Journalisme Audiovisuel, Supports Mobiles et Data-Journalisme

Master ProCAN, Semestre 3
Pr. Abdelhak Mahmoudi
abdelhak.mahmoudi@um5.ac.ma
Octobre 2019

02-Nettoyage Des Données

Plan

- Encodage des caractères
- Correction des fautes d'orthographes
- Correction des erreurs dans les données
- Crowdsourcing
- Bases de données

Plan

- Encodage des caractères
- Correction des fautes d'orthographes
- Correction des erreurs dans les données
- Crowdsourcing
- Bases de données

Encodage des caractères

- **ASCII** (American Standard Code for Information Interchange) :
  - début de l'informatique aux USA : 128 codes à 7 bits,
- **UTF-8** (Universal character set Transformation Format) :
  - en 2016 adopté par 90% des utilisateurs : 1 à 4 octets (8 à 32 bits)
- Voir d'avantage :
  - http://sdz.tdct.org/sdz/comprendre-les-encodages.html

65	1000001	A
66	1000010	B
90	1011010	Z
58	0111010	:
59	0111011	;
60	0111100	<
61	0111101	=

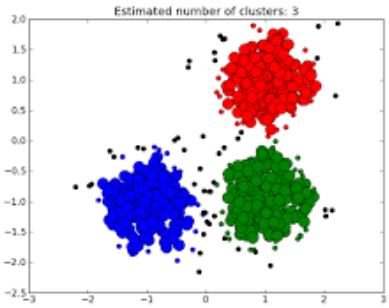
U+0633	س	1101100010110011
U+0634	ش	1101100010110100
U+0635	ص	11011000101101010000101010
U+0636	ض	1101100010110110

Plan

- Encodage des caractères
- Correction des fautes d'orthographes
- Correction des erreurs dans les données
- Crowdsourcing
- Bases de données

Correction des fautes d'orthographes

- Outil: Openrefine
- Eliminer les espaces
- Majuscules, Minuscules
- Concaténation
- Méthodes automatiques de clustering

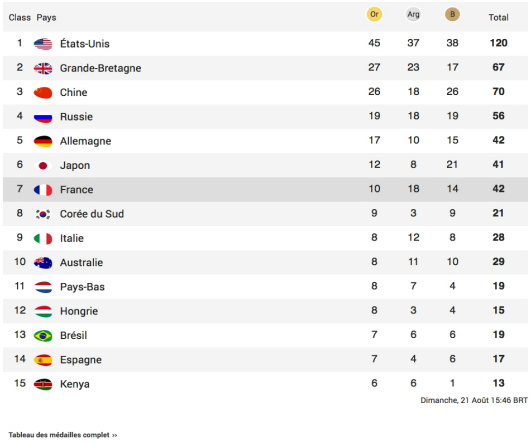


Plan

- Encodage des caractères
- Correction des fautes d'orthographes
- Correction des erreurs dans les données
- Crowdsourcing
- Bases de données

Correction des erreurs dans les données

- Outil: Openrefine
- Trouver les valeurs invalides (textes, nombres) avec Facets
- Trouver les valeurs aberrantes (histogramme)
- Trouver les doublants
  - méthode de clustering
  - Voir demo de kNN



Class	Pays	Or	Arg	B	Total
1	États-Unis	45	37	38	120
2	Grande-Bretagne	27	23	17	67
3	Chine	26	18	26	70
4	Russie	19	18	19	56
5	Allemagne	17	10	15	42
6	Japon	12	8	21	41
7	France	10	18	14	42
8	Corée du Sud	9	3	9	21
9	Italie	8	12	8	28
10	Australie	8	11	10	29
11	Pays-Bas	8	7	4	19
12	Hongrie	8	3	4	15
13	Brésil	7	6	6	19
14	Espagne	7	4	6	17
15	Kenya	6	6	1	13

Dimanche, 21 Août 15:46 BRT

Tableau des médailles complet ››

Plan

- Encodage des caractères
- Correction des fautes d'orthographes
- Correction des erreurs dans les données
- Crowdsourcing
- Bases de données

Crowdsourcing

- Demander aux gens de nettoyer les données existantes pour vous
- Expliquer avec des instructions claires et simples
- Outils: googleforms, web-apps, mobile-apps
- Exemple: ProPublica
  - https://www.propublica.org/atpropublica/propublica-launches-collaborate-tool-to-help-newsrooms-tackle-large-data-projects-together

Crowdsourcing

- PyBossa (https://pybossa.com/):
  - Vous permet de créer de petites tâches pour votre public
  - Exemple: NLP, transcription PDF, image classification, geocoding (transformer des adresse en coordonnée géo), etc.
- Amazon Mécanique Turc (https://www.mturk.com)
  - Payer les gens pour leurs tâches
- Attentions:
  - Les volontaires ne trouveront pas toujours les valeurs aberrantes, les valeurs incorrectes ou non valides
  - Ne devrait pas être votre option par défaut

Plan

- Encodage des caractères
- Correction des fautes d'orthographes
- Correction des erreurs dans les données
- Crowdsourcing
- Bases de données

Bases de données

- Une Base de données:
  - Constituée de plusieurs tables reliées entre elles
- Dans le cas où vous avec une tables très grande,
  - Divisez en autant de tables que possible
  - Isolez les différents éléments afin d'avoir le moins d'éléments à nettoyer
- Outils:
  - Mysql, Sql server, Oracle, MongoDB, etc

À lire

- OpenRefine with Arabic: https://www.linkedin.com/pulse/cleaning-arabic-data-openrefine-abed-khooli/