

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

$$f(x) = \frac{1}{C_m} \prod_{j=1}^m \frac{1}{\mu_j} \exp\left(-\frac{x_j}{\mu_j}\right) \left(\frac{x_j}{\mu_j}\right)^{k_j - 1},$$
 $x \in \mathbb{R}_+^m, \quad \sum_{j=1}^m \frac{1}{\mu_j} = 1, \quad \mu_j \in \mathbb{R}_+^m$ - среднее направление, $\sum_{j=1}^m \mu_j = 1, \quad k_j \geq 0$ - мера концентрации, $m \geq 2, C_m = \sum_{k=0}^{\infty} \frac{1}{k!} \prod_{j=1}^m k_j!$ - нормализующий множитель.

Дискриминантная функция

$$D(x) = \frac{1}{C_m} \sum_{k=0}^{\infty} \frac{1}{k!} \prod_{j=1}^m \frac{x_j^{k_j}}{\mu_j^{k_j}} + \sum_{j=1}^m \frac{x_j}{\mu_j} - \sum_{j=1}^m \frac{x_j^{k_j}}{\mu_j^{k_j}}$$

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} , $i = 1, \dots, n$, для рубрик r_{jk} , $j = 1, \dots, m$ находятся путем использования следующей формулы

$$w_{ij} = \frac{w_{ij}}{\sum_{k=1}^m w_{ik}},$$

где w_{ij} – мера соответствия блока $i = 1, \dots, n$ рубрике r_{jk} , $j = 1, \dots, m$

$$w_{ij} = \frac{w_{ij} \prod_{k=1}^m \sigma_{r_{jk}}}{\sum_{k=1}^m \sigma_{r_{jk}} \prod_{k=1}^m \sigma_{r_{jk}}^2}.$$

Выделение фрагментов путем классификации иерархического покрытия

Вес предложения $\mathcal{L}_{i,j}$, $i = 1, \dots, n$, для рубрики \mathcal{L}_j , $j = 1, \dots, m$, вычисляется как

$$\mathcal{L}_{i,j} = \frac{\mathcal{L}_j \cdot \mathcal{L}_{i,j}}{\sum_{k=1}^m \mathcal{L}_k \cdot \mathcal{L}_{i,k}}$$

где \mathcal{L}_j множество векторов признаков, соответствующих всем непрерывным фрагментам предложений в данном документе, т.е.

$$\mathcal{L}_j = \mathcal{L}_1 \cup \mathcal{L}_2 \cup \dots \cup \mathcal{L}_n \quad | \quad 1 \leq n \leq n_2 \leq n_1.$$

$$\mathcal{L}_1 = \mathcal{L}_2$$

Выделение фрагментов с использованием ОПТИМИЗАЦИОННЫХ МЕТОДОВ

Веса $w_{ij} = w_{i1}, \dots, w_{in}$ предложений для рубрики $R_j, j = 1, \dots, n$ находятся путем нахождения

$$\max_{\lambda_j} \sum_{j=1}^n \lambda_j f_j(\mathbf{x}),$$

где f_j – дискриминантная функция, \mathbf{x} – вектор документа,

$$w_{ij} = \frac{\sigma_{i=1}(\mathbf{x})}{\sum_{j=1}^n \sigma_{j=1}(\mathbf{x})}.$$

Схема итерационного обучения отдельной рубрики

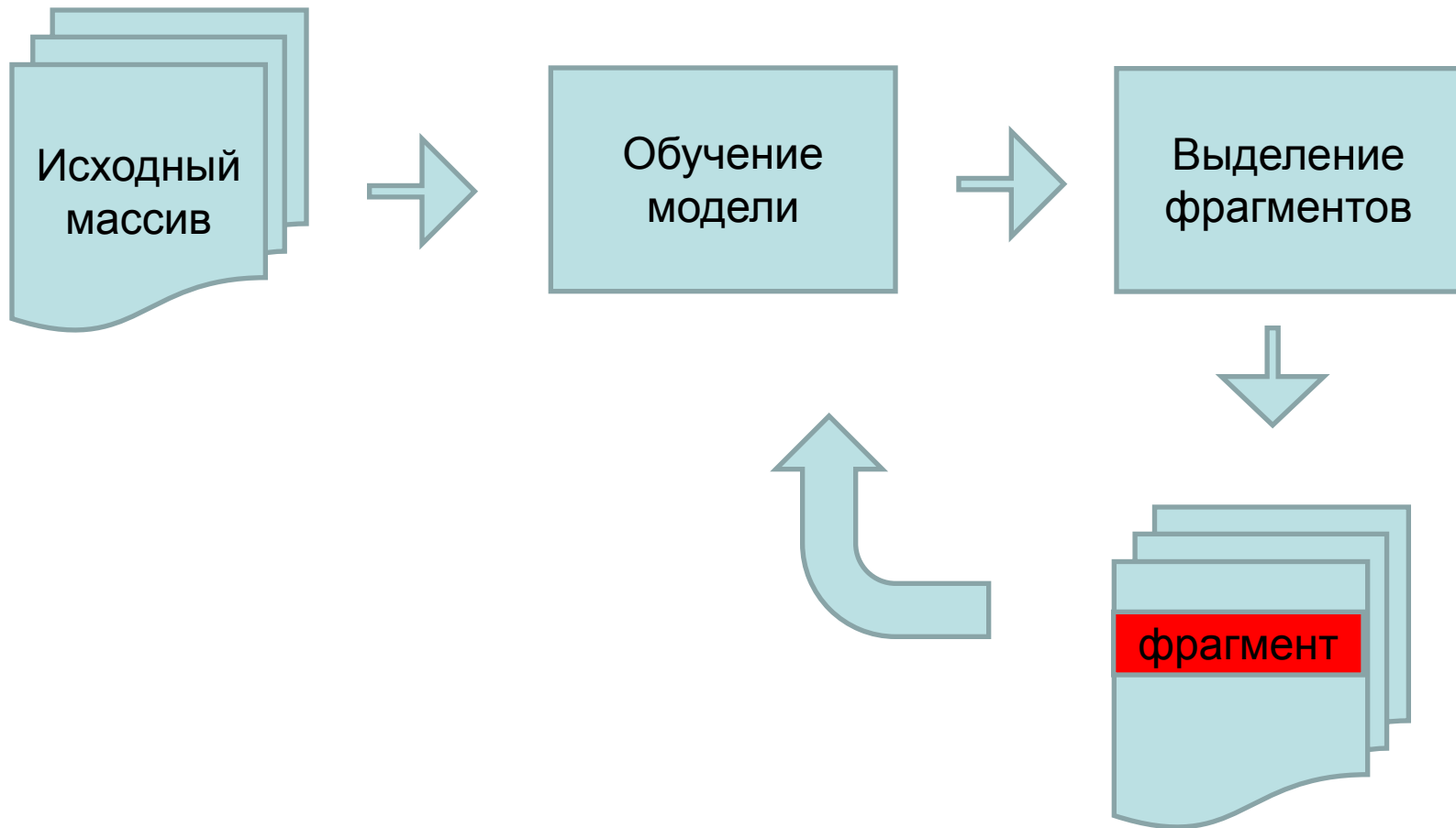
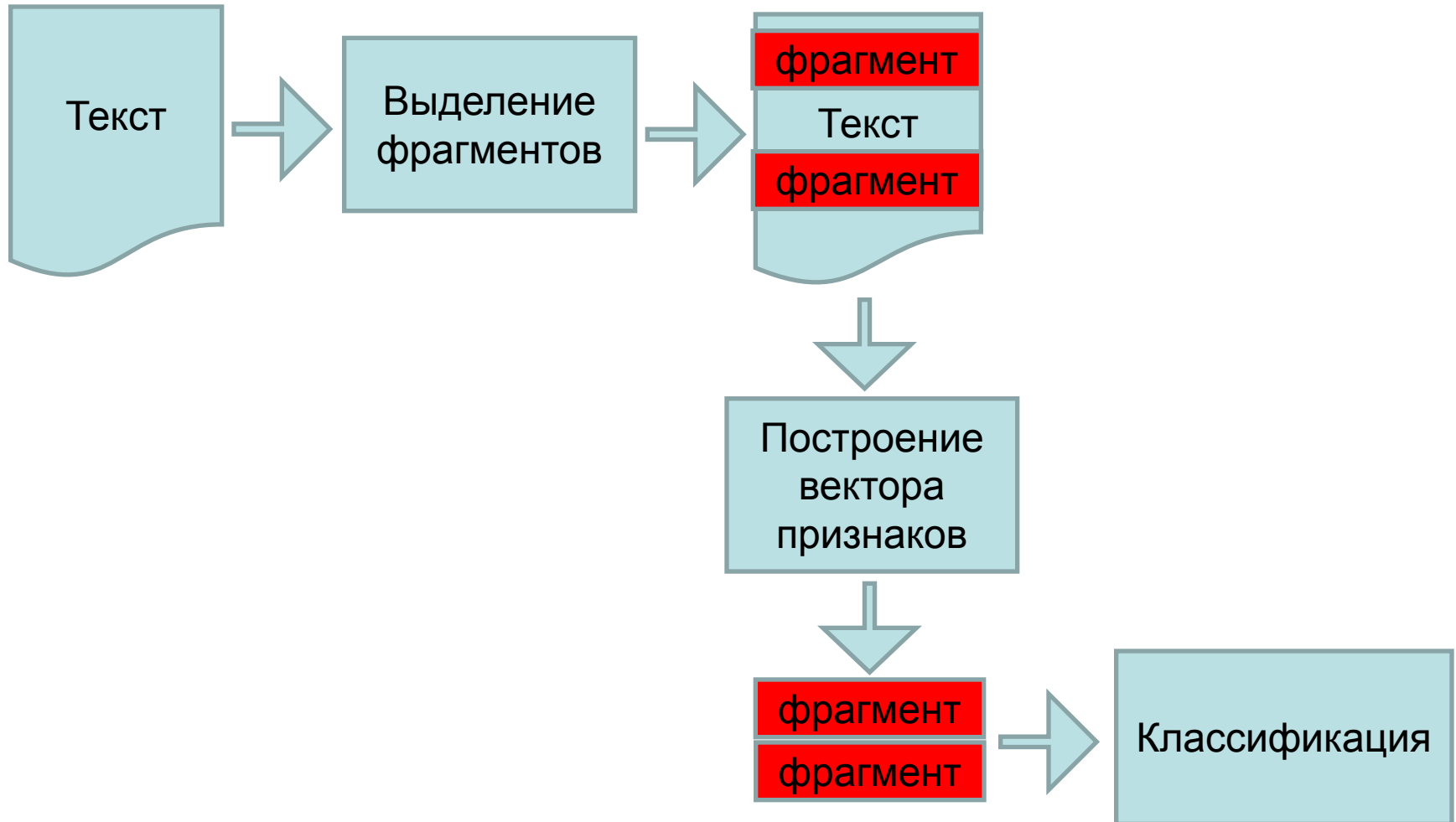


Схема классификации текстов с использованием фрагментов



Характеристики массивов текстов

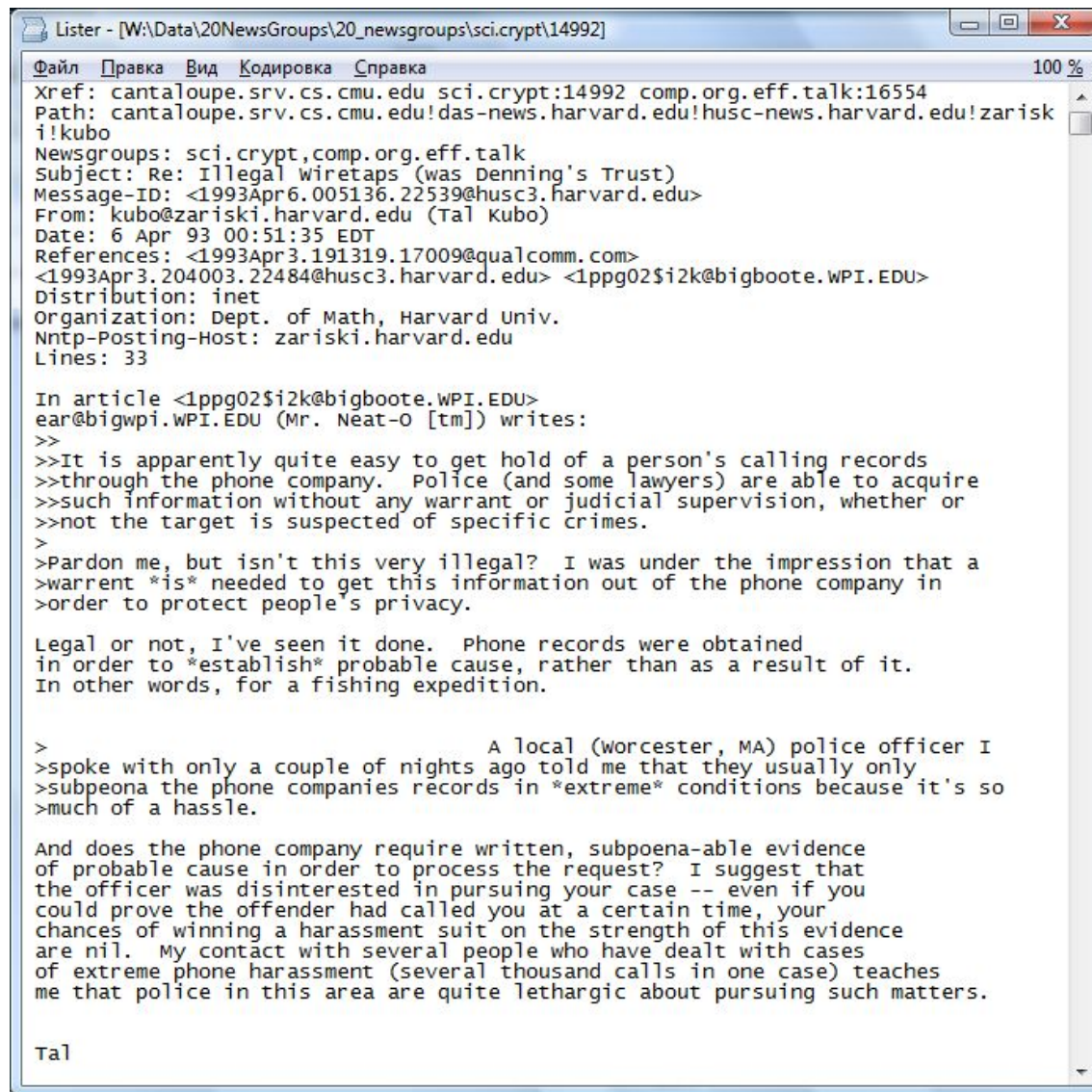
Полные массивы

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	113Mb

Пример текста из массива 20NG



```
Listner - [W:\Data\20NewsGroups\20_newsgroups\sci.crypt\14992]
Файл  Правка  Вид  Кодировка  Справка  100 %
Xref: cantaloupe.srv.cs.cmu.edu sci.crypt:14992 comp.org.eff.talk:16554
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu!husc-news.harvard.edu!zarisk
i!kubo
Newsgroups: sci.crypt,comp.org.eff.talk
Subject: Re: Illegal wiretaps (was Denning's Trust)
Message-ID: <1993Apr6.005136.22539@husc3.harvard.edu>
From: kubo@zariski.harvard.edu (Tal Kubo)
Date: 6 Apr 93 00:51:35 EDT
References: <1993Apr3.191319.17009@qualcomm.com>
<1993Apr3.204003.22484@husc3.harvard.edu> <1ppg02$i2k@bigboote.WPI.EDU>
Distribution: inet
Organization: Dept. of Math, Harvard Univ.
Nntp-Posting-Host: zariski.harvard.edu
Lines: 33

In article <1ppg02$i2k@bigboote.WPI.EDU>
ear@bigwpi.WPI.EDU (Mr. Neat-O [tm]) writes:
>>
>>It is apparently quite easy to get hold of a person's calling records
>>through the phone company.  Police (and some lawyers) are able to acquire
>>such information without any warrant or judicial supervision, whether or
>>not the target is suspected of specific crimes.
>
>Pardon me, but isn't this very illegal?  I was under the impression that a
>warrent *is* needed to get this information out of the phone company in
>order to protect people's privacy.

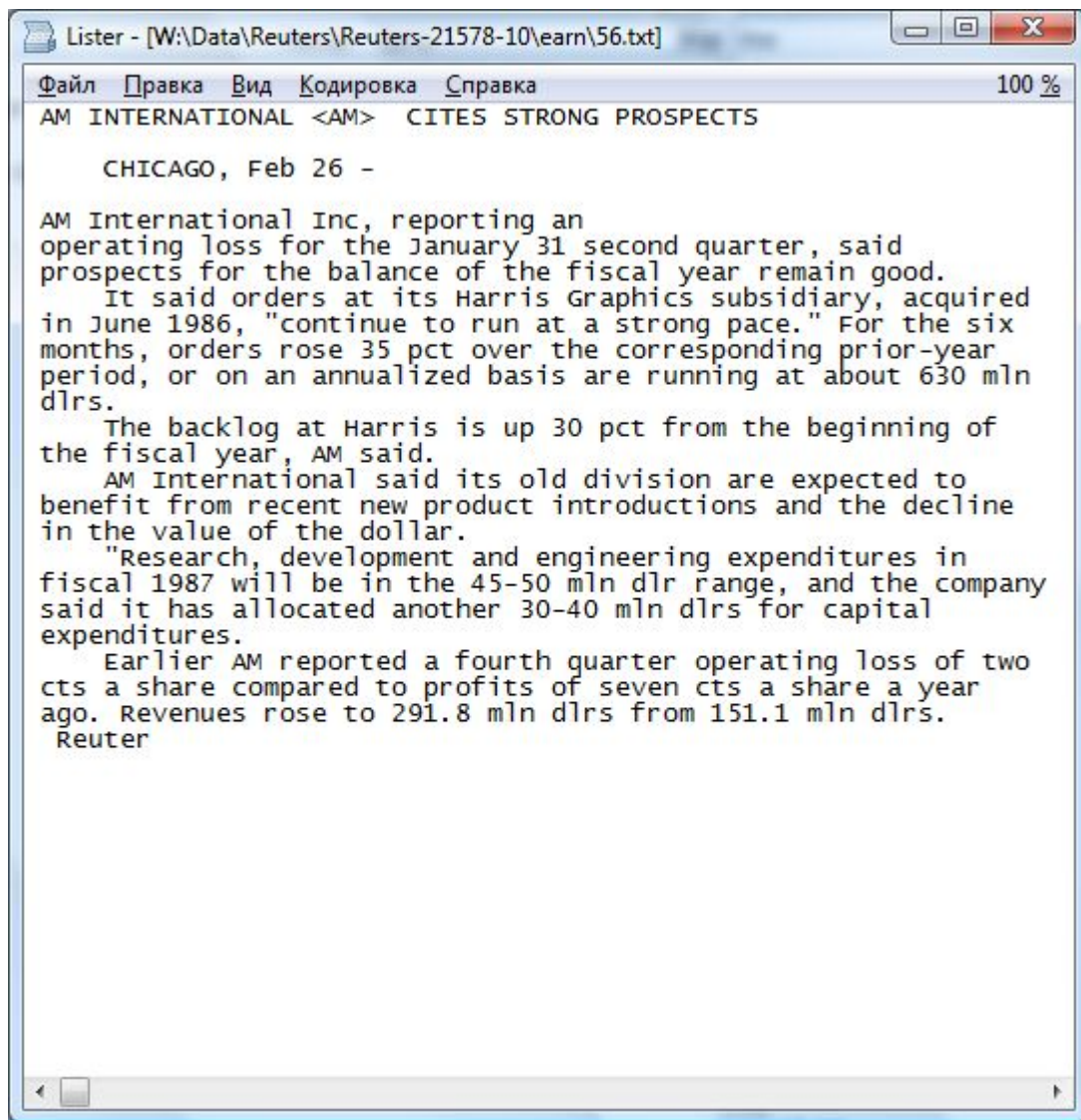
Legal or not, I've seen it done.  Phone records were obtained
in order to *establish* probable cause, rather than as a result of it.
In other words, for a fishing expedition.

>
>A local (worcester, MA) police officer I
>>spoke with only a couple of nights ago told me that they usually only
>>subpeona the phone companies records in *extreme* conditions because it's so
>>much of a hassle.

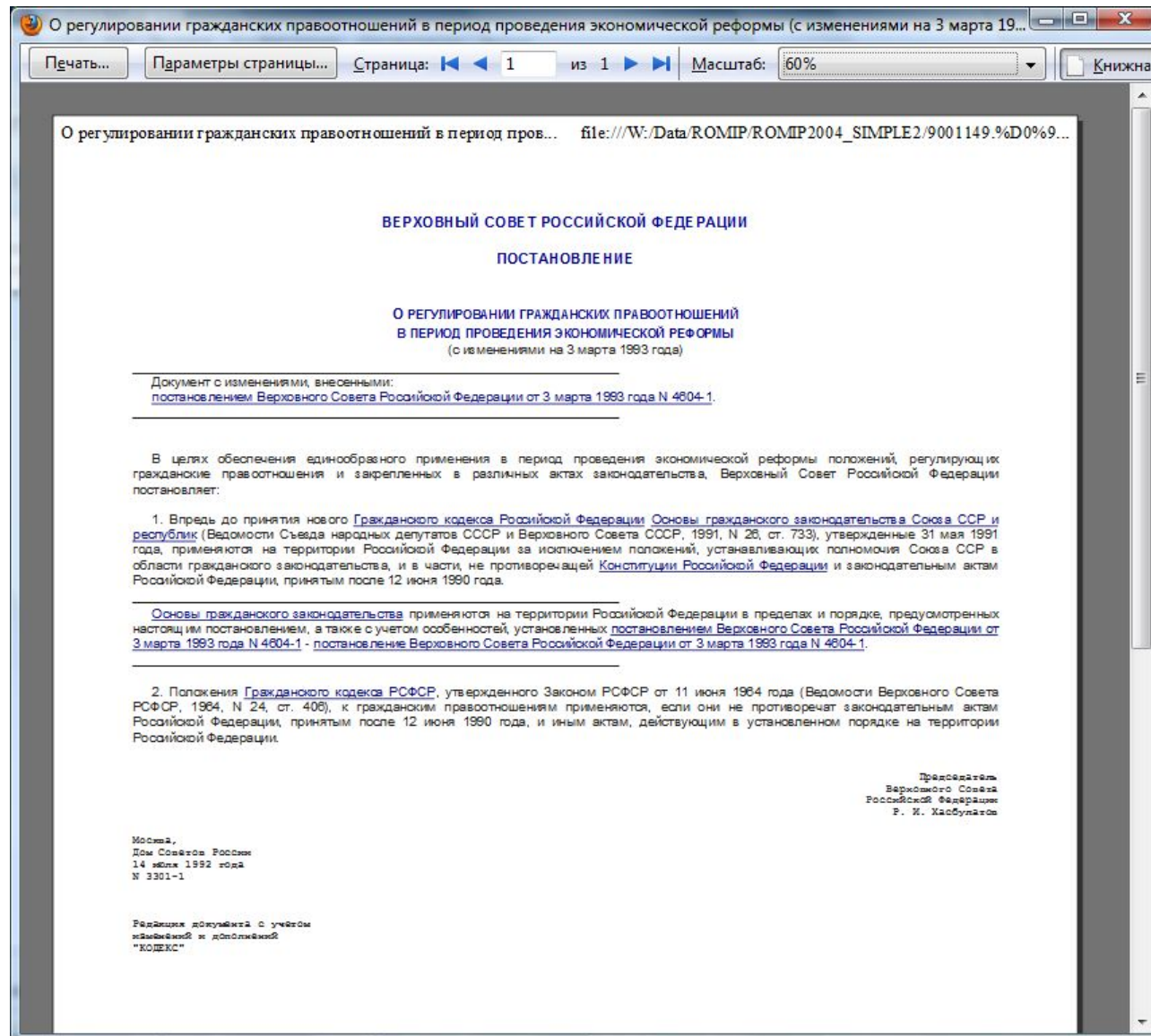
And does the phone company require written, subpoena-able evidence
of probable cause in order to process the request?  I suggest that
the officer was disinterested in pursuing your case -- even if you
could prove the offender had called you at a certain time, your
chances of winning a harassment suit on the strength of this evidence
are nil.  My contact with several people who have dealt with cases
of extreme phone harassment (several thousand calls in one case) teaches
me that police in this area are quite lethargic about pursuing such matters.

Tal
```

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

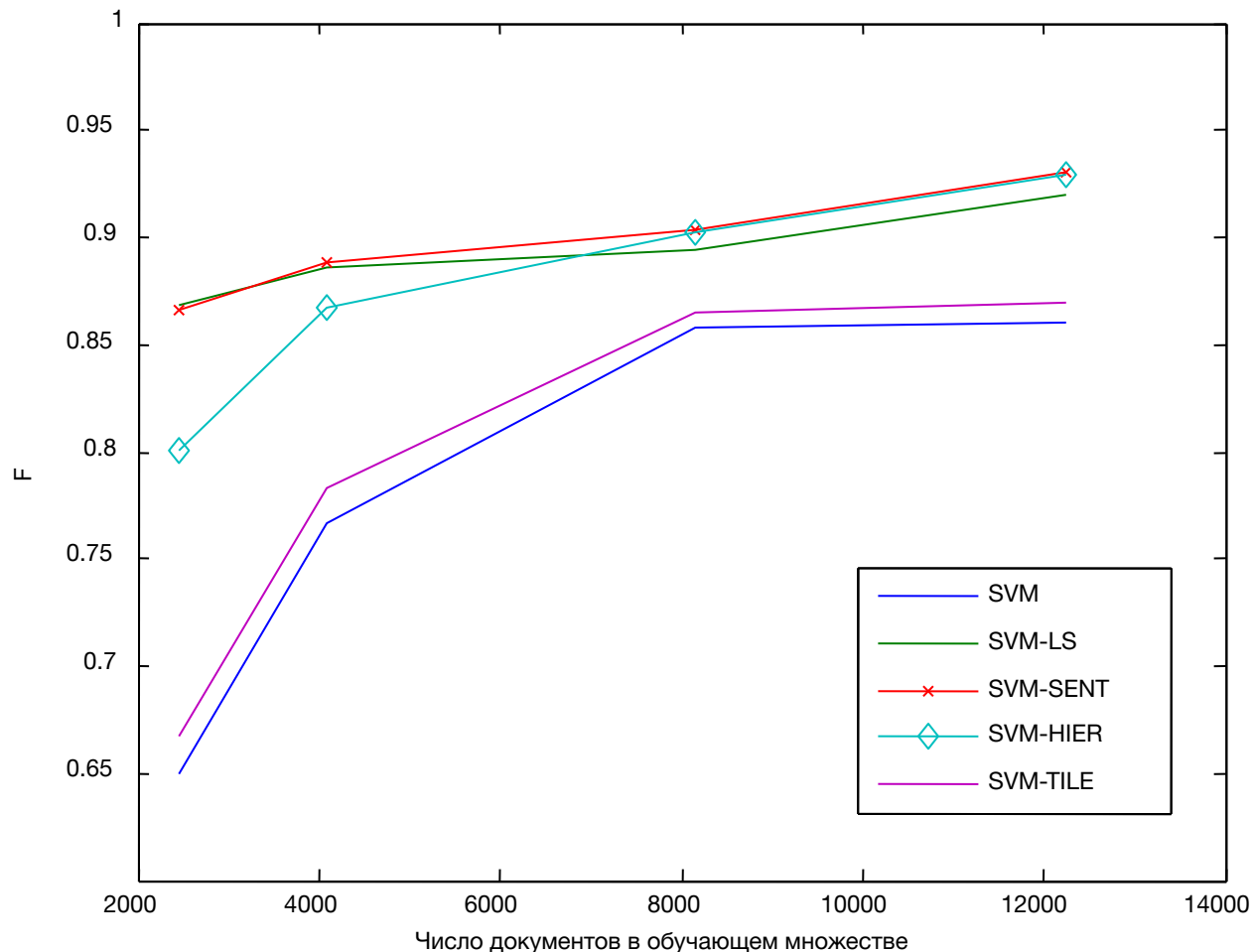
20 News Group Mini

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

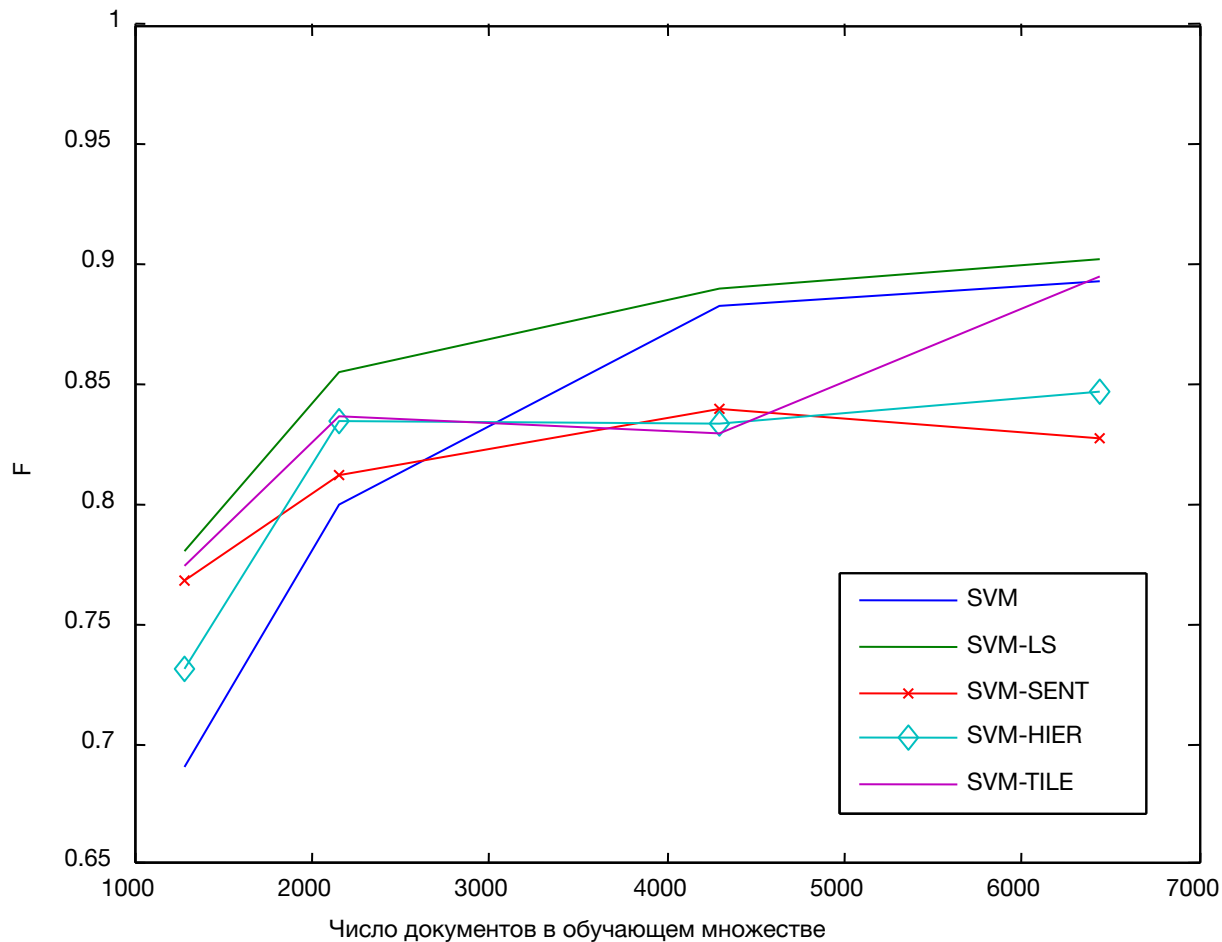
Romip 2004 Legal Mini

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

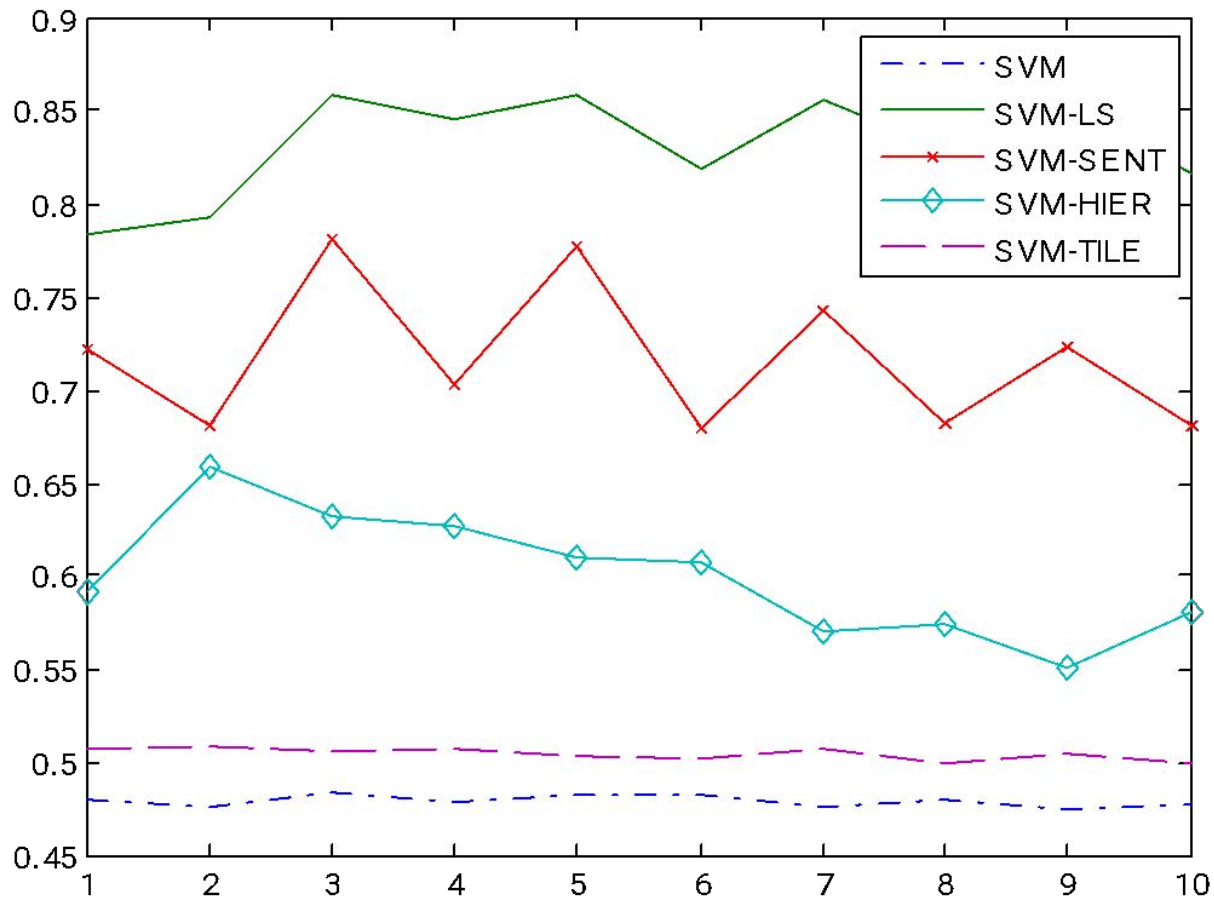
Качество классификации для массива 20 NG от размера обучающего множества



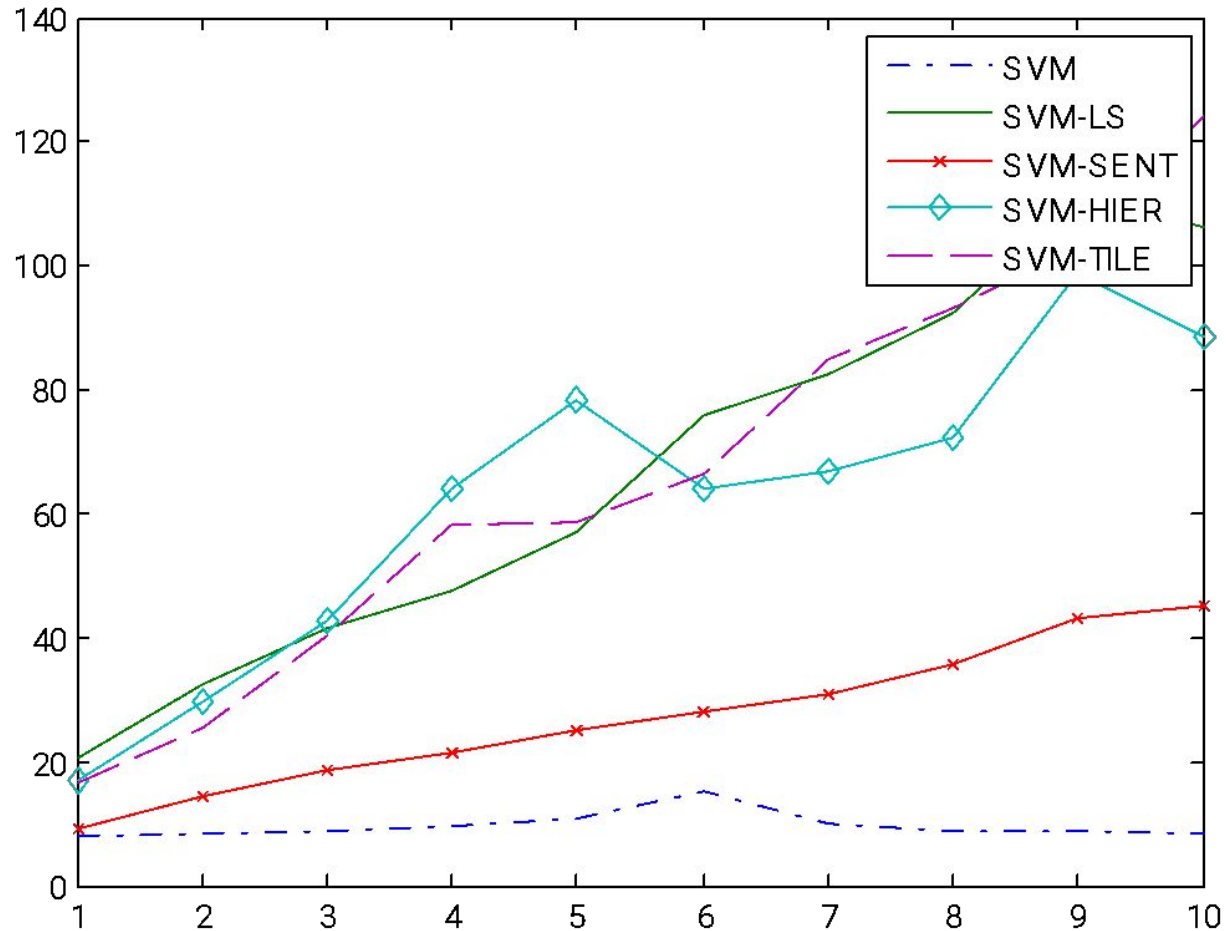
Качество классификации для массива Reuters-21578-10 от размера обучающего множества



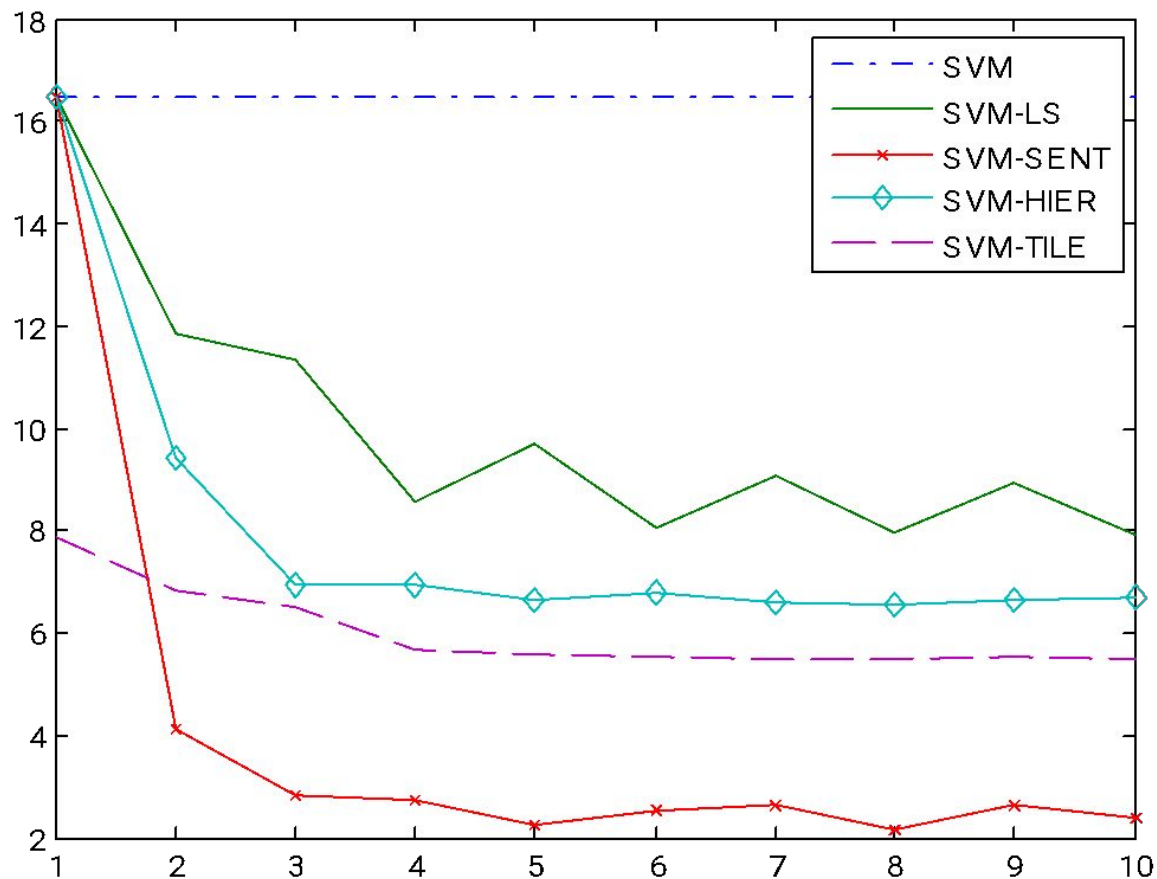
Качество классификации для массива 20 NG от числа итераций



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