



Относительная релевантность сообщений в информационных корпусах спама и СМИ

*Дмитрий Владимирович ЛАНДЭ,
ЦИ “ЭЛВИСТИ”,
Андрей Александрович СНАРСКИЙ,
НТУУ “КПИ”*

Задачи

- **Достаточно точное выявление спама**
- **Извлечение полезной информации**

α -подобие сообщений

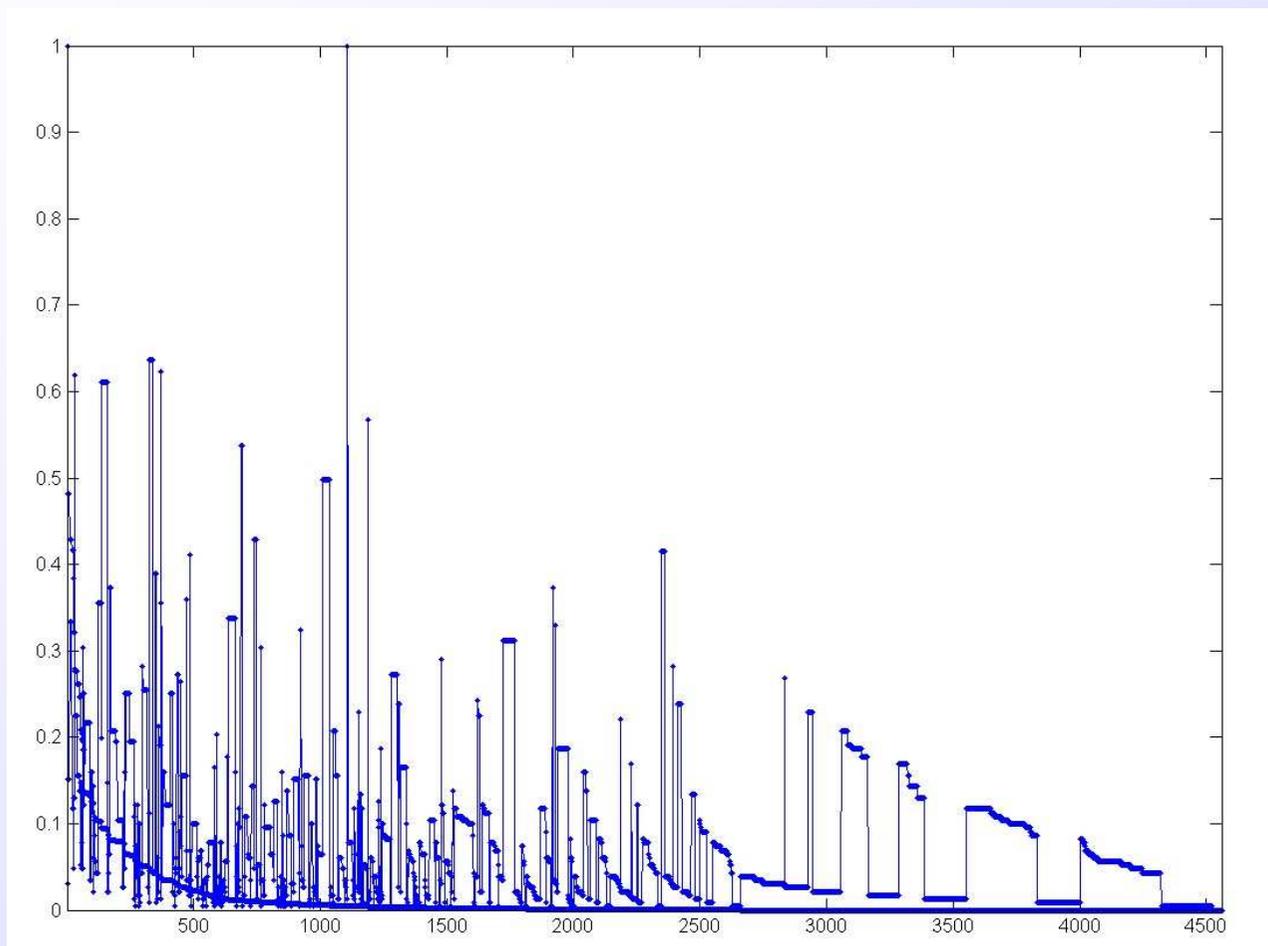
В системе *InfoStream* сообщение считается подобным исходному, если содержит некоторое количество (α) значимых слов из него (α -подобие)

Спам- и СМИ-популярность

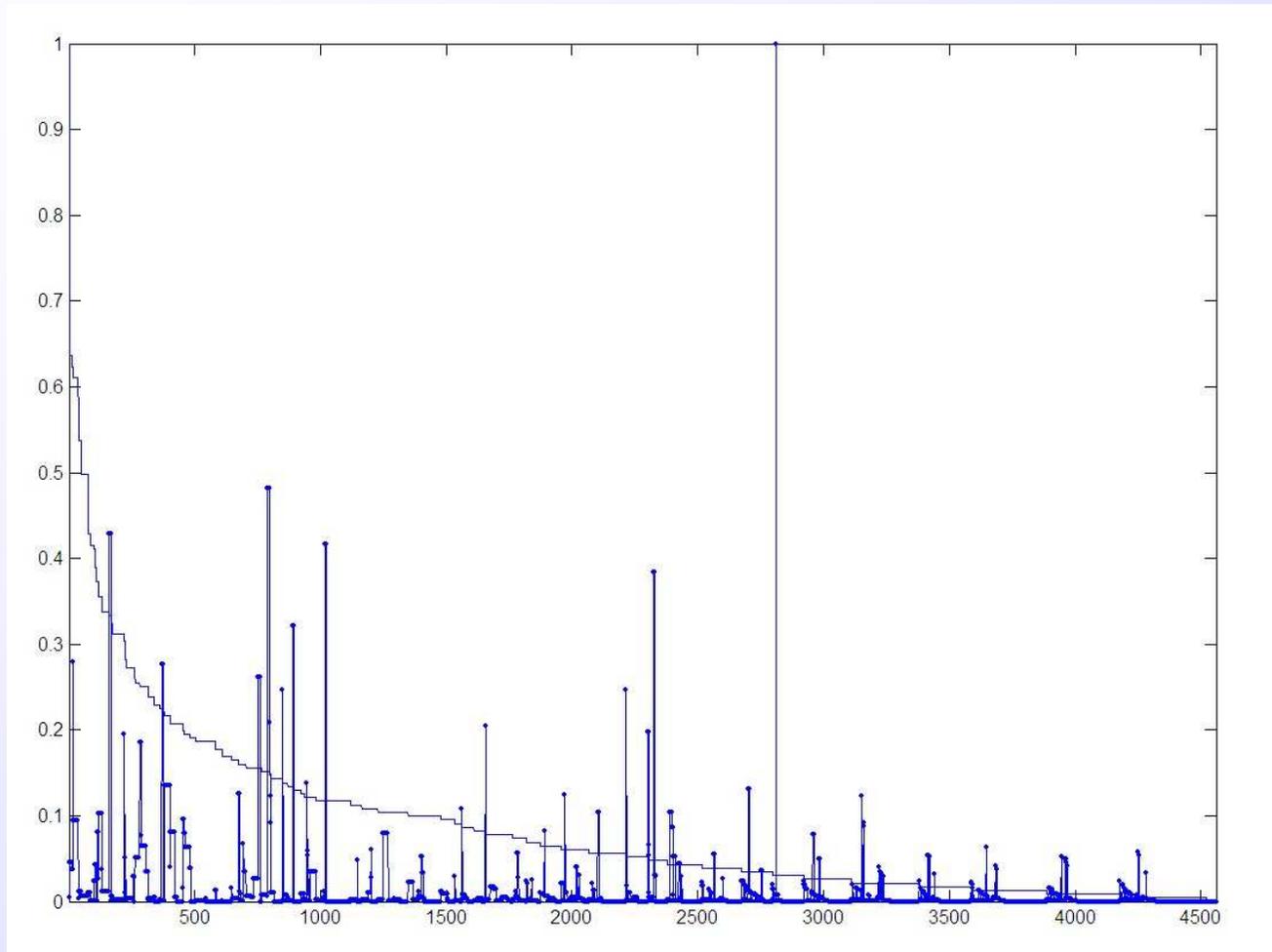
Под спам-популярностью сообщения будем понимать количество α -подобных ему сообщений в текстовом корпусе спама.

Под СМИ-популярностью понимается количество α -подобных сообщений в ретроспективной базе электронных СМИ.

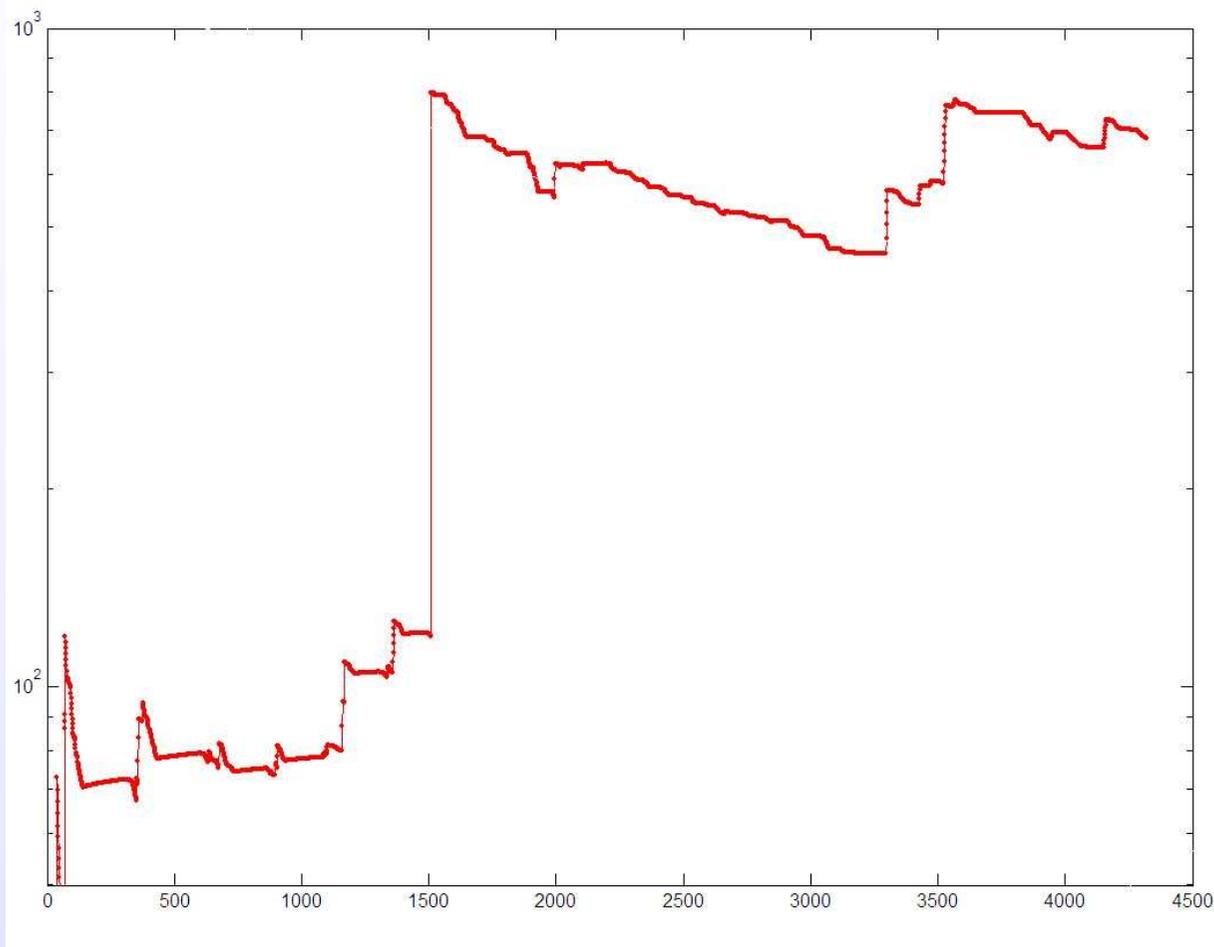
Массив, ранжированный по СМИ-популярности



Массив, ранжированный по спам-популярности



Показатель Фано



Словосочетания пиковой спам-популярности

- ! КИЕВ КУРЬЕР МОСКВА НАЛИЧНЫЕ ОПЛАТА
- РУБ АРЕНДА БАНЯ ДОМ КВ.М
- ДИЗАЙН ПЕЧАТЬ БУМАГА ТАРИФНЫЙ ЭКЗ
- ! КИЕВ КУРЬЕР НАЛИЧНЫЙ ОПЛАТА ОФИС
- ! РУБ МОСКВА РУБ РОССИЯ ВСЯ
- ДЕФЕКТ ЖИР НАКОПЛЕНИЕ ПОХУДЕТЬ ПРИЧИН
- ! БЕСПЛАТНО МОСКВА ПРЕДОСТАВЛЯЕМ РАССЫЛКА БУХГАЛТЕРСКАЯ
- СЕМИНАР ИНТУИЦИЯ КИЕВ НЕОБХОДИМО ВИДЕН
- ! АДРЕС РУБ ФИЗ.ЛИЦО ЮР.ЛИЦО МОСКВА
- ! САЙТ GOOGLE RAMBLER YANDEX ПРОДВИЖЕНИЕ
- ! БАЗА БЕСПЛАТНО БЫСТРО ДЕШЕВО ДОВЕРЬТЕ
- ! РАССЫЛКА МАКЕТ РЕКЛАМНАЯ АДРЕС АНОНИМНАЯ
- ОБЪЕКТ БТИ ВАШ ВНУТРЕННИМ ДНЕПРОПЕТРОВСК
- ! ЛИСТОВКА ПРОДУКЦИЯ РЕКЛАМНАЯ БУМАГА ПЛОТНОСТЬ
- ! РАССЫЛКА МАКЕТ БАЗА ВАШ ФИЛЬТР

Спасибо за внимание!

***Дмитрий Владимирович ЛАНДЭ,
Информационный центр «ЭЛВИСТИ»,***

***Андрей Александрович СНАРСКИЙ,
НТУУ «КПИ»***

+380 (44) 239-9091

stream@visti.net

<http://infostream.ua>