / С. А. Нестеров - СПб.: Изд-во Политехи, ун-та, 2011. - 272 с.

Этап 1. Реальные данные. Магазин

1. Понимание и формулировка задачи анализа.
2. Подготовка данных для автоматизированного анализа (препроцессинг).
3. Применение методов Data Mining и построение моделей.
4. Интерпретация моделей человеком.

Расставание с иллюзией, что возможно обнаружения скрытых знаний, применяя методы автоматического анализа.

Этап 1. Магазин

Количество покупок – чеков 15469

Всего в покупках участвовало 4068 различных наименований.

Было продано наименований в количестве 66535 (единиц хранения) на общую сумму 3476326,87 руб.

Максимально покупаемыми товарами **по количеству** были выявлены:

1) Картофель 1кг - 1788,95
2) Арбуз 1кг - 1622,77
3) Сахарный песок 900г - 1362,00

Максимально покупаемыми товарами **по стоимости** оказались:

1) Сахарный песок 900г - 61232,80
2) Водка Беленькая люкс 40% 0,5л - 53849,98
3) Водка Старая марка классическая СТМ 40% 0,5л - 33233,40
4) Картофель 1кг - 31865,60

Этап 1. Магазин

Определены товары, которые **покупаются вместе.**

1) Хлеб Дарница Волжская горяч. 0,7кг В.Пекарь, Батон Нарезной горячий 0,4кг В.Пекарь - 177 раз
2) Огурцы тепличные 1кг, Томаты тепличные 1кг – 175 раз
3) Лук Репка 1кг, Картофель 1кг – 111 раз
4) Картофель 1кг, Батон Нарезной горячий 0,4кг В.Пекарь – 97 раз

Составлены рекомендации по **связанным покупкам**:

1) Хлеб Дарница Волжская горяч. 0,7кг В.Пекарь =>Батон Нарезной горячий 0,4кг В.Пекарь
2) Огурцы тепличные 1кг => Томаты тепличные 1кг
3) Петрушка 1кг => Укроп 1кг
4) Лапша Ролтон Говядина 60г пакет => Лапша Ролтон с куриным вкусом 60г

Этап 2

Реализуются программно алгоритмы поиска ассоциативных правил, кластеризации и классификации на Python, библиотеку Scikit-Learn.

Различные методы распределяются персонально и собираются в один проект, хранимый на https://github.com.

Проекты расположены по адресу: https://github.com/IljaNovo/ProjectOfDataMining/tree/newInterface_2.0

2018
https://github.com/klik0121/SklearnWrapper.

Этап 2

Задача	Алгоритм
Классификация	
	Support vector machines
	Stochastic gradient descent
	K-Nearest Neighbors
	Gaussian Processes
	Decision trees
	Naïve bayes classification
	C 4.5
Кластеризация	
	K-means
	Affinity Propagation
	Birch
	Mean Shift
	Hierarchical clustering
	DBSCAN
	Clustering performance evaluation
Поиск ассоциативных правил	
	AprioriTID

Этап 3

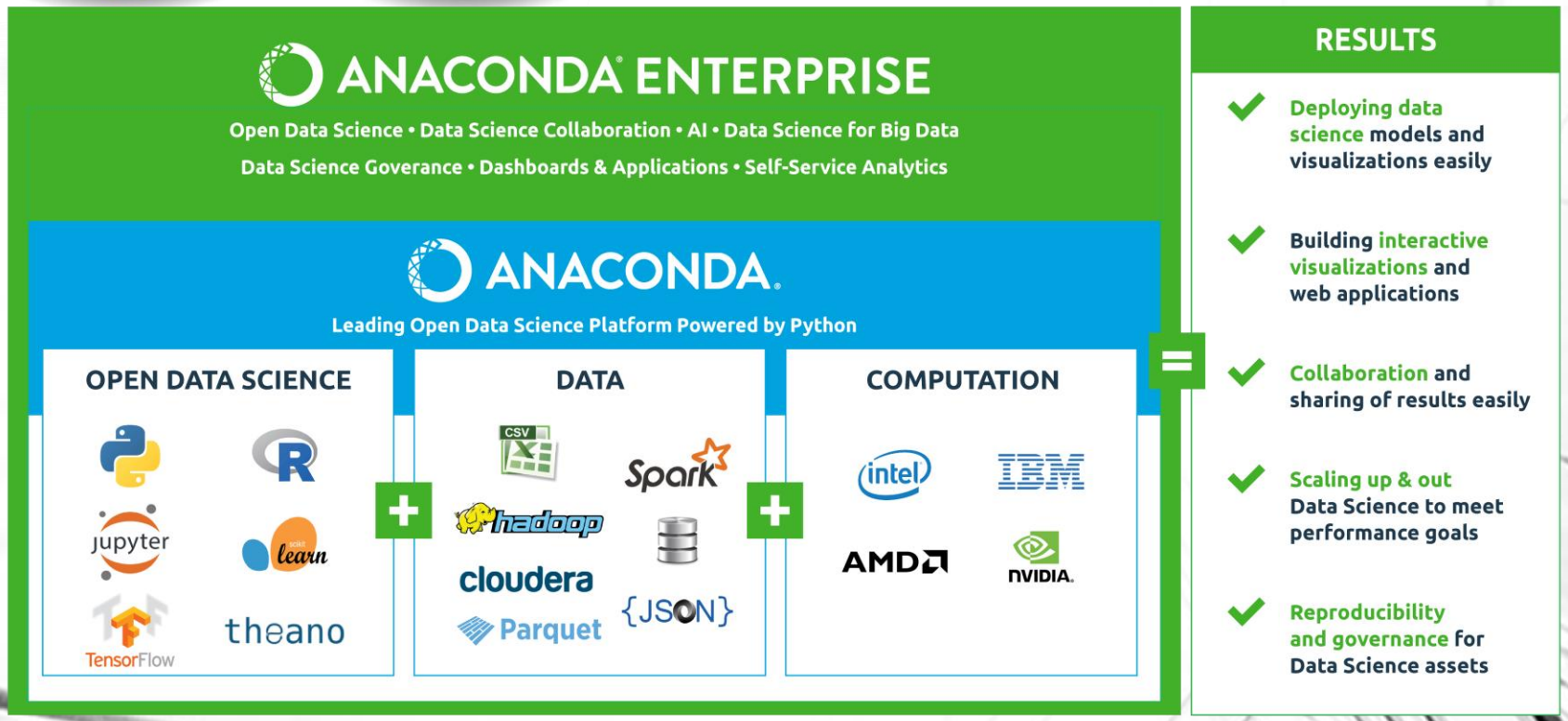
Изучение и анализ неструктурированных текстов методами Text Mining в приложении для технических текстов.

Программное обеспечение разработано с помощью языка Python и библиотек pyMorphy2, Qt-5, Scikit-Learn, NumPy, Pandas и представляет собой desktop-приложение.

Проект расположен в веб-сервисе для хостинга IT-проектов и их совместной разработке по адресу: https://github.com/mhyhre/TextStageProcessor. Предлагаем использовать наше приложение в Вашей работе.

Средства разработки

- PYTHON ANACONDA
- QT -5 NumPy
- Pandas Scikit-Learn,
- pyMorphy2



Основная функция TextStageProcessor

Реализация методов Text Mining **с возможностью поэтапной**

- обработки текстовых документов (вывод промежуточных результатов);
- **настройки параметров методов** анализа.

Основная функция TextStageProcessor

Реализация методов Text Mining **с возможностью поэтапной**

- обработки текстовых документов (вывод промежуточных результатов);
- **настройки параметров методов** анализа.

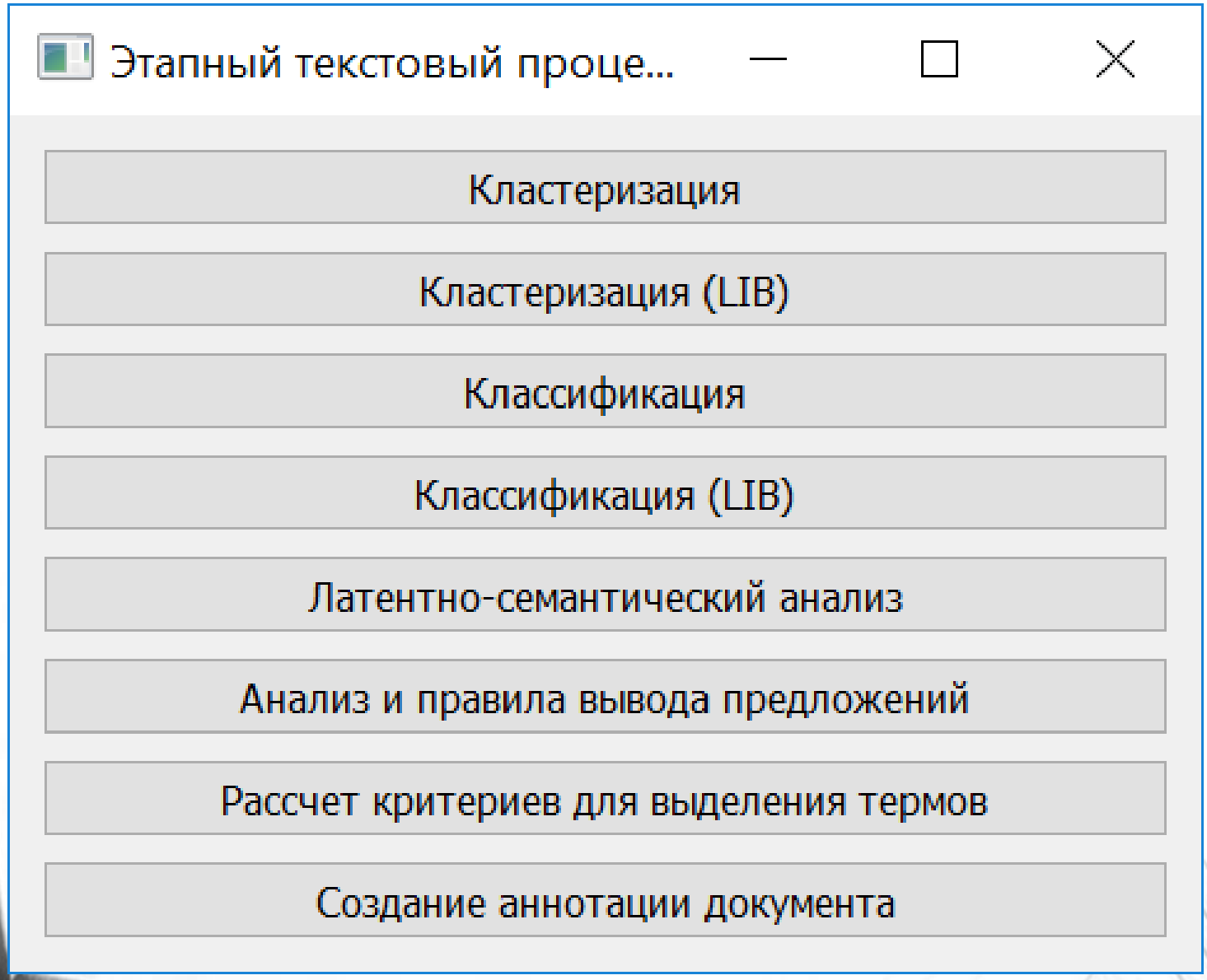
Реализованные методы Text Mining

1. Предварительная обработка текстовых документов
2. Классификация
3. Кластеризация
4. Выделение ключевых понятий и термов
5. Автоматическое аннотирование
6. Латентно-семантический анализ
7. Логический анализ и применение правил вывода
8. Вычисление критериев для выделения термов

Особенности предварительной обработки

- Многоступенчатая настраиваемая фильтрация
- Морфологический анализ (Нормальная форма слова)
- Удаление стоп слов из внешнего списка
- Настраиваемое отсечение по TD-IDF, IDF

Интерфейс

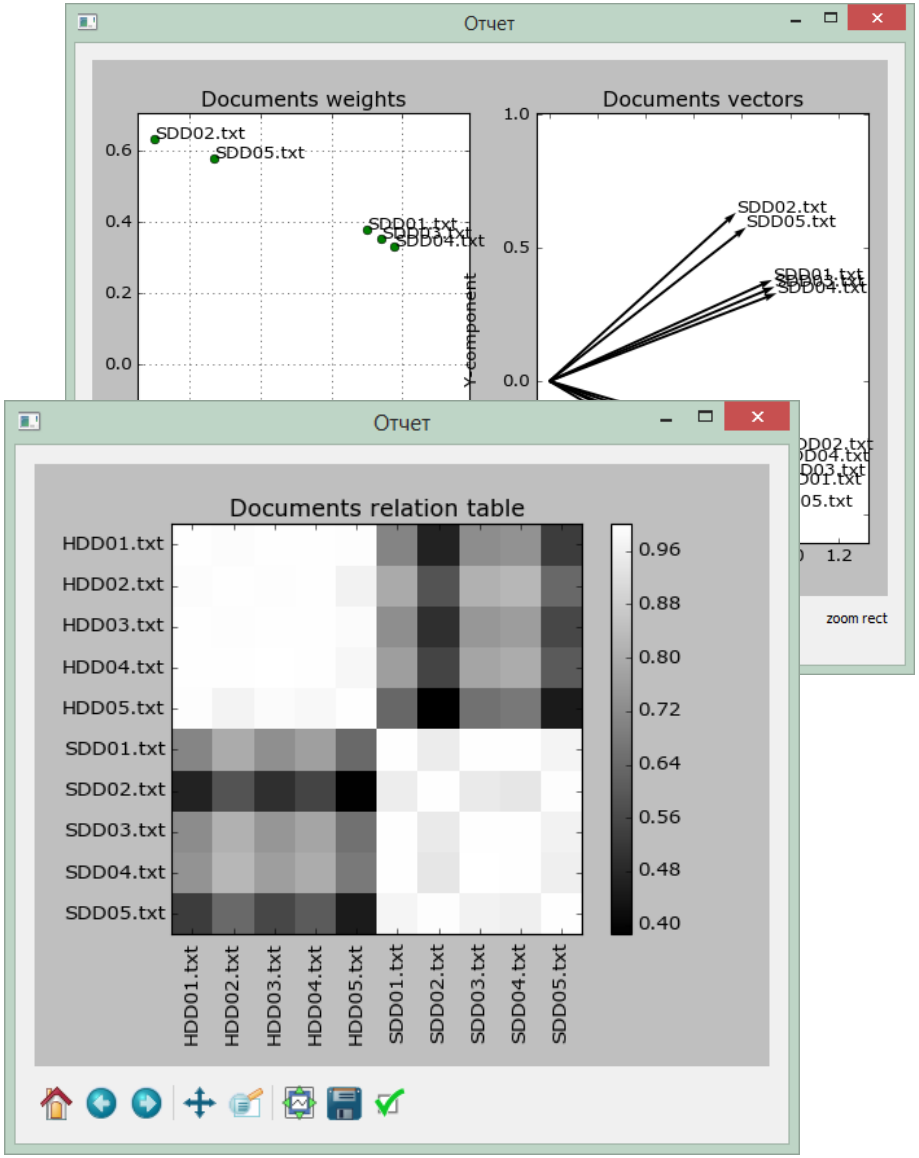

Этапный текстовый проце...
Кластеризация
Кластеризация (LIB)
Классификация
Классификация (LIB)
Латентно-семантический анализ
Анализ и правила вывода предложений
Рассчет критериев для выделения термов
Создание аннотации документа

Латентно-семантический анализ

Позволяет выявить скрытые семантические связи между документами или частями одного документа.

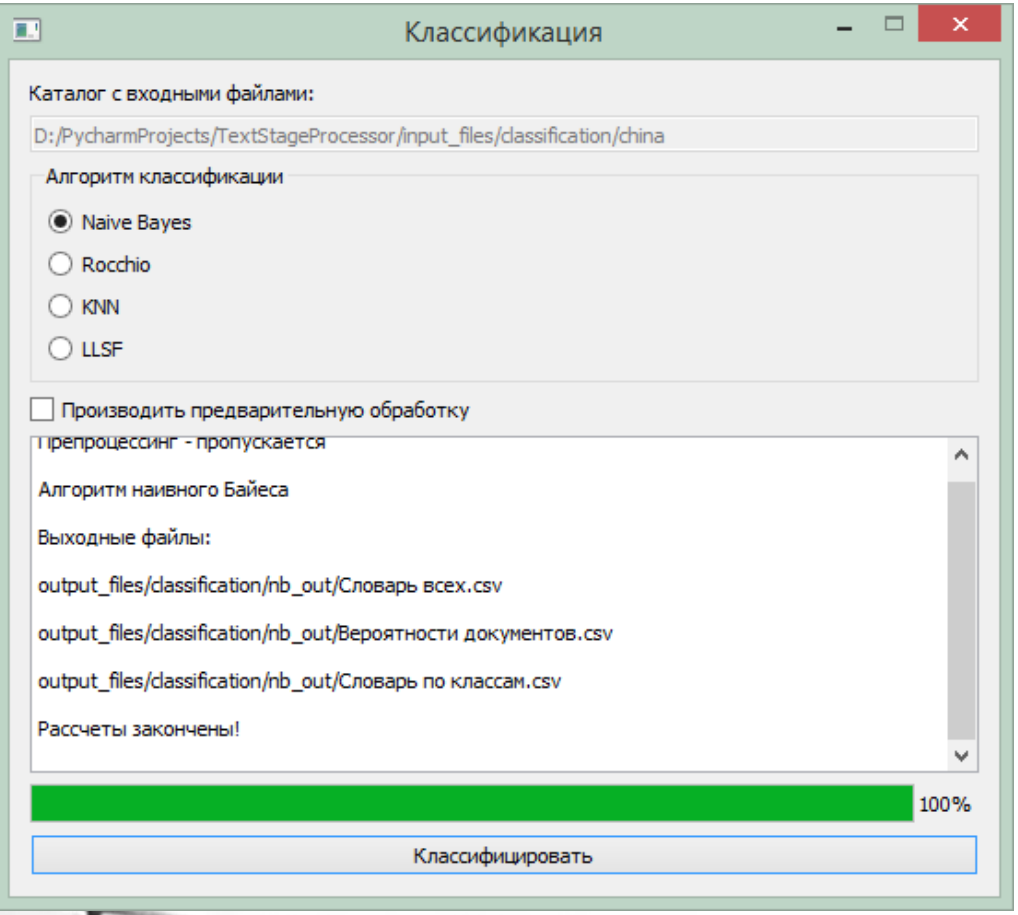
Также позволяет оценить степень схожести тематики документа.

Результаты выводятся в визуальной форме и в файлах формата .CSV.



Классификация – 4 алгоритма

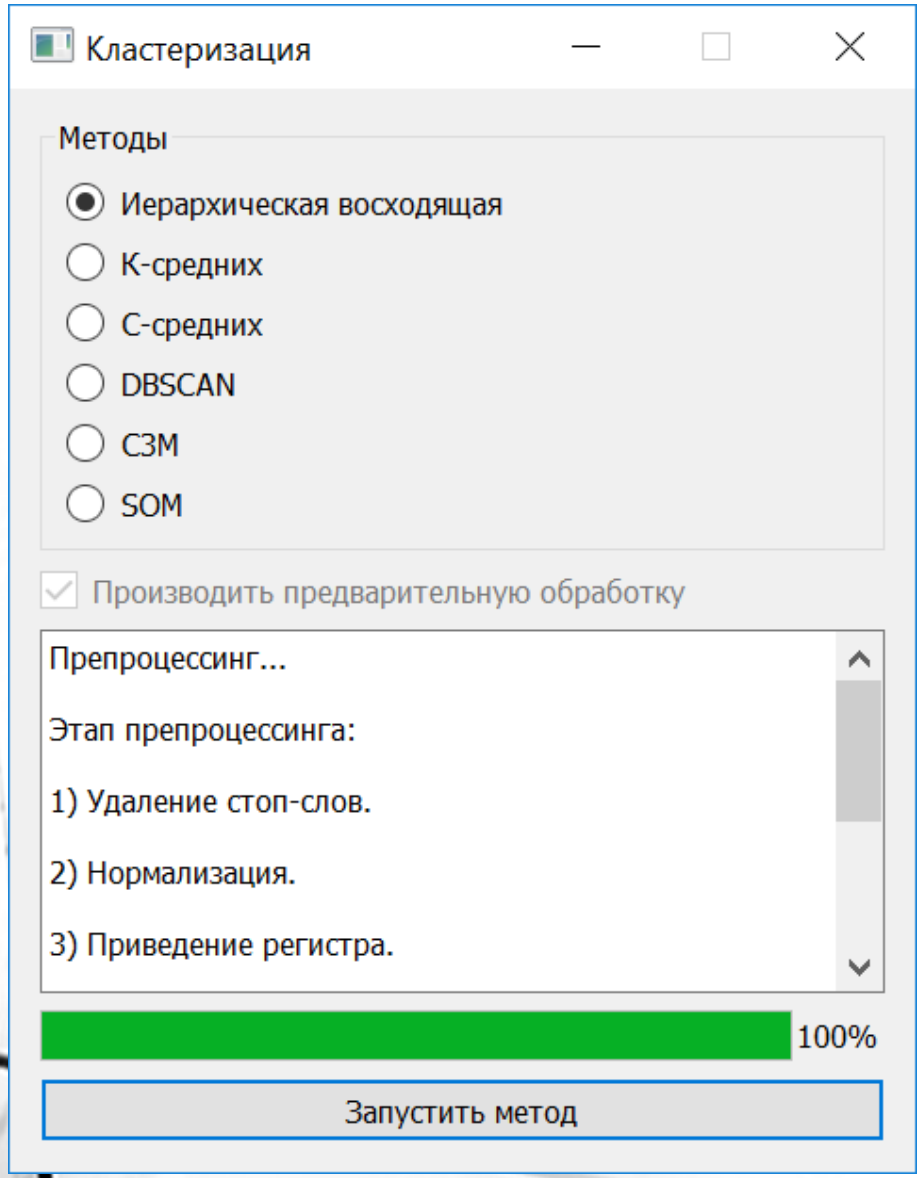
программная реализация Сябро Николай



1) k-Nearest Neighbor
2) Rocchio
3) Naive Bayes
4) Learning Label-Specific Features

Кластеризация 6 алгоритмов

программная реализация Кузнецов Дмитрий



1) Иерархический восходящий
2) K-Means
3) C-Means
4) DBScan
5) СЗМ
6) SOM

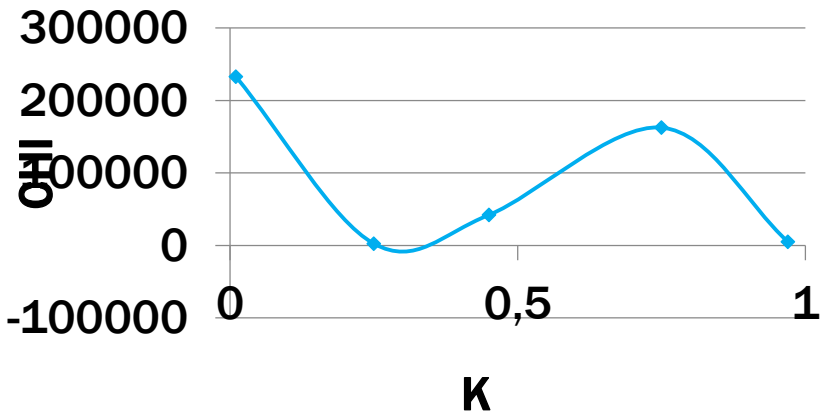
Критерии отбора значимых признаков

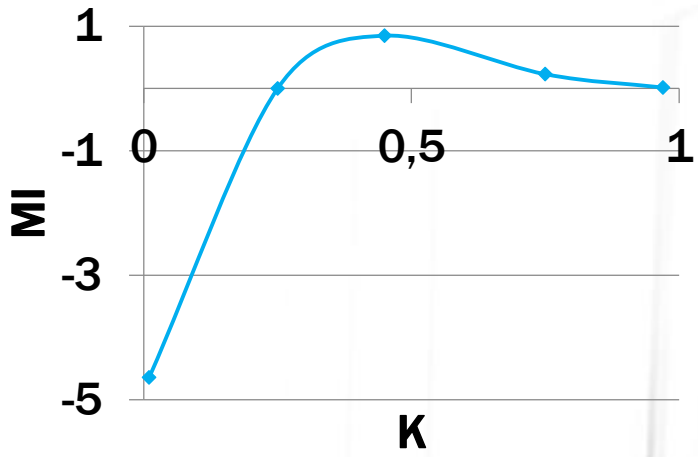
- Взаимная информация: $MI(t_k, c_j) = \log_2 \frac{A*|\Omega|}{(A+C)*(A+B)}$
- Информационная выгода: $IG(t_k, c_j) = \frac{A}{|\Omega|} * \log_2 \frac{A*|\Omega|}{(A+C)*(A+B)} + \frac{C}{|\Omega|} * \log_2 \frac{C*|\Omega|}{(A+C)*(C+D)} + \frac{B}{|\Omega|} * \log_2 \frac{B*|\Omega|}{(B+D)*(A+B)} + \frac{D}{|\Omega|} * \log_2 \frac{D*|\Omega|}{(D+C)*(D+B)}$
- Критерий ХИ-квадрат: $CHI(t_k, c_j) = \frac{|\Omega|*(A*D-C*B)^2}{(A+C)*(B+D)*(A+B)*(C+D)}$

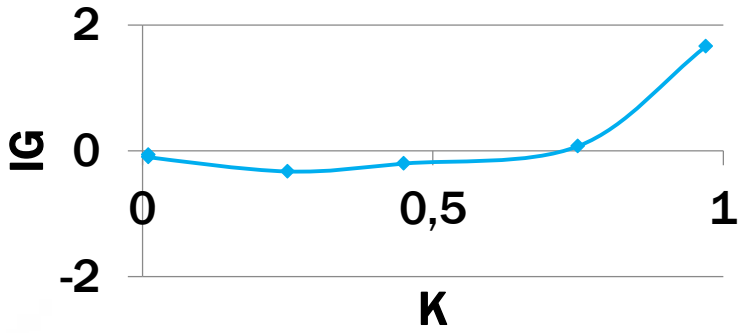
$|\Omega|$ - обучающее множество документов; A - обучающее множество документов принадлежит категории и содержит термин; B - обучающее множество документов не принадлежит категории и содержит термин; C - обучающее множество документов принадлежит категории и не содержит термин; D - обучающее множество документов не принадлежит категории и не содержит термин.

Используемые источники: Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с. ISBN 978-5-94506-294-8

Результаты отбора значимых терминов







K=A/N, A-документ принадлежит категории и содержит термин, N-обучающее множество документов

Отсутствие монотонности свидетельствует о неработоспособности этих критериев.

Критерии качества

$$P = \frac{a}{(a+b)} \quad (1)$$

$$R = \frac{a}{(a+c)} \quad (2)$$

$$E = \frac{b+c}{a+b+c+d} \quad (3)$$

$$A = \frac{a+d}{a+b+c+d} = 1 - E \quad (4)$$

$$I = P + R + A$$

P — точность, R — полнота, E — ошибка, A — правильность классификатора.
I – интегральный критерий,

Используемые источники: Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с. ISBN 978-5-94506-294-8

Кластеризация 4 алгоритма

without preprocessing							
	HDD		CPU		Error %	Error%	время
	True	False	True	False	HDD	CPU	увеличилось в
Real	16	0	15	0			with preprocessing
Hierarchical	16	0	20	0	0	0	8,1
KMiddle	**13**	**9**	**3**	**4**	**69**	**300**	6,8
SMiddle	12	2	12	5	17	17	7,3

K-Means – не работает

C-Means – работает плохо

Результаты оценки методов классификации

	NB	kNN	SVM_liner	SVM_rbf	LLSF	Rocchio
P	1	1	1	1	1	1
R	1	1	1	1	0,9	0,9
E	0,1	0,1	0,1	0,1	0,1	0,1
A	0,9	0,9	0,9	0,9	0,9	0,9
P+R+A	2	2,9	2,9	2,9	2,8	2,8

Экспериментальная оценки результатов

- шести методов **классификации** (NB — наивный байесовский метод классификации, kNN — метод k-ближайшего соседа, LLSF — метод наименьших квадратов, Rocchio — метод Роккио, SVM_linear — линейный метод опорных векторов, SVM_rbf — нейронный метод опорных векторов) полнотекстовых документов.
- Здесь « > » означает « эффективнее. Эксперименты показали, что
- {NB, kNN, SVM_linear, SVM_rbf} > {LLSF, Rocchio}
- девяти методов **кластеризации** (SMiddle — нечеткий метод с-средних, Spectral — спектральный метод, Ward — метод Варда, Mean shift — сдвиг среднего и развитие, Affinity — метод распространения близости, KMiddle — четкий метод k-средних, Birch — метод использующий иерархию, DBSCAN — плотностный метод) полнотекстовых документов.
- Иерархия ≫ SMiddle > {Spectral, Ward, Mean_shift} > Affinity > {KMiddle, Birch} ≫ DBSCAN

Структура системы интеллектуального анализа Web-сайтов

АИС содержит оболочку, которая по запросу пользователя добывает текстовые данные из сети, проводит их структуризацию, после чего отправлять данные в систему Text Mining, для интеллектуального анализа.

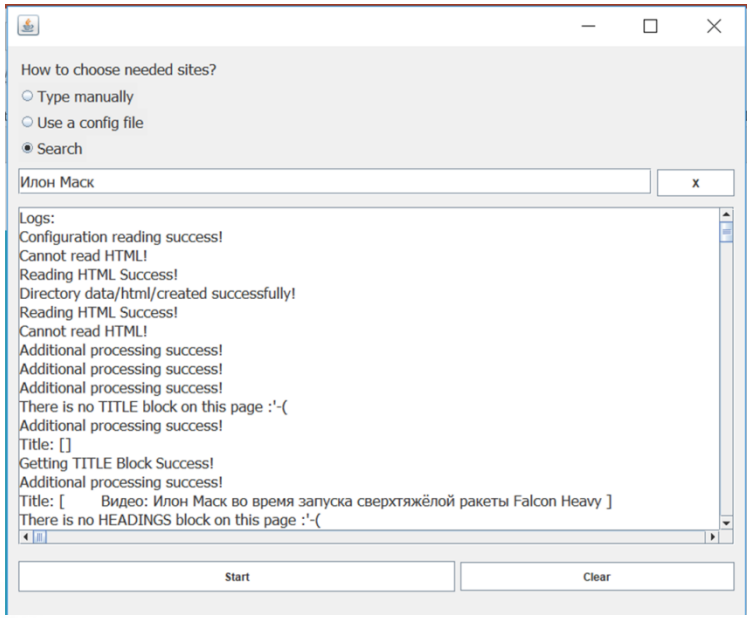
- Модуль поиска – по запросу пользователя ищет сайты, подходящие по запросу и извлекать из них HTML код основного текста.

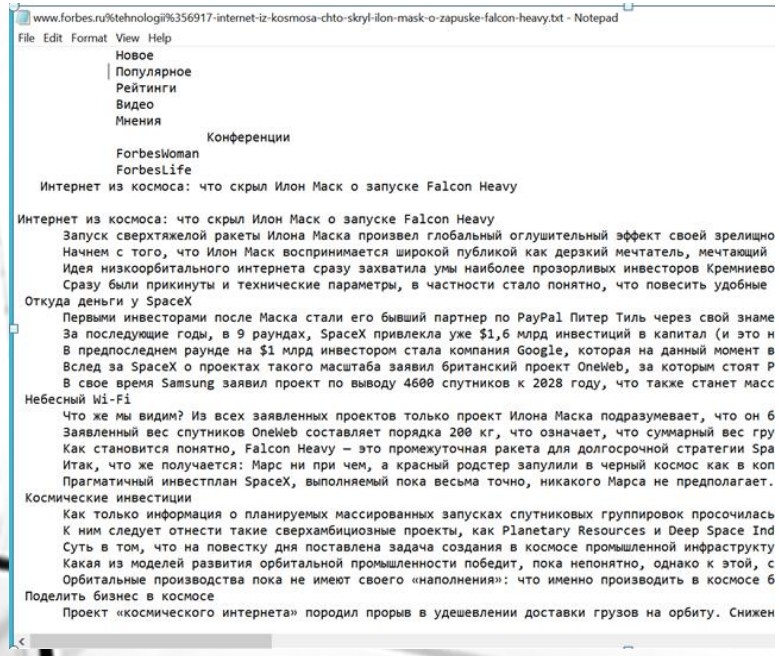
- Модуль обработки – парсит HTML код по тэгам и обрабатывает текст для передачи этого текст в систему Text Mining в более “чистом” виде.

- Модуль передачи –запускает Интеллектуальный Анализ Текстов(ИАТ) и передает туда необходимые данные

- Модуль вывода – данный модуль принимает данные от системы ИАТ и выводить их для пользователя.

Система интеллектуального анализа Web-сайтов. Демо.





www.novayagazeta.ru%articles%2018%02%09%75443-kosmos-kak-sredstvo-ot-skuki.txt
www.forbes.ru%tehnologii%356917-internet-iz-kosmosa-chto-skryl-ilon-mask-o-zapuske-falcon-heavy.txt
www.bbc.com%russian%features-43016528.txt
tass.ru%kosmos%4935688.txt
svpressa.ru%persons%ilon-mask%.txt
rusvesna.su%news%1518633323.txt
russian.rt.com%opinion%478500-sokolov-ilon-mask-i-sayrus-fild.txt
ru.wikipedia.org%wiki%%D0%9C%D0%B0%D1%81%D0%BA_%D0%98%D0%BB%D0%BE%D0%BD.txt
ru.euronews.com%2018%02%07%elon-musk-profile.txt
ria.ru%lenta%person_EHlon_Mask_predprinimatel%.txt
rb.ru%tag%musk%.txt
news.rambler.ru%person%mask-ilon%.txt
meduza.io%video%2018%02%07%ilon-mask-zapustil-tesla-v-kosmos-na-samoy-bolshoy-rakete-kak-eto-
life.ru%t%%D0%BD%D0%BE%D0%B2%D0%BE%D1%81%D1%82%D0%B8%1089177%ilon_mask_nazval_p
inosmi.ru%science%20180211%241424409.html.txt
hi-news.ru%tag%elon-mask.txt
file.liga.net%person%116615-mask-ilon.html.txt
daily.afisha.ru%person%ilon-mask%.txt
daily.afisha.ru%infoporn%8176-a-kak-tebe-takoe-ilon-mask-russkiy-otvet-izobretatelyu-spacex%.txt

Полезную информацию можно извлечь, примерно, из 50% найденных текстов

Оценка связанность текста по кластеризации его частей

- Каждую из часть текста представляем в виде отдельного текстового файла.
- Находим кластеры этих отдельных частей.

гипотеза

- если текст связан то будут разбиты кластеры состоящий из последовательных частей исследуемого текста.

Оценка связанность текста по кластеризации его частей

- Пример, допустим если у нас есть пять частей, то

в первый класс должно входить,

(1 2 3) или (1 2) части этого текста и

во второй кластер

(4 5) или (3 4 5) части этого текста.

(1 2 3 4 5) = (1 2 3) + (4 5) = (1 2) + (3 4 5)

Оценка связанность текста по кластеризации его частей

		Сон смешного человека			
Метод	результат				
	Кластер 0:	Кластер 1:	Кластер 2:	Кластер 3:	Кластер 4:
DBSCAN	Сон с ч_11.txt	Сон с ч_13.txt	Сон с ч_14.txt		
Hierarchical clustering	Сон с ч_12.txt		Сон с ч_15.txt		
C2ICM	Сон с ч_11.txt	Сон с ч_12.txt	Сон с ч_13.txt	Сон с ч_14.txt	Сон с ч_15.txt
SMiddle	Сон с ч_11.txt	Сон с ч_13.txt			
Spectral clustering	Сон с ч_12.txt	Сон с ч_14.txt			
Ward hierarchical clustering	Сон с ч_15.txt				
K-Means					
Birch					
Mean-shift	Сон с ч_11.txt	Сон с ч_13.txt	Сон с ч_14.txt		
Affinity propagation	Сон с ч_12.txt				
	Сон с ч_15.txt				
KMiddle	Сон с ч_11.txt	Сон с ч_13.txt			
	Сон с ч_12.txt				
	Сон с ч_14.txt				
	Сон с ч_15.txt				

Заключение

- Работа в команде
- Развитие коммутативности
- Навыки управления проектом
- Доведение проекта до работоспособности
- Изучение методов Data mining

Заключение

Работа в команде	20%
Развитие коммутативности	40%
Навыки управления проектом	15%
Доведение проекта до работоспособности	80%
Изучение методов Data mining	90%

Спасибо за внимание!