seance_1

September 28, 2020

1 Révisions DAC rentrée 2020

Les trois séances utiliseront le même jeu de données, à télécharger ici. Ce jeu de données contient tous les articles publiés sur arxiv (un site d'open-access pour article scientifique) de 2010 à 2017, avec les informations des auteurs, domaine, date de publication, titre et résumé, doi du journal de publication si disponible, etc.

Le programme des séances est le suivant : * Mardi : analyse exploratoire des données et étude du comportement des publications en fonction des domaines * Mercredi : apprentissage non supervisé sur le corpus à partir des résumés * Jeudi : apprentissage supervisé pour classifier dans la/les bonnes catégories à partir des résumés.

L'objectif de ces séances est de vous remettre à niveau à la fois en apprentissage statistique et en programmation python. N'hesitez pas à explorer vos propres pistes à partir des indications (très limitées) de l'énoncé, d'expérimenter et d'explorer les sujets sur lesquels vous n'êtes pas à l'aise.

Côté programmation, la consigne principale est : **ne jamais faire de boucle** (pour du calcul numérique). Vous pouvez utiliser tous les modules python qui vous semblent nécessaires (sauf contre-indication) : ***sklearn, pandas, nltk, matplotlib, seaborn ...***

Dans toute cette série de séances : * on considère comme date de publication de l'article la date correspondant à la première version (v1) * on ne prend en compte que les articles qui ont un champ ***doi*** non vide (articles publiés par ailleurs dans des conférences ou revues) * seules les informations sur la date de publication, les catégories des articles et le résumé seront considérés.

Pour cette première séance, on ne travaillera que sur la date de publication et les catégories des articles.

1.1 Analyse exploratoire des données

Proposez dans un premier temps des visualisations qui permettent d'appréhender les données de manière simple.

Choisissez trois catégories/communautés (par exemple **cs.AI**, **math.CA**, **physics.optics**) et observez le rythme des publications dans chacune en fonction du mois selon les années, puis en fonction du mois et du jour de la semaine en fonction des années.

On aimerait : * comprendre à quel point les publications sont faites de manière eparpillée ou si il existe des moments privilégiés où les scientifiques publient (en fonction des communautés et des

années); * comparer les habitudes de publications dans une même communauté en fonction des années; * comparer les habitudes de publications entre les communautés en fonction des années.

Quelques questions pour vous guider : * Quels outils visuels proposez vous (évaluation qualitative) ? * Quels outils pour une évaluation quantitative ? * Doit-on se satisfaire d'une modélisation discrète ou une modélisation continue est nécessaire ? * Comment régler les hyper-paramètres ?

[1]:
```
import json
from datetime import datetime
import sklearn
import numpy as np
from scipy.sparse import csr_matrix,lil_matrix
from calendar import monthrange
import gzip
```

[ ]:
```
FILENAME = "arxiv-2010-2017.json.gz"

def read_arxiv(f=FILENAME):
    res = dict()
    with  gzip.open(f) as fp:
        for l in fp:
            js = json.loads(l)
            res[js["id"]] = js
    return res

def get_dates(dic):
    return [datetime.strptime(x["created"],"%a, %d %b %Y %H:%M:%S GMT") for k␣
 ↪in dic.values() for x in k["versions"] if x['version']=="v1"]
def get_day(d):
    return (d.weekday()+(d.hour+(d.minute+d.second/60)/60)/24)/7
def get_month(d):
    return d.month+(d.day-1+(d.hour+(d.minute+d.second/60)/60)/24)/monthrange(d.
 ↪year,d.month)[1]
```