ОБУЧЕНИЕ КЛАССИФИКАТОРОВ НА ОСНОВЕ ВЫДЕЛЕНИЯ ФРАГМЕНТОВ

Васильев В.Г.

vvg_2000@mail.ru

Задачи выделения фрагментов

- Задача поиска фрагментов релевантных запросу (Passage Retrieval)
  - вычисление весов фрагментов
  - вычисление функции правдоподобия запроса
  - построение вероятностных моделей запросов
  - использование методов машинного обучения
  - использование скрытых марковских моделей
- Задача классификации фрагментов в соответствии с классификатором (Passage Recognition)
  - обучение на полных текстах , выделение фрагментов, классификация фрагментов
  - классификация текста целиком, поиск наиболее релевантного фрагмента
  - оценивание параметров скрытой марковской модели на полных текстах, выделение фрагментов

Методы классификации текстов

- Метод машин опорных векторов (SVM) – рубрика отделяется от других классов с помощью гиперплоскости
- Байесовский классификатор на основе модели смеси распределений фон Мизеса-Фишера (VMF) – рубрика описывается с помощью точки на гиперсфере единичного радиуса

Модель фон Мизеса-Фишера

Функция плотности

$$f(x \mid \mu_j, \kappa_j) = c_m(\kappa_j) e^{\kappa_j \mu_j^T x},$$

$x \in \mathbb{R}^m$, $\|x\|_2 = 1$, $\mu_j \in \mathbb{R}^m$ - среднее направление, $\|\mu_j\|_2 = 1$, $\kappa_j \geq 0$ - мера концентрации, $m \geq 2$, $c_m(\kappa)$ - нормализующий множитель.

Дискриминантная функция

$$g_j(x) = w_0^{-1} \left(\log \frac{\pi_j}{\pi_0} + \kappa_j \mu_j^T x - \bar{\kappa}_0 \bar{\mu}_0^T x \right).$$

Методы выделения фрагментов в текстах

- Выделение фрагментов путем классификации предложений (SENT)
- Выделение фрагментов путем классификации блоков текста (TILE)
- Выделение фрагментов путем классификации иерархического покрытия (HIER)
- Выделение фрагментов с использованием оптимизационных методов (LS)

Выделение фрагментов путем классификации блоков текста

Веса предложений w_{ij}, $i = 1, \ldots, n$, для рубрик r_j, $j = 1, \ldots, m$ находятся путем использования следующей формулы

$$w_{ij} = s_{ij},$$

где s_{ij} – мера соответствия блока $i = 1, \ldots, n$ рубрике r_j, $j = 1, \ldots, m$

$$s_{ij} = \frac{\sigma_{i=1,\ldots,n;\, j=\text{const}}}{\sigma_{i=1,\ldots,n;\, j=\text{const}}^2}.$$

Выделение фрагментов путем классификации иерархического покрытия

Вес предложения [illegible], [illegible] = 1, … , [illegible], для рубрики [illegible], [illegible] = 1, … , [illegible], вычисляется как

[illegible] = [illegible] [illegible] [illegible] [illegible] [illegible],

где [illegible] множество векторов признаков, соответствующих всем непрерывным фрагментам предложений в данном документе, т.е.

[illegible] = [illegible] [illegible] [illegible] |1 ≤ [illegible] ≤ [illegible] ≤ [illegible].

Выделение фрагментов с использованием оптимизационных методов

Веса $\lambda_j = \lambda_{j1}, \ldots \lambda_{jn}$ предложений для рубрики c_j, $j = 1, \ldots, k$ находятся путем нахождения

$$\max_{\lambda_j} K_{c_j}(d),$$

где K_{c_j} – дискриминантная функция, $d(s_i)$ – вектор документа,

$$w_{ji} = \frac{\sigma_{i=1} \lambda_{ji} s_i}{\| \sigma_{i=1} \lambda_{ji} s_i \|_2}.$$

Схема итерационного обучения отдельной рубрики

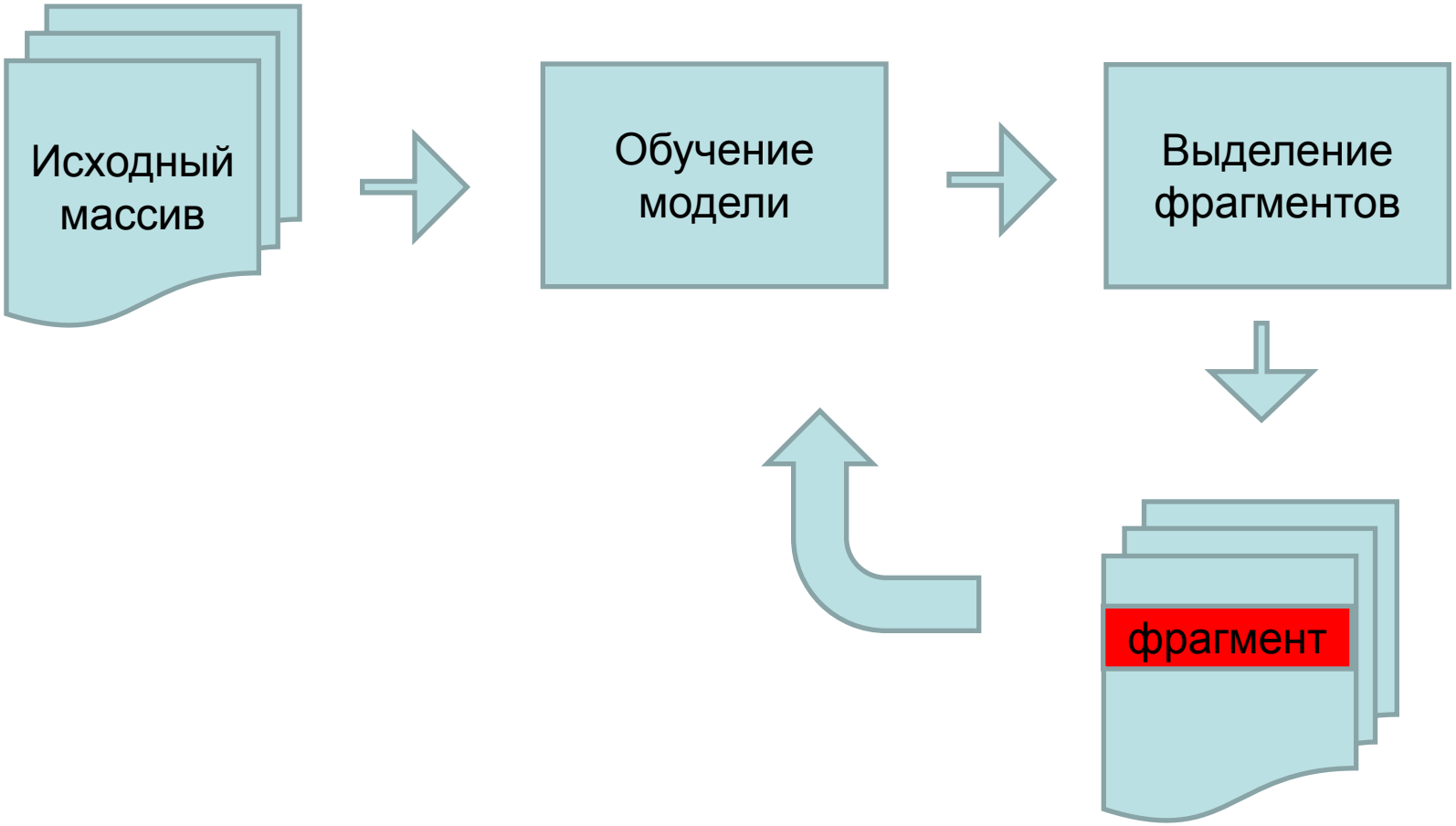
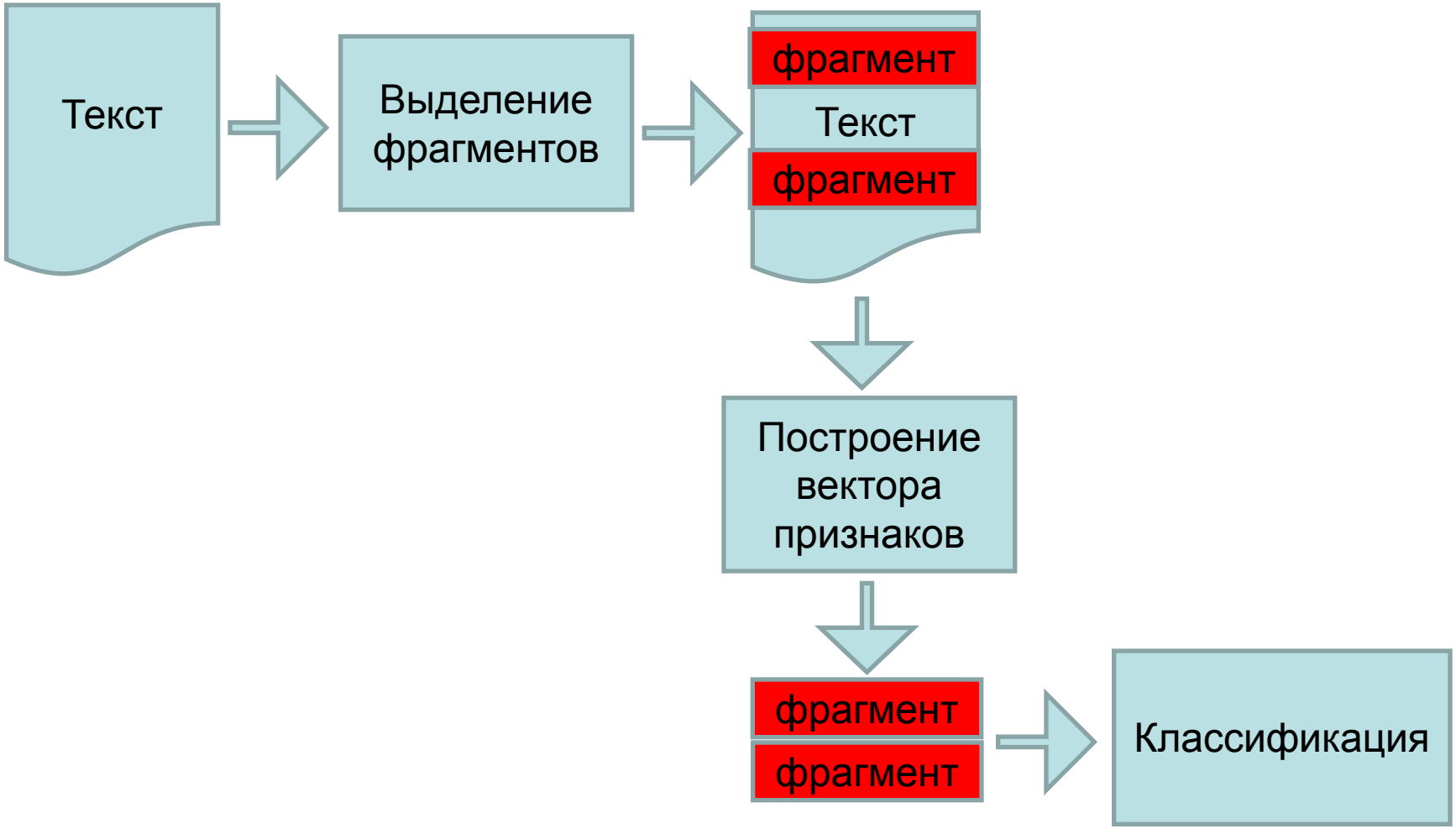


Схема классификации текстов с использованием фрагментов



Характеристики массивов текстов

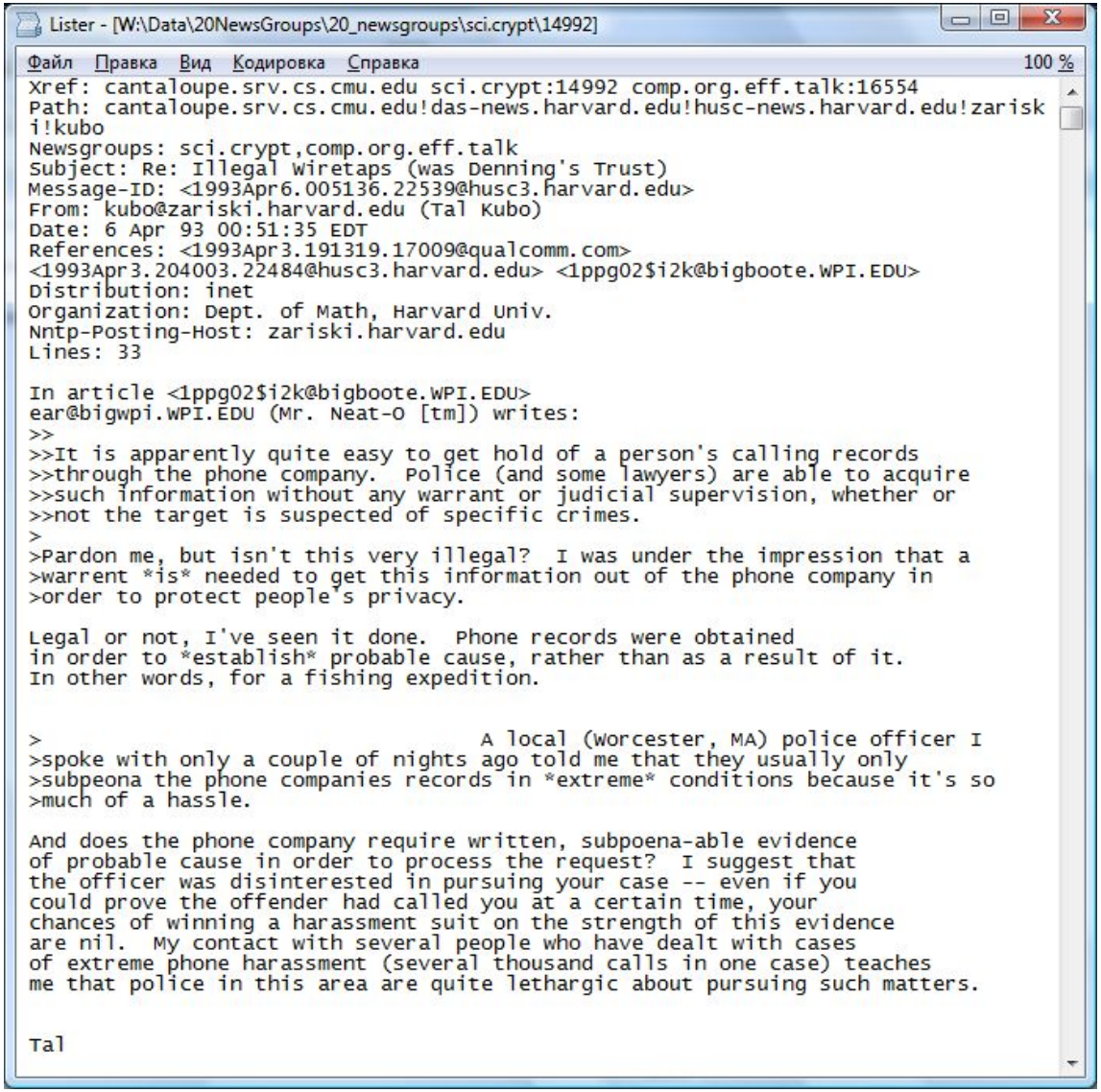
Полные массивы

Массив	Число документов	Число рубрик	Размер
Reuters-21578	13476	123	16.2Mb
20 News Groups	16330	20	46Mb
ROMIP 2004 Legal	6931	173	281Mb

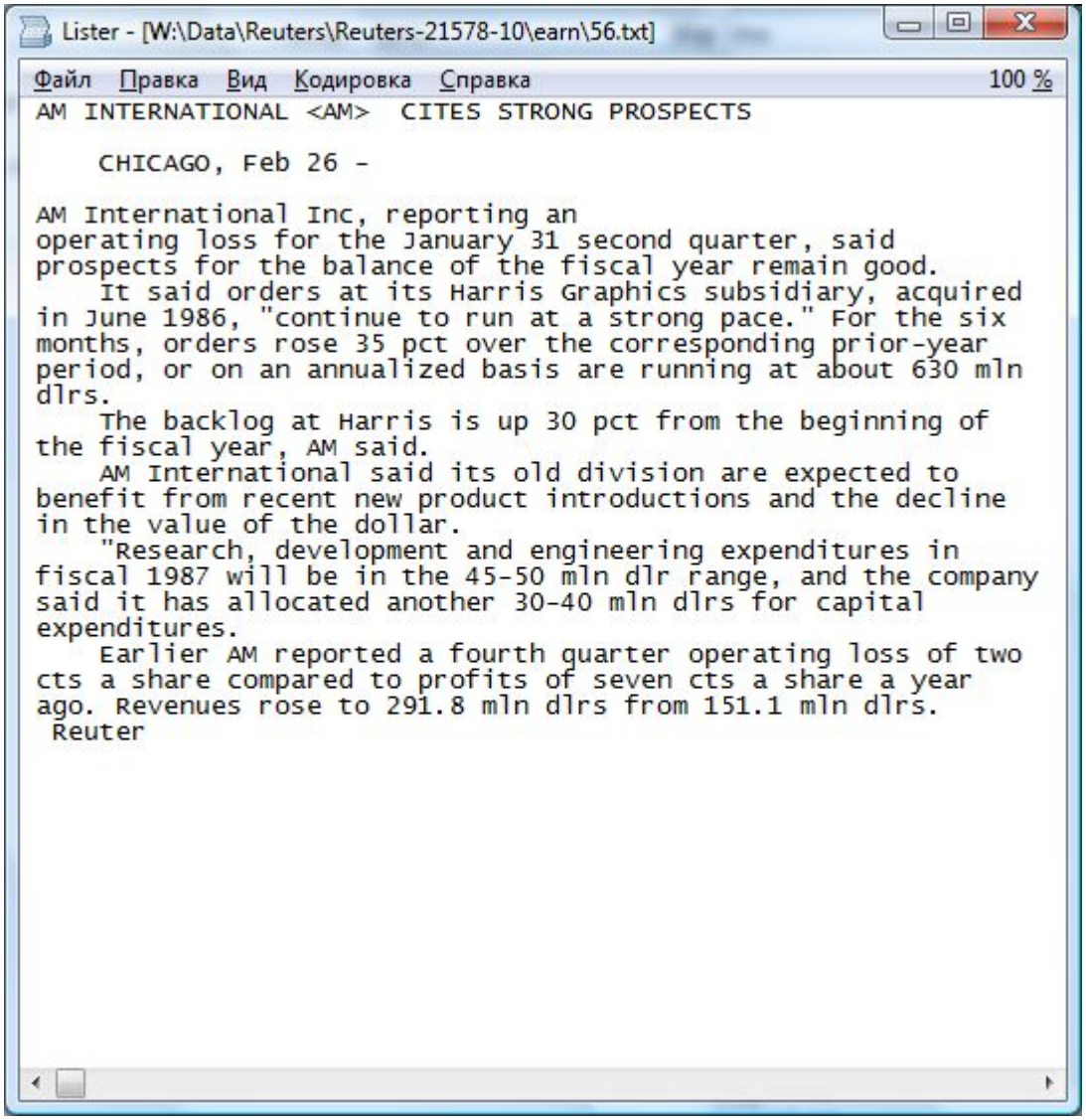
Сокращенные массивы текстов

Массив	Число документов	Число рубрик	Размер
Reuters-21578-10	8592	10	8.9Mb
20 News Groups Mini	1954	20	4.7Mb
ROMIP 2004 Legal Mini	1704	10	113Мб

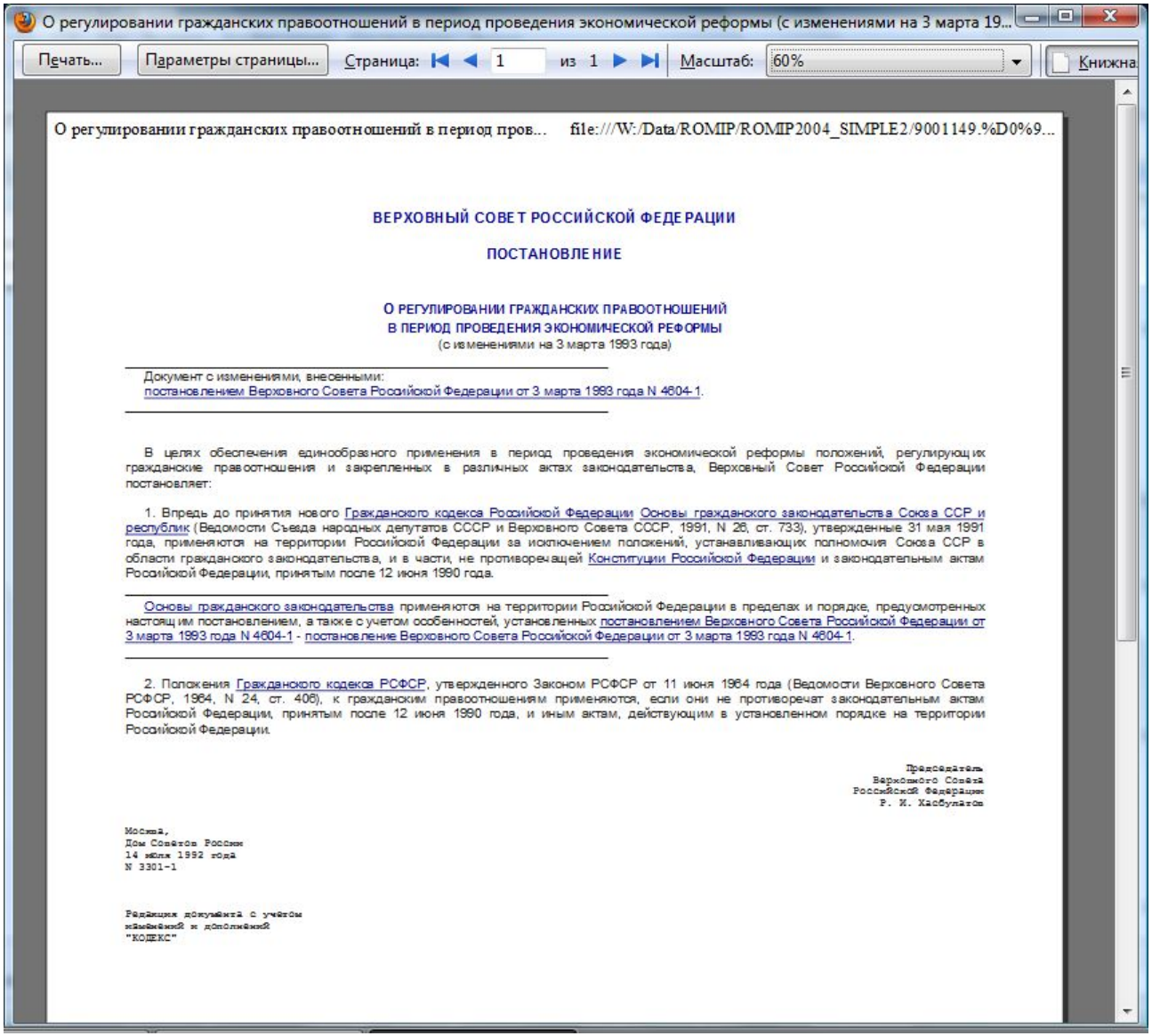
Пример текста из массива 20NG



Пример текста из массива Reuters-21578



Пример текста из массива ROMIP 2004 Legal



Оценка качества классификации для массива ROMIP 2004 Legal Mini

Метод	F-мера	Метод	F-мера
SVM	0.37	VMF	0.45
SVM-SENT	0.39	VMF-SENT	0.47
SVM-HIER	0.38	VMF-HIER	0.46
SVM-TILE	0.39	VMF-TILE	0.46
SVM-LS	0.50	VMF-LS	0.37

Оценка точности и полноты классификации

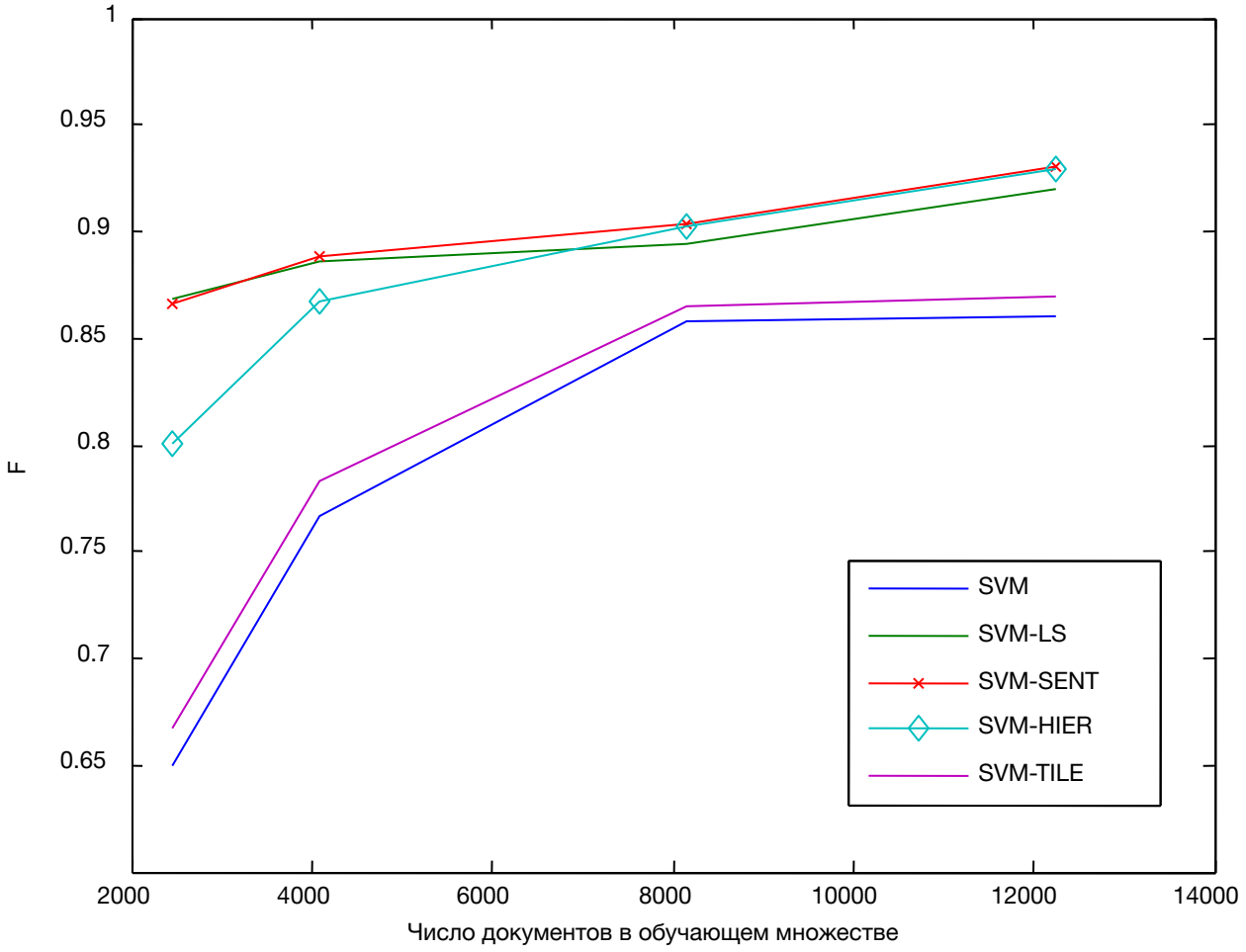
20 News Group Mini

Метод	Точность	Полнота	F-мера
SVM	0.94	0.27	0.40
SVM-LS	0.96	0.89	0.92

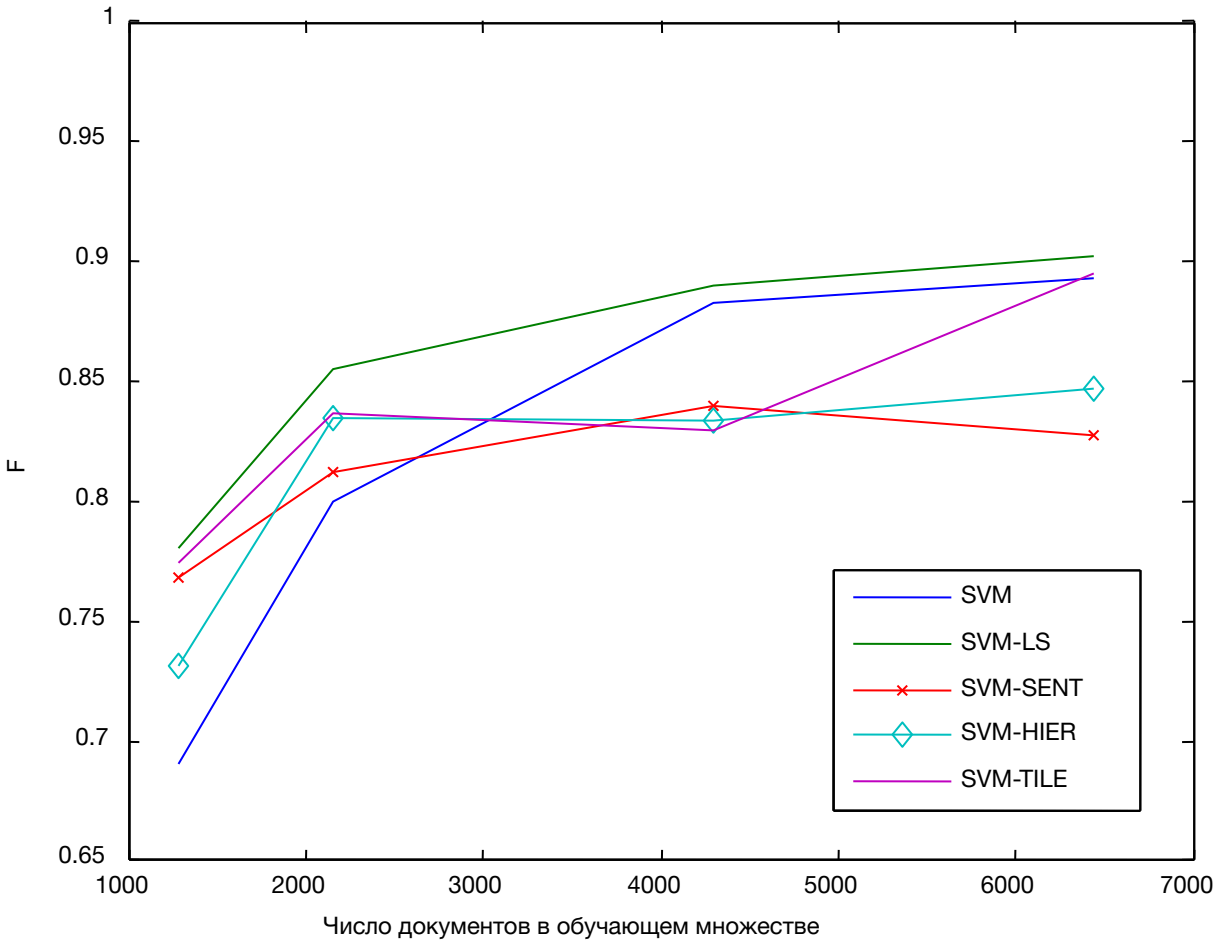
Romip 2004 Legal Mini

Метод	Точность	Полнота	F-мера
SVM	0.67	0.29	0.36
SVM-LS	0.76	0.39	0.50

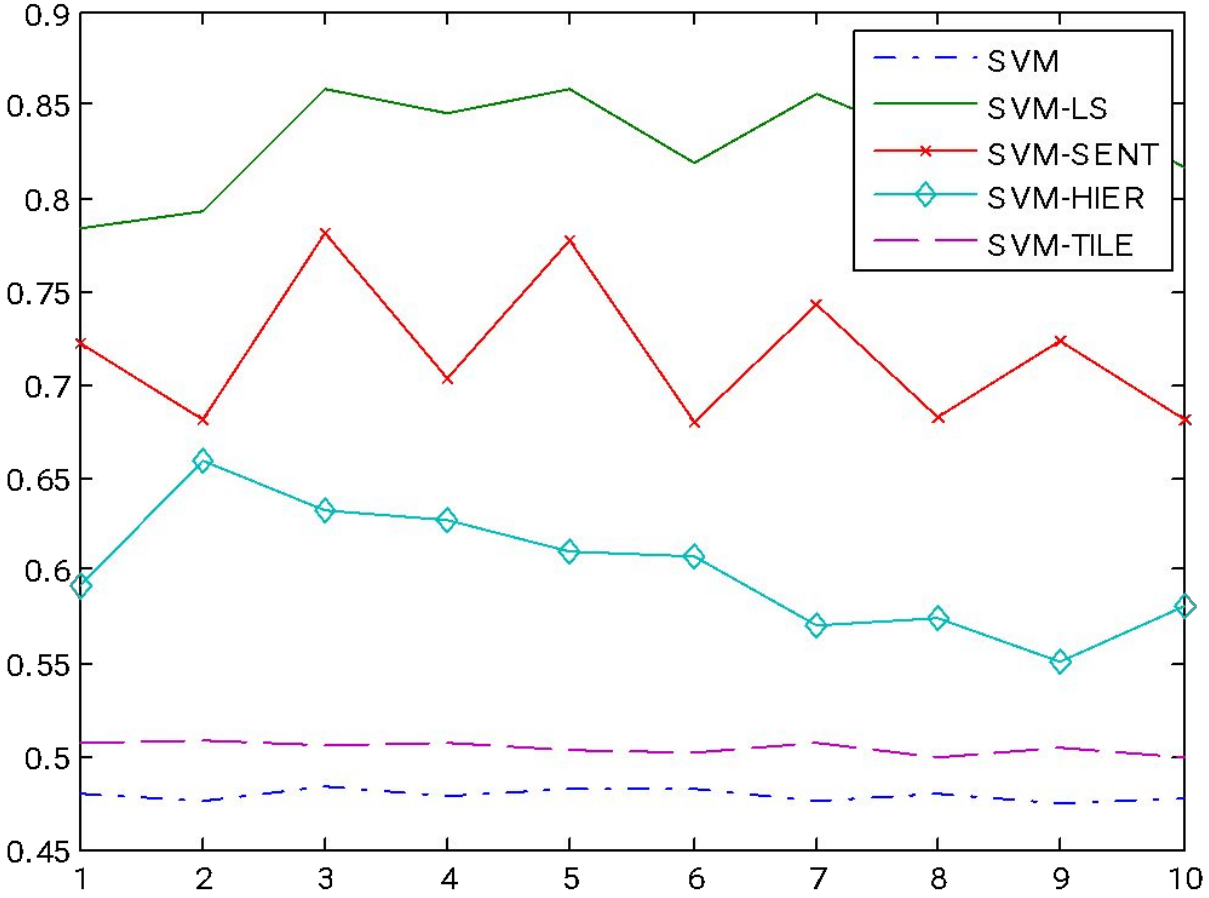
Качество классификации для массива 20 NG от размера обучающего множества



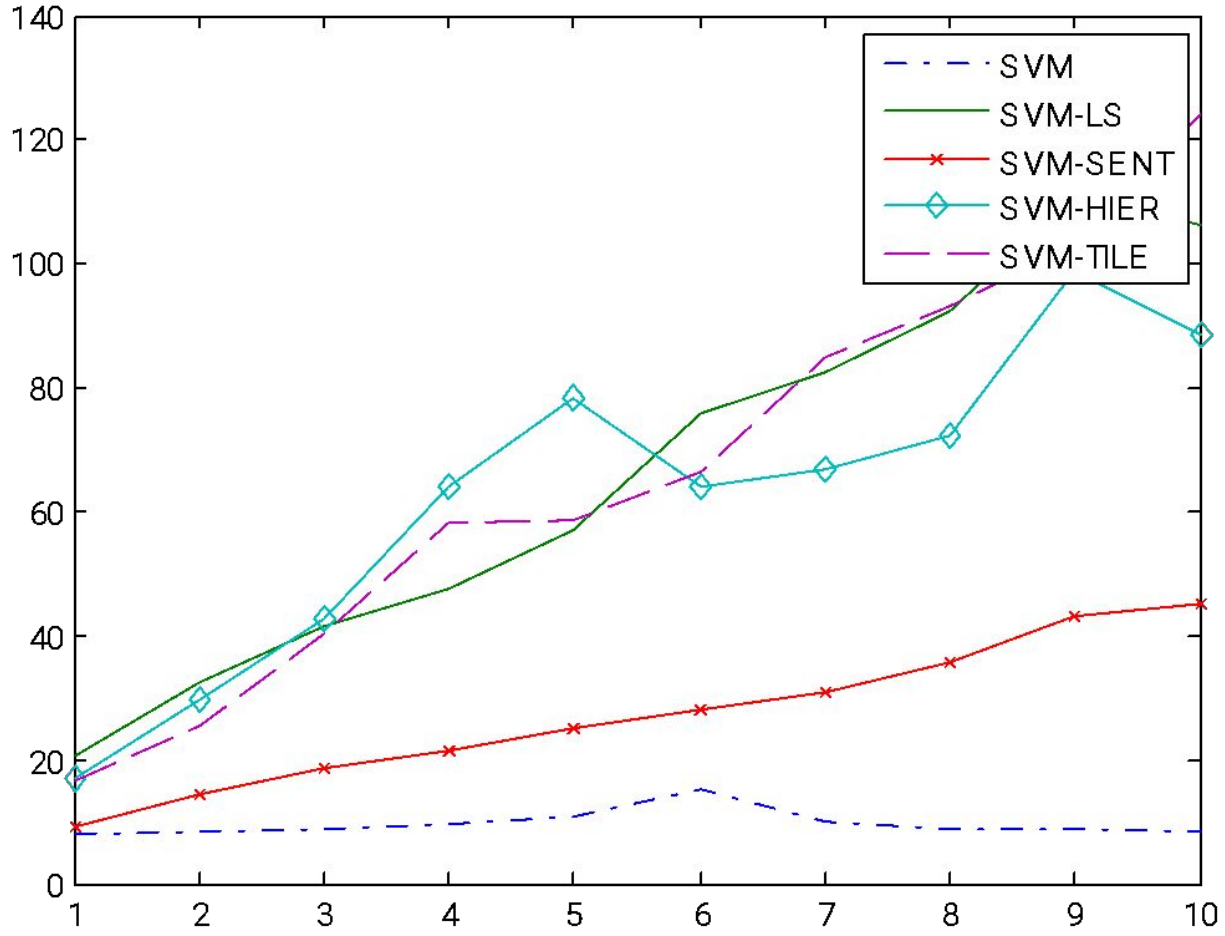
Качество классификации для массива Reuters-21578-10 от размера обучающего множества



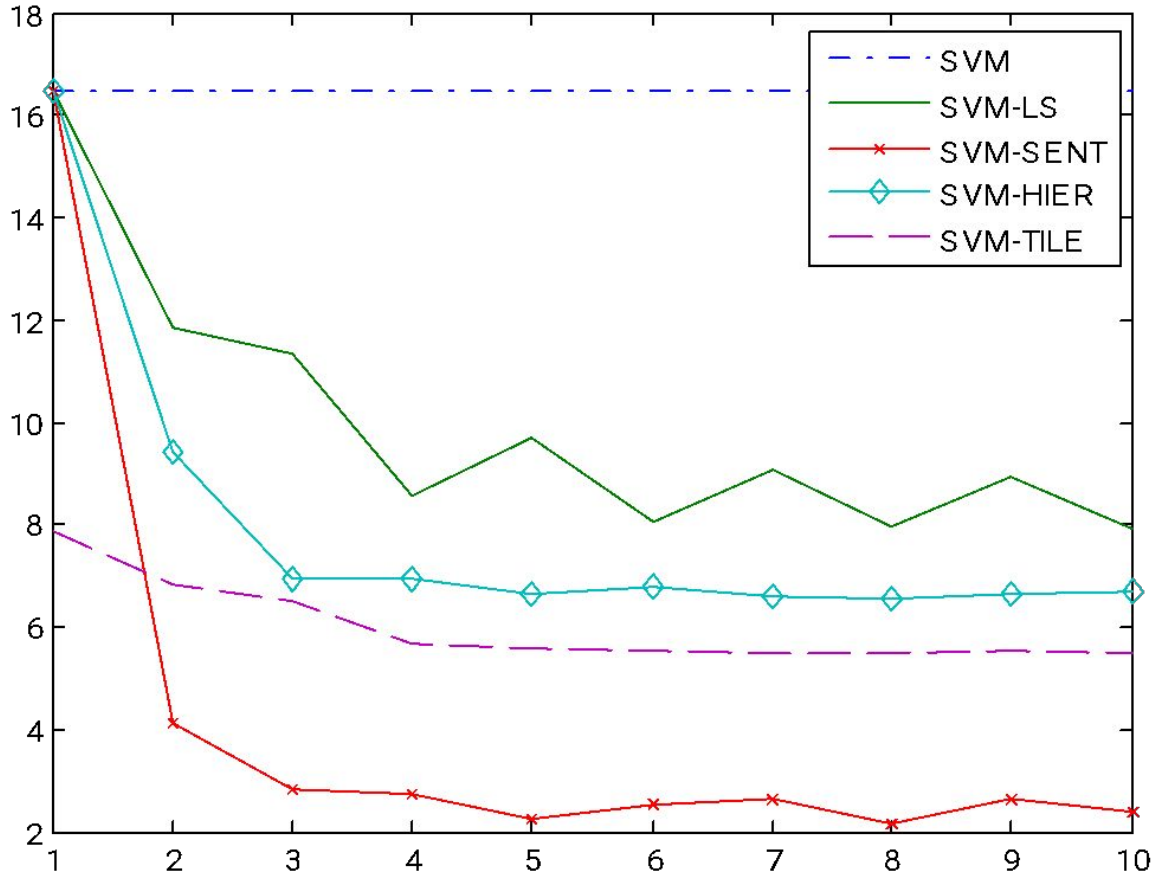
Качество классификации для массива 20 NG от числа итераций



Время обучения и классификации массива 20 NG от числа итераций



Среднее число выделяемых предложений для массива 20 NG в зависимости от числа итераций



Выводы

- Обучения классификаторов с использованием фрагментов более эффективно при маленьких размерах обучающих выборок
- В некоторых случаях использование данного метода может заметно улучшить полноту классификации
- Для сходимости метода достаточно выполнения нескольких итераций

Направления дальнейших исследований

- Анализ характера выделяемых фрагментов и областей применимости рассмотренного подхода
- Использование методов рандомизации (бутстреп метода) для расширения объема обучающих выборок
- Выделение фрагментов с использованием правил на специальном языке