Шаблоны (в … е-мейлах) и как их найти

История

Александр Слипченко
Booking.com B.V.

О чем будет доклад?

- TemplateFingerprint
  - В чем проблема?
  - «мыслительный процесс» и решение
  - Как довести до продакшена
  - Что со всем этим делать дальше?
- Вопросы (и, возможно, даже ответы ☺)
- «О космических кораблях, бороздящих просторы …» (на самом деле: о том, где и для чего еще у нас практикуется AI / ML / etc.)

Проблема

- Много почты (не как у Гугла, но все-равно много – сотни тысяч сообщений в день, каждый день)
- Фильтровать по отправителю не получается
- Большая часть почты – «хорошая», но есть и «плохая»
  - Более-менее безобидный спам
  - «Серая зона»
  - Фишинг всех цветов и оттенков
- «Теряем время (и деньги)»

Вот как это выглядит (фишинг)

Dear Jonas,

Please remit the 20% deposit (of 782 EUR total, all inclusive) for Sesimbra Apartment via **this link** (http://tinyurl.com/[redacted]) to make your reservation (booking.com 275291823, OLB279571) final (and to receive check-in instructions - make sure you get them)! You have 24hrs to complete this payment!

Your credit card has not been charged!

We are looking forward to having you stay with us!

Yours

Casa Sesimbra

А что там пишут в интернетах?

- Что-нибудь вроде soundex, pHash …
- Определение «неменяющихся кусков»
  - «мы выгружаем с сайта страницу, как бы нам вытянуть из нее только полезные для нас данные»
- Анализ метаданных (заголовки письма)
- Спам-фильтры (чаще всего некий классификатор + много эвристики)

А чего хотим мы

- Чтобы работало (хорошо и быстро ☺)
  (aka «Нажал на кнопку – и … магия»)
- «Хорошо»
  - Приемлемое количество ложных срабатываний
  - Понятно для для пользователя (не слишком сложно)
  - Работает для всех / любого языка
- «Быстро»
  - Не слишком требовательно к вычислительным ресурсам
    - Обработка сообщений
    - Поиск \ группировка

Так что будем делать?

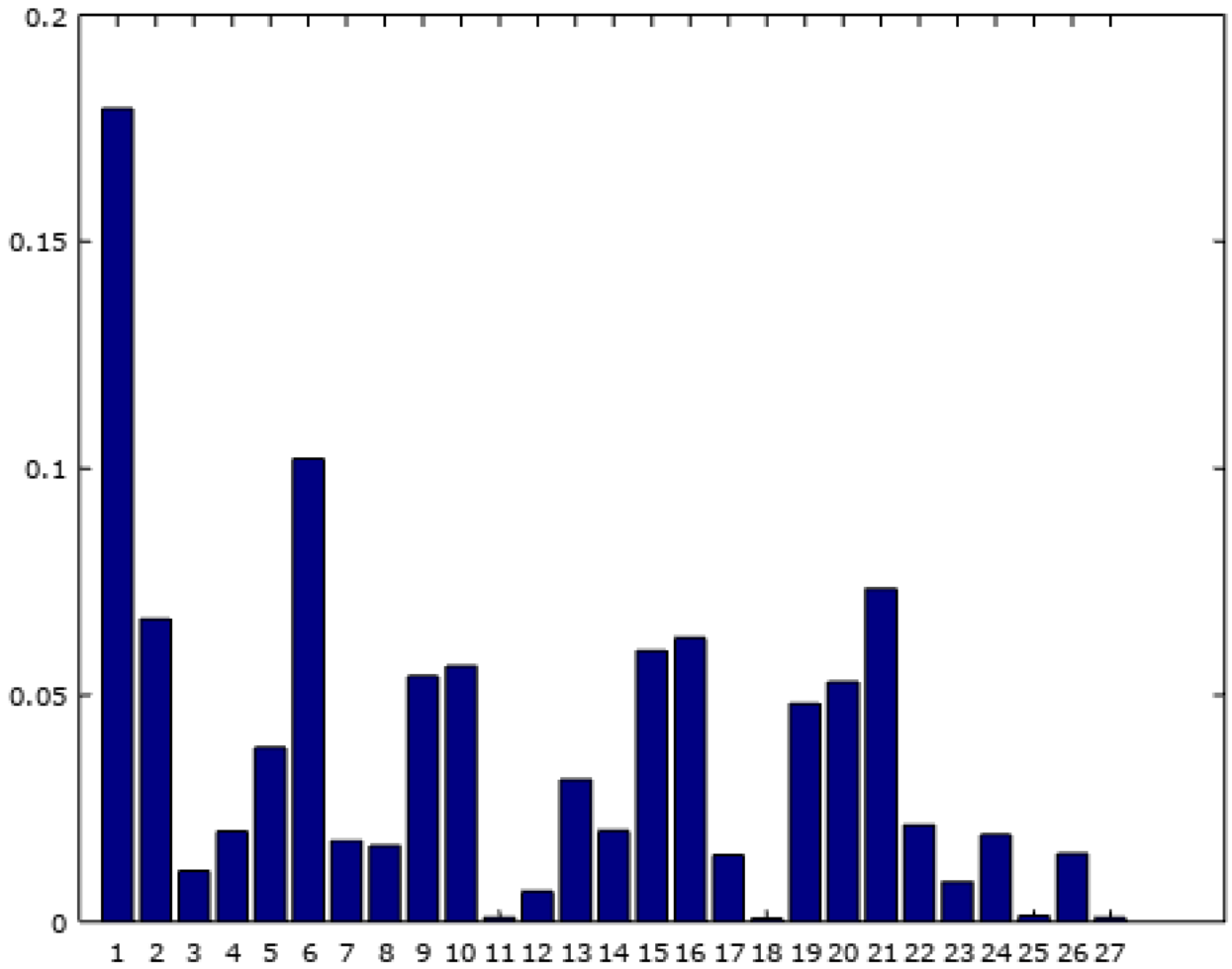
- Что-нибудь … эдакое ☺
- Какая-нибудь частотная характеристика …
- N-gram ?
  - В какой форме?
  - Как сравнивать (муторно …)
  - Как хранить (неудобно …)
  - А это будет работать?

N-gram(s) + частота

- unigram – будет мерять «среднюю температуру по больнице» (т.е. язык сообщения)
- bi-gram
- …
- infini-gram
- Символы или лексемы

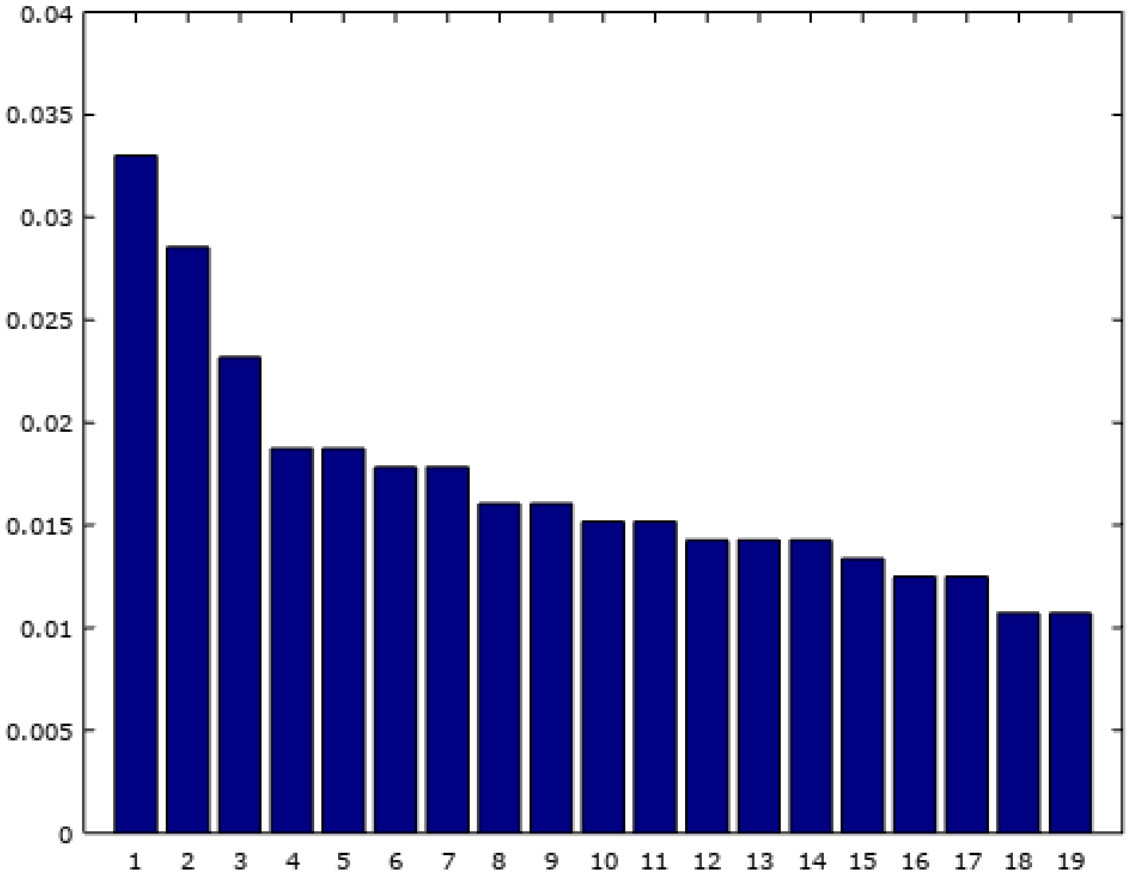
Предварительные мысли

```
  =  0.17940
a =  0.066678
b =  0.011244
c =  0.019882
d =  0.038392
e =  0.10219
f =  0.017816
g =  0.016657
h =  0.054213
i =  0.056392
j =    8.3571e-04
k =  0.0066260
l =  0.031327
m =  0.020007
n =  0.059771
o =  0.062588
p =  0.014659
q =    7.5620e-04
r =  0.048053
s =  0.052867
t =  0.073359
u =  0.021236
v =  0.0087320
w =  0.019215
x =  0.0013173
y =  0.015016
z =    7.7470e-04
```

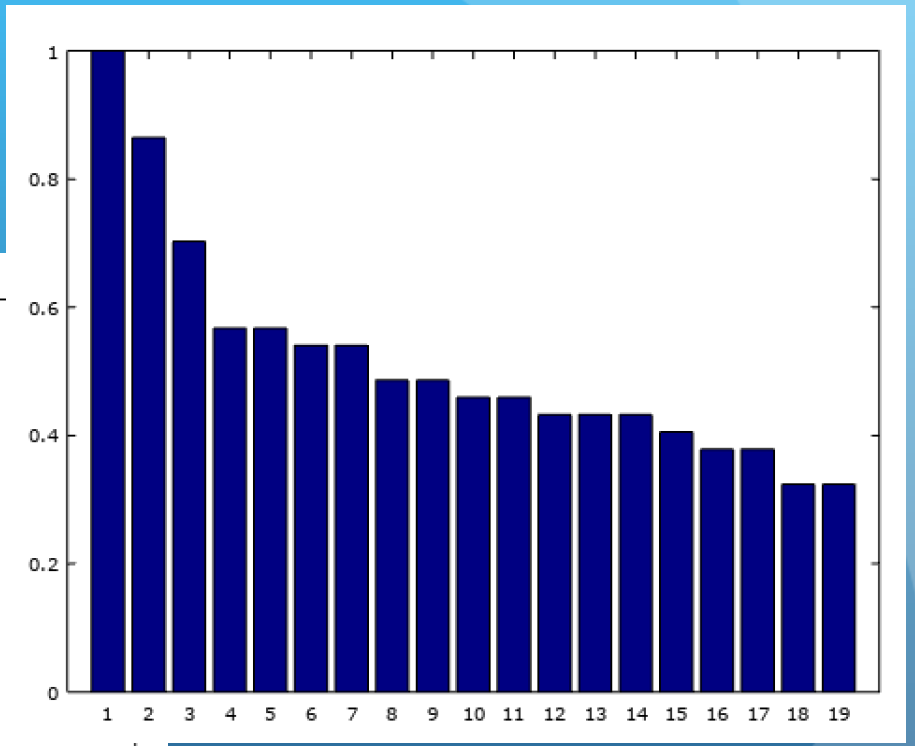


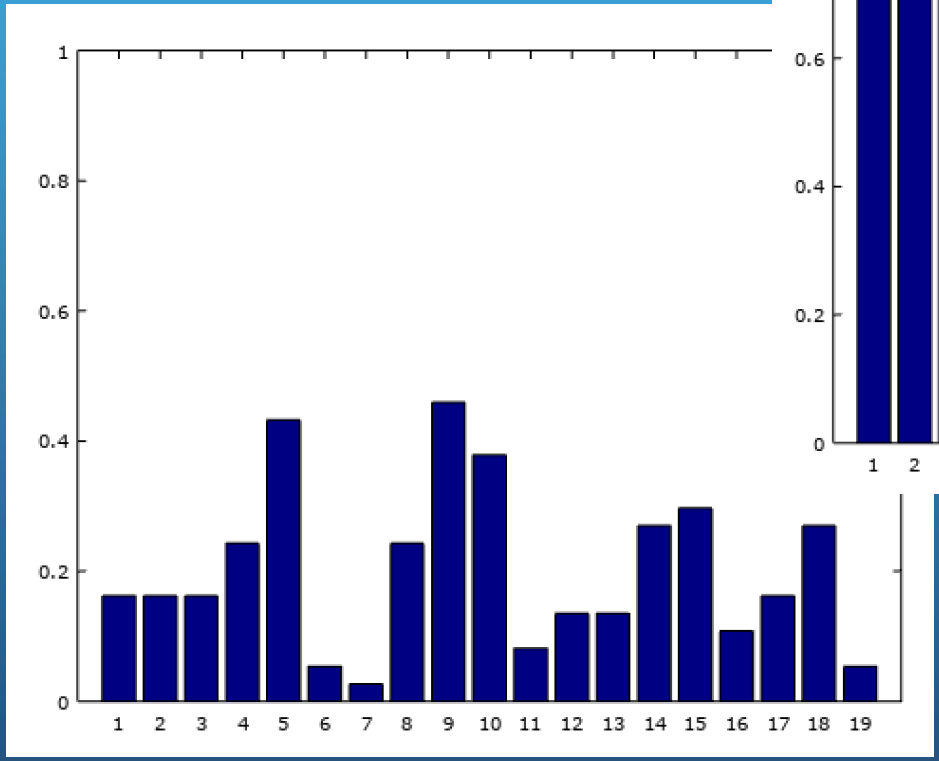
Bi-gram-ы в картинках и цифрах

```
e  =  0.033006
in =  0.028546
s  =  0.023194
d  =  0.018733
re =  0.018733
 t =  0.017841
 i =  0.017841
es =  0.016057
th =  0.016057
te =  0.015165
he =  0.015165
 a =  0.014273
n  =  0.014273
 r =  0.014273
an =  0.013381
er =  0.012489
y  =  0.012489
 d =  0.010705
ed =  0.010705
is =  0.010705
```



Bi-gram-ы в картинках и цифрах





«N...», «T...» и немного магии

- «N-gram» + «Top X» + Hash = TemplateFingerprint
  - 8_50_8bcd70378b27537e15c1e65b26062b1b4347f2960ca8ffd6adc7d3b8be7124d9
  - 8-gram, top 50
- Фиксированная длинна
  - Удобно хранить
  - Удобно сравнивать и группировать
  - В некоторой степени «user readable» ☺

TemplateFingerprint

- Разбиваем документ на n-gram-ы
- Вычисляем частоту n-gram-ов
- Сортируем в порядке убывания частоты
- Оставляем “Top X” n-gram-ов
- Записываем в строку «самый частый – в начале»
- Вычисляем хэш полученной строки
- Profit ☺

Полезно знать

- Оказалось, что не существует единственной, самой лучше комбинации параметров (а было бы хорошо …)
- Если данных много - следи за производительностью!
  - Обработка / вычисление n-gram-ов
  - Хранилище
  - Поиск

Вопросы?

- Буду рад услышать ваше мнение о докладе
  - Скучно / нескучно
  - Интересно / неинтересно
  - Свежий взгляд / я знал(а) это и раньше
  - … (ваш вариант ☺)
- Пожелания / предложения / есть что обсудить?
- slipchenko@gmail.com / FB: Alexander Slipchenko