Compte-rendu d'activité sur la plateforme OSIRIM

Année 2018

Les responsables des projets hébergés par la plateforme sont invités à fournir un bref bilan d'activité annuel. Ce document rassemble les éléments qui ont été retournés au titre de l'année 2018. Il n'est que partiel dans la mesure où tous les rapports attendus n'ont pas pu être collectés au moment de produire ce compte-rendu.

Equipe Pyramide

Thèse de Mohamed Mehdi KANDI : Allocation de ressource élastique pour l'optimisation de requêtes - Responsables du projet : Abdelkader Hameurlain (Abdelkader.Hameurlain@irit.fr), Shaoyi Yin (Shaoyi.Yin@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

L'objectif des travaux de thèse de Mohamed Mehdi KANDI est de proposer des méthodes d'allocation de ressources pour l'interrogation des bases de données dans le Cloud. L'allocation de ressources englobe le dimensionnement, la placement et l'ordonnancement, d'abord dans un cadre statique (au moment de la compilation de la requête) puis dans un cadre dynamique (durant l'exécution de la requête). Le but c'est de minimiser le coût économique côté fournisseur et même temps respecter les contrats SLA établies avec les clients.

Dans ce contexte, la plateforme OSIRIM est exploitée pour : 1) lancer des requêtes sur Hive afin d'observer la structure du plan d'exécution représentés par des graphes orientés acycliques (DAG), 2) résoudre des modèles ILP (Integer-Linear programming) via l'outil GLPK, 3) lancer des programmes java qui simulent les méthodes d'allocation de ressources proposées et quelques méthodes de l'état de l'art.

Résultats obtenus grâce aux services de la plateforme OSIRIM :

- Validation de nos modèles ILP.

- Comparaison des coûts économiques de nos méthodes de placement-ordonnancement basées sur les ILP et quelques méthodes existantes.

- Comparaison de plusieurs stratégies de dimensionnement avec un ou plusieurs agents parallèles.

Publications :

- **Mohamed Mehdi Kandi**, **Shaoyi Yin**, **Abdelkader Hameurlain**
  An Integer Linear-programming based Resource Allocation Method for SQL-like Queries in the Cloud (regular paper).
  ACM Symposium on Applied Computing (SAC 2018), Pau, France, 09/04/2018-13/04/2018, **ACM**, avril 2018

Equipe SAMOVA

Travaux de l'équipe (Julien PINQUIER, julien.pinquier@.fr)

L'équipe SAMOVA travaille sur la modélisation de l'audio et de la vidéo. Les travaux actuels concernent la réalisation de moteurs de reconnaissance automatique de la parole grand vocabulaire, la classification d'événements et d'environnements sonores et la mesure automatique de l'intelligibilité de la parole.

Nous nous intéressons à la réalisation de modèles profonds adaptés à ces tâches, mais également à concevoir des architectures de réseaux profonds plus adaptés pour ces tâches.

https://www.irit.fr/recherches/SAMOVA/

Activités scientifiques réalisées sur la plateforme OSIRIM :

Réalisation de modèles acoustiques pour la reconnaissance automatique de la parole, classification en événements et environnements sonores, réalisation de traitements automatiques pour la mesure de l'intelligibilité de la parole.

Résultats obtenus grâce aux services de la plateforme OSIRIM :

Utilisation multi-GPU qui permet d'avoir des modèles calculés en 20 jours au lieu d'un an (1 seul GPU) pour l'apprentissage de modèles acoustiques sur 1600h d'enregistrements audio.

Pour la classification en événements sonores, 12h de calcul sur CPU ne prennent plus que 1h30 sur GPU.

Equipe SIG

Projet “Détection automatique de la déforestation” (Josiane MOTHE, mothe@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

Segmentation et classification d'images Copernicus. Une étudiante en thèse étudie l'utilisation du deep learning dans le cadre de la détection de la déforestation.

Deux groupes d'étudiants de Master ont travaillé pour une initiation au deep learning

Résultats obtenus grâce aux services de la plateforme OSIRIM :

Premiers traitements relatifs à la tâche de détection automatique de déforestation

Publications :

- Duy Huynh, Josiane Mothe, Nathalie Neptune.
  Automatic image annotation : the case of deforestation (regular paper). Dans / In : Rencontres Jeunes Chercheurs en Recherche d'Information (RJCRI 2018), Rennes, 14/05/2018-18/05/2018
  Association Francophone de Recherche d'Information et Applications (ARIA), (support électronique), mai / may 2018.

- Nathalie Neptune, Josiane Mothe, Julius Akinyemi
  Fouille de publications scientifiques pour une analyse bibliométrique de l'activité de recherche sur la déforestation (regular paper). Dans / In : EGC - Atelier Fouille du Web, Paris, 23/01/2018
  Association Internationale Francophone d'Extraction et de Gestion des Connaissances (EGC), (support électronique), janvier / january 2018.
  URL : https://www.irit.fr/publis/SIG/2018_EGC_AMN.pdf

- Nathalie Neptune, Josiane Mothe.
  Analyse bibliométrique : Une aide pour l'évaluation des unités de recherche (regular paper). Dans / In : Colloque Veille Stratégique Scientifique et Technologique (VSST 2015), Université de Grenade (Espagne), 11/05/2015-13/05/2015
  Veille Stratégique Scientifique et Technologique(VSST), (support électronique), 2015.
  Résumé / Abstract URL : https://www.irit.fr/publis/SIG/2015_VSST_NNMJ.pdf - http://oatao.univ-toulouse.fr/15267/

- Julius Akinyemi, Josiane Mothe, Nathalie Neptune.
  Fouille de publications scientifiques pour une analyse bibliométrique de l'activité de recherche sur la déforestation. novembre / november 2017. Poster à la Journée commune AFIA - ARIA
  URL : https://www.irit.fr/publis/SIG/2017_AFIA-ARIA_JA_JM_NN.pdf

- William Lefiot (william.lefiot@wanadoo.fr), Victoria Yinka Adeoye (yinkaadeoye6@gmail.com), Niko Sahradyan (niko.sahradyan@gmail.com)
  MONITORING DEFORESTATION USING REMOTE SENSING DATA FROM COPERNICUS
  Rapport de projet de Master

Equipe SIG

Projet “ Aide au diagnostic du cancer” (Josiane MOTHE, mothe@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

Apprentissage supervisé (deep learning) pour détecter les parties cancéreuses sur une image

Résultats obtenus grâce aux services de la plateforme OSIRIM :

Comparaison de différentes architectures et évaluation des résultats en faisant varier différents paramètres

Equipe SIG

Projet “ VICTORIA” (Florence SEDES, sedes@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

Extraction automatique des objets (piétons et véhicules) présents au sein des vidéos du corpus TOCADA [1] via l'utilisation du réseau YOLOv2[2]

Expérimentation sur le transfert d'apprentissage pour la ré-identification des véhicules et des piétons dans les systèmes de vidéos surveillances (copus VeRI[3] et TOCADA[3])

Résultats obtenus grâce aux services de la plateforme OSIRIM :

Constitution de meta-données de contenu pour le filtrage négatif des vidéos dans les systèmes de vidéo-surveillances.

Reproduction de l'état de l'art sur la ré-identification des véhicules du corpus VeRI.

Résultats préliminaires encourageant sur le fine-tuning inductif des réseaux de l'état de l'art par encodage parcimonieux pour la ré-identification des véhicules du corpus VeRI.

Equipe SIG

Thèse CIFRE de Nabil El Malki : (Olivier Teste, olivier.teste@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

Nabil utilise OSIRIM pour lancer une succession d'expérimentations consistant à comparer deux algorithmes :

- kmeans standard

- kmeans modifié par ces propositions de thèse

Pour cela il génère des données artificielles (vecteur de valeurs numériques faisant varier la dimension/taille ainsi que la distribution des données dans l'espace) et fait varier différents paramètres qu'il mesure. L'expérimentation en cours nécessite environ 2000 runs.

Résultats obtenus grâce aux services de la plateforme OSIRIM :

Nous avons identifié les caractéristiques des données en fonction desquelles il est souhaitable (accélération) d'utiliser notre algorithme par rapport à l'algorithme kmeans standard

Publications :

En cours de rédaction, soumission prévue d'ici fin 2018

Equipe MELODI

Travaux de l'équipe (Tim Van De Cruys, tim.vandecruys@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

L'équipe MELODI utilise le cluster OSIRIM pour effectuer diverses tâches de traitement automatique du langage naturel (TAL). Une première tâche consiste à l'annotation syntaxique automatique (syntactic parsing) de corpora massifs. Le cluster a été utilisé pour annoter automatiquement jusqu'à 3 milliards de mots de textes en anglais et jusqu'à 2 milliards de mots de textes français. En raison du fait que le processus est facilement parallélisable, le cluster est parfaitement adapté à la tâche.

Une deuxième tâche consiste à la construction d'espaces de mots vectorielles. Dans un modèle espace-mot, les différent mots d'une langue sont encodés dans un modèle d'espace vectoriel selon les contextes dans lesquelles ces mots apparaissent. Un tel modèle permet de comparer (calculer la similarité entre) les différents mots, ce qui est utile pour les applications sémantiques. De nouveau, la construction est fortement parallélisable, ce qui le rend avantageux d'utiliser le cluster.

Une troisième tâche concerne l'éntraînement d'architectures neuronales pour effectuer des tâches de traitement automatique de langues. Il s'agit de la construction de plongements de phrases pour le calcul de similarité, ainsi que des modèles de langue pour la génération de textes.

Finalement, le cluster, et notamment le base de tweets français, a été utilisé pour effectuer de l'apprentissage non-supervisé.

Publications :

- Julien Hay, Tim Van de Cruys, Philippe Muller, Bich-Liên Doan, Fabrice Popineau, Lyes Benamsili. 2018. In Extraction et Gestion des Connaissances, EGC 2018, Paris, France, pp. 179-190

- Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. Computational Linguistics

- Damien Sileo, Tim Van de Cruys, Philippe Muller et Camille Pradel. 2018. Concaténation de réseaux de neurones pour la classification de tweets. DEFT 2018, Rennes, France, pp 259-264.

Equipe IRIS

Projet LISTIC (Guillaume CABANAC, guillaume.cabanac@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

LisTIC (« Liens socionumériques et Technologies (mobiles) de l'Information et de la Communication ») est un projet de recherche interdisciplinaire coordonné par des chercheurs en sociologie du LISST de l'université de Toulouse 2 en partenariat avec des chercheurs en sciences de l'information et de la communication et en informatique affiliés au LERASS et à l'IRIT de l'université de Toulouse 3. Ce projet est financé de 2016 à 2020 par l'Agence Nationale de la Recherche (projet ANR-16-CE26-0014-01) et a obtenu le label du GDRI Web Science du CNRS. Les différentes enquêtes réalisées dans le cadre de ce projet, à l'aide des méthodes des humanités numériques, ont été mises en place à l'aide de la plateforme OSIRIM.

Le projet LisTIC explore les usages contemporains des téléphones mobiles et des applications développées par les principaux médias sociaux dans le but de mieux cerner leurs effets sur la composition et la morphologie des réseaux relationnels, sur les sociabilités tissées au travail et sur les formes médiatisées de la participation politique, que ce soit en France durant les élections présidentielles ou dans le cadre des mouvements sociaux actuels au Brésil.

Pour cartographier ces pratiques, le projet LisTIC propose de développer les méthodologies des humanités numériques en exploitant les ressources offertes par les smartphones afin d'en documenter les usages, notamment les formes nomades de la participation aux applications mobiles des réseaux socionumériques (mSNS). L'objectif est de faire converger des protocoles de recherche susceptibles de nous livrer un aperçu exhaustif de l'appropriation contemporaine de ces technologies et de la manière dont elles infléchissent, voire renouvellent, nos structures sociales et relationnelles. Pour relever cet ambitieux défi sociologique, des méthodologies innovantes sont développées en lien étroit avec les méthodes classiques des SHS et leurs acquis théoriques pour favoriser au maximum la cumulativité des connaissances.

L'hypothèse de ce projet revient à percevoir les smartphones et les déclinaisons nomades des réseaux socionumériques comme le creuset technologique où se forge la propagation d'une forme sociale qui tend à être dominante aujourd'hui, celle d'un « individualisme en réseaux », soit une manière d'être et de tisser des liens sociaux de plus en plus travaillée de manière individualisée sous l'amplification des sollicitations sociales et relationnelles émanant des TIC. Cette hypothèse sociologique est explorée dans le cadre de quatre recherches qui ont été conçues pour appréhender trois modes d'appropriation des smartphones :

- mSNS & Réseaux : la manière dont les usages nomades des réseaux socionumériques viennent s'articuler avec les pratiques téléphoniques et leurs effets sur la morphologie des réseaux de relations personnels

- E-participation en France et au Brésil : les formes nomades de la e-participation aux mouvements sociaux (un cas en France, un cas au Brésil) et la manière dont elles favorisent une hybridation des arènes publiques à l'interface entre les territoires urbains et numériques

- mSNS & Travail : l'inscription des usages des réseaux socionumériques dans l'écologie des activités professionnelles et la manière dont la disponibilité permanente des affinités relationnelles

personnelles vient concurrencer les sociabilités professionnelles d'autant plus vécues comme « contraintes ».

Dans ce cadre, OSIRIM héberge depuis novembre 2016 une machine virtuelle sur laquelle les services suivants sont déployés par les partenaires IRIT :

- site web du projet https://listic.irit.fr
- moissonneur Twitter DMI-TCAT (https://github.com/digitalmethodsinitiative/dmi-tcat/wiki)
- moissonneur Facebook : développement interne basé sur Netvizz (https://apps.facebook.com/netvizz)
- plugin d'enquêtes de sociologie (développement interne) basé pour LimeSurvey (https://listic.irit.fr/limesurvey)
- gestionnaire de code SVN

Résultats obtenus grâce aux services de la plateforme OSIRIM :

La première phase du projet a consisté à moissonner des données depuis novembre 2016, liées à l'élection présidentielle française de 2017. Afin de questionner l'usage des médias sociaux par des activistes en politique, ces données sont collectées à partir de :

- Twitter : 42 millions de tweets (72 Go) relatifs,
- Facebook : 450 000 posts (4 Go) liés à 81 groupes et 214 pages.

L'analyse des données moissonnées a permis d'informer la constitution de l'échantillon des personnes enquêtées par entretien individuel. Elles permettent de confronter les discours des différents partis politiques (par analyse lexicométrique), d'identifier les individus influenceurs (par analyse résiliaire), les moments de rupture dans la campagne (par analyse de séries temporelles), les pratiques des activistes (par analyses de pratiques égocentrées), etc.

Nos résultats de recherche sont publiés dans les actes des conférences :
- ACM Hypertext & Social Media (HT 2018),
- AAAI International Conference on Weblogs and Social Media (ICWSM 2018),
- International Conference on the Theory of Information Retrieval (ICTIR 2017)

et ont fait l'objet de plusieurs communications, notamment au Datapol organisé par Sciences Po Paris : http://www.medialab.sciences-po.fr/projets/datapol/ .

Publications :

---------- conférences internationales ----------

- Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besancon, Mohand Boughanem. Stance Classification through Proximity-based Community Detection (regular paper). Dans : ACM Conference on Hypertext & Social Media (HT 2018), Baltimore, Maryland, USA, 09/07/2018-12/07/2018, ACM, p. 220-228, juillet 2018.

https://doi.org/10.1145/3209542.3209549
https://www.irit.fr/publis/IRIS/2018_HT_FCPBB.pdf

- Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besancon, Mohand Boughanem. #Élysée2017fr: The 2017 French Presidential Campaign on Twitter (regular paper). Dans : International Conference on Weblogs and Social Media (ICWSM 2018), Stanford, California, États-Unis, 25/06/2018-28/06/2018, AAAI Press, juin 2018 (à paraître).
https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17821
https://www.irit.fr/publis/IRIS/2018_ICWSM_FCPBB.pdf

- Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besancon, Mohand Boughanem. Uncovering Like-minded Political Communities on Twitter (short paper). Dans : International Conference on the Theory of Information Retrieval (ICTIR 2017), Amsterdam, 01/10/2017-04/10/2017, Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke (Eds.), ACM, p. 261-264, octobre 2017.
https://www.irit.fr/publis/IRIS/2017_ICTIR_FCPBB.pdf
http://dx.doi.org/10.1145/3121050.3121091
https://www.irit.fr/publis/IRIS/2017_ICTIR_FCPBB_poster.pdf

--------- communications orales ----------

- Pascal Marchand, Pierre Ratinaud, Julien Figeac, Guillaume Cabanac, Ophélie Fraisier, Gilles Hubert, Xavier Milliner, Yoann Pitarch, Tristan Salord, Nikos Smyrnaios, Thibaut Thonet. La campagne présidentielle 2017 sur les réseaux socionumériques. Diffusion scientifique. mars 2018. Colloque du Labex Structuration des Mondes Sociaux (SMS), Toulouse, 27-29 mars 2018
Accès : https://www.irit.fr/~Guillaume.Cabanac/docs/sms2018.pdf

- Bilel Benbouzid, Nikos Smyrnaios, Julien Figeac, Ophélie Fraisier, Anne L'Hôte, Benjamin Loveluck, Alexis Perrier, Pierre Ratinaud, Tristan Salord. Twitter vs. Facebook : réseaux et discours de l'extrême droite. Présentation orale. décembre 2017. Participation du projet Listic au Sprint Datapol organisé par Sciences Po : http://www.medialab.sciences-po.fr/projets/datapol/
https://www.irit.fr/publis/IRIS/2017_Datapol_BSFFLLPRS.pdf
https://www.irit.fr/publis/IRIS/2017_Datapol_BSFFLLPRS_audio.mp4

- Nathalie Paton, Héloïse Prévost, Tristan Salord, Julien Figeac, Guillaume Cabanac. Comment analyser une mobilisation collective dans les réseaux socionumériques : l'exemple des groupes Facebook brésiliens. Diffusion scientifique. septembre 2017. Séminaire « Pragmatic » du Labex Structuration des Mondes Sociaux, Université Toulouse 2, 28 septembre 2017
Accès : http://bit.ly/pragmaticListic2017

--------- vulgarisation ----------

- Julien Figeac, Nathalie Paton, Guillaume Cabanac. L'importance des réseaux sociaux au Brésil : le projet LisTIC. Diffusion scientifique. juillet 2017. Magazine « Nouvelles CNRS Rio : Sciences Brésil & Cône Sud », n°1, pages 24-26

Résumé Accès : https://web.archive.org/web/201707/http://www.cnrs-brasil.org/uploads/Numéro-1-Juillet-20171.pdf

Equipe IRIS

Thèse de Gia-Hung Nguyen : apprentissage de représentation pour la recherche d'information (Lynda Tamine Lechani, lynda.lechani@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

Poursuite des activités sur l'apprentissage de représentations (voir rapport 2017, rappelé ci-bas). Extension à l'utilisation des ressources sémantiques pour l'apprentissage de représentation des documents.

L'objectif des travaux de thèse de Gia est d'exploiter, adapter et développer des modèles d'apprentissage profond basé sur les réseaux de neurones pour 1) améliorer les représentations des textes en intégrant les connaissances issues de ressources sémantiques externes, exemples : WordNet, UMLS; 2) améliorer l'efficacité des modèles d'ordonnancement de documents en réponse à une requête utilisateur en se basant sur des architecture de type 'end to end'.

Dans ce cadre, la plateforme OSIRIM est exploitée pour : 1) indexer des collections de documents volumineuses (eg. Gov2, PubMed) et des ressources (WordNet,, UMLS); 2) eploiter les librairies (eg. Indri, Lucene, Tensor Flow, Torch) offertes pour implémenter les modèles neuronaux; 3) lancer les programmes d'apprentissage, test et les analyses des sorties associées

Résultats obtenus grâce aux services de la plateforme OSIRIM :

- Modèles d'ordonnancement end-to-end avec des appariements combinant les mots et concepts
-Techniques d'apprentissage joint des concepts, documents et mots
-Analyses spatio-temporelles d'événements détectés sur Twitter (en cours)

Publications :

- Gia Hung Nguyen, Lynda Tamine, Laure Soulier, Nathalie Souf.
  A Tri-Partite Neural Document Language Model for Semantic Information Retrieval (regular paper). Dans : European Semantic Web Conference (ESWC 2018), Crète (Grèce), 03/06/2018-07/06/2018, juin 2018

Equipe IRIS

Thèse de Paul Mouset dans le cadre du projet CIFRE IRIT-ATOS (2016-2019) : Valorisation de media par annotation événementielle générées à partir de flux d'informations hétérogènes (Lynda Tamine Lechani, lynda.lechani@irit.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

Lancement en 2018 d'une nouvelle activité liée à la mise en oeuvre d'un modèle de réseau de neurones profond pour l'appariement de tweets avec des objets spatiaux comme les POI (Place Of Interest). Dans ce cadre, la plate-forme OSIRIM est exploitée pour 1) implémenter le réseau de neurones profond; 2) Exploiter les librairies (lucene, Keras); 3) Lancer les programmes d'apprentissage/test et analyses

L'objectif des travaux de thèse de Paul Mousset est de développer des modèles de résumés spatio-temporels d'événements incluant : des techniques d'annotation spatiale de flux d'informations dans les réseaux sociaux; des techniques d'agrégation spatio-temporelle de l'information

Résultats obtenus grâce aux services de la plateforme OSIRIM :

- Modèles d'ordonnancement end-to-end avec des appariements combinant les mots et concepts
-Techniques d'apprentissage joint des concepts, documents et mots
-Analyses spatio-temporelles d'événements détectés sur Twitter (en cours)

Publications :

- Paul Mousset, Yoann Pitarch, Lynda Tamine.
  Studying the Spatio-Temporal Dynamics of Small-Scale Events in Twitter (regular paper).
  Dans : *ACM Conference on Hypertext & Social Media (HT 2018), BALTIMORE, MARYLAND, 09/07/2018-12/07/2018*, 2018

Equipe REVA

Apprentissage profond pour la segmentation sur des données bruitées

(Axel Carlier, axel.carlier@enseeiht.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

- Apprentissage automatique sur la segmentation sémantique d'images
- Entraînement et test d'algorithmes existant de l'état de l'art avec l'objectif de reproduire les résultats
- Ré-entraînement des algorithmes sur données altérées

Résultats obtenus grâce aux services de la plateforme OSIRIM :

Les résultats obtenus sont pour le moment préliminaires et n'ont pas donné lieu à publication

CLLE – ERSS

Travaux d'équipe (Franck Sajous, sajous@univ-tlse2.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

CLLE-ERSS recourt à la plateforme pour l'analyse syntaxique automatique de divers corpus de grande taille. Ces corpus analysés sont ensuite utilisés dans le cadre de travaux portant sur la sémantique distributionnelle. Les analyses syntaxiques produites sur Osirim sont aussi utilisées pour l'enrichissement de lexiques électroniques.

D'autres travaux portent sur la construction et l'évaluation d'embeddings.

Résultats obtenus grâce aux services de la plateforme OSIRIM :

Plusieurs travaux autour de la construction et l'évaluation de word embeddings.

Mise au point d'un système de prédiction de la liaison en français par dess GAM (modèles additifs généralisés).

Publications :

- B. Pierrejean and L. Tanguy :
  Predicting Word Embeddings Variability. Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM), 2018.

- B. Pierrejean and L. Tanguy :
  Towards Qualitative Word Embeddings Evaluation: Measuring Neighbors Variation. Proceedings of NAACL: Student Research Workshop, 2018.

- B. Pierrejean et L. Tanguy :
  Étude de la reproductibilité des word embeddings : repérage des zones stables et instables dans le lexique. Actes de TALN, 2018.

Projet MUSK (Nicolas Turenne, nturenne@u-pem.fr)

Activités scientifiques réalisées sur la plateforme OSIRIM :

Mes activités 2017-2018 portent sur l'extraction terminologique et la distribution lexicale à travers l'analyse des medias sociaux. Le media qui focalise mon attention est Youtube.

L'idée consiste à étudier l'activisme: comme le zadisme ou le lancement d'alerte.

Pour cela il faut mettre en place une plateforme de traitement dont les modules principaux sont :

- le crawling

- le filtrage

- l'extraction lexicale

- la classification automatique

Dans ce projet on cherche à comprendre la structuration du domaine, les catégories mises en jeu, la portée politique.

Résultats obtenus grâce aux services de la plateforme OSIRIM :

L'espace de données et l'espace des caractéristiques peuvent être assez importants, ce qui nécessite une parallélisation des traitements. C'est ce que j'ai fait jusqu'à présent à travers une configuration technique R/Hadoop/MongoDB. La quantité de videos extraites peut avoisiner 500,000. Il en va de même pour l'espace de description terminologique (entités nommées, dépendances relationnelles, groupes nominaux, bigrammes). Avec de tels espaces de descriptions sur-dimensionnés, les approches de structuration nécessitent aussi une parallélisation. L'utilisation de la plateforme OSIRM (slurm/hadoop) a permis de produire une chaîne de traitements capables de procéder à des traitements quantitatifs prometteurs.