FeedUS

Julien Filion – 07 177 770
Simon Renaud – 07 149 640
Marc-Alexandre Côté – 07 166 997

Présentation

- FeedUS: Classificateur de flux RSS
- Réalisé par l’équipe Hocus
- Dans le cadre du cours IFT603

Contenu

- Introduction au flux RSS
- Travaux existants
- Notre application de classification
- Classification
  - Méthodes
  - Résultats
- Conclusion
- Démonstration & Questions?

Flux RSS

- Fichier dont le contenu est généré suite à une mise à jour
- Site d’actualité, blog
- S’appuie sur XML (Titre, Desc., Source, Date)

Notre application

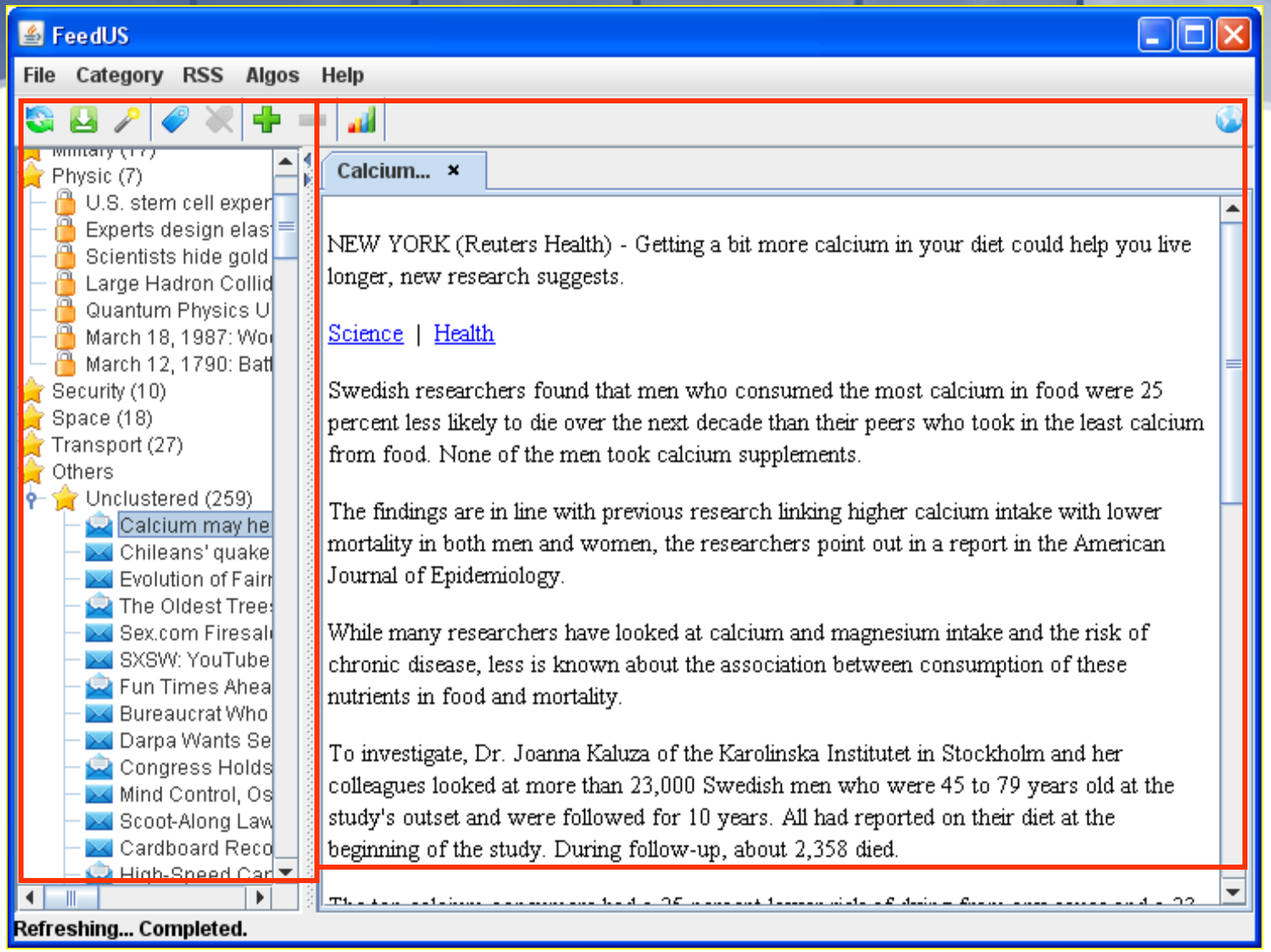
- Lecteur et classificateur de flux RSS

- Permets aux utilisateurs:
  - Lire des flux RSS de diverses sources
  - Regrouper les articles non-classifiés
  - Créer ses propres catégories
  - Classer automatiquement les nouveaux articles
  - Comparer les classifications

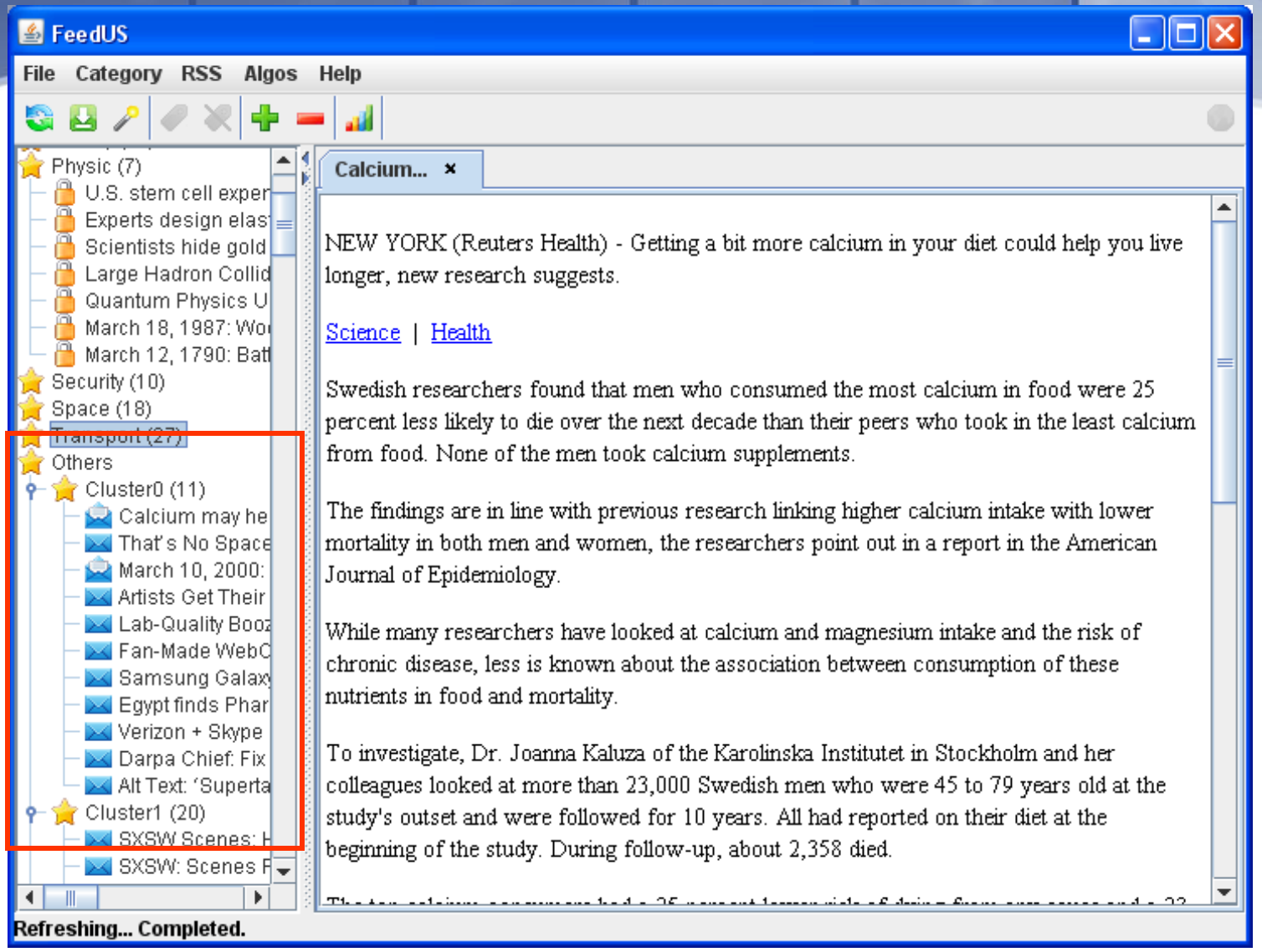
Travaux existants

- Plusieurs logiciels permettent la gestion de flux RSS
- Aucun ne semble offrir la classification automatique

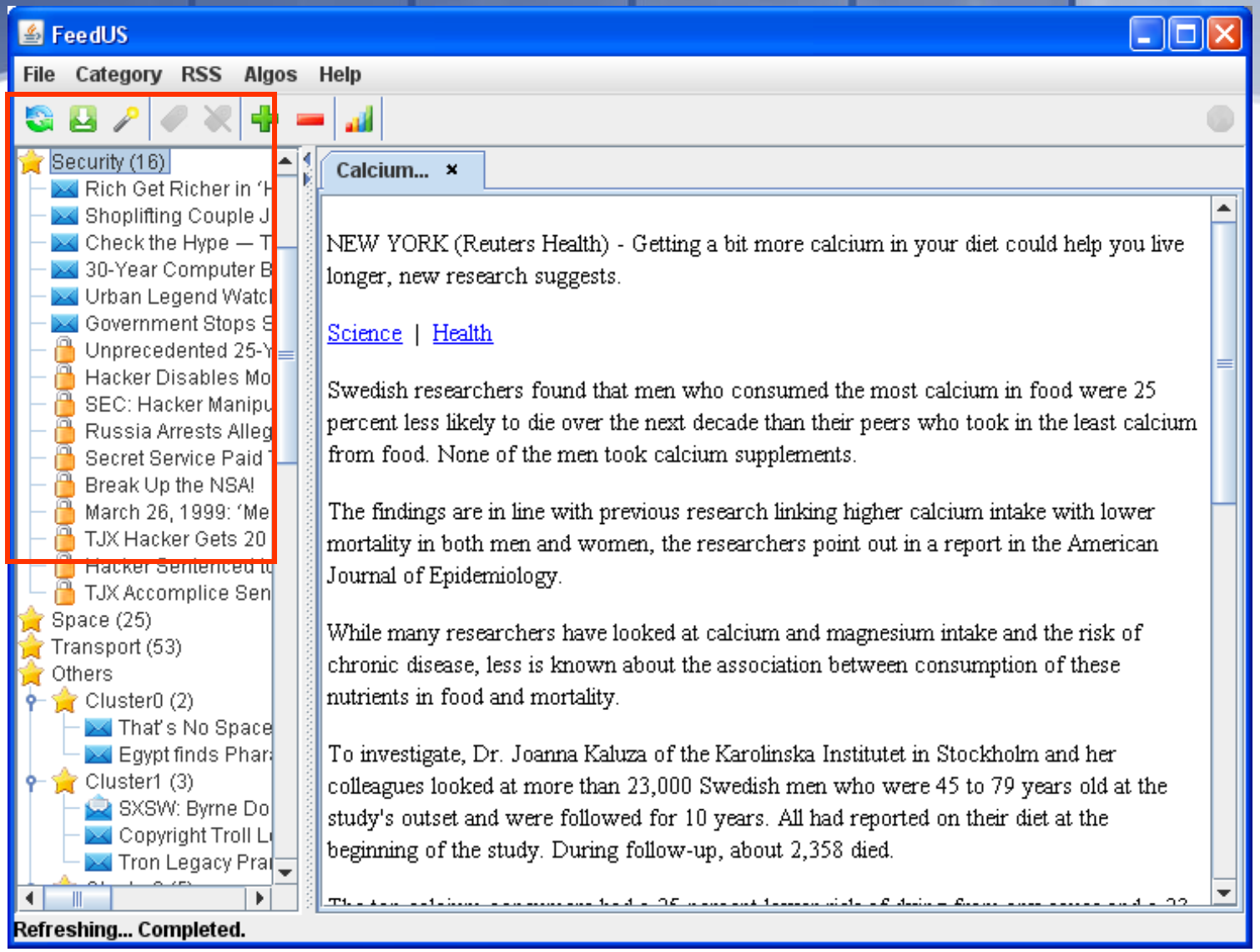
FeedUS
Interface



FeedUS
Interface



FeedUS
Interface



FeedUS

Interface

Documents	KNN	Naive Bayes	Random Forest Tree	User
Alan Moore's Arty Came...		Cinema	Gaming	
Amanda Palmer Freaks...	Cinema	Computer Science	Gadget	
New Godzilla Movie Will...	Cinema	Cinema	Cinema	Cinema
IMAX puts 3D spin on s...	Space	Space	Space	
U.S.-Russian crew blas...	Space	Space	Space	
Bolivia, China team up ...	Computer Science	Environnement	Cinema	
Scientists find aging ge...	Biology	Biology	Biology	
CERN tackles glitches, ...	Space	Space	Gaming	
Songbird genome may ...	Biology	Biology	Cinema	
Mega-flood triggered co...	Environnement	History	Space	
Rare cancer cells captu...	Biology	Biology	Biology	
Scientists open way for ...	Biology	Biology	Biology	
Special Report: Fast m...	Biology	Biology	Biology	
Why Volcanic Eruptions ...	Space	Environnement	Transport	
Inca Skeletons Show Ev...	History	History	Security	
Video: Tortoises Learn ...	Computer Science	Cinema	Cinema	
Think You're Good at Dr...	Transport	Gadget	Military	
End of Gene Patents Wi...	Biology	Biology	Biology	
Hunt for Missing Geneti...	Biology	Biology	Biology	
10 Years on, 'The Geno...	Biology	Biology	Military	
Skydiver Aims to Jump ...	Transport	Computer Science	Space	
New Evidence of Ice Ag...	Environnement	History	Computer Science	
Bats Use Sun to Calibr...	Environnement	Biology	Cinema	
iPad Could Boost Intera...	Gadget	Gadget	Gaming	
For iFixit, iPad Is D-Day ...	Gadget	Gadget	Cinema	
Google, China and Cen...	Computer Science	Computer Science	Gadget	
Comcast Rolls Out Bro...	Computer Science	Computer Science	Computer Science	
The New Media Mall: W...	Gadget	Gadget	Computer Science	
Verizon + Skype Not Mo...	Gadget	Cinema	Gadget	
Broadcast Spectrum or ...	Gadget	Gadget	Computer Science	

Difficulté

- Manque d’informations (flux RSS)
- Standardisation (reddit, google news)
- Recueillir seulement le texte
- Article parfois très court (image)

Classification
Méthodes

- Algorithme non-supervisé:
  - Bisecting K-Mean
- Algorithme supervisé:
  - K-NN (Modifié)
  - Naive Bayes (Rapid Miner)
  - Random Forest (Rapid Miner)

Classification
K-NN

- Ajout d’un seuil déterminant si un document peut ne pas être classé

Classification
Résultats

- K-NN:
  - Erreur de généralisation : 27%
  - Erreur de classification : 15%
- Naive Bayes:
  - Erreur de généralisation : 38%
  - Erreur de classification : 0%
- Random Forest
  - Erreur de généralisation : 63%
  - Erreur de classification : 0 %

Conclusion

- Avoir un système multi-utilisateur pour faire de la corrélation de données
- Utiliser des algorithmes de classification qui n’ont pas besoin de garder les documents pour l’entraînement

Démonstration

Questions ???