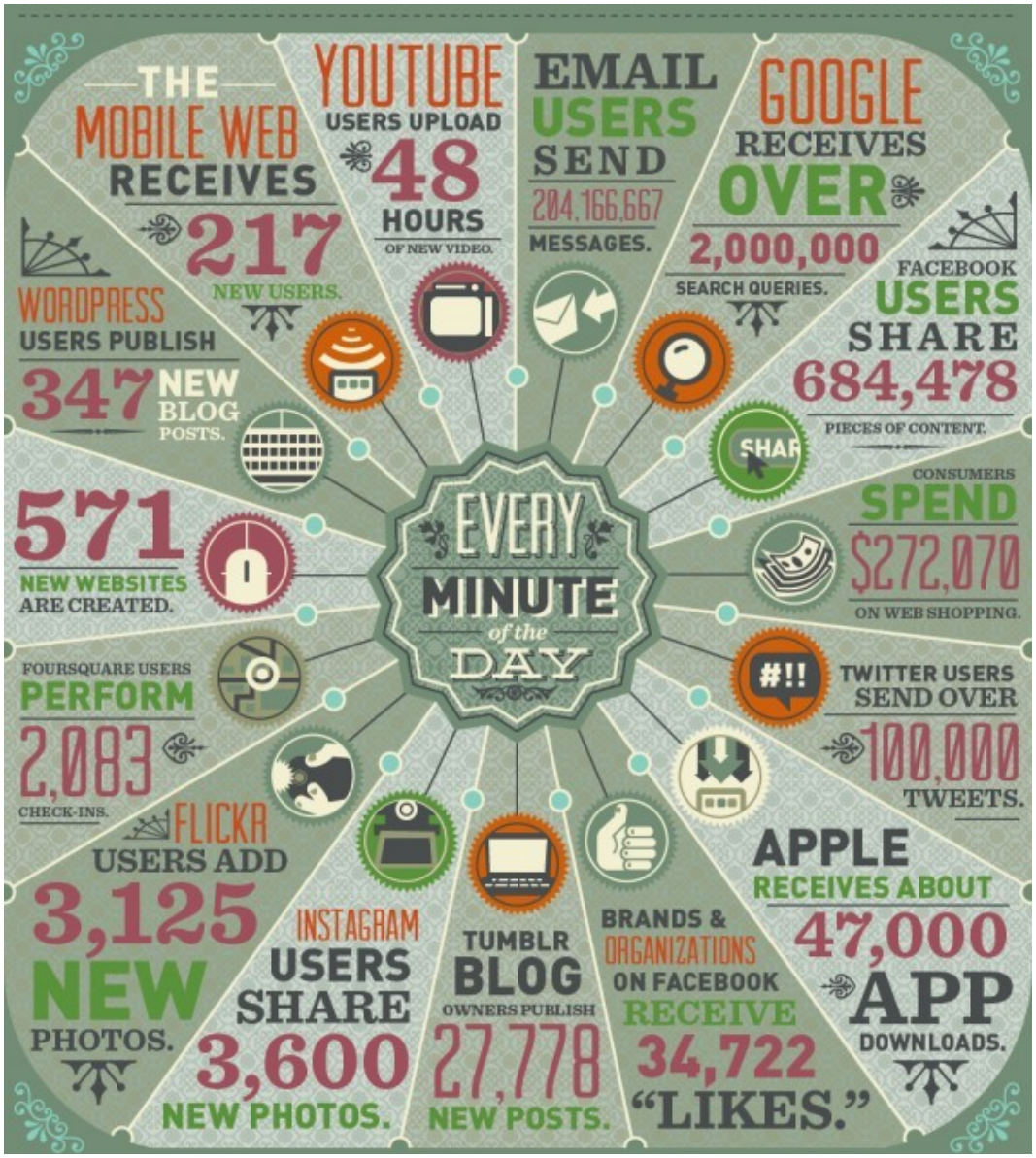
Data Mining

Exposés logiciels, systèmes et réseaux.

Damien Jubeau IR3

Lundi 19 novembre 2012


EVERY MINUTE of the DAY
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.

Plan

Data mining : définition, utilisations et concepts

Wolfram|Alpha : extraction de données d'un compte Facebook

Le problème du « badge » avec Weka

I – Data mining : définition, utilisations et concepts

I – Data mining : définition, utilisations et concepts

Définition

se traduit par

Exploration de données

ou encore

Extraction de connaissances à partir de données

« Processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données »

(Fayyad, 1996)

I – Data mining : définition, utilisations et concepts

Définition

Analyse de données et statistiques exploratoires existent depuis plus de 30 ans.

On peut voir le Data mining comme un prolongement :
- ajout de techniques d'IA
- données non structurées
- caractère business marqué

Big Data : 4ème révolution industrielle ?

I – Data mining : définition, utilisations et concepts

Domaines d'application

Scroring (réduction du coût d'acquisition ou de conservation d'un client – assurances, banques, opérateurs téléphoniques...)

Prévention du crime

Reconnaissance vocale

Recherche médicale

Détection de fraudes

Poker !

...etc

I – Data mining : définition, utilisations et concepts

Domaines d'application

Les recherches Google :

- Google spell checker
- Autocomplétion
- Recherche locale

« Storing and analyzing logs of user searches is how Google's algorithm learns to give you more useful results. Just as data availability has driven progress of search in the past, **the data in our search logs will certainly be a critical component of future breakthroughs.** »

Log
IP – Cookie – Recherche – Date & heure

http://www.google.org/flutrends

I – Data mining : définition, utilisations et concepts

Démarche

Mode projet

+ importance de l'objectif

+ Préparation des données en vue du traitement
structurées (base de données, fichiers tabulaires...) ou non structurées (textes, images...)

+ Élaboration, choix des modèles à appliquer (monde des statistiques ou de l'apprentissage automatique)

+ Évaluation et validation des connaissances extraites

I – Data mining : définition, utilisations et concepts

Des outils issus des statistiques et de l'IA

2 familles d'algorithmes :

Les **méthodes descriptives** permettent d'organiser, de simplifier et d'aider à comprendre l'information à partir des sources de données.

Recherche d'associations / Recherche de séquences similaires ...etc

Les **méthodes prédictives** visent à expliquer ou prévoir plusieurs phénomènes observables et effectivement mesurés. On cherche à prédire la valeur d'une **variable cible** à partir des valeurs de prédicteurs.

Régression linéaire multiple / Réseaux de neurones / Arbres de régression...

I – Data mining : définition, utilisations et concepts

Problèmes et limites

Hétérogénéité des données

→ importance de la préparation des données

Volume des données

→ calcul distribué (Hadoop)

Risque d'apporter une réponse hors scope

→ problème éthique

Danger pour la vie privée

→ explosion des données personnelles sur Internet

II – Extraction de données d'un compte Facebook

II – Extraction de données d'un compte Facebook

Présentation de Wolfram|Alpha

Outil de calcul en langage universel – lancé 2009

Analyse des données (10 milliards d'informations) issues de disciplines très variées (mathématiques, physique, chimie, nouvelles technologies, données socio-écononomiques...)

Utilisé par Bing, DuckDuckGo et Siri.

II – Extraction de données d'un compte Facebook

Présentation de Wolfram|Alpha

Outil de calcul en langage universel – lancé 2009

Analyse des données (10 milliards d'informations) issues de disciplines très variées (mathématiques, physique, chimie, nouvelles technologies, données socio-écononomiques...)

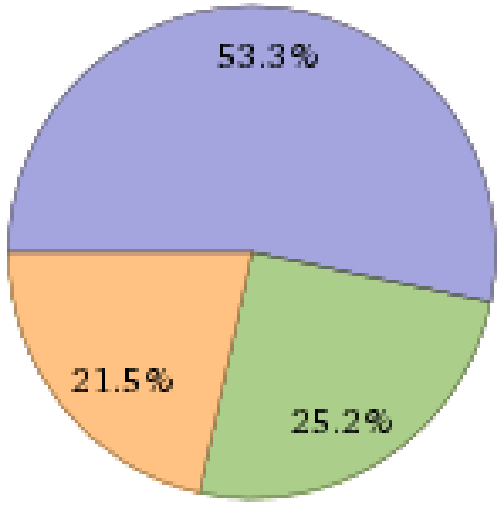
Utilisé par Bing, DuckDuckGo et Siri.

II – Extraction de données d'un compte Facebook

Publications de l'utilisateur

Types:

type	count	ratio	
statuses	245		
posted links	116		
uploaded photos	99		

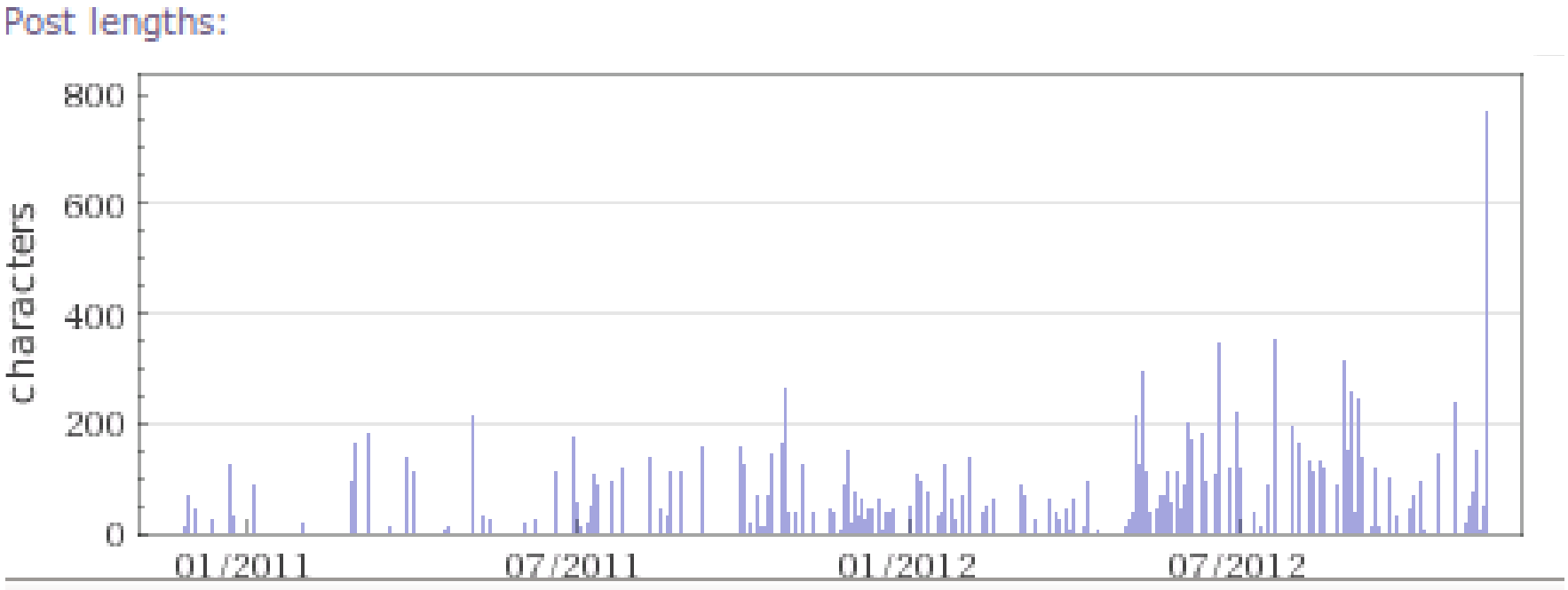


(based on 460 total activities)

Clip 'n Share

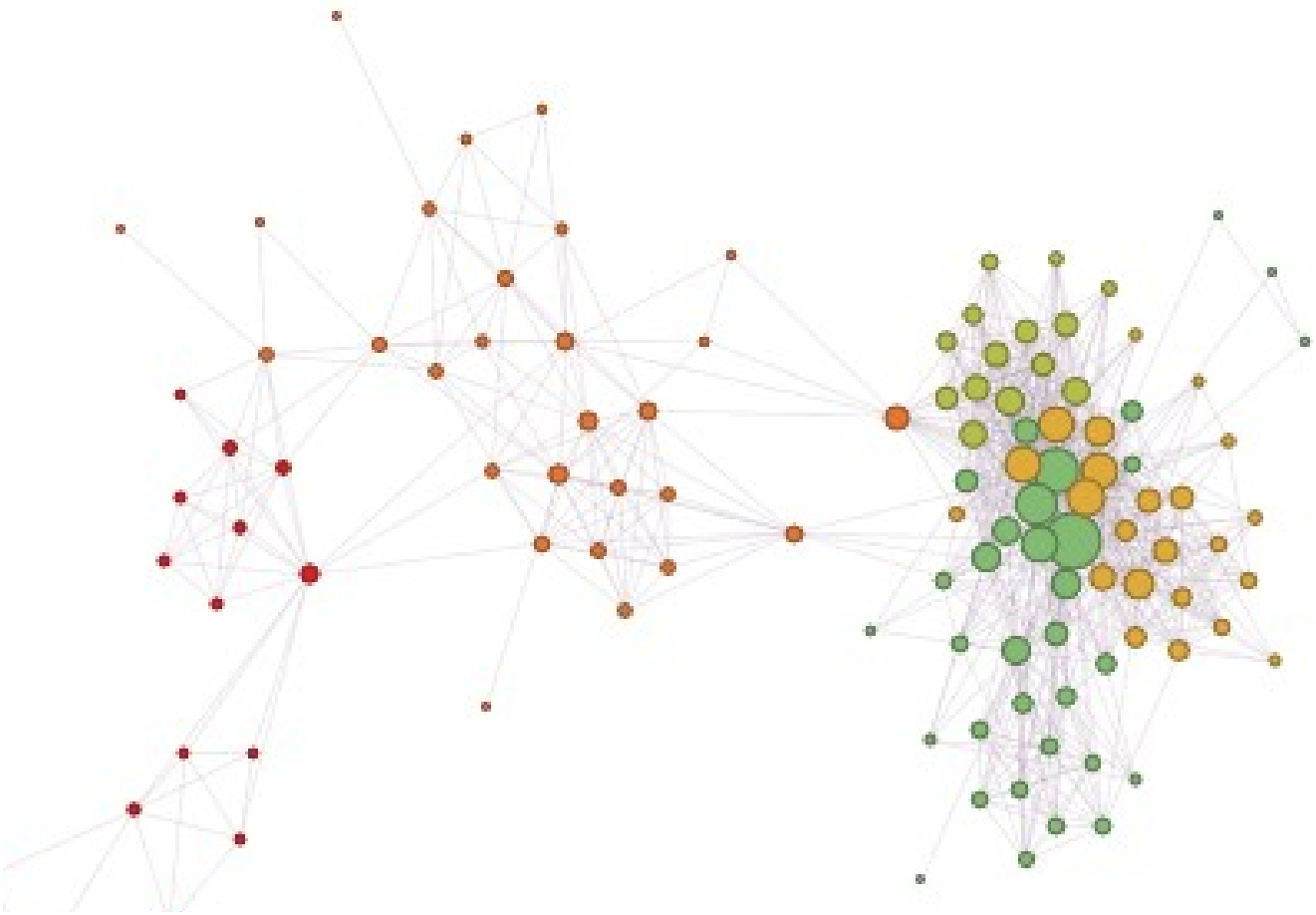
II – Extraction de données d'un compte Facebook

Taille des statuts



II – Extraction de données d'un compte Facebook

Réseau de connaissances



II – Extraction de données d'un compte Facebook

Analyse d'un compte Facebook

Et bien d'autres : opinions politiques, mots les plus utilisés...

Facebook possède des données similaires
pour plus d'un milliard d'utilisateurs...
...et si le modèle de l'outil intégrait des recoupements ?

Problématique Data mining / vie privée évidente

Utilisation commerciale ? Surveillance abusive des gouvernements ?

III – Le problème du « badge »

III – Le problème du « badge »

Contexte

En 1994, chaque participant à une conférence sur l'apprentissage automatique (Machine Learning) s'est vu remettre un badge « + » ou « - ».

Le but était pour un participant de découvrir la fonction d'attribution des badges, sachant qu'elle était basée uniquement sur les noms des participants. Pour cela la liste des participants et des badges attribués étaient fournie.

Utilisation de Weka, logiciel libre de Data Mining

III – Le problème du « badge »

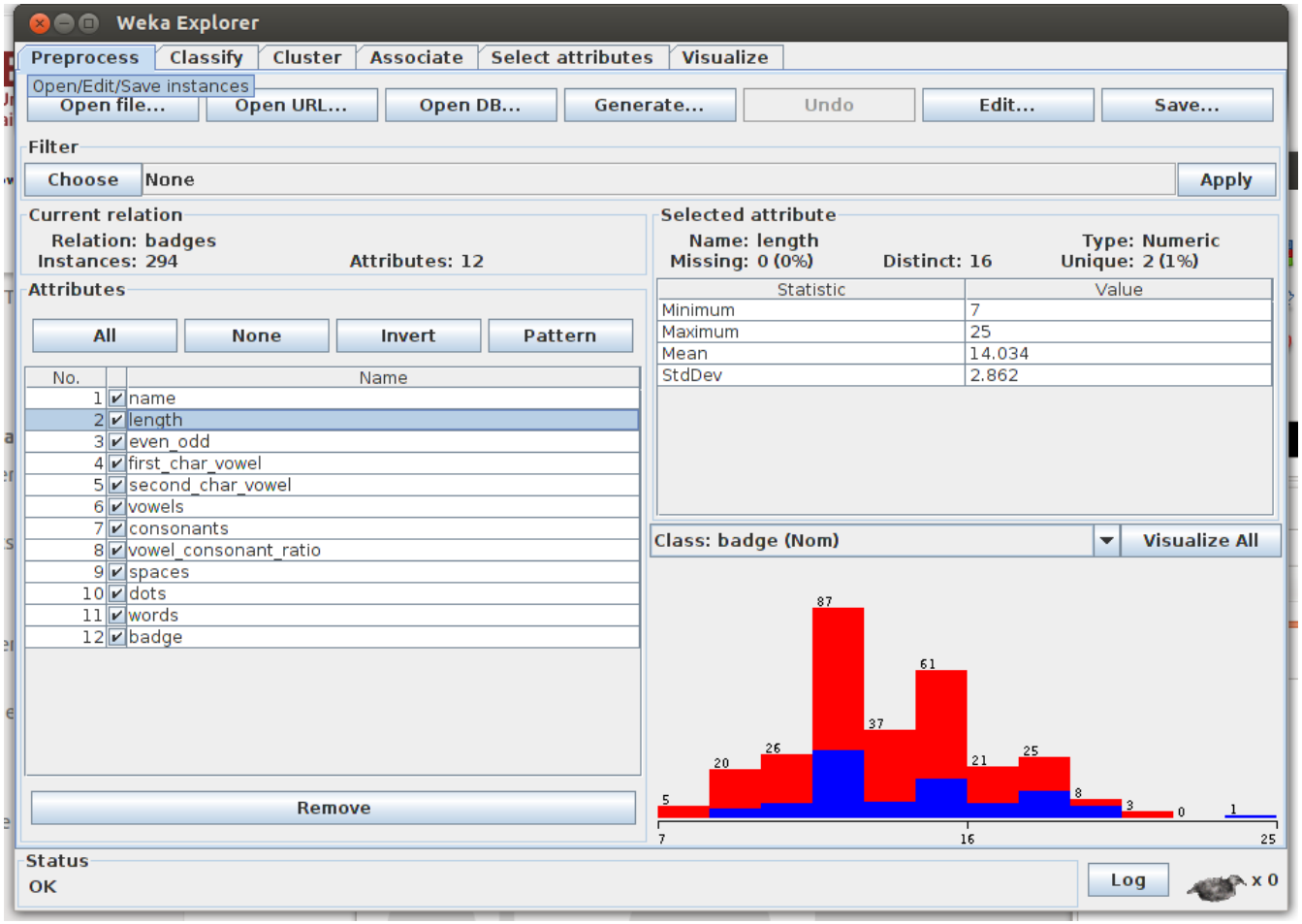
Préparation des données

Attribute name, and type	**Explanation**
name {...}	all the names (given in the original)
length numeric	length of name
even_odd {0,1}	length of name even or odd?
first_char_vowel {0,1}	is first character a vowel?
second_char_vowel {0,1}	is second character a vowel?
vowels numeric	number of vowels in the name
consonants numeric	number of consonants
vowel_consonant_ratio numeric	the ratio of vowels / consonant
spaces numeric	number of spaces
dots numeric	number of "." in the name, i.e. name initials
words numeric	number of words, i.e number of names
class {+,-}	the badge labels (given in the original)

Ex : Ameur Foued, 11 , 0 , 1 , 0 , 6 , 4 , 1.50 , 1 , 0 , 2 , -

III – Le problème du « badge »

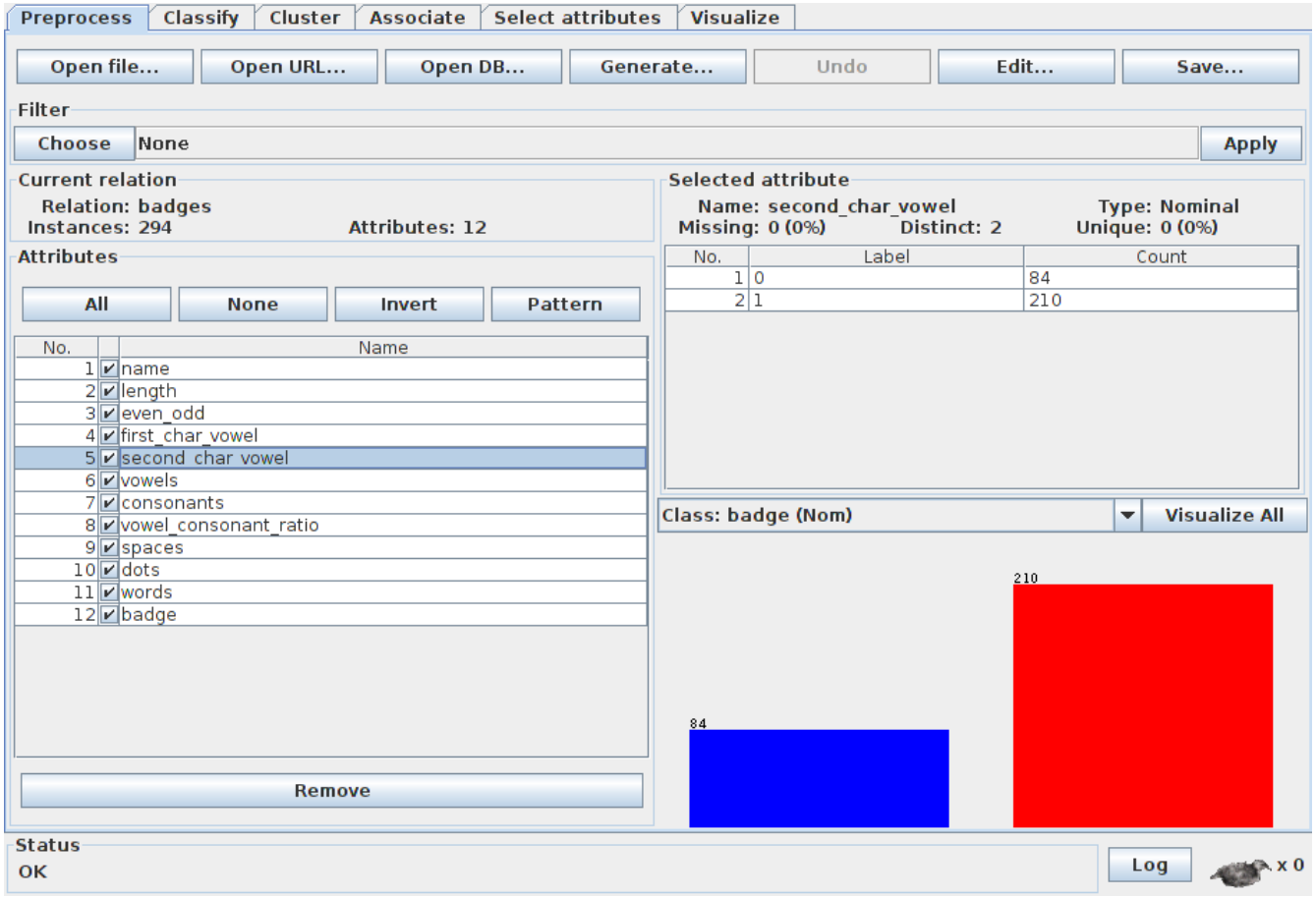
Utilisation du logiciel Weka



Classification Simple : + ou -

III – Le problème du « badge »

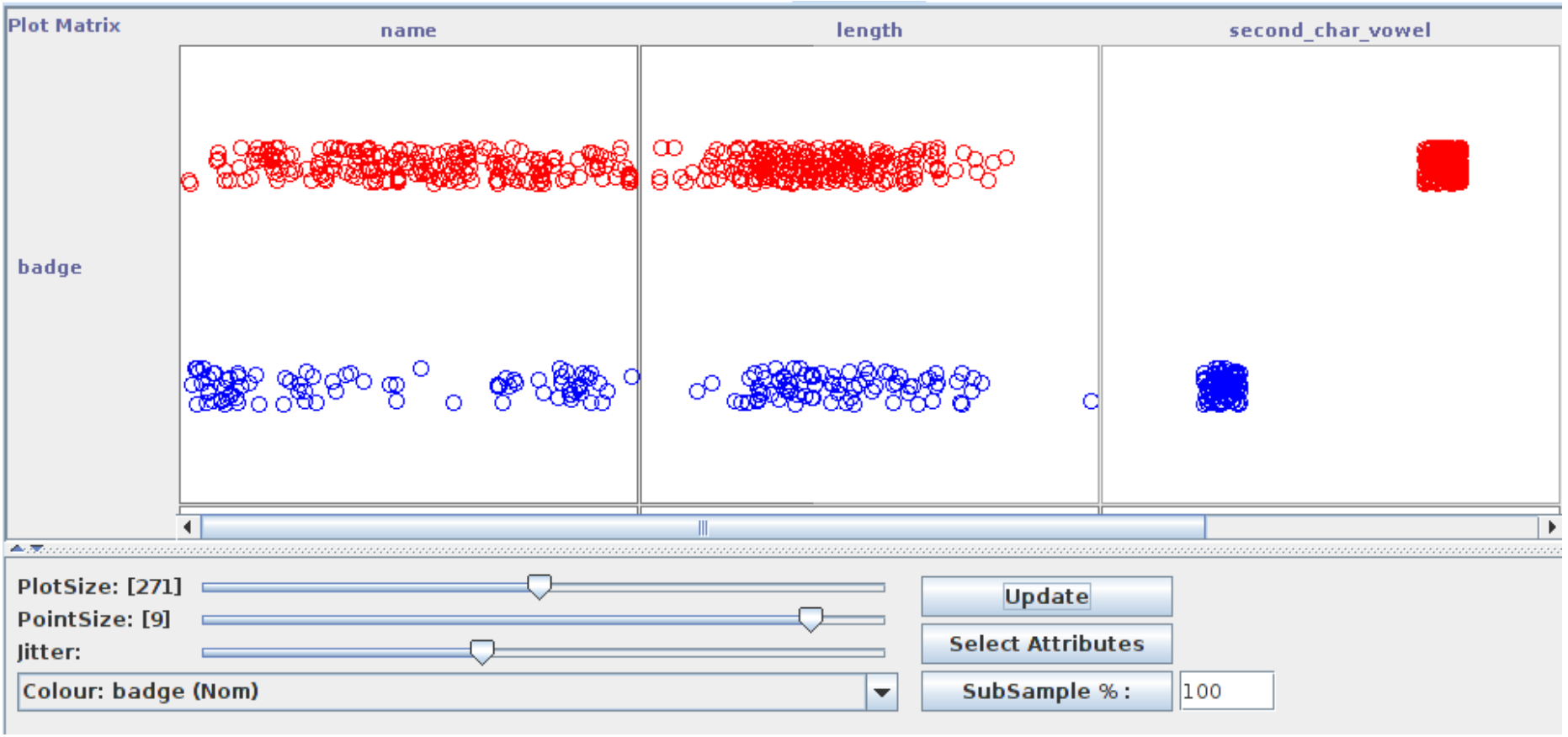
Utilisation du logiciel Weka



Cas simple : on cherche une association forte : elle est systématique

III – Le problème du « badge »

Visualisation des résultats



III – Le problème du « badge »

D'autres cas d'application

http://ftp.ics.uci.edu/pub/machine-learning-databases/

Conclusion

De nombreux logiciels mais nécessitent une expertise

Des techniques qui vont se développer et se démocratiser

De grandes avancées potentielles (médicales...)

Des dangers sans cadre légal

Data Mining

Exposés logiciels, systèmes et réseaux.

Damien Jubeau IR3

Lundi 19 novembre 2012