Московский государственный технический университет им. Н.Э. Баумана
кафедра "Системы обработки информации и управления"

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ИССЛЕДОВАНИЯ ЯЗЫКА ПЕЧАТНЫХ ИСТОЧНИКОВ XVIII – НАЧ. XIX ВВ.

Расчетно-пояснительная записка
курсовой работы по дисциплине
"Семиотика информационных технологий"
студент группы ИУ5-95

Алексеев Владимир Николаевич

Шифр:95_01

Преподаватель: к.т.н., доц. А.Ю.Филиппович

Москва, 2010 г.

Оглавление

Введение

Цель работы

Целью предлагаемых учебно-практических занятий является изучение современных информационных технологий и инновационных разработок для сохранения исторических и культурных ценностей России на примере задачи исследования языка печатных источников XVIII – нач. XIX вв. Выполнение заданий позволяет на практике изучить особенности рассматриваемых текстов, проанализировать факторы, влияющие на эффективность их распознавания с помощью современных OCR-систем.

Задачи курсовой работы:

1. Подготовка материалов для выполнения задания: установка ПО, анализ и фиксация параметров ПО. Анализ характеристик источника.
2. Ввод и распознавание текстового фрагмента. Предварительная оценка эффективности работы OCR-системы.
3. Формирование шрифтовых эталонов фрагментов, используя технологию обучения. Оценка эффективности использования технологии распознавания, включающей шрифтовые эталоны.
4. Оценка эффективности использования технологии распознавания, включающей дополнительный словарь языка текста.
5. Корректура текста и анализ лексики и типов ошибок.
6. Квантитативные исследования текста, формирование словников фрагментов, построение функций распределения частот. Выводы. Формирование отчета.

Основная часть

Обработка исходного документа при помощи FineReader

Описание и анализ источника

Исходным файлом для последующей обработки в программе FineReader является произведение Николая Степановича Ильинского «Историческое описание города Пскова и его древних пригородов с самаго их основания»

Краткая биографическая справка об авторе произведения

Николай Степанович Ильинский (1760-1846) - русский писатель и историк, При Александре I состоял в комиссии для составления законов и юрисконсультом министерства юстиции. В 1781 году был произведен в губернские секретари и направлен в Псковское наместничество. В Пскове он прослужил, продвигаясь по служебной лестнице, семнадцать лет. Будучи человеком пытливым и любознательным, он не смог не увлечься городом, в который его занесла судьба. Получив возможность знакомиться с историческими документами, Ильинский кропотливо собирает сведения, касающиеся псковской истории. Он изучает рукописные и печатные материалы, сверяя их с летописями, сочинениями В.Н. Татищева, Ф.И. фон Страленберга, М.М. Щербатова, И.Н. Болтина, А.П. Сумарокова, Екатерины II и др. авт.

Настоящим раритетом является изданное 220 лет назад и отнесенное к книжным памятникам федерального значения «Историческое описание города Пскова и его древних пригородов с самого их основания, заключающее в себе многие достойные любопытства происходимости, составленное из древних летописцев, надписей, записок и Российской истории Николаем Ильинским».

Другие сочинения Ильинского: «В память славному мужу нижегородскому купцу Козьме Минину» — стихотворение, которое Ильинский поднёс вместе с прошением Екатерине II, в котором просил об установке памятника знаменитому патриоту; «Описание жизни и пр. купца Козьмы Минина» (СПб. 1799), «Житие Франца Яковлевича Лефорта и описание жизни нижегородского купца Козьмы Минина» (СПб. 1799; первая часть принадлежит И. Виноградову); «Мысль о человеке», с стихотворным посвящением М. А. Ладыженскому; «Изображение человека», в стихах, «На взятие Очакова» (СПб. 1790) и «Разные стихотворения», сочиненные Ильинским во Пскове.

Анализ палеографических характеристик источника

Первое издание датируется 1795 годом, выпущенное в Санкт-Петербурге в типографиях Б.Л. Гека и Ф. Мейера. Также было издание в тот же период в в типографии при губернском правлении Нижнего Новгорода.

Состоит из: Ч. 1. - В Санктпетербурге : В типографии Б.Л. Гека, 1790. - 67, [4] с.
Ч. 2. - В Санктпетербурге : В типографии Б.Л. Гека, 1791. - 71, [4] с.
Ч. 3. - В Санктпетербурге, 1793. - 52 с.
Ч. 4. - В Нижнем Нове-Граде, 1794. - 64 с.
Ч. 5. - В Санктпетербурге : В типографии Ф. Мейера, 1795. - 62 с.
Ч. 6. - В Санктпетербурге : В типографии Ф. Мейера, 1795. - 60 с.

Особенностью данного источника является тот факт, что это первое крупное исследование по истории г. Пскова. Издание обладает статусом книжного памятника федерального уровня. В данном старопечатном издании используется гражданский шрифт XVIII века.

Анализ бумаги

В конце XVIII века в России насчитывалось уже свыше 60 бумажных мануфактур, расположенных в двадцати четырех губерниях. Растущий спрос на бумагу, ее недостаточное количество и дороговизна способствовали поискам новых видов сырья. На рубеже XVIII-XIX веков в бумажном производстве начинают использовать отходы текстильного и канатного производства, асбест, осоку, камыш, водоросли, солому и пр. Происходят изменения и в цвете бумаги. До середины XVIII века оттенок бумажного листа зависел в основном от цвета исходного сырья. Применение в бумажном производстве красителей во второй половине столетия позволило вырабатывать бумагу различных цветов и оттенков.

Можно предположить, что из-за широкого распространения железистых чернил в XVIII веке, они были использованы и в данном издании. Важно отметить, то кроме выцветания чернила оставляли след на обратной стороне листа даже на плотной бумаге. Этот недостаток визуально не портит старинные письма и документы, а наоборот придает оттенок времени.

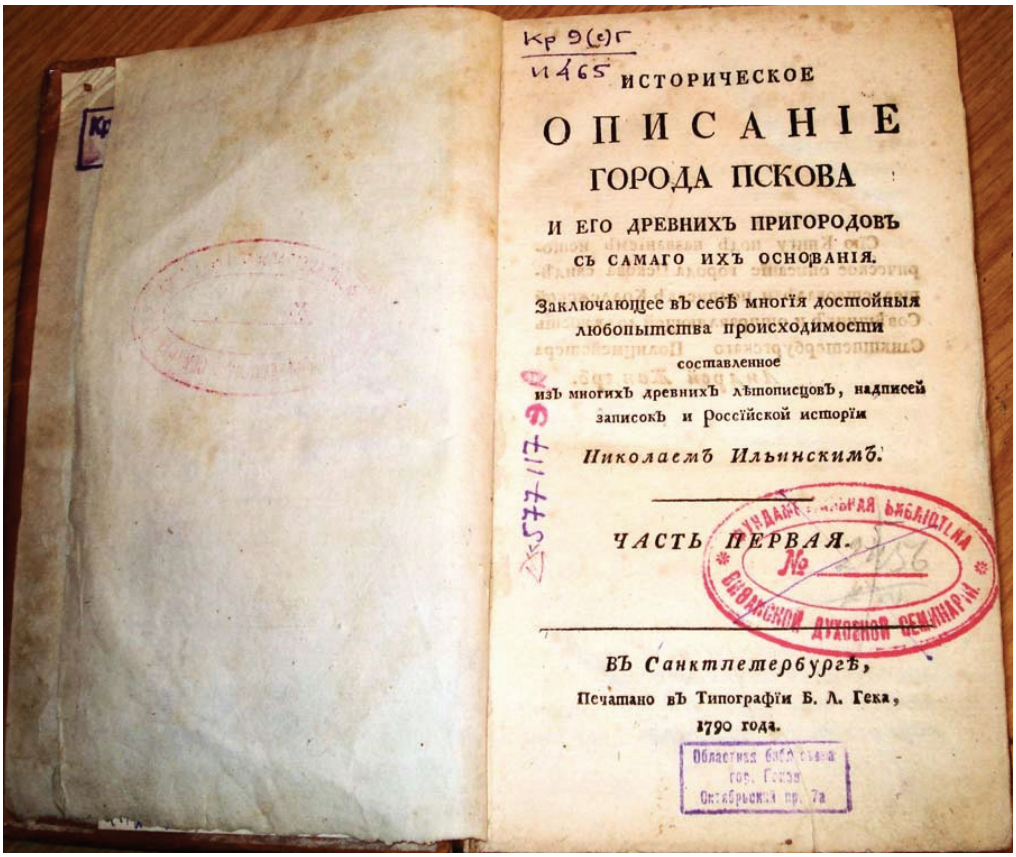


Рисунок 1. Фотография оригинала источника

Ввод фрагмента источника

Для обработки взяты главы 1-4 (25 страниц)

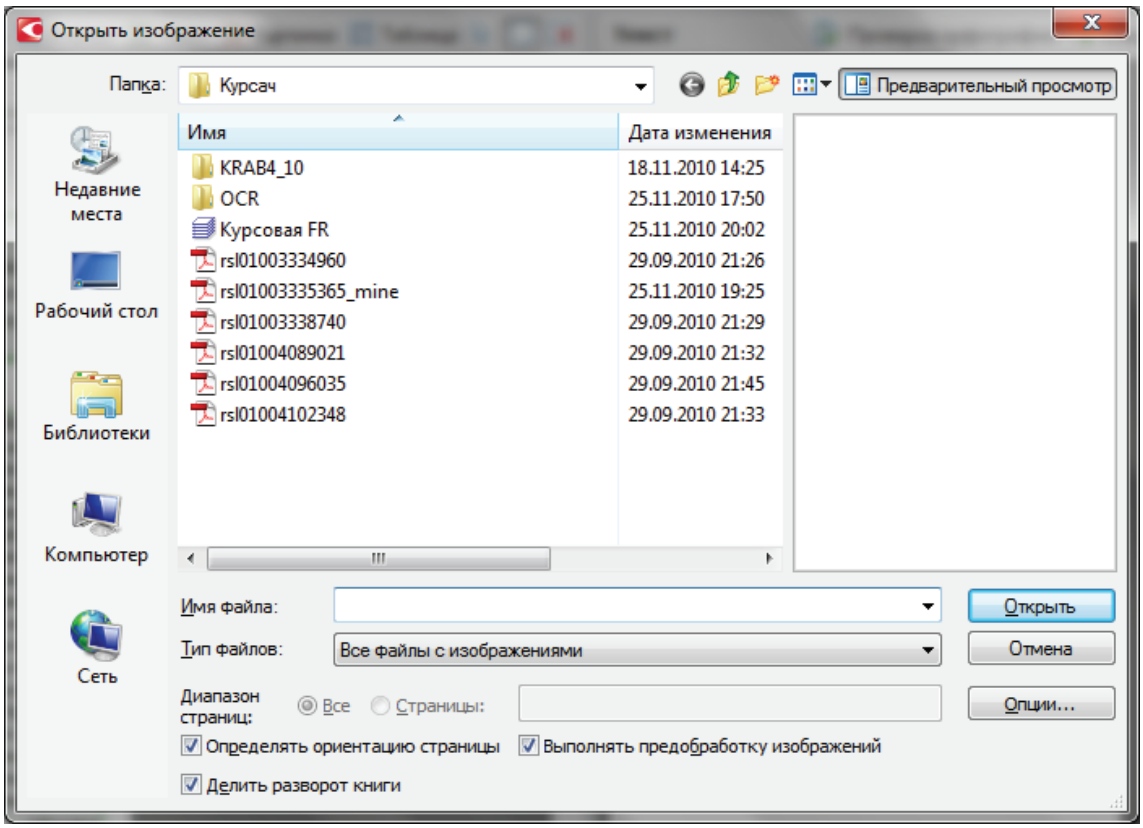


Рисунок 2. Загрузка файлов в систему

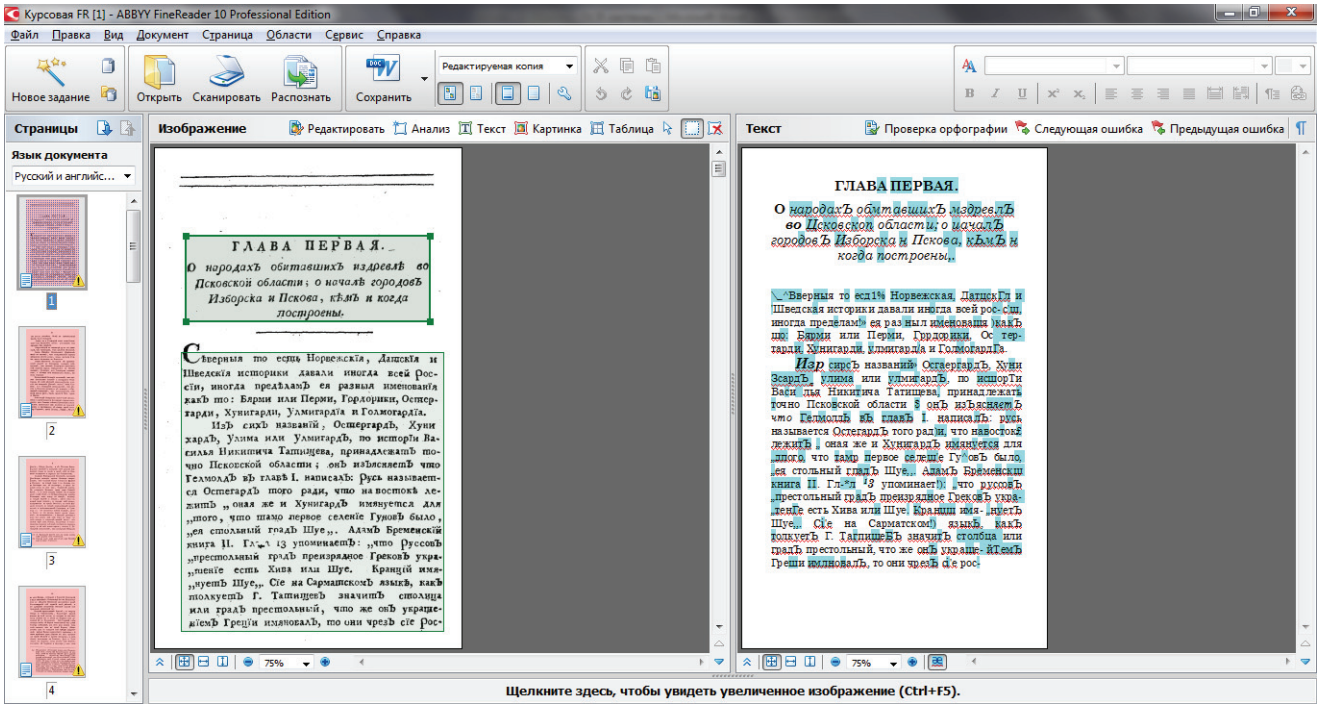


Рисунок 3. Загруженный файл в FineReader

Алгоритм ввода источника в систему

Рисунок 4. Алгоритм загрузки источника в систему

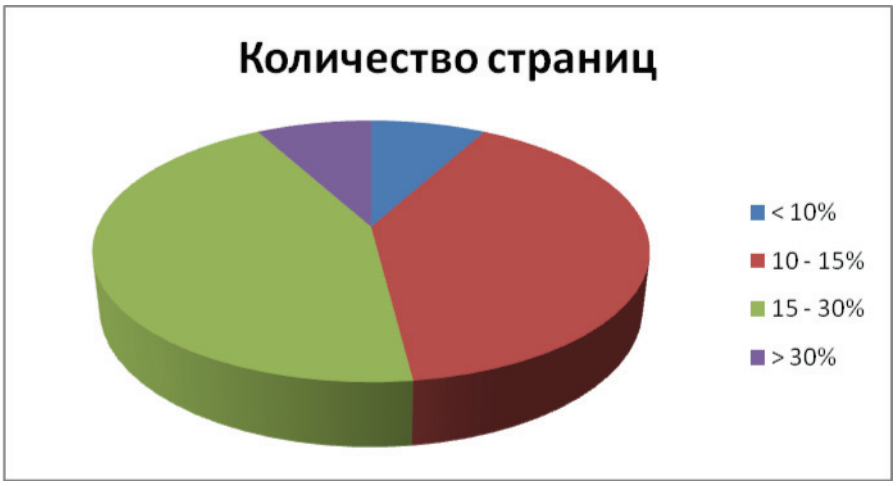
Анализ качества распознавания

Таблица 1. Распознавание без обучения

Стр	Количество символов	Количество неуверенно распознанных символов	Количество неуверенно распознанных символов (%)	Точность распознавания (%)
1	1013	310	31	69
2	1277	418	33	67
3	1325	255	19	81
4	1402	298	21	79
5	1435	318	22	78
6	1410	191	14	86
7	1370	263	19	81
8	1406	427	30	70
9	1399	349	25	75
10	1342	285	21	79
11	1487	348	23	76
12	399	74	19	81
13	1032	170	16	84
14	1230	189	15	86
15	1130	134	12	88
16	1300	181	14	86
17	1339	190	14	86
18	1345	320	24	76
19	1368	198	14	86
20	1325	142	11	89
21	**1322**	**121**	**9**	**91**
22	**1388**	**145**	**10**	**90**
23	1212	159	13	87
24	1063	142	13	87
25	538	64	12	88

Необходимо отметить, что из 25 страниц только в двух количество неуверенно распознанных символов меньше 10% от общего числа символов на странице, а еще на 10 страницах – меньше 15.

На диаграмме ниже представлено разделение по количеству неуверенно распознанных символов (в %).


Количество страниц
< 10%
10 - 15%
15 - 30%
> 30%


Количество неуверенно распознанных символов на страницах
450
400
350
300
250
200
150
100
50
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Среднее значение количества неуверенно распознанных символов = 227,64

Среднее значение точности распознавания = 81,84

Распознавание с обучением

Для использования режима распознавания с обучением необходимо подключить данную опцию в программе (Рисунок 9)

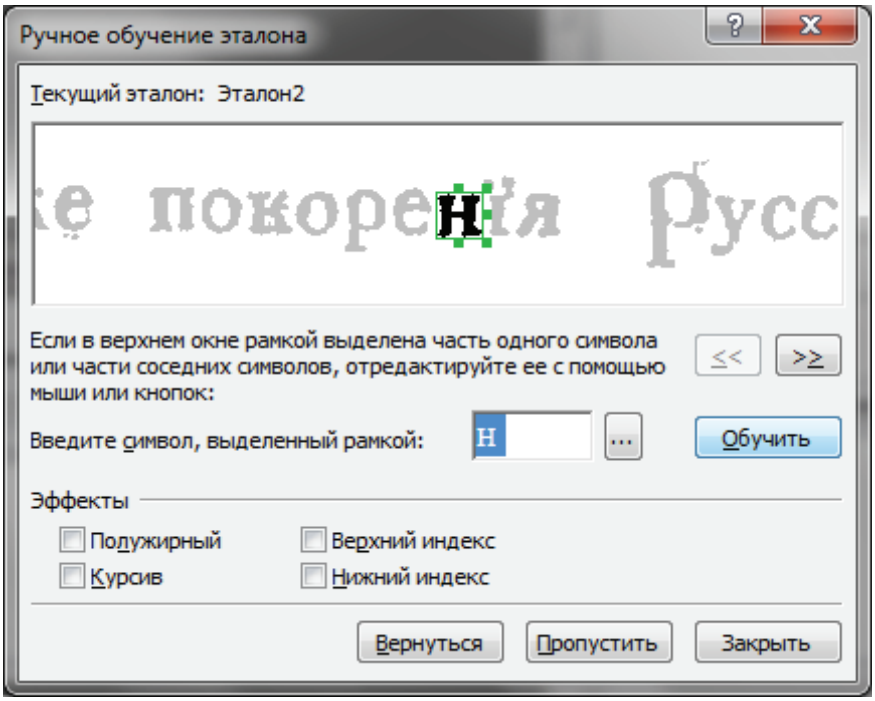


Рисунок 5. Распознавание с обучением

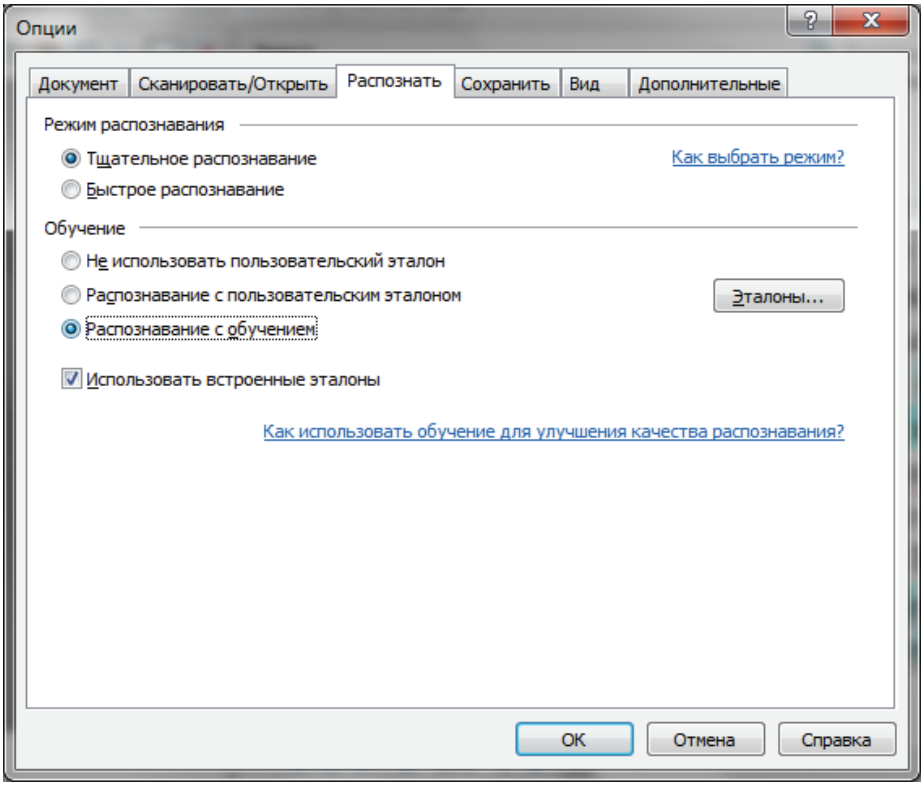


Рисунок 6. Настройка распознавания с обучением

Анализ качества распознавания

Данное распознавание проводится при обучении на 2 страницах документа

Таблица 2. Распознавание с обучением на 2 страницах

Стр	Количество символов	Количество неуверенно распознанных символов	Количество неуверенно распознанных символов (%)	Точность распознавания (%)
1	1019	106	10	90
2	1262	228	18	82
3	1309	195	15	85
4	1395	247	18	82
5	1434	291	20	80
6	**1405**	**132**	**9**	**91**
7	1364	221	16	84
8	1401	376	27	73
9	1395	301	22	78
10	1326	214	16	84
11	1482	316	21	79
12	398	59	15	85