

Projet

Nous disposons d'un fichier texte contenant des informations sur des articles scientifiques parus dans des revues et conférences. Ces informations comprennent :

- Le titre de l'article,
- le ou les auteurs,
- l'année de publication,
- le nom de la revue (ou de la conférence),
- les citations entre articles.

Partie 1. Analyse textuelle

- 1) Télécharger le jeu de données (fichier DBLP_Subset.txt sur le lien suivant <http://up5.fr/yMZbV>). Le jeu de données est un fichier texte sous le format suivant:

```
#* --- paperTitle
#@ --- Authors
#t --- Year
#c --- publication venue
#index 00---- index id of this paper
#% ---- the id of references of this paper (there are multiple lines, with each indicating a reference)
#! --- Abstract
```

The following is an example:

```
#*Information geometry of U-Boost and Bregman divergence
#@Noboru Murata,Takashi Takenouchi,Takafumi Kanamori,Shinto Eguchi
#t2004
#cNeural Computation
#index436405
#%94584
#%282290
#%605546
#%620759
#%564877
#%564235
#%594837
#%479177
#%586607
```

```
#!We aim at an extension of AdaBoost to U-Boost, in the paradigm to build a stronger classification machine from a set of weak learning machines.
```

Chaque information est introduite par un caractère spécifique comme dans l'exemple ci-dessus.

- 2) Charger le fichier texte sous python.
- 3) Réaliser les traitements de données (Analyse textuelle) afin d'extraire les informations séparément sous forme d'un dataframe (figure ci-dessous). En considérant les éléments suivants:
 - a. Les revues qui ont moins de 50 articles ne seront pas considérées.
 - b. Le ground truth (soit les vrais labels) dans notre cas sont les revues ou ont été publiés les articles
- 4) Appliquer sur les *abstracts* les différents traitements vus en cours , à savoir:
 - a. Suppression des ponctuation,
 - b. Suppression des stopwords,
 - c. Suppression des chiffres,
 - d. Lemmatisation,

- 5) Construire la matrice Documents-termes et documents-Entités (après extraction des entités nommées) pour les résumés (abstracts) articles.
- 6) Construire la matrice Documents-auteurs, qui vous aidera par la suite à analyser les résultats du clustering.

	Venue	Year	Authors	Title	NbrAuthor	Id	ListCitations	NbrCitations
1	#cSIAM Journal of Applied Mathematics	2005	.Christian Schmeiser,Yasmin Dolak	The Keller-Segel Model with Logistic Sensitivity Func...	2	1057374		0
2	#cSIAM Journal of Applied Mathematics	1977	.David S. Johnson,M. R. Garey	The Rectilinear Steiner Tree Problem in NP Complete.	2	1057375		0
3	#cSIAM Journal of Applied Mathematics	2006	.F. O'Doherty,James P. Gleeson	Non-Lorentzian Spectral Lineshapes Near a Hopf Bif...	2	1057376		0
4	#cSIAM Journal of Applied Mathematics	1969	.Ronald L. Graham	Bounds on Multiprocessing Timing Anomalies.	1	1057377		0
5	#cSIAM Journal of Applied Mathematics	1995	.Blake Temple,Eli Isaacson	Convergence of the s2'times 2s Godunov Method f...	2	1057378		0
6	#cSIAM Journal of Applied Mathematics	1995	.Antonin Chambolle	Image Segmentation by Variational Methods: Mumf...	1	1057379		0
7	#cSIAM Journal of Applied Mathematics	1995	.Vijay K. Samalam,Thomas M. Chen	Time-Dependent Behavior of Fluid Buffer Models wit...	2	1057380		0
8	#cSIAM Journal of Applied Mathematics	1995	.David A. Edwards	Constant Front Speed in Weakly Diffusive Non-Fickia...	1	1057381		0
9	#cSIAM Journal of Applied Mathematics	1995	.David R. Kassoy,Meng Wang	Nonlinear Oscillations in a Resonant Gas Column: A...	2	1057382		0
10	#cSIAM Journal of Applied Mathematics	1995	.Eric Henderson,James Vesenka,Richard Miller	Tip Reconstruction for the Atomic Force Microscope.	3	1057383		0
11	#cSIAM Journal of Applied Mathematics	1995	.Robert T. Tranquillo,Richard B. Dickinson	Transport Equations and Indices for Random and Bia...	2	1057384		0
12	#cSIAM Journal of Applied Mathematics	1995	.Michael Landman,Nancy Kopell	Spatial Structure of the Focusing Singularity of the ...	2	1057385		0
13	#cSIAM Journal of Applied Mathematics	1995	.P. A. Martin,C. J. Luke	Fluid-Solid Interaction: Acoustic Scattering by a Smo...	2	1057386		0
14	#cSIAM Journal of Applied Mathematics	1995	.Zbigniew Galias	On a Discrete-Time Nonlinear System Associated wit...	1	1057387		0
15	#cSIAM Journal of Applied Mathematics	1995	.Michael Vogelius,Fadil Santosa	Erratum to the Paper: First-Order Corrections to the ...	2	1057388		0
16	#cSIAM Journal of Applied Mathematics	1995	.Robert Almgren	Crystalline Saffman-Taylor Fingers.	1	1057389		0
17	#cSIAM Journal of Applied Mathematics	1995	.Stanley Ocken	Recognizing Convergent Orbits of Discrete Dynamic...	1	1057390		0
18	#cSIAM Journal of Applied Mathematics	1995	.Tzy-Wei Huang,Sze-Bi Hsu	Global Stability for a Class of Predator-Prey Systems.	2	1057391		0
19	#cSIAM Journal of Applied Mathematics	1995	.Aslak Tveito,Nils Henrik Risebro,Helge Holden	Maximum Principles for a Class of Conservation Laws.	3	1057392		0
20	#cSIAM Journal of Applied Mathematics	1995	.R. H. Tew,John R. Ockendon,J. R. King,S. J. Chapman...	Stokes Phenomenon and Matched Asymptotic Expa...	5	1057393		0
21	#cSIAM Journal of Applied Mathematics	1995	.Amitabha Bose	Symmetric and Antisymmetric Pulses in Parallel Cou...	1	1057394		0
22	#cSIAM Journal of Applied Mathematics	1995	.Wolfgang Ring	Identification of a Core from Boundary Data.	1	1057395		0
23	#cSIAM Journal of Applied Mathematics	1995	.Brian T. R. Wetton,John M. Stockie	Stability Analysis for the Immersed Fiber Problem.	2	1057396		0
24	#cSIAM Journal of Applied Mathematics	1995	.Vladimir A. Sharafutdinov	On Emission Tomography of Inhomogeneous Media.	1	1057397		0
25	#cSIAM Journal of Applied Mathematics	2005	.George Christakos,Hwa-Lung Yu	Porous Media Upscaling in Terms of Mathematical E...	2	1057398		0
26	#cSIAM Journal of Applied Mathematics	1995	.E. A. Spiegel,L. N. Howard,N. J. Balmforth	Instability of Rapidly Rotating Polytropes.	3	1057399		0
27	#cSIAM Journal of Applied Mathematics	1995	.Phoebus Rossakis	An Equal Area Rule for Dissipative Kinetics of Propa...	1	1057400		0
28	#cSIAM Journal of Applied Mathematics	1995	.Thomas P. Witelski,Andrew B. White Jr.,Donald S. Co...	Shock Formation in a Multidimensional Viscoelastic ...	3	1057401		0

Partie 2. Analyse exploratoire des données

- 7) En utilisant les mesures et graphiques statistiques vus en cours, réaliser une analyse descriptive des données:
 - Les auteurs les plus productifs.
 - Les articles les plus populaires (nombre de citations).
 - Les années avec une grande productivité.
 - Distribution du nombre d'articles par revue/conférence.
 - Distribution du nombre de citations.
 - Les termes les plus utilisés (nuage de mots, barplots, etc).

Vous pouvez proposer une multitude d'analyses en croisant ces différentes variables !

Partie 3. Classification non supervisée (Clustering)

- 8) Appliquer K-means avec un nombre de classes choisi sur la matrice documents-termes et documents-entités.
- 9) Interpréter les classes (clusters) obtenues.
 - Les termes les plus fréquents pour chaque cluster.
 - Les entités les plus fréquentes pour chaque cluster.
 - Le type de revues pour chaque cluster.
 - Évolution des clusters selon le temps (les années).
 - Analyse des clusters selon le nombre de citations des articles contenus dans chaque cluster.

- 10) Refaire l'analyse en utilisant Spherical K-means.

Vous pouvez proposer une multitude d'analyses en croisant ces différentes variables !

Date de remise: 12/11/2021 à 23h59 pour les AMDS et 25/11/2012 à 23h59 pour les MLDS. Le projet sera remis sous format de notebook (code + commentaire) et envoyé à l'adresse e-mail suivante rafika.boutalbi@ipvs.uni-stuttgart.de