МЕТОД ОПРЕДЕЛЕНИЯ МАССОВО ПОРОЖДАЕМЫХ НЕЕСТЕСТВЕННЫХ ТЕКСТОВ

Павлов А.С. Факультет вычислительной математики и кибернетики МГУ имени М.В. Ломоносова

Добров Б.В. Научно-исследовательский вычислительный центр МГУ имени М.В. Ломоносова

План доклада

- Поисковый спам
  - Неестественные тексты
  - Цепи Маркова
- Предлагаемый метод
  - Читаемость
  - Статистические характеристики
  - Авторство
  - Стиль
- Эксперименты
- Заключение

Поисковый спам

- Дорвеи – сайты и страницы, не содержащие полезной информации, созданные для перенаправления пользователей, пришедших с поисковой системы
- Тексты на дорвеях:
  - На популярную тематику
  - Уникальные
  - Должны создаваться массово
  - Трудно-обнаружимы поисковой системой

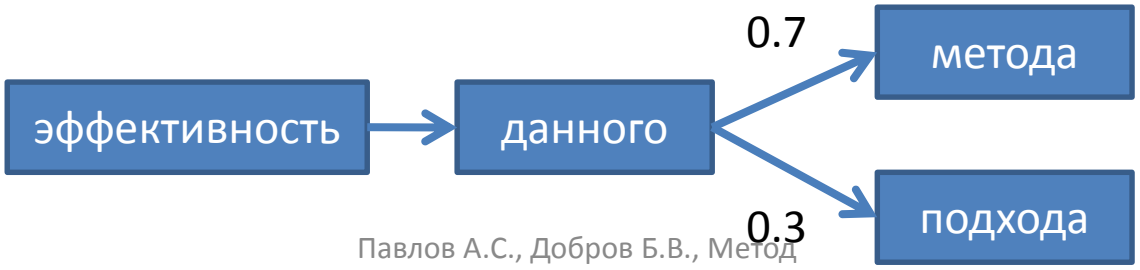
Неестественные тексты

- Спамеры применяют методы автоматического порождения текстов:
  - Цепи Маркова
  - Предложения из различных текстов
  - Вставки запросов в существующие тексты
- Проблемы обнаружения:
  - Локальная связность
  - Сохраняется общая тематика документов

Цепи Маркова

- Порождаем последовательность слов, где каждое слово зависит только от N предыдущих
- Вероятности порождения собираем по тестовой коллекции

- *Проверялась эффективность данного метода опорных векторов позволяет формулировать критерии принадлежности спаму или неспаму.*



Предлагаемый метод

- Естественным текстам свойственны различные уровни связности:
  - Локальная связность
  - Единство стиля
  - Читаемость
- Гипотеза: не существует метода порождения текстов, который не нарушал бы некоторые из этих условий
- Метод обнаружения:
  - Выделяем статистические характеристики текстов
  - С помощью машинного обучения строим автоматический классификатор

Читаемость текстов

- Чем длиннее слова в тексте и длиннее предложения, тем сложнее восприятие текста
- «Сложность» текста можно измерить:
  - Например, индекс Колмана-Лиау:

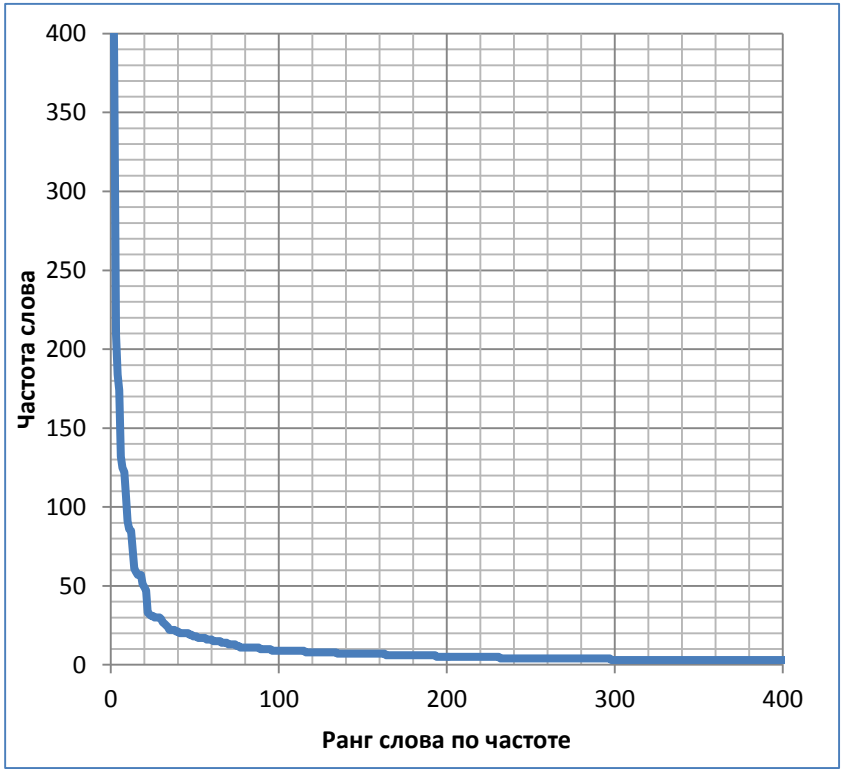
$$R = 5.89\left(\frac{\#символов}{\#слов}\right) - 0.3\left(\frac{\#слов}{\#предложений}\right) - 15.8;$$

  - Применяется в США при оценке уровня владения школьниками письменной речью
- Признаки, связанные с читаемостью:
  - Средняя длина слов в символах/слогах
  - Средняя/максимальная длина предложений
  - И т.п.

Глобальные статистические закономерности

- Естественным текстам свойственны повторы
- Повторы приводят к выполнению глобальных статистических законов:
  - Закон Ципфа
  - Закон Хипса
- Оцениваем отклонение параметров для конкретного текста
- Сжимаемость текста алгоритмами gzip, bz2

Закон Ципфа: $Freq(i) \approx \frac{\alpha}{i^{\beta}};$



Авторство

- Статистические методы определения авторства:
  - Доли частей речи и служебных слов
- Порожденные тексты объединяют характеристики нескольких авторов
- Характеристики:
  - Доля различных частей речи
  - Дисперсия долей по предложениям
- Для определения частей речи использовался парсер mystem

Стилистические характеристики

- Наличие экспрессивной пунктуации и частиц могут указывать на определенный стиль текста:

  «Ну и что же теперь делать?!»

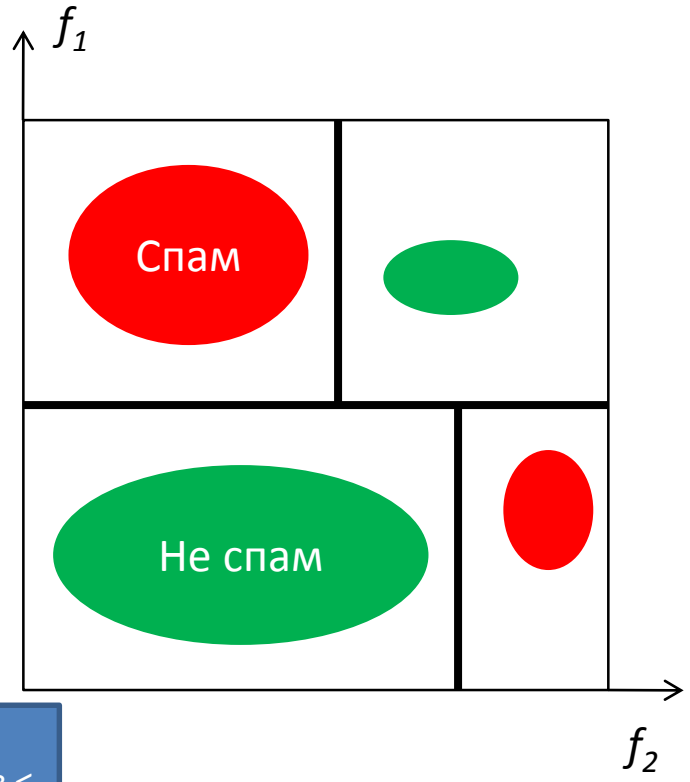
- Собираем статистику употребления:
  - Экспрессивной пунктуации (!,?,:))
  - Редких оборотов и частей речи

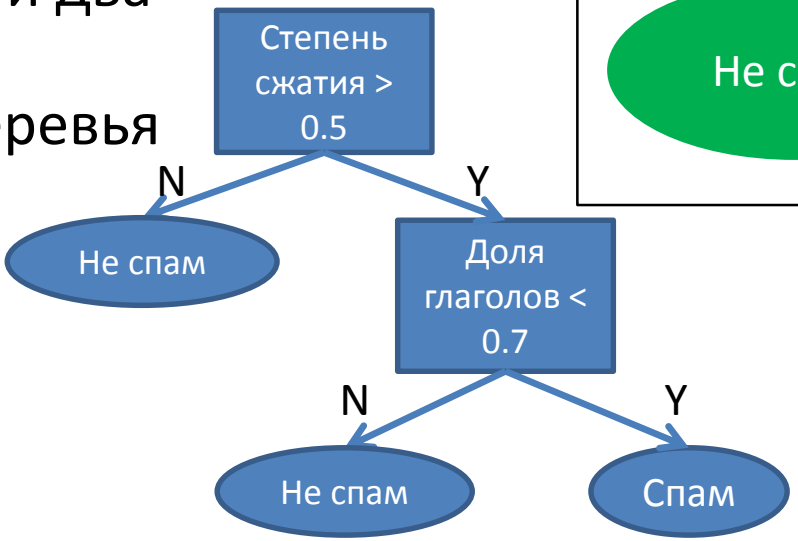
Английский язык

- Все рассуждения справедливы и для английского языка:
  - Вместо mystem применялся Stanford Part-of-Speech Tagger
- Большинство признаков без изменений переносится на другие языки

Машинное обучение

- Гипотеза: классы естественных и неестественных текстов разделимы в пространстве выделенных признаков
- Строим поверхность, разделяющую эти два класса
- Применялись деревья решений





Эксперимент по обнаружению неестественных текстов

- Наборы веб-документов:
  - Romip By.Web
  - WebSpam UK-2007
- На основе наборов были порождены по 10000 спам-документов цепями Маркова длины 2 и 3
- Измерялись точность, полнота и F-мера обнаружения

	Точность	Полнота	F-мера
Русский ЦМ-2	94,98%	95,71%	95,34%
Русский ЦМ-3	91,56%	95,02%	93,25%
Англ. ЦМ-2	96,19%	96,11%	96,15%
Англ. ЦМ-3	94,08%	92,29%	93,18%

Чем больше порядок цепи Маркова, тем бóльшие куски текста дублируются

Эксперимент по оценке силы признаков

- Признаки были распределены в 4 группы:
  - Характеристики разнообразия (Ципф, степень сжатия, …)
  - Глобальные статистические характеристики (средняя длина слов, пунктуация, …)
  - Статистика употребления частей речи (доля глаголов, …)
  - Редкие части речи, встречающиеся менее чем на 1% слов (доля местоименных наречий, …)
- Исследовалась возможность классификации с использованием только отдельных признаков и групп признаков

Эксперимент по оценке силы признаков

Английский язык

№	Название признака	F-мера, %	Тип признака
1	Степень сжатия gzip	**89,70**	Разнообразие
2	Степень сжатия bz2	**85,04**	Разнообразие
3	Параметр распределения Ципфа для существительных	**81,28**	Разнообразие
4	Доля слов повторяющихся в соседних предложениях	79,60	Разнообразие
5	Доля глаголов в прошедшем времени	74,49	Части речи

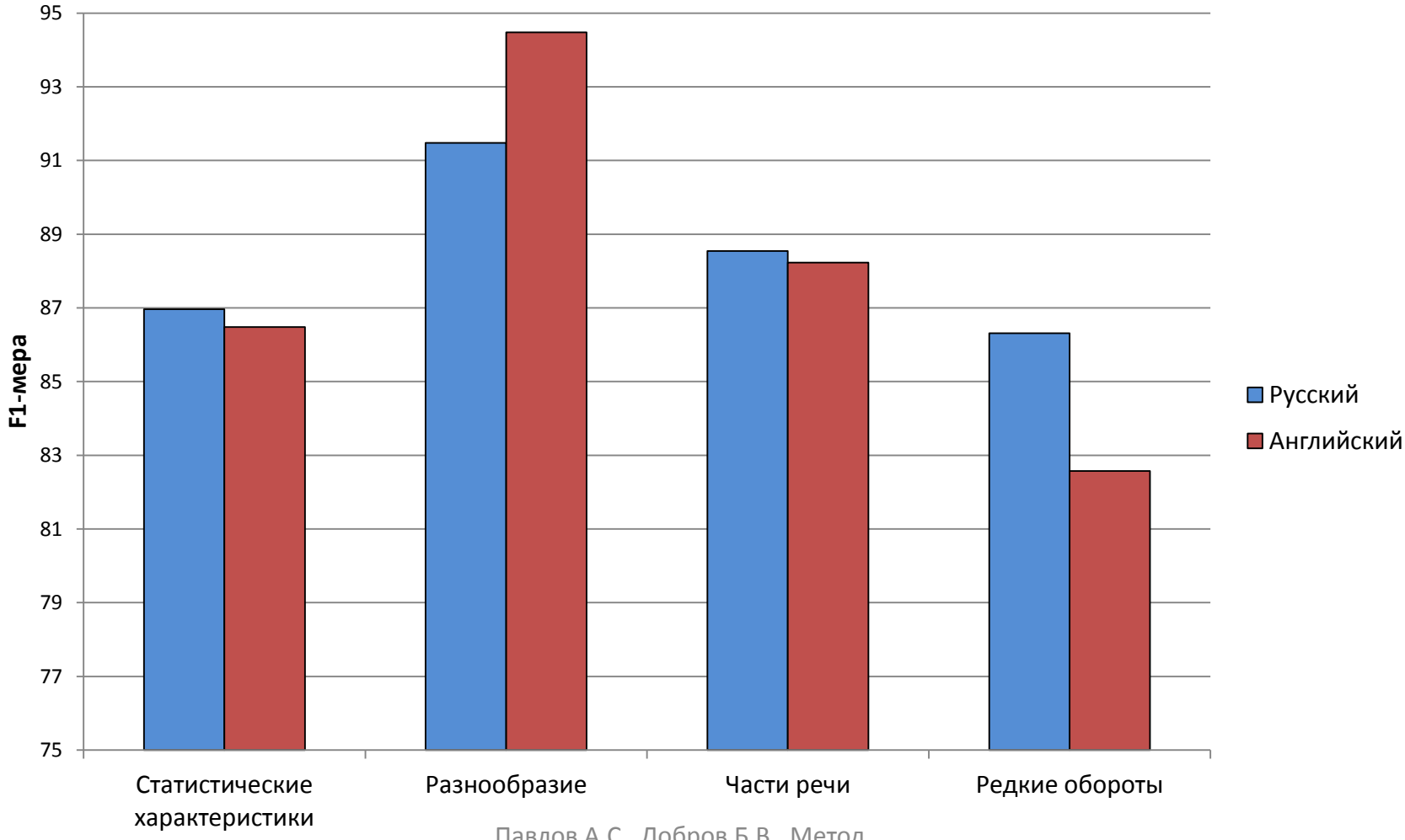
Русский язык

№	Название признака	F-мера, %	Тип признака
1	Степень сжатия gzip	78,87	Разнообразие
2	Степень сжатия bz2	77,92	Разнообразие
3	Параметр распределения Ципфа для существительных	77,67	Разнообразие
4	Дисперсия доли местоименных наречий по предложениям	75,64	**Редкие**
5	Дисперсия доли междометий по предложениям	75,23	**Редкие**

Метрики разнообразия дают больший вклад при классификации англоязычных текстов

Статистика употребления редких частей более важна дл русского языка

Эксперимент по оценке вклада групп признаков



Заключение

- Разработанный метод позволяет обнаруживать неестественные тексты
- Алгоритм можно адаптируется для английского языка:
  - Подготовка нового тренировочного набора
  - Обучение классификатора
- Вклад групп признаков зависит от языка
  - Устойчивость алгоритма к попыткам его обойти также зависит от языка

Литература

1. Павлов А.С., Добров Б.В. Методы обнаружения поискового спама, порожденного с помощью цепей Маркова // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск: 2009.
2. Ntoulas A., Manasse M., Detecting spam web pages through content analysis // In Proceedings of the World Wide Web conference, ACM Press, 2006. p. 83-92
3. Piskorski, J., Sydow, M., Weiss, D., Exploring Linguistic Features for Web Spam Detection: A Preliminary Study // In Proceedings of the 4th international workshop on Adversarial Information Retrieval on the Web, Beijing, China, 2008. p. 25-28.
4. Гречников Е.А., Гусев Г.Г., Кустарев А.А., Райгородский А.М., Поиск неестественных текстов // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2009, Петрозаводск: 2009.
5. Mishne, G., Carmel, D., and Lempel, R. Blocking blog spam with language model disagreement // In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, 2005.
6. Urvoy T., Chauveau E., Filoche, P. Tracking Web Spam with HTML Style Similarities // ACM Transactions on the Web, 2006. Vol. 2, n.1, Article 3.
7. Castillo C., Donato D., Murdock V., Silvestri F., Know your neighbors: Web spam detection using the web topology // In Proceedings of SIGIR, ACM, 2007.
8. Dubay W.H. The Principles of Readability // Costa Mesa, CA: Impact Information, 2004.
9. Фоменко В.П., Фоменко Т.Г., Авторский инвариант русских литературных текстов // В сб.: Методы количественного анализа текстов нарративных источников. - М.: АН СССР, Ин-т Истории СССР, 1983. с.86-109.
10. Парсер mystem (http://company.yandex.ru/technology/mystem/).
11. Braslavski P. Document Style Recognition Using Shallow Statistical Analysis // Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP, Nancy, France, 2004. p.1–9.
12. Quinlan J. R. C4.5: Programs for Machine Learning // Morgan Kaufmann Publishers, 1993.
13. Stanford Log-linear Part-Of-Speech Tagger (http://nlp.stanford.edu/software/tagger.shtml).
14. Marcus M.P., Marcinkiewicz M.A., Santorini B., Building a Large Annotated Corpus of English: the Penn Treebank // Computational Linguistics, 1993. Vol.19 n.2
15. Веб коллекция BY.Web, http://romip.ru/ru/collections/by.web-2007.html.
16. Yahoo! Research: "Web Spam Collections". (http://barcelona.research.yahoo.net/webspam/datasets/), Crawled by the Laboratory of Web Algorithmics, University of Milan, (http://law.dsi.unimi.it/).
17. Gelbukh A., Sidorov G. Zipf and Heaps Laws' Coefficients Depend on Language // Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, 2001. p.332-335.
18. Зеленков Ю.Г., Сегалович И.В., Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль: 2007.