Нижегородский государственный технический университет им. Р.Е. Алексеева
Институт радиоэлектроники и информационных технологий

Кафедра: «Вычислительные системы и технологии»

Выпускная квалификационная работа

Программная система детектирования спам сообщений

Выполнил: Егоров А.Д., 15-В-1

Научный руководитель: к.т.н., доцент Гай В.Е.

Нижний Новгород
2019 г.

Цель и задачи исследования

- Цель:
  - Разработка программной системы детектирования спам сообщений

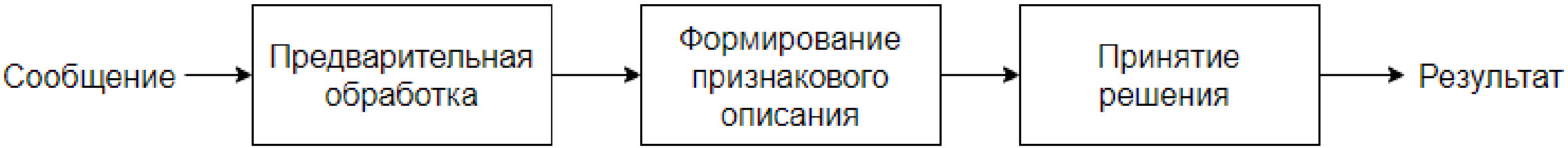
- Задачи:
  - Обзор существующих подходов детектирования спам сообщений
  - Разработка алгоритма классификации сообщений
  - Разработка программной реализации алгоритма
  - Проведение вычислительного эксперимента для установления корректности работы созданной системы

Объект и предмет исследования

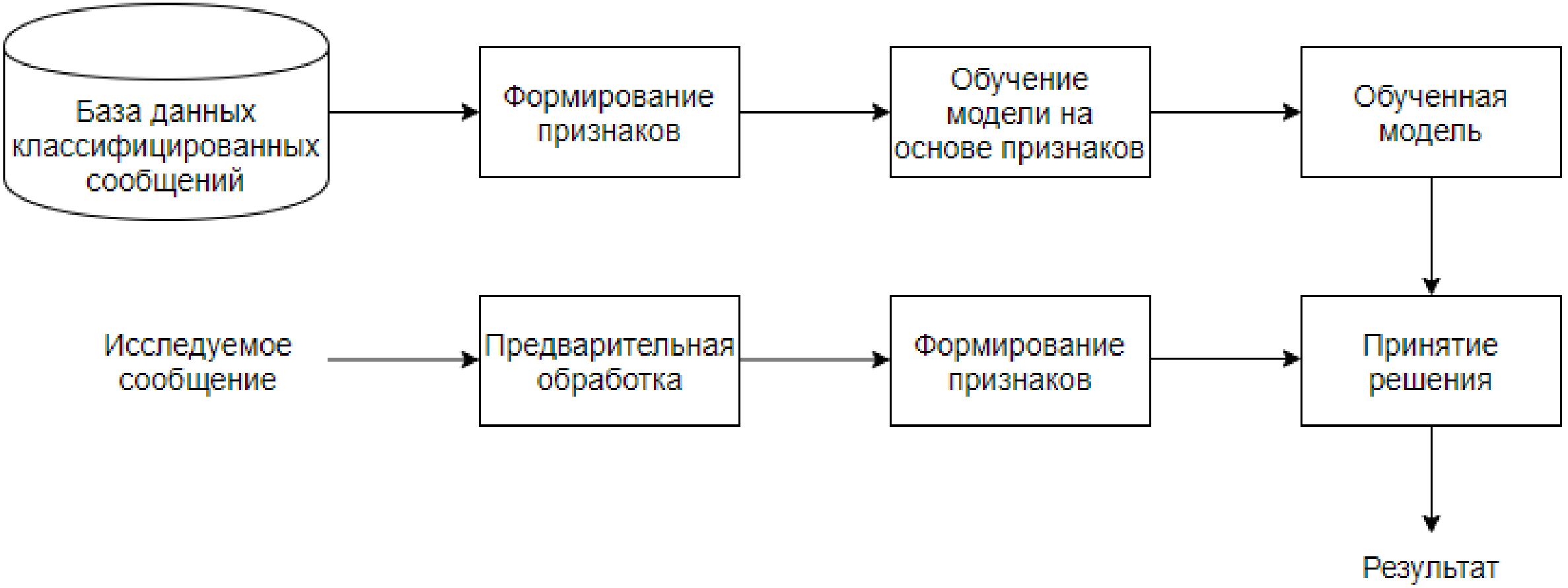
- Объект исследования:
  - Текстовые сообщения почтового трафика

- Предмет исследования:
  - Алгоритмы и методы классификации спам-сообщений

Этапы решения задачи



Информационная модель детектирования спам сообщений



Предварительная обработка сообщения

- Замена всех символов табуляции и тире на символ пробела
- Очистка сообщения от пунктуационных знаков
- Очистка сообщений от слов, которые не несут смысловой нагрузки (размером 2 и менее символов)

Признаковое описание

- На основе базы классифицированных и обработанных сообщений, создается словарь, включающий в себя слова и количество их вхождений в сообщения
- На основе словаря и базы классифицированных сообщений, создается частотный список, который содержит информацию о том сколько раз слово из словаря встречается в сообщениях и к какому классу эти сообщения относятся.

$$(\boldsymbol{x_1}, \boldsymbol{y_1}), \dots, (\boldsymbol{x_m}, \boldsymbol{y_m}), \quad \boldsymbol{x_i} \in \boldsymbol{R^n}, \quad \boldsymbol{y_i} \in \{-\mathbf{1}, \mathbf{1}\}$$

- Обучение модели на основании признаков полученных выше

$$\boldsymbol{F(x) = sign(\langle w, x \rangle + b)}$$

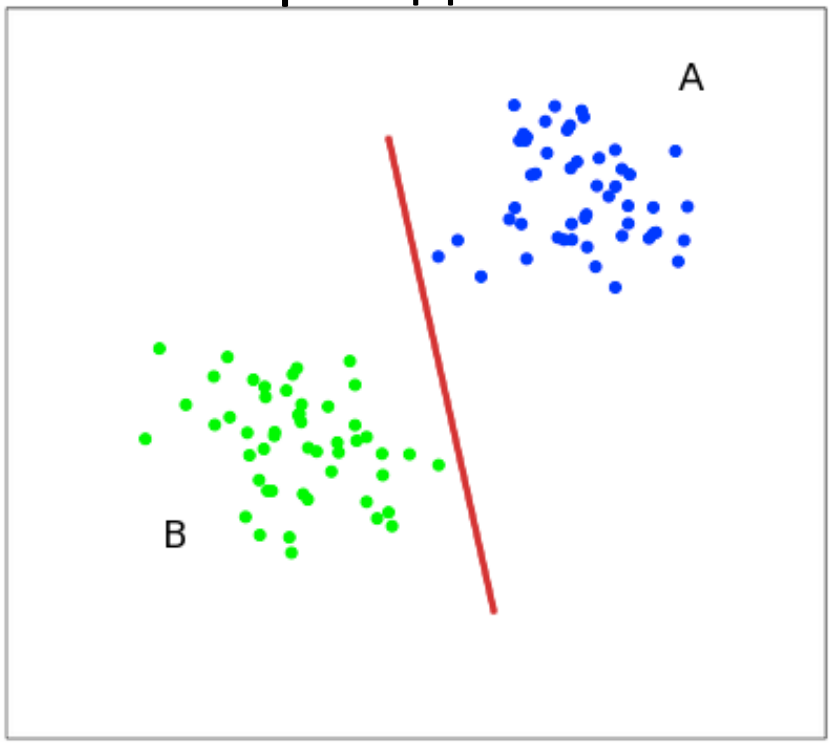
Принятие решения

- Признаковые описания классифицированных сообщений формируют два класса: «спам» и «не спам»
- Классификатор однозначно относит признаковое описание исследуемого сообщения к одному из классов
- Решение принимается на основе положения признака относительно гиперплоскости разделяющей классы

Метод опорных векторов

Основной задачей данного метода является поиск такой гиперплоскости размерности (**p**-1), расстояние от которой до ближайшей точки было бы максимальным.

Итак, математическая формулировка задачи классификации такова: пусть **X** — пространство объектов (например, R^n), ***Y*** — наши классы (например, ***Y*** = {-1,1}). Дана обучающая выборка $(x_1,y_1),\ldots,(x_m,y_m)$. Требуется построить функцию F: X -> Y (классификатор), сопоставляющий класс ***y*** произвольному объекту ***x.***



Вычислительный эксперимент

- База классифицированных сообщений: 5172 сообщения
  - 3672 – не спам сообщения
  - 1500 – спам сообщений

http://www2.aueb.gr/users/ion/data/enron-spam/

- Обучающая выборка: 3620 сообщений
- Выборка для тестирования: 1552 сообщения

- Программный продукт разработан в PyCharm (интерпретатор Anaconda) на языке Python 3.7
- Программные модули используемые при разработке:
  - NumPy
  - Pandas
  - MatPlotLib

Тестирование системы

		Верная гипотеза	
		«Не спам»	«Спам»
Результат применения критерия	«Не спам»	1049 верно принятых	29 неверно принятых (Ошибка второго рода)
	«Спам»	69 неверно отвергнутых (Ошибка первого рода)	405 верно отвергнутых

Итоги

- Выполнена разработка системы
- Проведено тестирование и отладка

Публикации

- В.Е. Гай, А.Е. Егоров Программная система детектирования спам-сообщений // Труды XXV Международной конференции «Информационные системы и технологии» ИСТ-2019, 19 апреля 2019 г., С. 849-852

Спасибо за внимание!