Projet Industries du TAL - Anonymisation d'émails

Victoria Musatova, Anca Boca

30/12/2015

1 Introduction

Dans les domaines du marketing et de la vente, utiliser des données personnelles des clients dans le but de developper des stratégies personnalisées, des réponses ou offres adaptées à leurs besoins, devient une pratique courante. Mais cette nouvelle manière de concevoir le client (à travers les données numériques qu'il met à disposition) impose, un défi aux nouvelles technologies : comment garantir la protection de celui-ci ?

Dans ce contexte, notre projet d'anonymisation des données s'impose comme une réponse évidente. Nous pouvons la définir comme :

un système de "cryptage/transformation" de données qui prend en entrée de données à caractère personnel et les transforme en données non-identifiables, tout en gardant le format et type initialles.

L'objectif de notre projet est de présenter et analyser un tel outil de transformation de données.

Ce rapport consistera en la présentation des principaux enjeux impliqués dans sa création. Après la présentation de s quelques observations et hypothèses sur notre corpus initial, nous présenterons les étapes mises en place lors du développement de notre propre système d'anonymisation. Nous finirons cet exposé, par la présentation des résultats obtenus, essayant d'évaluer, dans la limite du possible, ses performances aussi bien que ses limites.

2 Méthodes

2.1 Etude de corpus

Afin de réaliser notre tâche d'anonymisation, un extrait du corpus `ENRON` a été mis à notre disposition. Comprenant un ensemble de 8000 courriels, ce corpus se présente sous la forme des données textuelles non structurées, en anglais.

Quelles données à transformer ?

L'institut national des standards et de la technologie définit les données à caractère personnel comme suit :

Tracing of an individual or distinguishing of an individual : This is the information which by itself identifies an individual. For example, national insurance number, SSN, date of birth, and so on.

Quelques examples de données à caractère personnel sont :

- les données financières : numéro de la carte de crédit, numéro du compte bancaire, détails concernant le salaire.

- les données personnelles : photos, numéro de sécurité sociale, date de naissance, statut social, religion, nationalité, adresse.

- les détails sur l'éducation : nom de l'université, du lycée ou college, l'année d'obtention du diplôme.

- les données médicales : maladies, numéro d'identification du patient

Plus restreint, notre corpus de courriels ne contient pas tous les éléments mentionnés précédemment.

- les noms de personnes
- les noms d'organisations
- les sections FROM et TO contenant des adresses courriels

2.2 Traitement du corpus

Considérons un exemple de courriel du dossier `Inbox`. Nous observation une structuration de données spécifique aux courriels, structuration de laquelle nous essayerons de tirer profit dans le développement de notre système.

FIGURE 1 – exemple courriel (dossier inbox)

HEADER

```
Message-ID: <9825360.1075858929233.JavaMail.evans@thyme>
Date: Mon, 30 Apr 2001 14:36:51 -0700 (PDT)
From: amy.chandler@enron.com
To: sscott5@enron.com
Subject: hi
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: "Chandler, Amy" <Amy.Chandler@ENRON.com>@ENRON <IMCEANOTES-+22Chandler+2C+
X-To: sscott5@enron.com
X-cc:
X-bcc:
X-Folder: \SSCOTT5 (Non-Privileged)\Scott, Susan M.\Inbox
X-Origin: Scott-S
X-FileName: SSCOTT5 (Non-Privileged).pst
```

BODY MAIL

```
Hey Susan,

I don't know if you remember me or not, I went to UT and was a good
friend of Stephanie Spence.  I know I met you at one of her showers or
maybe as late has her wedding.  Anyway, several people have told me that
you work here so I thought that I would say hello.

How long have you worked here?  What do you do?  Do you like it?  I'm on
the 25th floor with Net Works.  I started in January and so far I like
it.  I'm still trying to get used to my job and Houston in general.

Well, I just wanted to let you know I'm here.  Talk to you later.

Amy
```

Nous avons choisi de distinguer entre l'en-tête et le corps du courriel. Nous avons appliqué des approches différentes pour chacune de ces deux parties, que nous allons détailler dans la section qui suit :

2.2.1 Traitement de l'en-tête du courriel

2.2.2 Traitement du corps du courriel

La tokenisation
Le tagging
La reconnaissance d' EN
L'anonymisation d'EN

2.3 Développement du système

(définir une hypothèse, des résultats attendus... autres idées ? ? !)

3 Discussion

Tests & résultats obtenus
Evaluation du système

4 Conclusion