

# МЕТОДЫ ОБНАРУЖЕНИЯ ПИСЕМ- ТРАНСФОРМЕРОВ

---

Ермакова Лиана

# Понятие спама

- **Спам** - это анонимные незапрошенные массовые рассылки электронной почты (Лаборатория Касперского)
- Но:
  - Спам в социальных сетях
  - Спам в IM

# Методы борьбы со спамом

- Black list
- White list
- Grey list
- Анализ заголовков
- Байесовская фильтрация по словам
- Генетические алгоритмы и ручное выставление весов
- Обнаружение повторов и признаков массовости
- Интегрирующие системы

# Сигнатурные подходы

- Синтаксические
  - Оперируют цепочками слов
  - «Шинглы»:
    - вычисление контрольных сумм для всех подцепочек текста
    - построение случайной выборки из полученного набора
- Лексические
  - Оперируют словарем
  - Метод опорных векторов

# Сообщения-трансформеры

- Сообщения, имеющие сходное содержание, но различные по форме
- Каждое отдельное письмо выглядит как обычный связный текст, и, только имея много копий сообщения, можно установить факт перефразировки

# Классификация спама

- По структуре:
  - спам, замаскированный под личную корреспонденцию
  - спам, замаскированный под легальные массовые рассылки
  - рекламный спам
- По тематике:
  - Нигерийские письма
  - Цепочечные письма
    - «страшилки»
    - письма счастья
  - Быстрый заработок
  - Реклама
    - Программное обеспечение
    - Медикаменты
    - Образование
    - Финансы
    - Страхование...



# Алгоритм выявления писем-трансформеров

- На основе количественных характеристик с применением машины опорных векторов новое сообщение относится к той или иной категории
- В качестве уточняющего признака используется триграммное сходство с учетом расстояния Дамерау-Левенштейна и выявленных правил замены символов

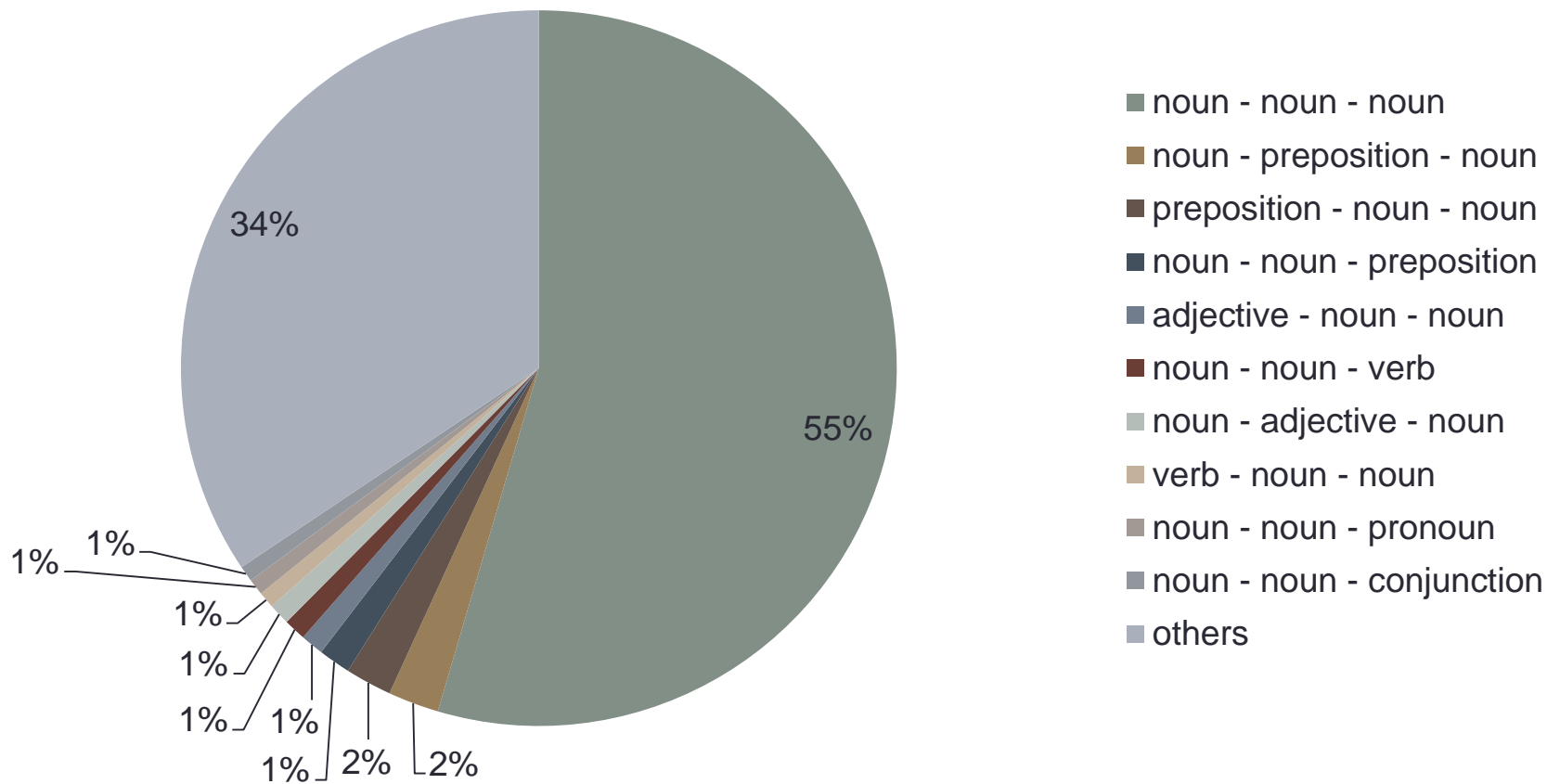


# Квантитативные характеристики

- доля полнозначных и служебных слов
- доля предложений, слов и абзацев определенной длины
- доля вхождения каждой части речи (краткие и полные формы прилагательных и причастий мы считали различными частями речи)
- количество знаков препинания
- встречаемость частей речи
- и т.д.
- **Общее число признаков – 135**

# Доли последовательностей частей речи

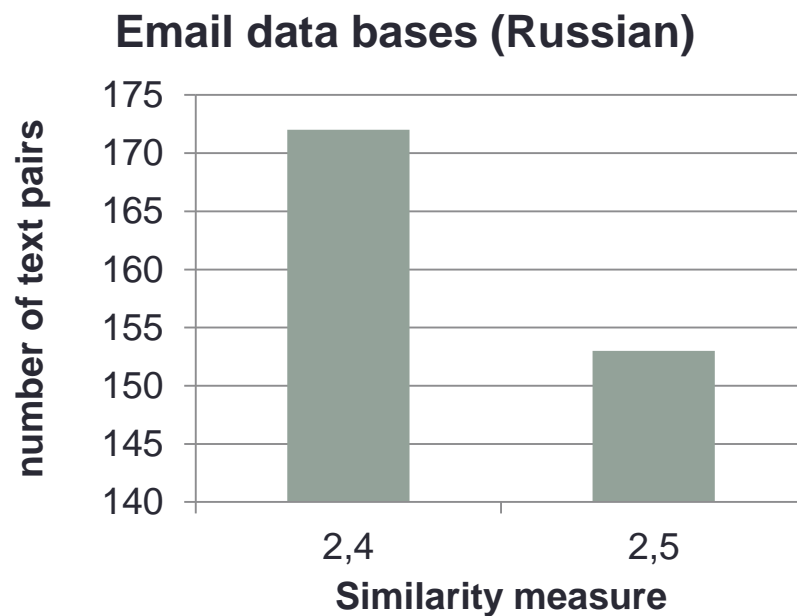
POS trigram share



# Email базы

- **sWEVIE email BAZY**  
pRODAVA BAZ email  
ADRESOW (ADRESA  
DLQ email  
RASSYLOK) <...>
- **aDRESA DLQ email  
RASSYLOK**  
pRODAVA BAZ email  
ADRESOW (ADRESA  
DLQ email  
RASSYLOK) <...>

Мера сходства,  
вычисленная при помощи  
триграмм



# ЕГРЮЛ

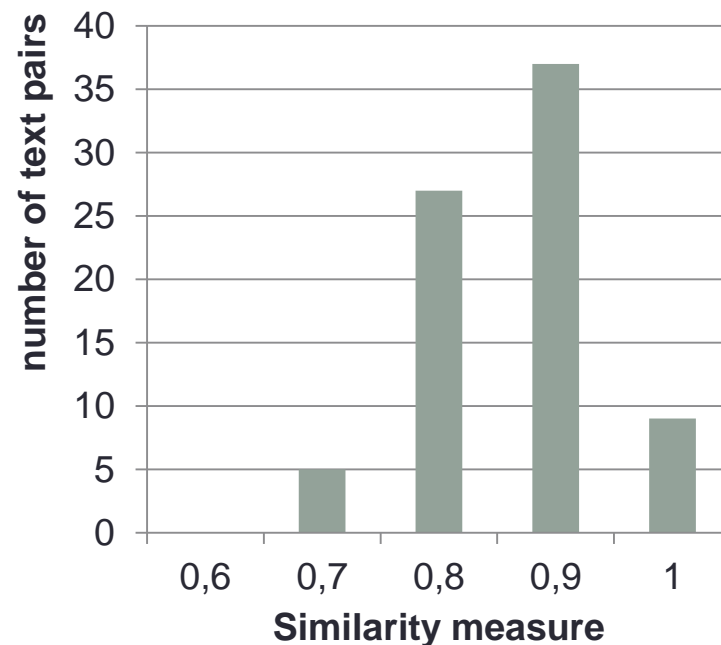
В начале года всегда возникает **потребность** в "свежих" выписках ЕГРЮЛ и справках Госкомстата. Предлагаем Вам: получение выписки ЕГРЮЛ за **1,200 рублей** справки Госкомстата за **1 200 руб.** заказ выписки ЕГРЮЛ + справки Госкомстата составит всего **2.000 рублей** Доставка курьером, оплата по факту.  
**Контактная информация + 7495 222+07.68**

В начале года всегда возникает **необходимость** в "свежих" выписках ЕГРЮЛ и справках Госкомстата. **Мы** предлагаем Вам: получение выписки ЕГРЮЛ за **1 200 рублей** справки Госкомстата за **1 тыс. 200 р.** заказ выписки ЕГРЮЛ + справки Госкомстата составит всего **2 тыс. 000 руб-й.** Доставка курьером, оплата по факту.  
**Телефон: + 7495 222\_07;68**

В начале года всегда возникает **потребность** в "свежих" выписках ЕГРЮЛ и справках Госкомстата. **Мы** предлагаем Вам: получение выписки ЕГРЮЛ за **1 тыс. 200 руб-й** справки Госкомстата за **1 200 рублей.** заказ выписки ЕГРЮЛ + справки Госкомстата составит всего **2,000 р.** Доставка курьером, оплата по факту. **Контакты + 7(495) 222-07-68**

Мера сходства,  
вычисленная при  
помощи триграмм

Unified State Register  
of Legal Entities (Russian)



## *Параметры машины опорных векторов для определения писем-трансформеров на русском языке*

- Sample size = 707 (Train), 236 (Test), 943 (Overall)
- Support Vector machine results:
- SVM type: Classification type 1 (capacity=10,000)
- Kernel type: Radial Basis Function (gamma=0,007)
- Number of support vectors = 118 (0 bounded)
- Support vectors per class: 94 (0), 16 (1), 8 (2)
- Class. accuracy (%) = 100,000(Train), 100,000(Test), 100,000(Overall)

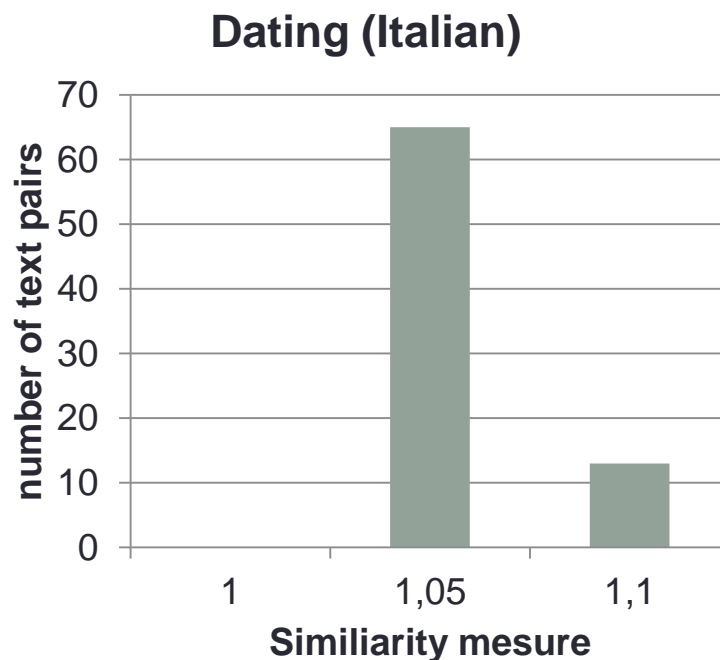
# Знакомства

<...> La preghiamo di rispondere solo alla mia personale e-mail: **khhaykanush**@yahoo.com Tua amica Haykanush.

<...>La preghiamo di rispondere solo alla mia personale e-mail: **haykanusharm**@yahoo.com Tua amica Haykanush.

<...>La preghiamo di rispondere solo alla mia personale e-mail: **khaykanush**@yahoo.com Tua amica Haykanush.

Мера сходства,  
вычисленная при помощи  
триграмм

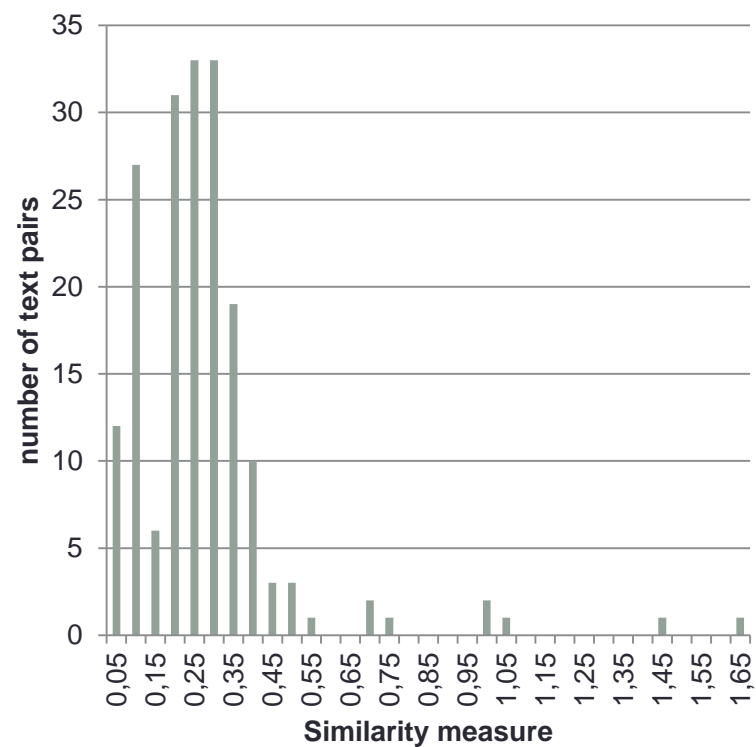


# Медикаменты

Тема	Текст
<b>Desire to impress and please your lover tonight</b>	The only bluepill you need to get bigger python. <a href="http://wanzulkifli.com/c6ave6lc.html">http://wanzulkifli.com/c6ave6lc.html</a>
<b>Gain in size and win your wife's addiction</b>	Desire to act like a pornstar? Bang a magicpilule! <a href="http://bpyasociados.com.ar/9vh6w3lf.html">http://bpyasociados.com.ar/9vh6w3lf.html</a>
<b>Wish to act like a porn-director Nail a blu colored med!</b>	0% amorous failure risk <a href="http://mikloswowmobile.com/uaagzeib.html">http://mikloswowmobile.com/uaagzeib.html</a>
<b>Dream to act like a porn-director Bang a blu colored pill!</b>	Long manliness is great <a href="http://antalyagunlugu.com/d4zz8qan.html">http://antalyagunlugu.com/d4zz8qan.html</a>

Мера сходства,  
вычисленная при помощи  
триграмм

Medicine (English)



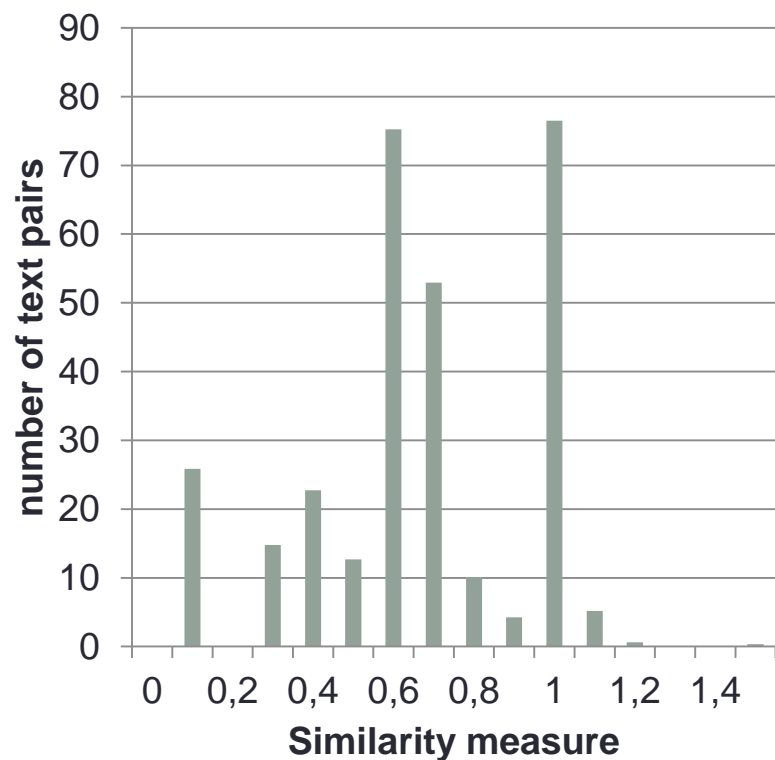
# Казино

Тема	Текст
<b>Comme Faire _200 de _20 - nous APPRENDRONS</b>	Bonne journee Jessikaparsons, { <a href="http://yxaqih983.o-f.com/kerizev.html">http://yxaqih983.o-f.com/kerizev.html</a> } Accueillez la fortune dans votre vie avec de grandes opportunit�s de gagner, avec l'assurance que vos informations personnelles sont prot�g�es et vos gains seront pay�s rapidement. Une demi-heure et �200 dans ta poche
<b>Gagner _100 pour une demi-heure c'est r�el</b>	Du jour reussi Shirley_patel, { <a href="http://gamingworldshop.ru">http://gamingworldshop.ru</a> } Il y a de grandes promotions auxquelles vous pouvez participer et qui vous promettent encore plus de plaisirs et de fa�ons de gagner. Faire �100 pour une demi-heure - Apprendre?
<b>Faire -100 pour une demi-heure Apprendre</b>	Bonne journee Nvshamshik, { <a href="http://beluwulod.maddsites.com/abimogek.html">http://beluwulod.maddsites.com/abimogek.html</a> } Il y a de grandes promotions auxquelles vous pouvez participer et qui vous promettent encore plus de plaisirs et de fa�ons de gagner. Gagner -100 pour une demi-heure c'est r�el
<b>Jouer ici, c'est le bonheur ! Telechargez maintenant</b>	{ <a href="http://opakypiwel.dreamstation.com/jededila.html">http://opakypiwel.dreamstation.com/jededila.html</a> } On ne peut pas faire plus simple, il suffit de vous inscrire, de faire un versement et vous recevez un fantastique bonus de bienvenue - alors foncez et gagnez ! La meilleure selection de jeu sur internet ! Jouez ici
<b>Jouez plus, gagnez plus</b>	Salut Shea.swan Des options bancaires s�res qui conviendront a tous sont disponibles. Relaxez-vous et soyez certains que vos informations confidentielles sont s�curis�es et ne seront p� #97;s divulgu�es. { <a href="http://durl.me/554k6">http://durl.me/554k6</a> } Comment aimeriez-vous commencer au mieux dans le jeu en ligne avec 1,200 Gratuits? Ils sont d�ja a vous, r�clamez-les, jouez et gagnez!

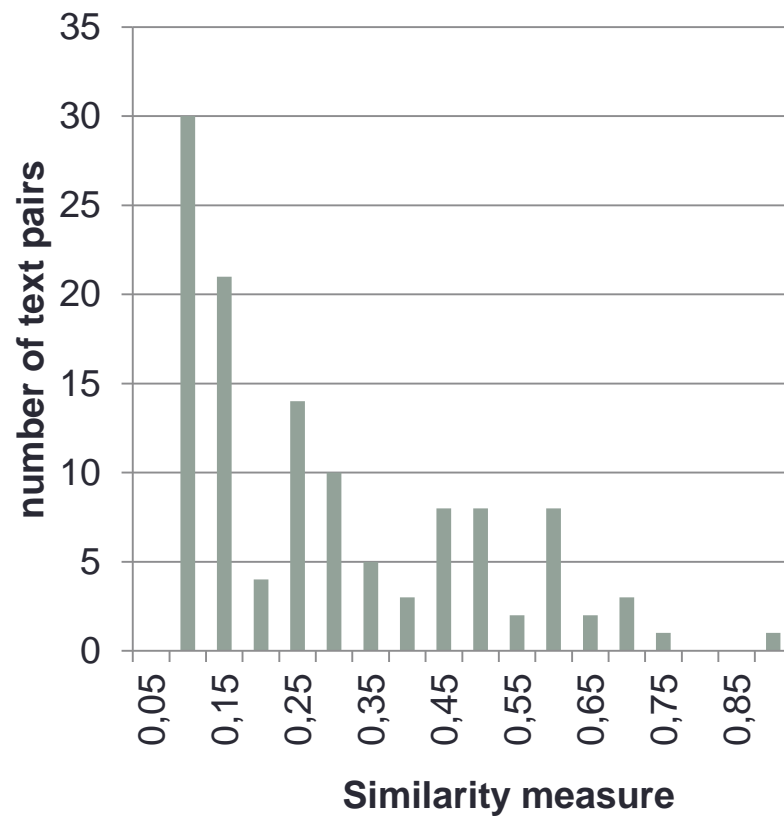


# Другие примеры

## Forex (French)



## Cigarette (French)



# Выводы

- Контентная фильтрация применима для различных систем обмена сообщениями (электронная почта, IM, социальные сети), т.к. она не опирается на служебную информацию
- Особо рассматривается проблема обнаружения намеренно искаженных фрагментов
- Предлагается метод детектирования массовых сообщений, фильтрация которых затруднительна из-за меняющегося контента. Рассматриваются возможности увеличения производительности за счет использования методов Монте-Карло
- Предложенный двушаговый метод обнаружения писем-трансформеров в почтовом потоке является улучшением сигнатурного метода

Спасибо за внимание!