

Améliorer la classification de documents par combinaison de descripteurs visuels et textuels

Olivier Augereau

**Gestform
38 Rue François Arago
33700 Mérignac**

 **oaugereau@gestform.com**

 **@oaugereau**

Nicholas Journet, Jean-Philippe Domenger

**LaBRI
351, cours de la Libération
F-33405 Talence Cedex**

Plan

Introduction

1. Classification texte et image
 1. Sac de mots
 2. Sac de mots visuels
 3. Fusion multimodales
2. Tests sur bases réelles
 1. Résultats
 2. Retours industriels

Conclusion

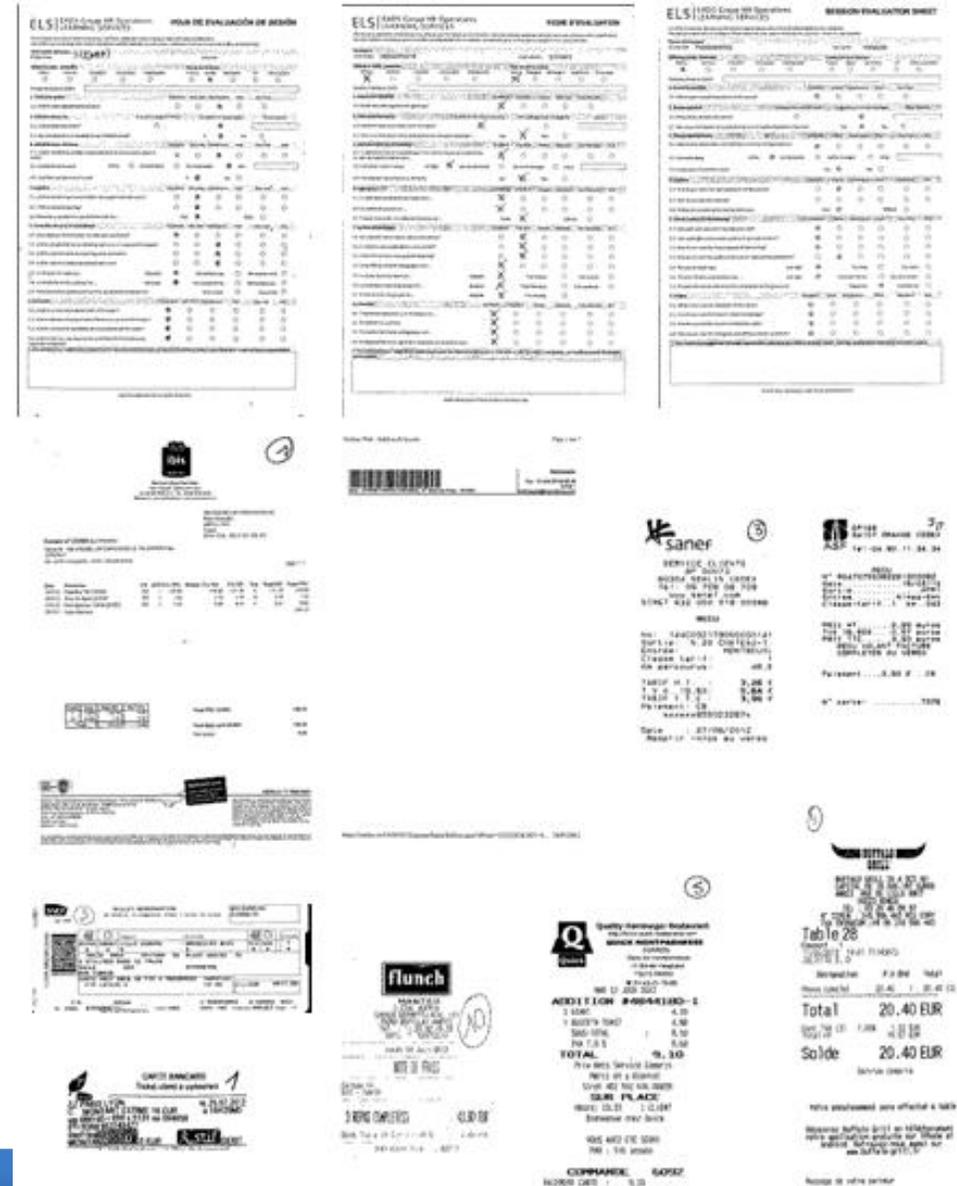
Introduction

Objectifs :

- Classifier des documents dans une base conséquente et variée
- Grande précision

Contexte industriel :

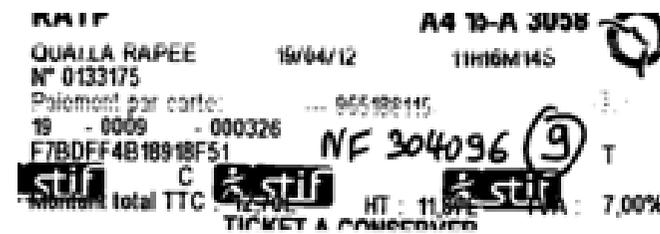
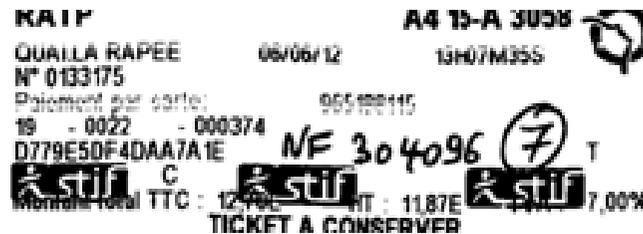
- Déformations géométriques
- Informations manquantes
- Qualités variables



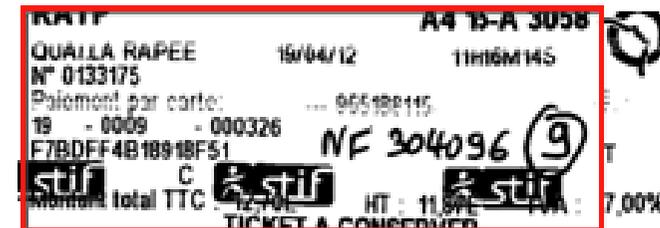
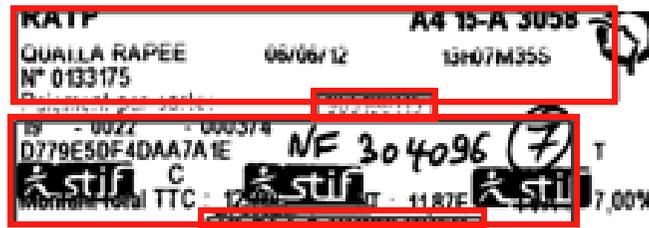
Introduction

Classification d'image de document : texte, image, structure.

image



structure



texte

KATP A4 13-A 3038
 QUAILLA RAPEE 06/06/12 13h07M35S
 N° 0133175
 Paiement par carte: 955188115
 19 - 0022 - 000374
 D779E5DF4DAA7A1E NF 304096 (7) T
 Montant total TTC : 12,70€ HT : 11,87€ TVA : 7,00%
 TICKET A CONSERVER

R H i r M-a-RWT
 QUAILLA RAPEE 19/04/12 TtKI6M14S
 N° 0133175
 Paiement par carte: --- 955188115
 19 - 0009 - 000326
 F7BDFF4B18918F51 NF 304096 (9) T
 Montant total TTC : 12,70€ HT : 11,87€ TVA : 7,00%
 TiriACT A/*nucco»

Introduction

⇒ Combiner les représentations texte et image [Rusiñol *et al.* 2012]

Texte : n-grams, sacs de mots (BoW), Latent Semantic Analysis (LSA)...

Image : textures, composantes connexes, points d'intérêts, sacs de mots visuels (BoVW), Spatial Pyramid Matching (SPM) ...

Combinaison : fusion tardive, précoce, apprentissage, ...

Apports :

- tester les BoVW sur les images de documents
- combinaison des BoW et BoVW.

⇒ Amélioration des performances globales par combinaison

Classification texte et image

Sacs de mots

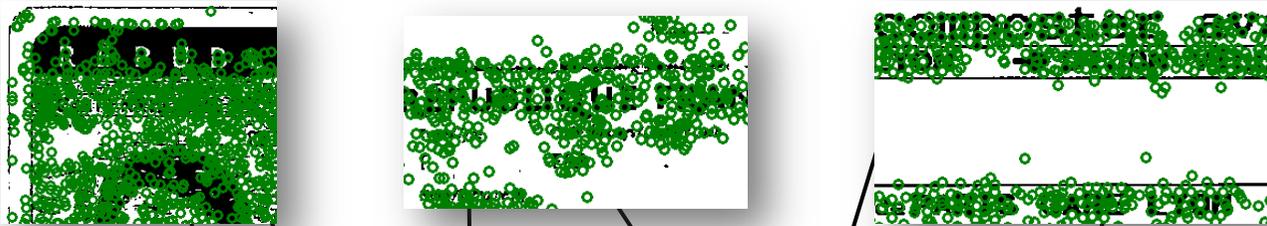
4 étapes :

1. Prétraitement du texte : filtrage, stop words, lemmatisation...
2. Création d'un dictionnaire à partir des K mots les plus fréquents (K=1000)
3. Chaque document est décrit par la fréquence d'apparition des mots qu'il contient qui sont compris dans le dictionnaire => vecteur
4. Utilisation des SVM pour apprentissage et classification supervisée.
(SVM à noyau radial ou polynomial > Bayes, Rocchio, C4.5 et k-NN avec distances cosinus [Joachims98])

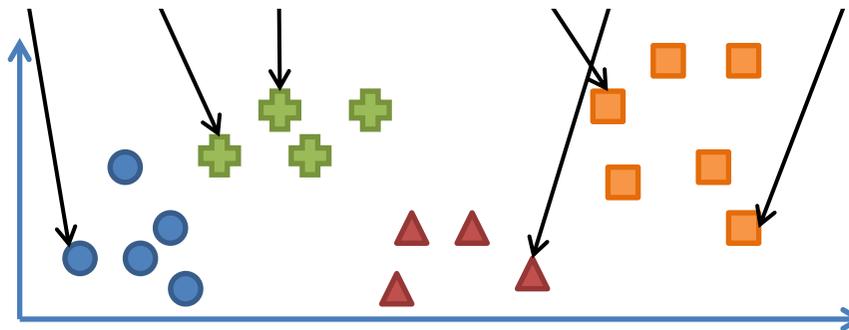
Classification texte et image

Sacs de mots visuels

1. Extraction des points d'intérêt



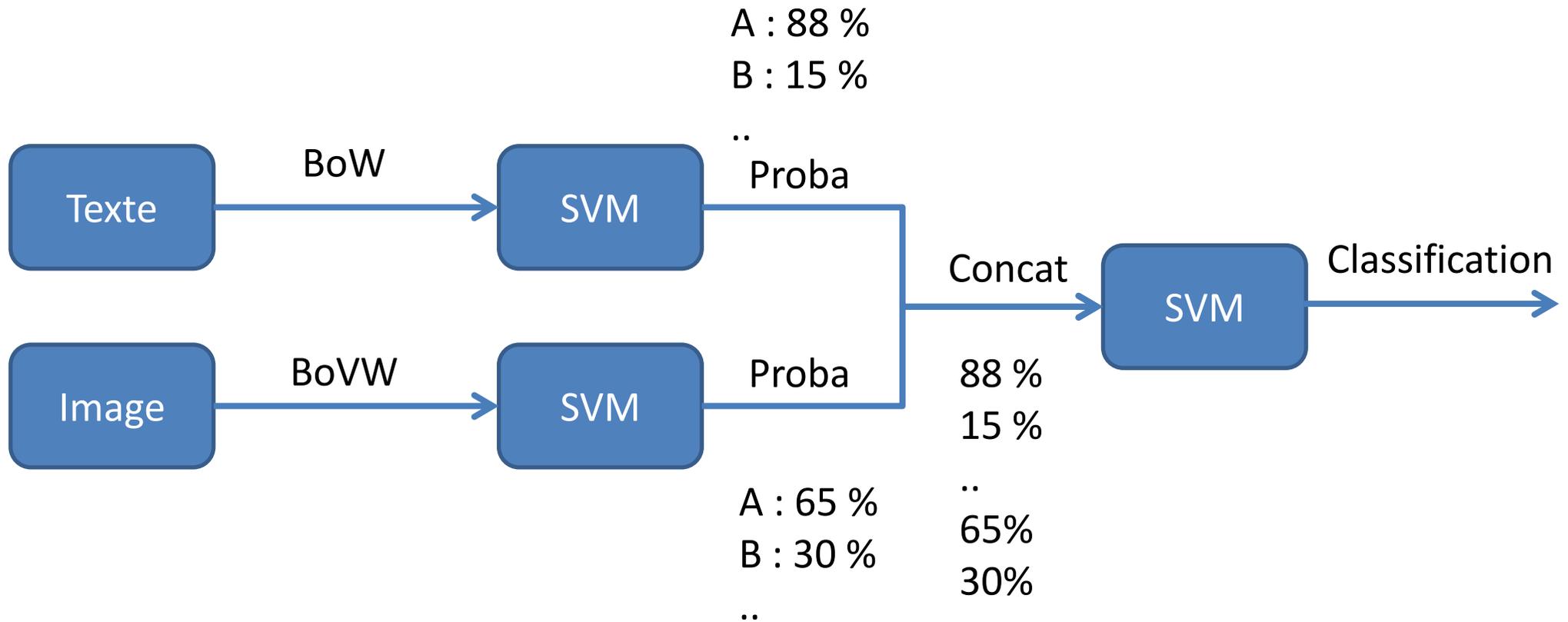
2. Partitionnement et création du dictionnaire visuel



3. Vecteurs de description



Classification texte et image



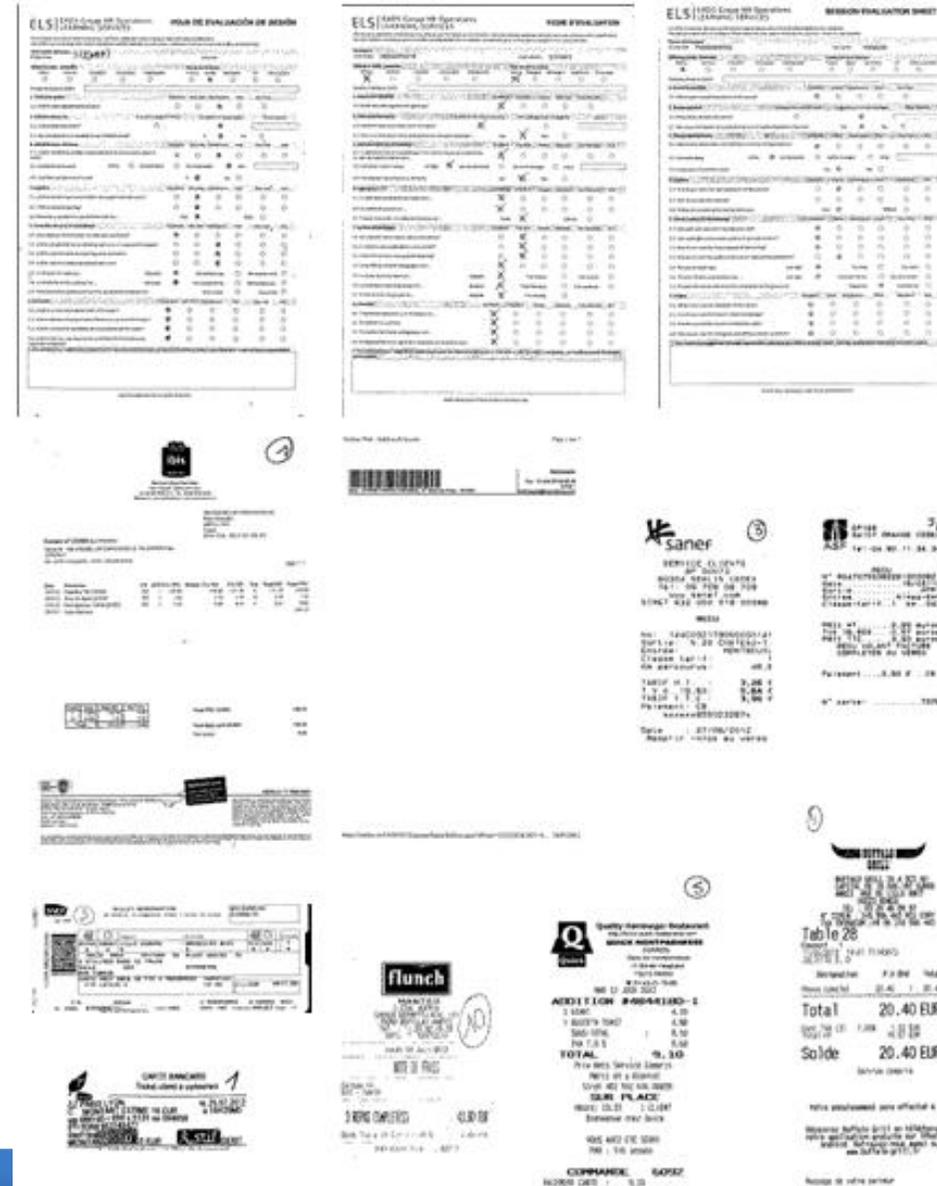
Tests sur bases réelles

Base de test :

- Documents numérisés par Gestform.
- 1985 pages choisies aléatoirement.
- 300 dpi, binarisation par le scanner.
- 12 classes de documents.

Classe	Sncf	ElsEs	ElsFr	ElsUk	Ibis	Lot
Nb doc	194	299	99	430	41	670

Asf	Sanef	Buffalo	Flunch	Quick	Ratp
31	22	13	4	12	110



Tests sur bases réelles

	Rappel	Précision
BoVW	87,1 %	86,9 %
BoW	98,2 %	97,7 %
Fusion Borda-Count	89,9 %	90,2 %
Fusion SVM	99,4 %	98,6 %

Rangs classifieur 1

A
B
C

Rangs classifieur 2

C
A
B



Rangs Borda

A
C
B

Tests sur bases réelles

		Rappel			Précision		
Classes	Nb docs	BoW	BoVW	Fusion	BoW	BoVW	Fusion
SNCF	194	1	0,984	0,995	0,964	0,954	0,990
ElsEs	299	1	0,677	1	0,987	0,709	0,997
ElsFr	99	1	0,816	1	1	0,626	1
ElsUk	430	1	0,782	1	1	0,802	1
IBIS	41	0,810	1	0,972	0,829	0,902	0,854
Lot	670	1	1	1	0,975	0,987	0,991
ASF	31	1	0,792	1	0,742	0,613	0,645
SANEF	22	1	0,167	1	0,955	0,682	1
Buffalo	13	1	1	1	1	1	1
Flunch	4	0,667	0,231	0,400	1	0,750	1
Quick	12	1	0,923	1	1	1	1
RATP	110	0,764	0,940	0,940	1	1	1
Moyenne		0,982	0,872	0,994	0,977	0,870	0,986



ELSI1000 Group 100 Operations
SESSION EVALUATION SHEET

POUR DE EVALUACIÓN DE RESULTADOS

ELSI1000 Group 100 Operations
SESSION EVALUATION SHEET



ELSI1000 Group 100 Operations
SESSION EVALUATION SHEET

POUR DE EVALUACIÓN DE RESULTADOS

ELSI1000 Group 100 Operations
SESSION EVALUATION SHEET



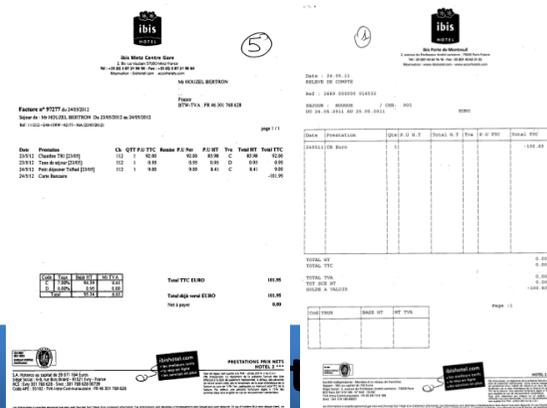
ELSI1000 Group 100 Operations
SESSION EVALUATION SHEET

POUR DE EVALUACIÓN DE RESULTADOS

ELSI1000 Group 100 Operations
SESSION EVALUATION SHEET

Tests sur bases réelles

Classes	Nb docs	Rappel			Précision		
		BoW	BoVW	Fusion	BoW	BoVW	Fusion
SNCF	194	1	0,984	0,995	0,964	0,954	0,990
ElsEs	299	1	0,677	1	0,987	0,709	0,997
ElsFr	99	1	0,816	1	1	0,626	1
ElsUk	430	1	0,782	1	1	0,802	1
IBIS	41	0,810	1	0,972	0,829	0,902	0,854
Lot	670	1	1	1	0,975	0,987	0,991
ASF	31	1	0,792	1	0,742	0,613	0,645
SANEF	22	1	0,167	1	0,955	0,682	1
Buffalo	13	1	1	1	1	1	1
Flunch	4	0,667	0,231	0,400	1	0,750	1
Quick	12	1	0,923	1	1	1	1
RATP	110	0,764	0,940	0,940	1	1	1
Moyenne		0,982	0,872	0,994	0,977	0,870	0,986



The screenshot shows a search results page from Ibis. It features a table with columns for 'Date', 'Presentation', and other search-related metrics. The page also includes the Ibis logo and some navigation elements like 'Page 1/1'.

Tests sur bases réelles

Classes	Nb docs	Rappel			Précision		
		BoW	BoVW	Fusion	BoW	BoVW	Fusion
SNCF	194	1	0,984	0,995	0,964	0,954	0,990
ElsEs	299	1	0,677	1	0,987	0,709	0,997
ElsFr	99	1	0,816	1	1	0,626	1
ElsUk	430	1	0,782	1	1	0,802	1
IBIS	41	0,810	1	0,972	0,829	0,902	0,854
Lot	670	1	1	1	0,975	0,987	0,991
ASF	31	1	0,792	1	0,742	0,613	0,645
SANEF	22	1	0,167	1	0,955	0,682	1
Buffalo	13	1	1	1	1	1	1
Flunch	4	0,667	0,231	0,400	1	0,750	1
Quick	12	1	0,923	1	1	1	1
RATP	110	0,764	0,940	0,940	1	1	1
Moyenne		0,982	0,872	0,994	0,977	0,870	0,986

BARBES-ROCH.
M4 33-A 5/99 158606
07/06/2012 09:57:11
00585199 5549
MONTANT TTC:12,70EUR
HT:11,87EUR TVA:7,00%

RATP
CARTE BANCAIRE
0133176 30

CARTE:-----956909893-
3099008723878228
A000000022100

stif

C T TICKET A CONSERVER

7

SAINI-LAZARE A4 16-B 5489 57803
30/04/2012 13:00:51
01331443 FF 6492
MONTANT TTC:25,40EUR
HT:23,74EUR TVA:7,00%

RATP
CARTE BANCAIRE
0133166 42

CARTE:-----016896020-
296FA4A82F182F40
A000000022100

stif

C T TICKET A CONSERVER

2

RATP
QUAILLA RAPEE 19/04/12
N° 0133175 11H16M14S
 Paiement par carte: --- 965188115
19 - 0009 - 000326
E7BDFF4B18918F51 NF 304096
stif C
montant total TTC: 12,70E HT: 11,87E TVA: 7,00%

stif

TICKET A CONSERVER

Tests sur bases réelles

- Paramétrage simple.
 - $200 > k$ (k-means et dictionnaire) > 2000
 - Seuillage de la matrice Hessienne ≈ 500 (doc. binaires)

- BoW, BoVW ou combinaison ?
 - BoW si OCR fiable, documents « propres »
 - BoVW si peu de texte et problème d'OCR
 - Combinaison si documents inconnus ou contenu variable

Conclusion

Conclusion :

Sur une base industrielle (notes de frais, formulaires, documents RH) :

- Les BoVW donnent des résultats intéressants.
- Combinaison BoW et BoVW > BoW.

Perspectives :

- Tester d'autres descripteurs texte/image et combinaisons.
- Rendre public des bases d'images (anonymes) et/ou de descripteurs.

Améliorer la classification de documents par combinaison de descripteurs visuels et textuels

Merci de votre attention

 oaugereau@gestform.com

 [@oaugereau](https://twitter.com/oaugereau)

Présentation à retrouver sur : www.olivier-augereau.com