ceres
Centre d'Expérimentation en méthodes numériques pour les REcherches en Sciences humaines et sociales

Sens Texte Informatique Histoire

Scraping : Définition, Motivation, Problèmes (et solutions ?)

Gaël Lejeune, STIH/CERES Sorbonne Université

20 janvier 2022

Définition

Scraping/scraper

- "Gratter", "Détacher", "Capturer, Nettoyer" . . . Raffiner

Définition

Scraping/scraper

- "Gratter", "Détacher", "Capturer, Nettoyer" . . . Raffiner
- Extraire et préparer

Définition

Scraping/scraper

- "Gratter", "Détacher", "Capturer, Nettoyer" . . . Raffiner
- Extraire et préparer
- Conserver les propriétés sémiotiques
- $\rightarrow$ des données structurées VS un sac

Définition

Scraping/scraper

- "Gratter", "Détacher", "Capturer, Nettoyer" . . . Raffiner
- Extraire et préparer
- Conserver les propriétés sémiotiques
- $\rightarrow$ des données structurées VS un sac

Input/output

- Sur du Web VS sur des applis

Définition

Scraping/scraper

- "Gratter", "Détacher", "Capturer, Nettoyer" . . . Raffiner
- Extraire et préparer
- Conserver les propriétés sémiotiques
- $\rightarrow$ des données structurées VS un sac

Input/output

- Sur du Web VS sur des applis
- Web Dynamique VS web statique

Définition

Scraping/scraper

- "Gratter", "Détacher", "Capturer, Nettoyer" . . . Raffiner
- Extraire et préparer
- Conserver les propriétés sémiotiques
- $\rightarrow$ des données structurées VS un sac

Input/output

- Sur du Web VS sur des applis
- Web Dynamique VS web statique
- Problématiques connexes :
  - PDF/Images, Fichiers Excel
  - Encodage, Méta-données

Outline

Introduction

Motivation (I) : Localisation/typage de la donnée

"Traiter automatiquement/mécaniquement c’est passer une donnée d’une forme (d’un format) à un autre"

Motivation (I) : Localisation/typage de la donnée

"Traiter automatiquement/mécaniquement c’est passer une donnée d’une forme (d’un format) à un autre"

Motivation (I) : Localisation/typage de la donnée

"Traiter automatiquement/mécaniquement c'est passer une donnée d'une forme (d'un format) à un autre"

Données structurées ou non structurées

Traitables/ générées par la machine/l'humain

	Structurées	Non-structurées
Human generated	Evaluation AIRBNB, QCM	Texte, tweet ...
Computer generated	Panier sur AMAZON, GPS	$\emptyset$?

Motivation (I) : Localisation/typage de la donnée

"Traiter automatiquement/mécaniquement c'est passer une donnée d'une forme (d'un format) à un autre"

Données structurées ou non structurées

Traitables/ générées par la machine/l'humain

	Structurées	Non-structurées
Human generated	Evaluation AIRBNB, QCM	Texte, tweet ...
Computer generated	Panier sur AMAZON, GPS	$\emptyset$?

Concepts liés

- Formats (PNG, BMP, ...PCX) et

Motivation (I) : Localisation/typage de la donnée

"Traiter automatiquement/mécaniquement c'est passer une donnée d'une forme (d'un format) à un autre"

Données structurées ou non structurées

Traitables/ générées par la machine/l'humain

	Structurées	Non-structurées
Human generated	Evaluation AIRBNB, QCM	Texte, tweet ...
Computer generated	Panier sur AMAZON, GPS	$\emptyset$?

Concepts liés

- Formats (PNG, BMP, ...PCX) et (*Digital Obsolescence*)
- Growth hacking (P.Paperon `https://tinyurl.com/mediumGrowthHacking`)

Motivation (II)

Traiter plus de données plus vites

- Reproduire des procédés

Motivation (II)

Traiter plus de données plus vites

- Reproduire des procédés
- Un pipeline end-to-end (ou presque)

Motivation (II)

Traiter plus de données plus vites

- Reproduire des procédés
- Un pipeline end-to-end (ou presque)
- La place des données dans le pipeline

Motivation (II)

Traiter plus de données plus vites

- Reproduire des procédés
- Un pipeline end-to-end (ou presque)
- La place des données dans le pipeline
- Pourquoi a-t-on du bruit ou du silence ?

Pourquoi dont on se soucier du traitement de Pages Web ?

Pas d'informatique sans données

Il faut aller chercher les données soi même sauf si on a :

- des données **bien structurées** (propres)
- en **quantité** suffisante
- **représentatives**
- **adaptées** à la tâche visée

On a plein de données en ligne, c’est facile

On a plein de données en ligne, c'est facile

Récupérer des Données Textuelles en ligne : facile ?

Différence entre langue naturelle et artificielle ?

On a plein de données en ligne, c'est facile

Récupérer des Données Textuelles en ligne : facile ?

Différence entre langue naturelle et artificielle ?

- HTML...

On a plein de données en ligne, c'est facile

Récupérer des Données Textuelles en ligne : facile ?

Différence entre langue naturelle et artificielle ?

- HTML...est une sorte de langue naturelle :
  - Les navigateurs sont sympas, ils font des efforts
  - On a plusieurs manières d'obtenir le même rendu
  - $\rightarrow$ non-bijectivité

Le lieu de données ?

Les méta-données rendent la collecte de textes facile ?

1. C.Doctorow (2001), Seven facts about *metacrap* `https://people.well.com/user/doctorow/metacrap.htm`

Le lieu de données ?

Les méta-données rendent la collecte de textes facile ?

- On ne peut pas totalement se fier aux méta-données [1]
  - Les gens peuvent être fainéants, bêtes, menteurs . . .

1. C.Doctorow (2001), Seven facts about *metacrap* `https://people.well.com/user/doctorow/metacrap.htm`

Le lieu de données ?

Les méta-données rendent la collecte de textes facile ?

- On ne peut pas totalement se fier aux méta-données [1]
  - Les gens peuvent être fainéants, bêtes, menteurs . . .
  - Ils ne connaissent pas bien leurs limites

1. C.Doctorow (2001), Seven facts about *metacrap* `https://people.well.com/user/doctorow/metacrap.htm`

Le lieu de données ?

Les méta-données rendent la collecte de textes facile ?

- On ne peut pas totalement se fier aux méta-données [1]
  - Les gens peuvent être fainéants, bêtes, menteurs . . .
  - Ils ne connaissent pas bien leurs limites
  - Les schéma ne sont pas neutres
  - Non plus que les métriques

1. C.Doctorow (2001), Seven facts about *metacrap* `https://people.well.com/user/doctorow/metacrap.htm`

Le lieu de données ?

Les méta-données rendent la collecte de textes facile ?

- On ne peut pas totalement se fier aux méta-données [1]
  - Les gens peuvent être fainéants, bêtes, menteurs . . .
  - Ils ne connaissent pas bien leurs limites
  - Les schéma ne sont pas neutres
  - Non plus que les métriques
  - Les points de vue influencent les descriptions

1. C.Doctorow (2001), Seven facts about *metacrap* `https://people.well.com/user/doctorow/metacrap.htm`

Le lieu de données ?

Les méta-données rendent la collecte de textes facile ?

- On ne peut pas totalement se fier aux méta-données [1]
  - Les gens peuvent être fainéants, bêtes, menteurs . . .
  - Ils ne connaissent pas bien leurs limites
  - Les schéma ne sont pas neutres
  - Non plus que les métriques
  - Les points de vue influencent les descriptions

Les données ne sont pas "monosémiques" hors contexte :

- 3.7s
- 12/11/10. . .

1. C.Doctorow (2001), Seven facts about *metacrap* `https://people.well.com/user/doctorow/metacrap.htm`

Input et défis

Ce que l'on a en entrée

Des pages web, des données sources :

- Qu'il faut donc collecter (et stocker)
- Avec des problèmes de nommage (en entrée et e sortie)
- De complétude (pagination, ascenseur infini)
- Et d'erreurs 404 . . .

Ce que l'on a en entrée

Des pages web, des données sources :

- Qu'il faut donc collecter (et stocker)
- Avec des problèmes de nommage (en entrée et e sortie)
- De complétude (pagination, ascenseur infini)
- Et d'erreurs 404 . . .

Des pages, des sources qui sont

- Destinées à l'œil humain (cf read first)
- Structurées par l'esprit humain (règles variables)
- Adaptées aux inférences (formats de date, abréviations)
- Construites selon des contraintes de mise en page (par ex. tableaux)

Les défis faciles

On peut éviter de passer par le web, tout va bien (modulo les problèmes de formats, RDF, JSON, XML . . .)

Les défis faciles

On peut éviter de passer par le web, tout va bien (modulo les problèmes de formats, RDF, JSON, XML . . .)

- Il existe des archives, dumps :
  - Wikipedia, Geonames

Les défis faciles

On peut éviter de passer par le web, tout va bien (modulo les problèmes de formats, RDF, JSON, XML . . .)

- Il existe des archives, dumps :
  - Wikipedia, Geonames
  - Open Data
  - Dépôts Github

Les défis faciles

On peut éviter de passer par le web, tout va bien (modulo les problèmes de formats, RDF, JSON, XML . . .)

- Il existe des archives, dumps :
  - Wikipedia, Geonames
  - Open Data
  - Dépôts Github
- Il existe une/des API (parfois en plus des dumps) :
  - Twitter, Instagram par ex.
  - `api.gouv.fr`

Les défis pas trop durs

$\rightarrow$ On exploite du web mais ça va bien se passer

Les défis pas trop durs

$\rightarrow$ On exploite du web mais ça va bien se passer

- Il y a une seule source

Les défis pas trop durs

$\rightarrow$ On exploite du web mais ça va bien se passer

- Il y a une seule source
- Les pages sont indépendantes

Les défis pas trop durs

$\rightarrow$ On exploite du web mais ça va bien se passer

- Il y a une seule source
- Les pages sont indépendantes
- Il estf acile de recenser le spages (robots, sitemaps)

Les défis pas trop durs

$\rightarrow$ On exploite du web mais ça va bien se passer

- Il y a une seule source
- Les pages sont indépendantes
- Il estf acile de recenser le spages (robots, sitemaps)
- Les concepteurs n'ont pas mis en place (consciemment ou non) de mesures anti-scraping

Les "vrais" défis

- pas de données en téléchargement
- pas d'API
- des paginations
- de la variété dans les sources (hétérogénéité)
- des barrières/protections contre le scraping

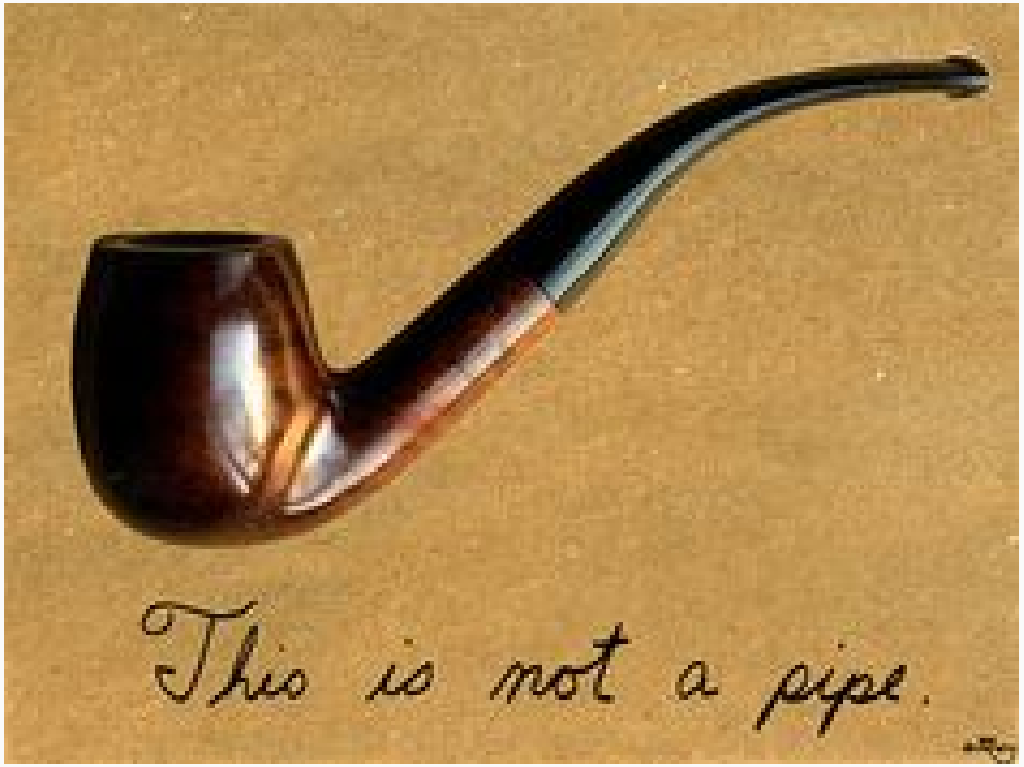
Les "vrais" défis

- pas de données en téléchargement
- pas d'API
- des paginations
- de la variété dans les sources (hétérogénéité)
- des barrières/protections contre le scraping

A suivre, un exemple "semi-facile"

La donnée et son contexte : exemple des données de presse

Phénomène et Représentation (I)



Phénomène et Représentation (II)

Ceci n'est pas un paragraphe

→ *twitter* : *title*

Phénomène et Représentation (III)

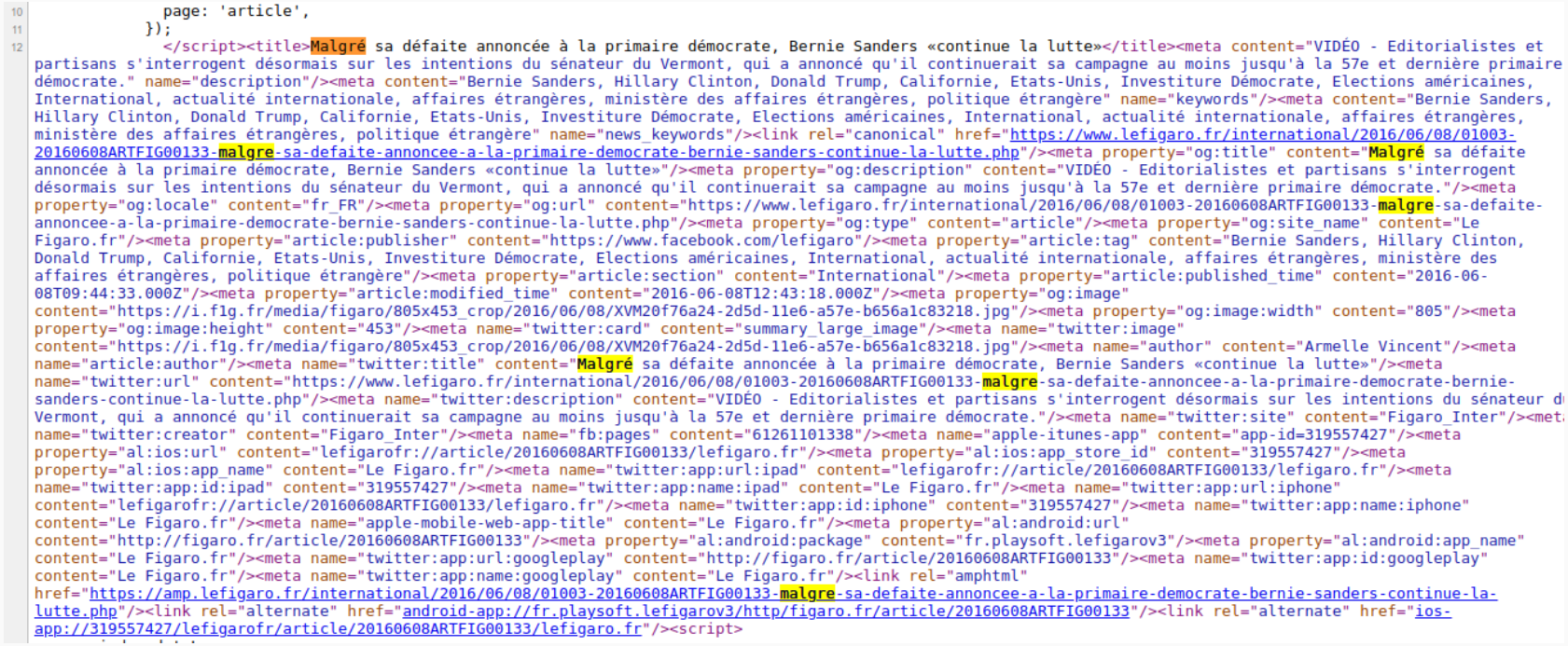


FIGURE 1 – Source : `www.lefigaro.fr` (2016)

Respecter les formats c'est compliqué

#	Site Name	HTML Validation Score	CSS Validation Score
1	Google	28 Errors, 4 warning(s)	Sorry! We found the following errors (10)
2	Facebook	36 Errors, 5 warning(s)	Sorry! We found the following errors (40)
3	YouTube	295 Errors, 2 warning(s)	Sorry! We found the following errors (746)
4	Yahoo!	272 Errors, 8 warning(s)	Sorry! We found the following errors (485)
5	Baidu	3 Errors, 4 warning(s)	Congratulations! No Error Found.
6	Amazon.com	997 Errors, 6 warning(s)	Sorry! We found the following errors (423)
7	Wikipedia	Passed, 21 warning(s)	Sorry! We found the following errors (4)
8	Taobao	Sorry! This document cannot be checked.	Sorry! We found the following errors (141)
9	Twitter	24 Errors, 4 warning(s)	Sorry! We found the following errors (79)
10	Tencent QQ	Sorry! This document cannot be checked.	Sorry! We found the following errors (89)

FIGURE 2 – Validité HTML/CSS des sites les plus fréquentés (2015)

https://theseosystem.com/html-css-validation-statistics-10-biggest-websites-world/

Comment Extraire les données et Évaluer la qualité

Méthodes d'extraction (I)

On veut de la donnée (ici du texte) mais comment l'extraire ?

Méthodes d'extraction (I)

On veut de la donnée (ici du texte) mais comment l'extraire ?

Quatre grands types

Site : comparer plusieurs pages d'un même site Web

Rendu : simuler le rendu du navigateur

HTML : exploiter le DOM ou le Xpath

Contenu : statistiques sur les phrases, les mots…

Méthodes d'extraction (II)

Trois exemples d'outils

	Site	Rendu	HTML	Contenu
`Boilerpipe`		★	★	★★★
`NCleaner`		★		★★★
`Justext`			★★	★★★

Deux types d'évaluation :

Intrinsèque : contenu textuel

Extrinsèque : impact sur une tâche réalisée en aval

Corpus de référence

Données	NB lignes	NB tokens	NB caractères
Html	1385 ($\pm$1303)	4726 ($\pm$3921)	75015 ($\pm$51924)
Clean	13 ($\pm$10)	321 ($\pm$323)	2296 ($\pm$1982)

Tableau 1 – Statistiques sur le corpus

Mesures d’évaluation

Vérité de terrain VS prédiction

Mesures d'évaluation

Vérité de terrain VS prédiction

Alignement de séquences (mots et/ou balises)

Précision, Rappel et F-Mesure

- VP : Séquence correcte
- FP : Insertion d'une séquence
- FN : Séquence manquante

Mesures d'évaluation

Vérité de terrain VS prédiction

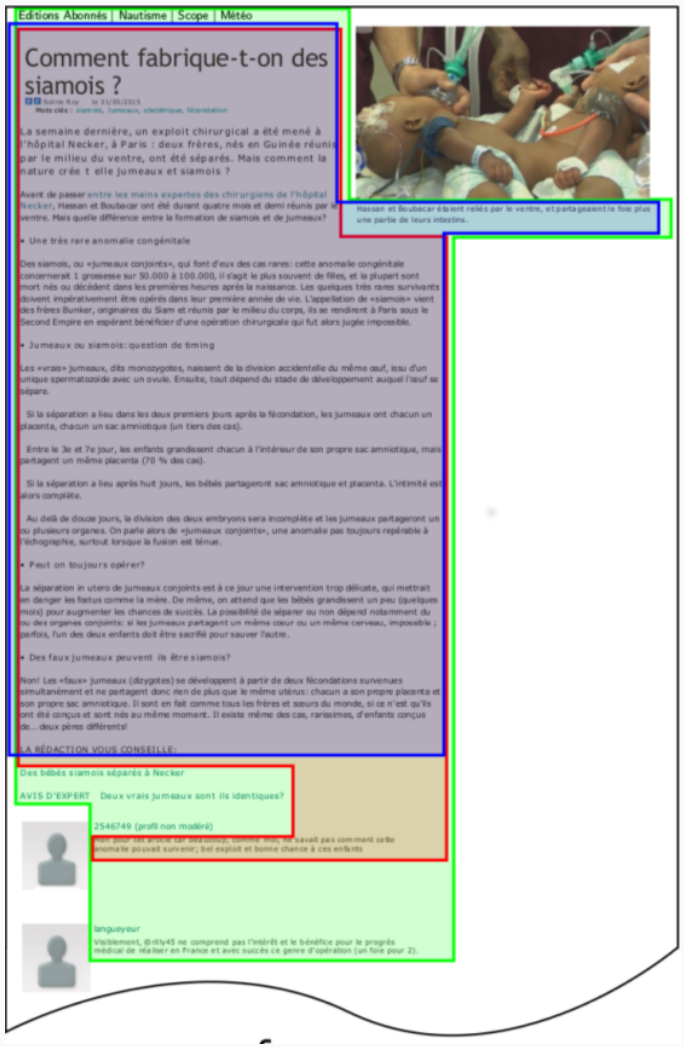
Alignement de séquences (mots et/ou balises)

Précision, Rappel et F-Mesure

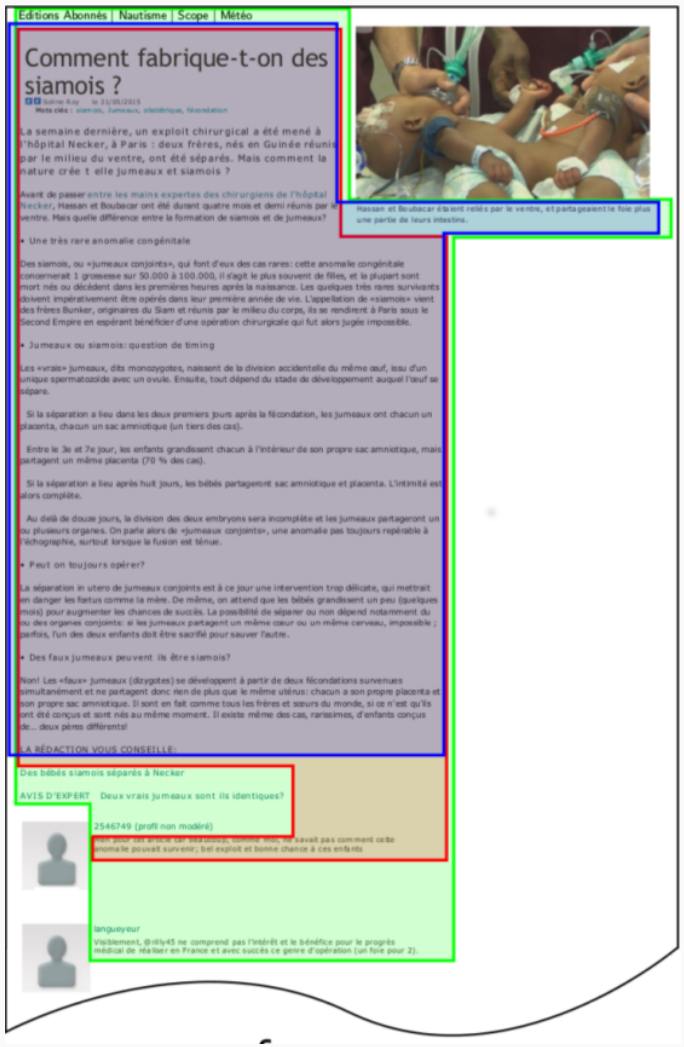
- VP : Séquence correcte
- FP : Insertion d'une séquence
- FN : Séquence manquante

Trois Configurations

- TO : *Text Only*
- TM : *Text and Mark-up*
- CAR : *Caractères*



page 2	TO	TM	CAR
BoilerPipe	91,24	94,94	33,33
JustText	86,93	34,68	43,75
NCleaner	68,43	23,26	21,05



page 2	TO	TM	CAR
BoilerPipe	91,24	94,94	33,33
JustText	86,93	34,68	43,75
NCleaner	68,43	23,26	21,05

Comment interpréter les chiffres ?

Résultats (Zhu et Lejeune 2018)

1600 documents (5 langues)

	Text Only (TO)			Caractères (CAR)			*Text & Markup*(TM)		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
BP	**81.80**	**88.89**	**85.20**	**76.93**	**81.12**	**78.97**	**64.47**	**85.42**	**73.48**
JT	68.75	83.41	75.37	63.79	67.03	65.37	61.94	63.23	62.58
NC	56.14	25.70	35.26	53.25	18.73	27.72	45.11	21.77	29.36

Résultats (Zhu et Lejeune 2018)

1600 documents (5 langues)

	Text Only (TO)			Caractères (CAR)			*Text & Markup*(TM)		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
BP	**81.80**	**88.89**	**85.20**	**76.93**	**81.12**	**78.97**	**64.47**	**85.42**	**73.48**
JT	68.75	83.41	75.37	63.79	67.03	65.37	61.94	63.23	62.58
NC	56.14	25.70	35.26	53.25	18.73	27.72	45.11	21.77	29.36

Sous-corpus anglais

	TO			CAR			TM		
BP	**85.97**	**92.02**	**88.89**	**84.92**	**91.03**	**87.87**	**69.29**	**93.59**	**79.63**
JT	68.41	85.38	75.96	69.98	82.79	75.85	67.17	79.69	72.90

Sous-corpus chinois

	TO			CAR			TM		
BP	**61.32**	**52.90**	**56.80**	**77.12**	**63.55**	**69.68**	**84.48**	**67.99**	**75.34**
JT	23.25	11.72	15.58	71.31	32.05	44.23	49.27	16.78	25.03

Résultats (Zhu et Lejeune 2018)

Sous-corpus grec

	TO			CAR			TM		
BP	**91.84**	**96.48**	**94.10**	**87.58**	**91.59**	**89.54**	66.74	**91.67**	**77.24**
JT	88.10	90.07	89.08	73.39	73.95	73.67	**76.65**	74.68	75.65

Sous-corpus polonais

	TO			CAR			TM		
BP	**83.27**	**85.28**	**84.26**	**80.76**	**82.08**	**81.42**	**63.27**	**85.57**	**72.75**
JT	67.78	82.63	74.47	67.64	78.53	72.68	62.74	73.05	67.51

Sous-corpus russe

	TO			CAR			TM		
BP	**58.79**	79.33	**67.53**	**51.67**	**70.16**	**59.51**	37.92	**85.50**	52.54
JT	52.72	**81.78**	64.11	41.28	63.35	49.98	**42.94**	75.04	**54.62**

Sur le papier, nous avons un vainqueur évident

Résultats (Zhu et Lejeune 2018)

Sous-corpus grec

	TO			CAR			TM		
BP	**91.84**	**96.48**	**94.10**	**87.58**	**91.59**	**89.54**	66.74	**91.67**	**77.24**
JT	88.10	90.07	89.08	73.39	73.95	73.67	**76.65**	74.68	75.65

Sous-corpus polonais

	TO			CAR			TM		
BP	**83.27**	**85.28**	**84.26**	**80.76**	**82.08**	**81.42**	**63.27**	**85.57**	**72.75**
JT	67.78	82.63	74.47	67.64	78.53	72.68	62.74	73.05	67.51

Sous-corpus russe

	TO			CAR			TM		
BP	**58.79**	79.33	**67.53**	**51.67**	**70.16**	**59.51**	37.92	**85.50**	52.54
JT	52.72	**81.78**	64.11	41.28	63.35	49.98	**42.94**	75.04	**54.62**

Sur le papier, nous avons un vainqueur évident

- Mais dans la vraie vie ?
- Hors des "conditions de laboratoire ?

Évaluation Extrinsèque

Impact on Classification

- Même Corpus, mais maintenant on s'en sert !
- Quel impact sur les résultats ?

Évaluation Extrinsèque

Impact on Classification

- Même Corpus, mais maintenant on s'en sert !
- Quel impact sur les résultats ?

	En			Zh			El			Pl			Ru			All		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BP	.60	.25	.36	.78	.93	**.85**	.85	.35	.50	.76	.48	.59	.76	.55	.64	.74	.47	.58
JT	.55	.45	**.50**	.66	.37	.48	.59	.76	**.66**	.59	.59	.59	.82	.79	**.80**	.64	.59	**.61**
NC	.50	.25	.33	.83	.31	.45	.2	.05	.09	.82	.51	**.63**	.61	.27	.38	.62	.29	.40
Ref	.68	.88	**.77**	.8	1.0	**.88**	.68	.76	**.72**	.61	.77	**.68**	.72	.82	**.77**	.69	.84	**.76**

- BP est faible sur le sous-corpus anglais
- JT a un bon rappel sauf en chinois
- NC n'est pas si mauvais, sauf en grec

Pour aller plus loin

Les dangers de l'évolution (Barbaresi et Lejeune 2020)

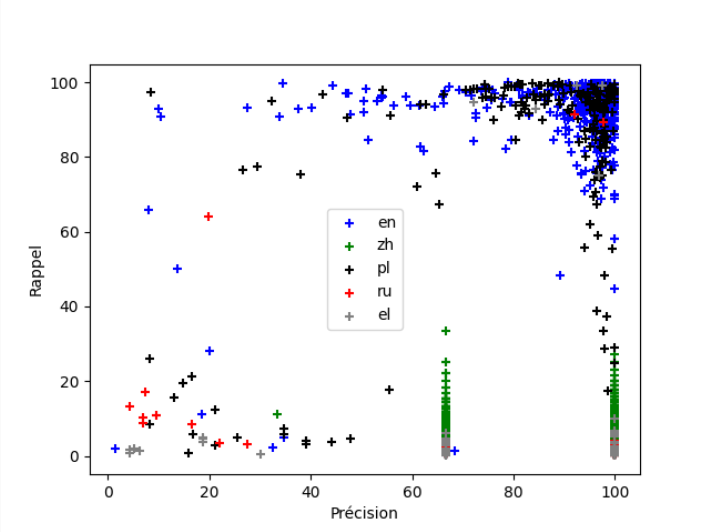
Catégorie	Outil	Version
Orientés rappel	HTML2TEXT	2020.1.16
	INSCRIPTIS	1.0
Orientés lisibilité	NEWSPAPER3K	0.2.8
	NEWS-PLEASE	1.4.25
	READABILITY	0.7.1
Dédiés à la tâche	BOILERPY3	1.0.2
	DRAGNET	2.0.4
	GOOSE3	3.1.6
	JUSTEXT	2.2.0
	TRAFILATURA	0.4.1

Sur 5 langues : vue globale

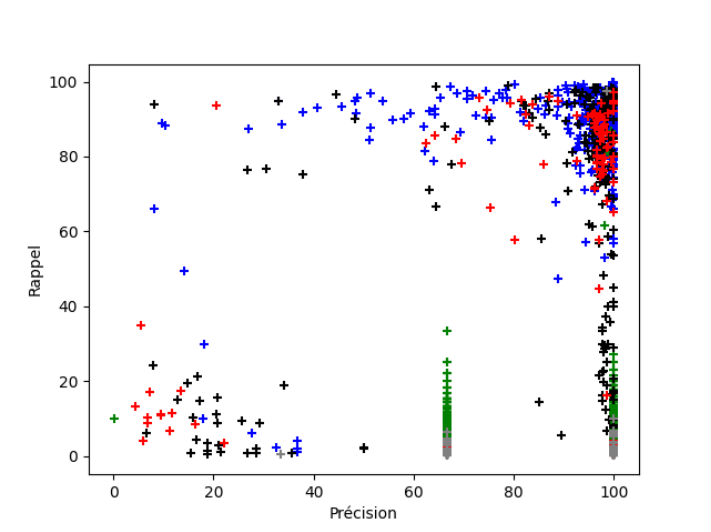
Outil	Macro F	Micro F	Micro P	Micro R
BP3_Art	70,41	**76,38**	80,60	72,57
JT	67,7	74,13	81,36	68,08
READ	**71,01**	73,25	72,43	74,09
TRAF	68,63	72.89	65.02	82.93
DRAGNET	56,12	67,09	86,82	54,67
NPLEASE	50,92	66,64	**92,03**	52,23
GOOSE	41,72	57,74	89,42	42,64
NPAPER	36,18	54,78	88,68	39,63
INSCRI	34,98	37,10	23,22	**92,22**
HTML2T	30,95	33,45	20,56	89,80

Tableau 2 – Mesure `occ_eval` sur 5 langues (grec, anglais, polonais, russe, chinois)

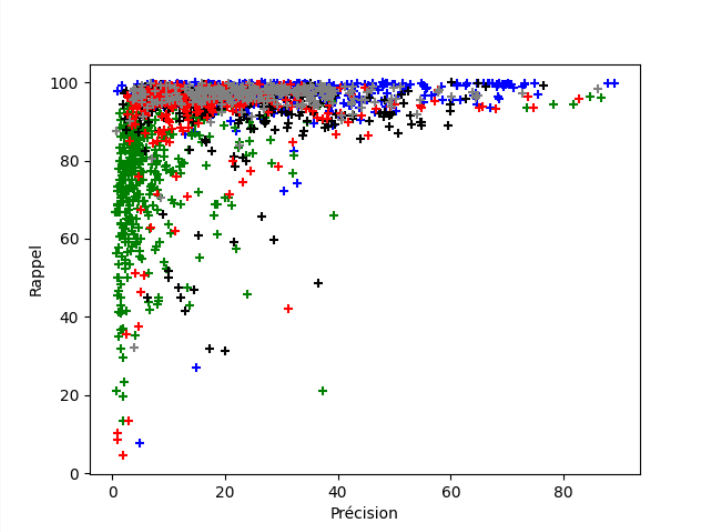
Les dangers de l'agrégation (I)



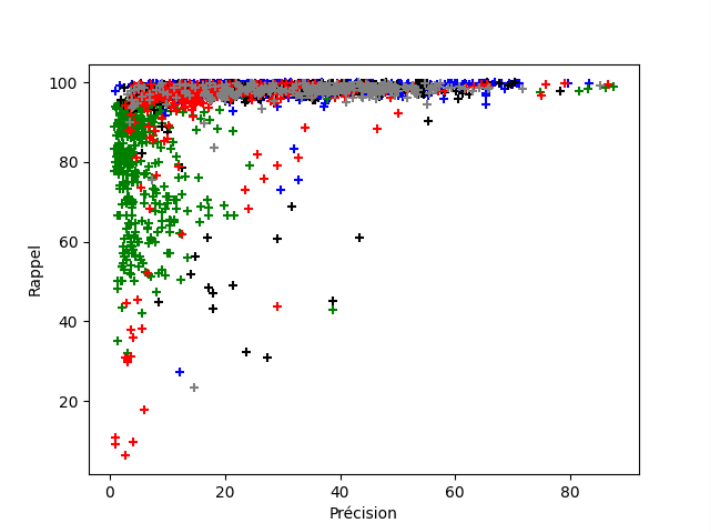
(a) `occ_eval` NEWSPAPER



(b) `occ_eval` GOOSE

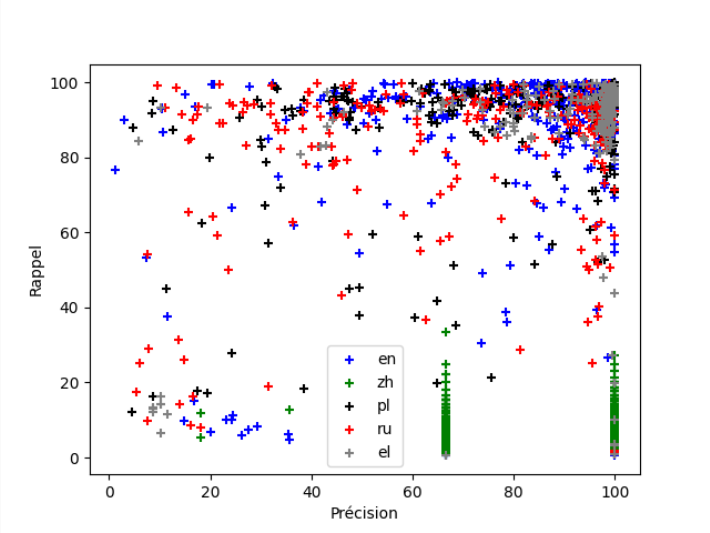


(c) `occ_eval` HTML2TEXT

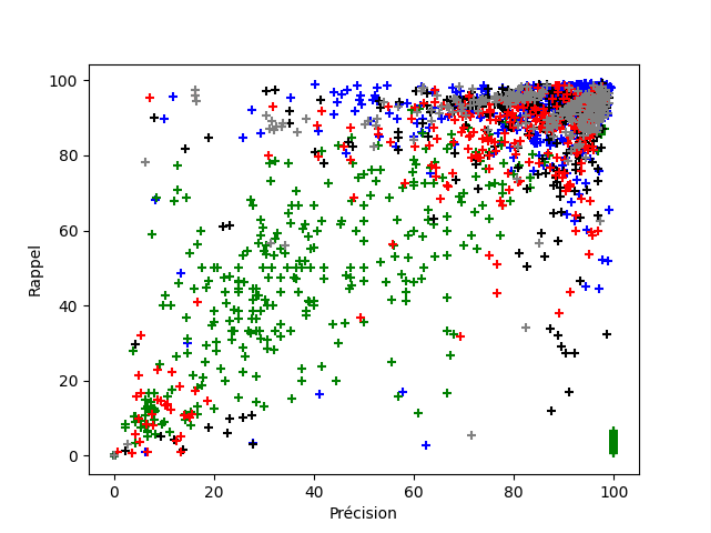


(d) `occ_eval` INSCRIPTIS

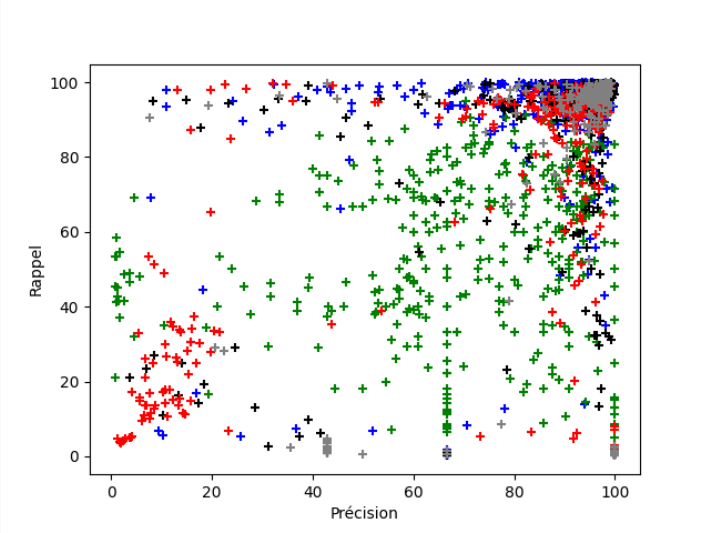
Les dangers de l'agrégation (II)



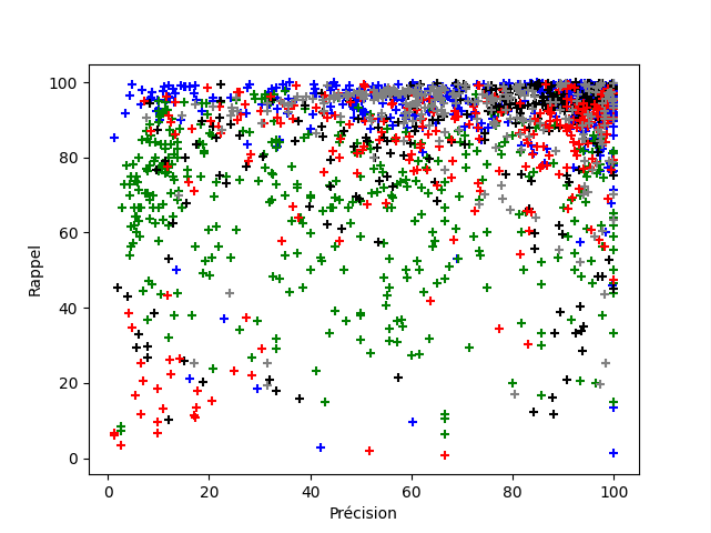
(a) `occ_eval` JUSTEXT



(b) `occ_eval` READABILITY



(c) `occ_eval` BP3_ART



(d) `occ_eval` TRAFILATURA

Les dangers du passage à l'échelle

Outil	Temps de Traitement	Ratio VS le plus rapide
INSCRI	**19,7**	x1
DRAG	24,0	x1,2
BP3_KeepE	37,5	x1,9
BP3_Larg	37,7	x1,9
JT_english	41,5	x2,1
READ	56,8	x2,9
HTML2T	71,0	x3,6
NPAPER	105,5	x5,5
JT_langid	112,6	x5,7
GOOSE	191,3	x9,7
JT	322,0	x16,3
NPLEASE	3755,6	x190

Tableau 3 – Test de vitesse sur les 1600+ documents du corpus

Conclusion

Conclusion

Que peut(on en déduire ?

- Évaluation Intrinsèque difficile à comprendre
  - La distance d'édition entre séquences insatisfaisante
  - Doit-on évaluer par document ou par source ?

Conclusion

Que peut(on en déduire ?

- Évaluation Intrinsèque difficile à comprendre
  - La distance d'édition entre séquences insatisfaisante
  - Doit-on évaluer par document ou par source ?
- Évaluation Extrinsèque est intéressante
  - Elle contredit la hiérarchie *a priori* entre les outils
  - Mais n'est pas exempte de biais (quid d'une autre tâche ?)

Conclusion

Que peut(on en déduire ?

- Évaluation Intrinsèque difficile à comprendre
  - La distance d'édition entre séquences insatisfaisante
  - Doit-on évaluer par document ou par source ?
- Évaluation Extrinsèque est intéressante
  - Elle contredit la hiérarchie *a priori* entre les outils
  - Mais n'est pas exempte de biais (quid d'une autre tâche ?)

Se méfier de l'évaluation . . .et des évaluateurs

Exemples d’outils

Quelques solutions pour le scraping tout de même

- Pour des articles de presse ou forums, des outils génériques existent tels que TRAFILATURA Tutoriel en ligne :
- `https://github.com/rundimeco/waddle/tree/master/Tutoriel_X-COTE`[2]
- Outil générique MAIS il faut programmer (un peu)

- Pour paramétrer à la main mais sans vraiment coder :
- `https://webscraper.io/`
- Pas besoin de programmer MAIS paramétrage coûteux

2. Documentation : `https://trafilatura.readthedocs.io/`