Построение обучаемого алгоритма распознавания научного контента в сети Интернет

Сергей Воронов

Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики

56-я научная конференция МФТИ

30 ноября 2013 г.

Постановка задачи

- **Цель**: создать систему тематического поиска научного контента в Интернете
- **Задача**: разработать обучаемый алгоритм распознавания научных документов
- **Сложность задачи**:
  - Большой размер коллекции
  - Сильно несбалансированная выборка документов (процент научных документов среди всех не превосходит 2%)
  - Изначально нет системы признаков
  - Изначально нет представительной обучающей выборки

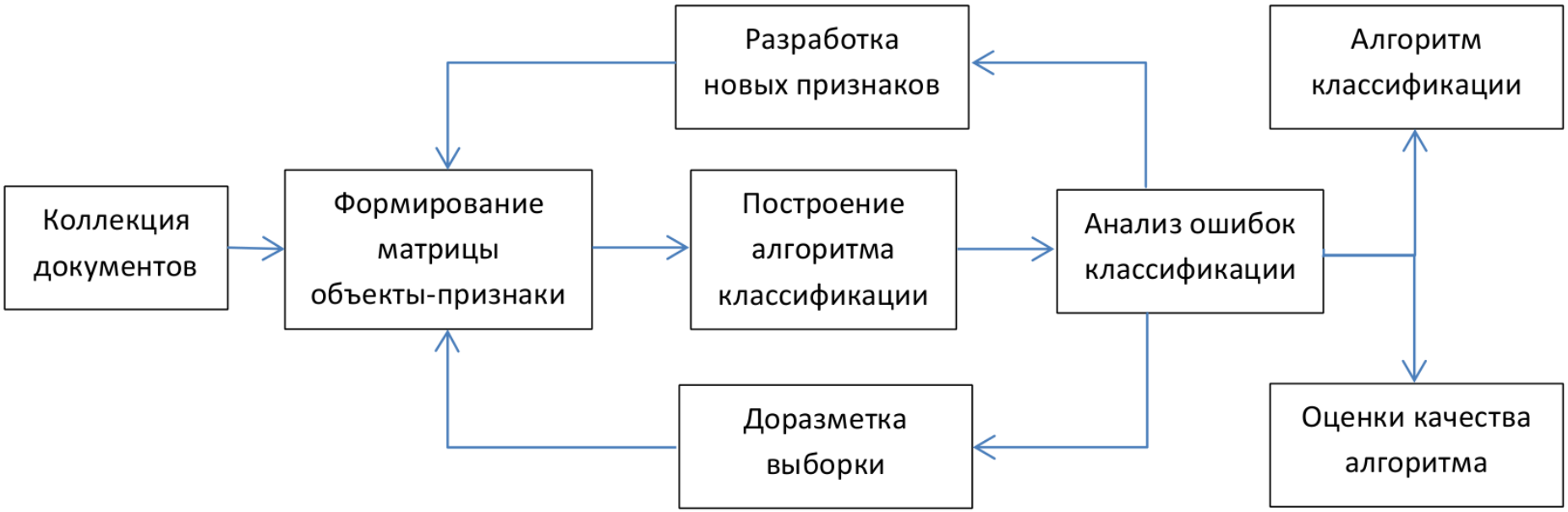
Ненаучный и научный документы

11:04 / ПОНЕДЕЛЬНИК 17 августа 2009 года
Прочитано 2390 раз
Студентов посвятили в чардымовцев
Текст: Альвина ЧАЛОВА , Елизавета ЗИНИНА Фото: Наталия САЛИЙ
В пятницу, 14 августа, в спортивно-оздоровительном лагере СГУ «Чардым» прошло закрытие 4 смены.
По традиции отдыхающих посвятили в «чардымовцев». Для этого студентам необходимо было пройти испытания. Вначале они с закрытыми глазами, держась за руки, ходили по острову, а воспитатели смены рассказывали им чардымские легенды о происхождении лагеря. Студенты услышали историю о красивой девушке и божестве, которые полюбили друг друга. Божество подарило девушке остров «Чардым» и сделало её богиней этого места. Также отдыхающие узнали легенду о чудесном чертополохе, который цветёт разными цветами только один раз в году. После рассказа легенд студентов привели в клуб лагеря. Отдыхающие участвовали в конкурсах на ловкость, скорость и сообразительность. Одним из самых запоминающихся испытаний стало соревнование на лучший танец трёх пар, которые должны были двигаться под музыку с бутылкой, находящейся между ними. Следом за конкурсами отдыхающие стали звать богиню «Чардыма». Её роль исполняла начальник смены Надежда Владимировна Васягина. Она совершила обряд пос вящения в «чардымовцев», подарив каждому по веточке чертополоха. Затем самых активных студентов наградили бейсболками и футболками с эмблемой лагеря. Подарок также сделали студенткам ИФиЖ, обучающимся по специальности «Журналистика», которые каждый день выпускали для отдыхающих стенгазету «В Чардыме».
После награждения студенты произнесли клятву, в которой обещали не употреблять алкоголь, делиться ужином с другом и

УДК 159.9.316.77
ПРЕДСТАВЛЕНИЯ О СОЦИАЛЬНО-ПСИХОЛОГИЧЕСКИХ ХАРАКТЕРИСТИКАХ СУБЪЕКТОВ ЗАТРУДНЕННОГО ОБЩЕНИЯ
Т.В. Бескова
Институт социального образования (филиал) Российского государственного социального университета в г. Саратове
E-mail: beskova-t@yandex.ru
В статье анализируются результаты эмпирического исследования представлений студентов о социально-психологических характеристиках субъектов затрудненного общения на примере представителей профессий социономического типа. Устанавливаются сходства и различия представлений по ряду стимулов (социальный работник, преподаватель, медицинский работник). Выделяются обобщенные портреты трудных партнеров по общению.
Ключевые слова: представления, субъект, общение, трудности, социально-психологические характеристики, ролевые позиции.
The Imaginations about Social-Psychological Descriptions of the Characters of Difficult Communication
T.V. Beskova
The article analyses the results of empirical research on students' imaginations about socialpsychological descriptions of the characters of difficult communication of the representatives of the social-directional occupations, as the example. It underlines the likenesses and the differences of the

Процесс разработки алгоритма классификации

Автоматизация процесса постепенного наращивания обучающей выборки, разработки признаков и улучшения классификатора.



Коллекция документов

- Размер: $\sim 10^6$ документов
- Источник: Страницы с 2000 сайтов российских ВУЗов
- Доля научных документов не превосходит 2%
- Метрика качества: контроль по 10 блокам (10-fold CV)

Оценка качества алгоритма классификации q-fold CV

Выборка случайным образом разбивается на q непересекающихся блоков почти одинаковой длины $k_1, \ldots, k_q$:

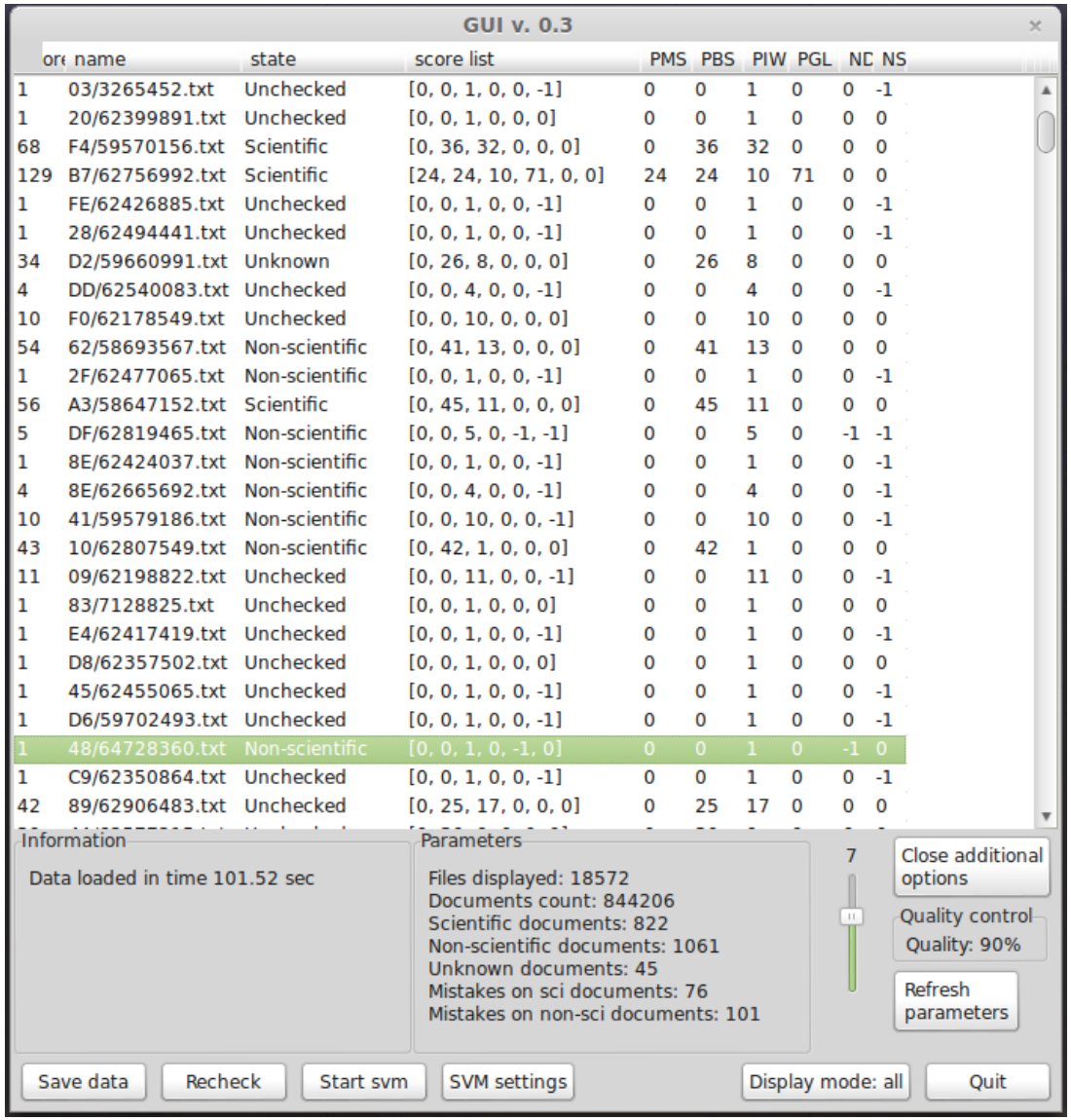
$$X^L = X_1^{k_1} \cup \cdots \cup X_q^{k_q},$$

$$k_1 + \cdots + k_q = L.$$

Каждый блок по очереди становится контрольной подвыборкой, при этом обучение производится по остальным $q - 1$ блокам. Критерий определяется как средняя ошибка на контрольной подвыборке:

$$CV(\mu, X^L) = \frac{1}{q} \sum_{n=1}^{q} Q\big(\mu(X^L \setminus X_n^{k_n}), X_n^{k_n}\big).$$

Программа для формирования обучающей выборки



Базовый необучамый классификатор

Линейный алгоритм классификации:

$$a(x, w) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right),$$

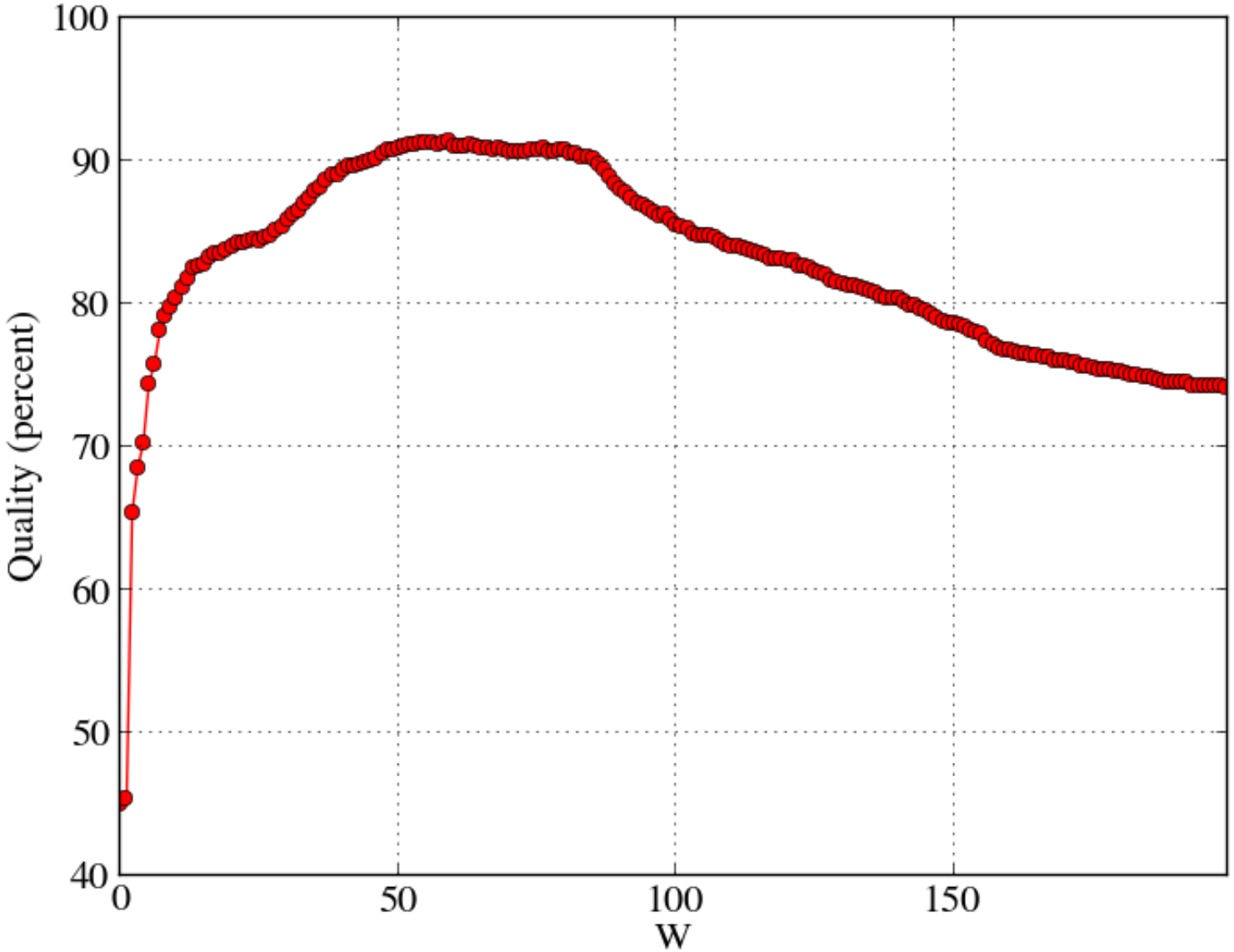
где w_j — вес j-го признака, w_0 — порог принятия решения, $w = (w_1, \ldots, w_n)$ — вектор весов признаков.

В базовой необучаемой версии w_i задаются экспертом.

Начальные признаки:

- Длина текста (длина текста больше заданного порога)
- Число греческих букв в тексте
- Число математических символов в тексте

Зависимость качества базового классификатора от w_0



Классификатор

Итоговый классификатор: SVM.

$$a(x, w) = \operatorname{sign}\left(\sum_{j=1}^{n} w_j f_j(x) - w_0\right),$$

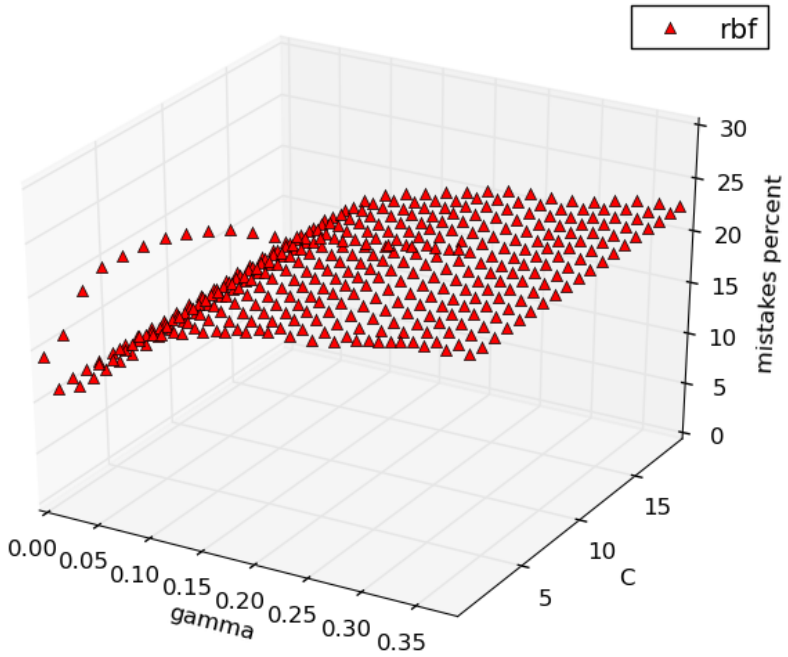
w_i определяются из условия:

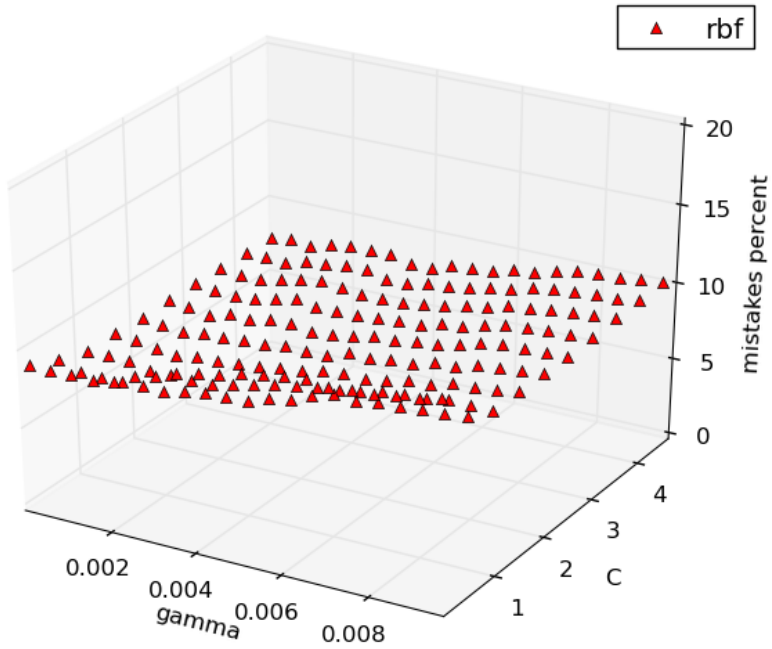
$$\begin{cases} \frac{1}{2}\langle w, w\rangle + C\sum_{i=1}^{n}\xi_i \to \min\limits_{w, w_0, \xi} \\ y_i(\langle w; x_i\rangle - w_0) \geq 1 - \xi_i, \quad i = 1 \ldots n \\ \xi_i \geq 0 \end{cases}$$

Выбор оптимальных параметров:

- ядро rbf (радиальные базисные функции, $e^{-\gamma(x-x_0)^2}$)
- γ и C (штраф за неверную классификацию) — по скользящему контролю

Оптимизация параметров SVM

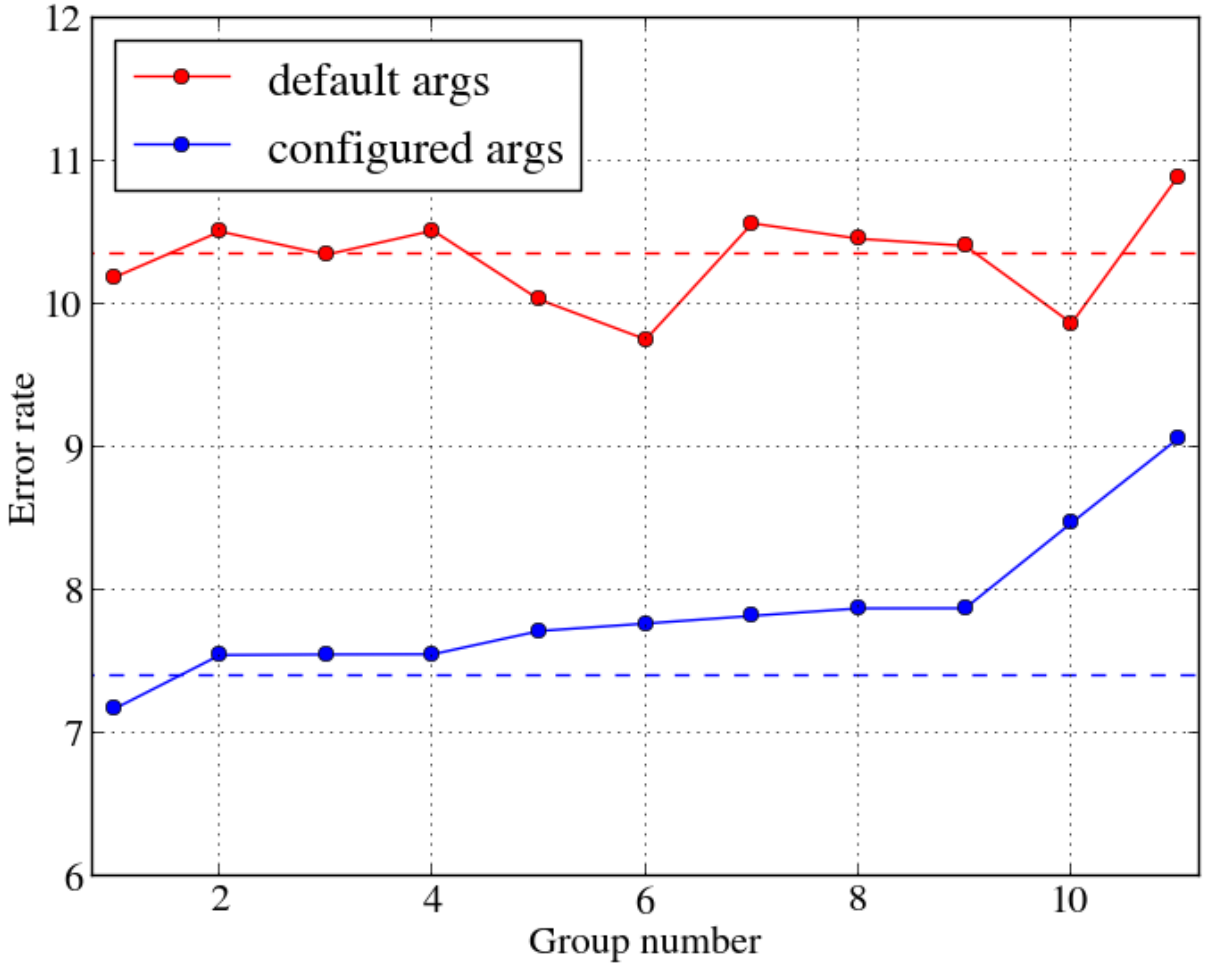




Признаковое описание документов

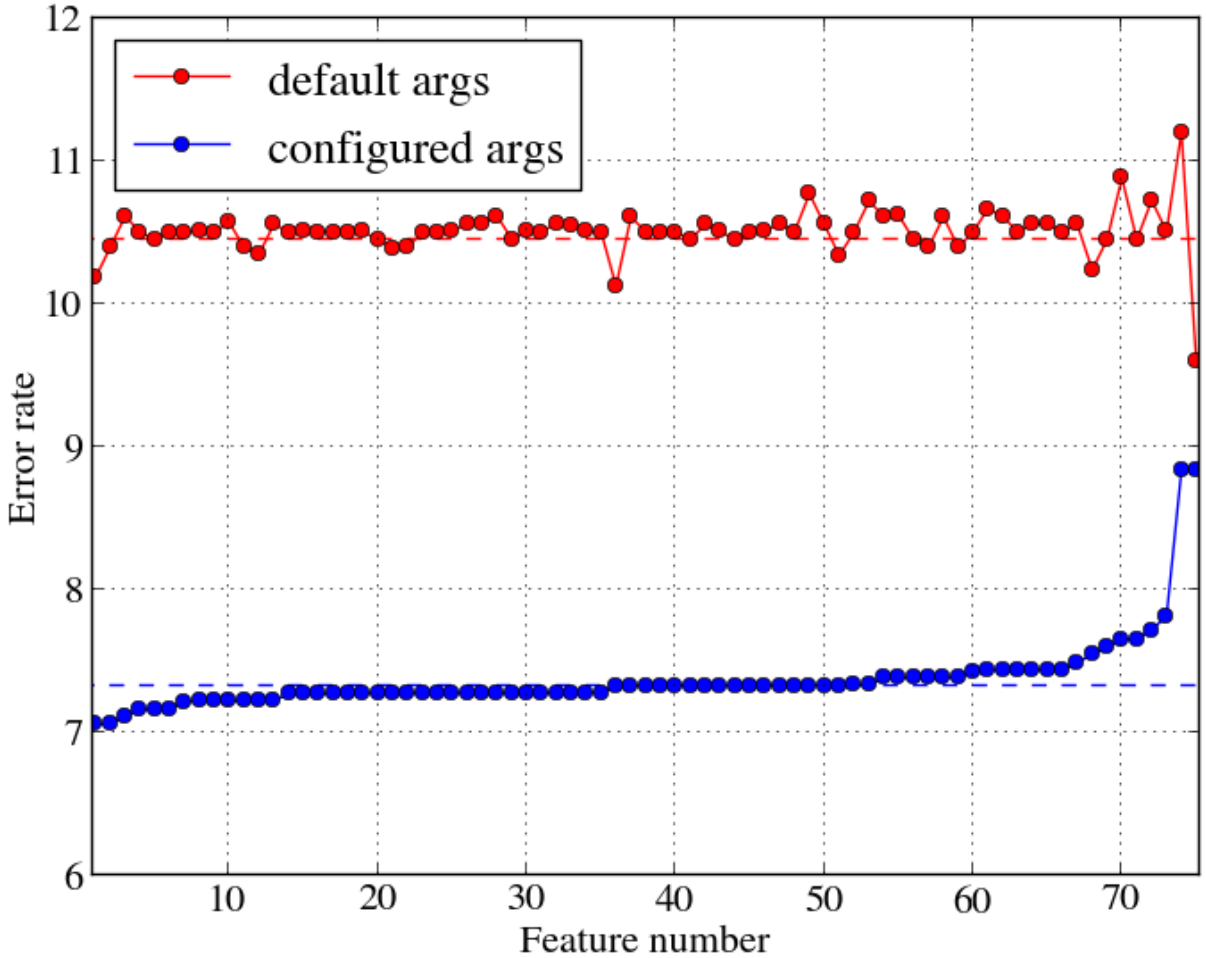
1. Количество ключевых слов в тесте (все слова разбиты на группы)
2. Наличие цифр в документе (их количество более 5 или нет)
3. Длина текста (длина текста больше заданного порога)
4. Минус-слова, присутствие которых понижает “научность” (количество часов, профком, экзамен)
5. Остальные слова, не попавшие в предыдущие группы
6. Слова, относящиеся к оформлению (редактор, редколлегия)
7. Слова, которые должны встречаться не более трех раз (Труды)
8. Число математических символов в тексте
9. Научные термины
10. Размер текста (логарифм размера текста в символах)
11. Число греческих букв в тексте

Отбор наиболее значимых признаков



Результаты проверки групп признаков на значимость (удаляется по одной группе)

Отбор наиболее значимых признаков



Результаты проверки всего набора признаков на значимость (удаляется по одному признаку)

Результаты

Стадия алгоритма	Ошибка 1 рода	Ошибка 2 рода
Базовый классификатор	4,9%	6,5%
SVM	3.2%	7.0%
Настройка параметров	4.2%	3.8%
Слова из оформления	4.2%	3.7%
Научные термины	4.2%	3.3%
Стоп слова	4.2%	3.2%

- Ошибка первого рода: научный классифицирован как ненаучный.
- Ошибка второго рода: ненаучный классифицирован как научный.