Plongements temporels et lexicaux

Expériences sur des microblogs culturels

CLEF 2017 MC2 lab

Projet ANR GaFes

Cassandre Ollivier[(1)], Mathias Quillot[(2)], Eric SanJuan[(1,2)]

1. Inst. Univ. Tech. (IUT) Statistique Informatique Décisionnelle (STID)

2. Laboratoire d'Informatique (LIA)

Université d'Avignon

Sommaire

I. Introduction

Le projet ANR Galerie des Festivals

Etude des publics des Festivals :

- ➔ GECE Cabinet d'études et de sondages : enquêtes public par questionnaires,
- ➔ CNE-UAPV Centre Norbert Elias : études sociologiques,
- ➔ EURECOM: Analyse des images partagées,
- ➔ LIA-UAPV Laboratoire Informatique d'Avignon : Sciences des données,
- ➔ SYLLABS : Analyse sémantique.

I. Introduction

Lexiques proposés par sociologues

Listes de mots clés choisi par les sociologues pour réaliser des études et répondre à des hypothèses :

Auteur
Avignon
Billetterie
Cour d'honneur
Ecole du spectateur
Festival
Festivaliers
Jean Vilar...

Béatrice Macé
Jean Louis
Brossard
Publics
Numérique
Affiche
Rennes
Punk...

Avignon
Cannes
Cinéma
Cool
Fest
Jazz
Music
Rock...

I. Introduction

Lexiques issus du WikiPedia

Les plus agréables à vivre :
- Berlin, Copenhagen...

Festivals de musique :
- Miami, Boom...

Festivals de théâtre :
- Edinburgh, Avignon...

Festivals de cinéma :
- Hollywood, Cannes...

Thèmes théâtre :
- International Theatre
- Teatri
- Thespian Festival...

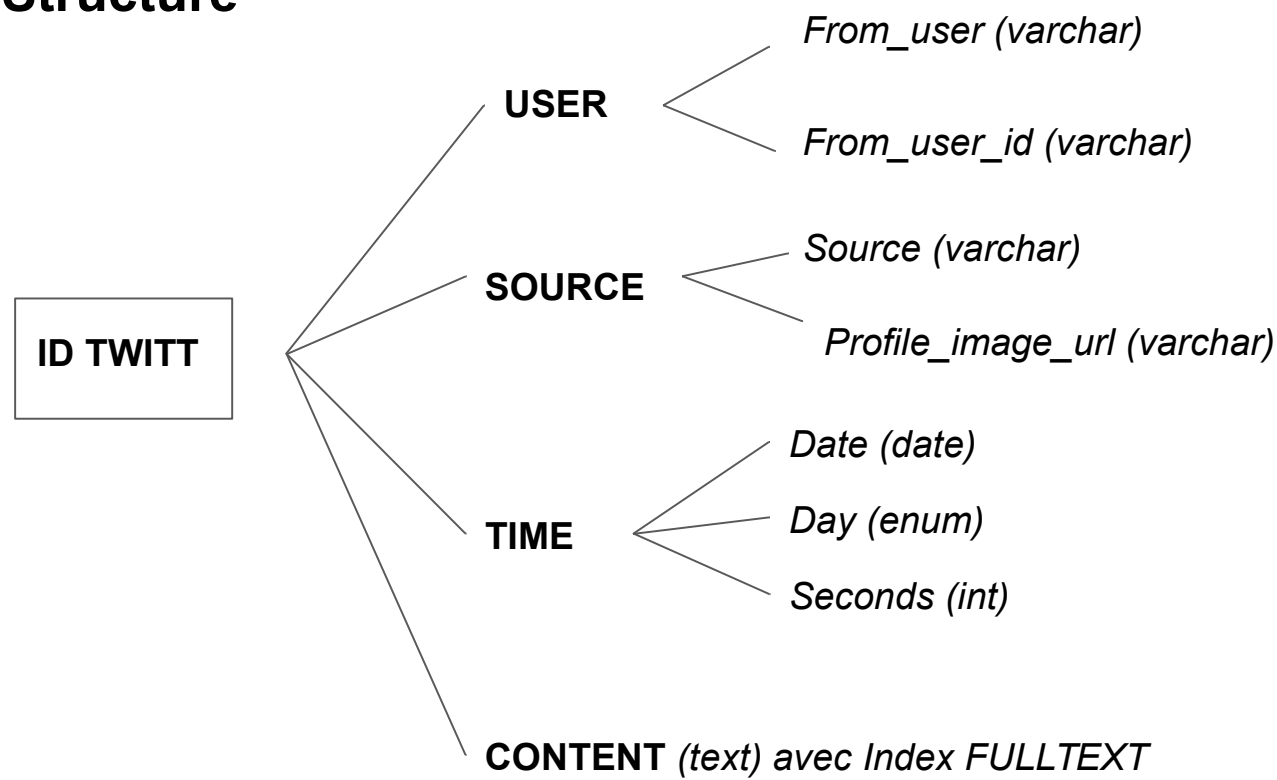
Thème de la musique :
- musique
- historic music
- electronic...

Thème du cinéma :
- Film
- international
- showcasing...

Sommaire

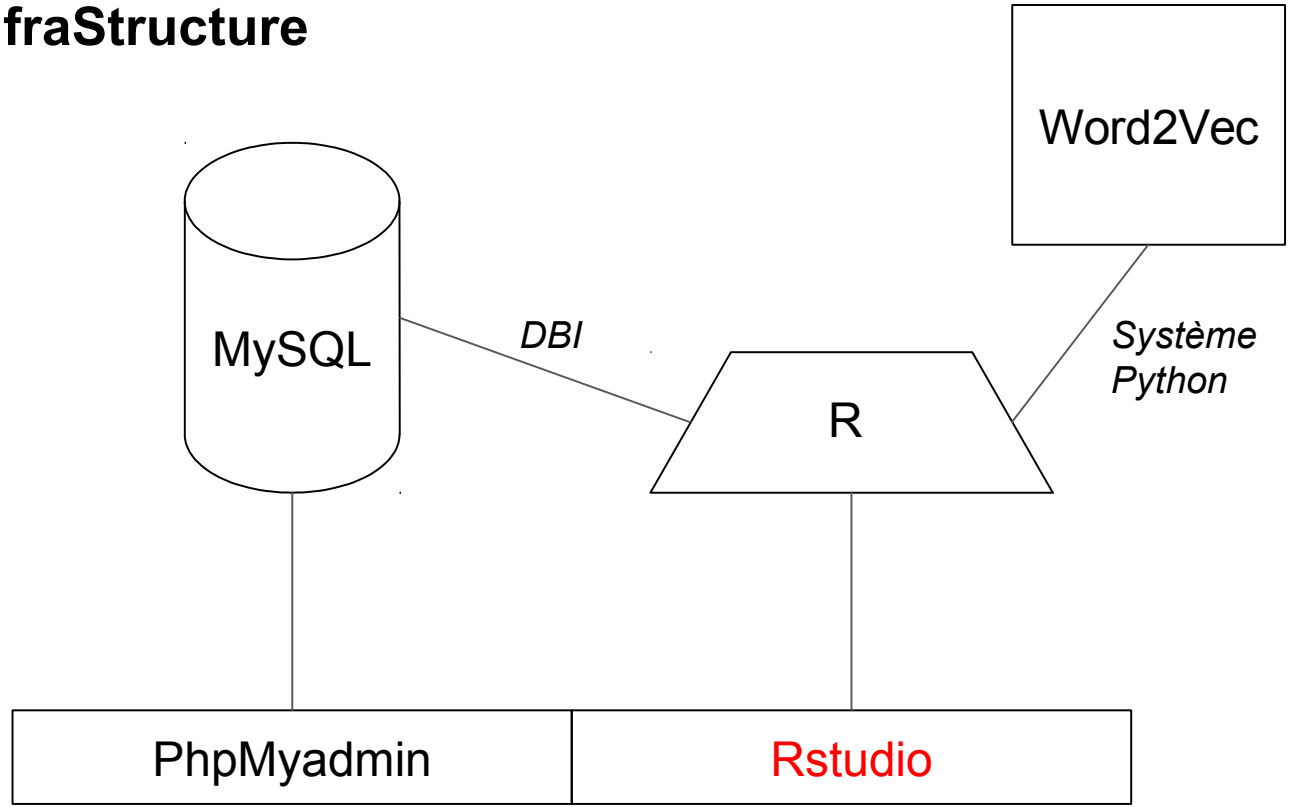
II. Données Microblogs

A) Structure



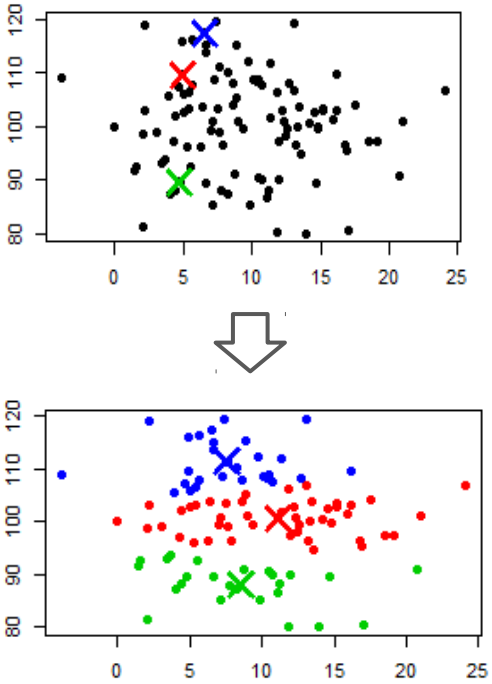
II. Données Microblogs

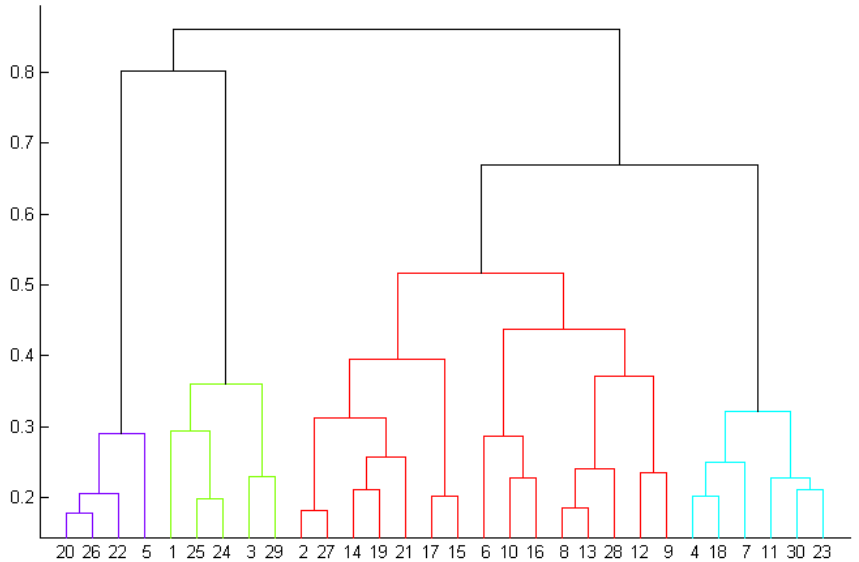
B) InfraStructure



II. Données Microblogs

C) Visualization relations sémantiques entre termes du lexique





Sommaire

III. Plongements

A) Temporels

Mots clés

147162 twitts comprenant “London” en septembre 2015

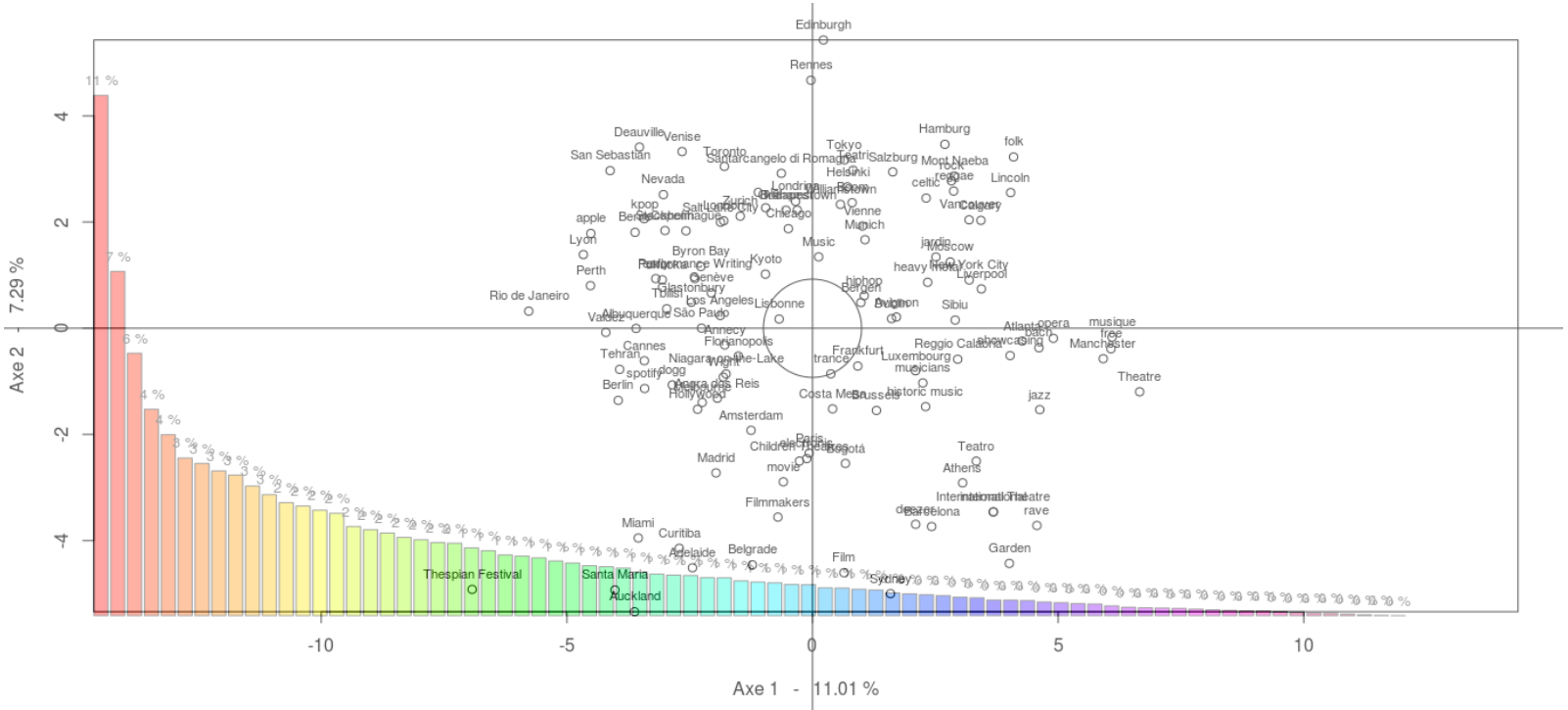
Mois

	Los Angeles	Venise	Cannes	Hollywood	London	Berlin	San Sebastián	Annecy	Deauville
mai 2015	1769	68	1696985	13564	6392	2510	25	115	70
juin 2015	12787	40	417999	4455	37305	11451	5	4022	524
juillet 2015	6582	215	135047	3215	48372	4764	13	368	248
aout 2015	22502	103	99887	2494	75568	5046	72	241	2928
septembre 2015	4459	665	100574	6086	147162	5906	227	365	7451
octobre 2015	5399	74	148400	4315	43998	8869	16	242	296
novembre 2015	2654	14	272246	2789	16790	2769	3	168	281
decembre 2015	2014	13	71524	2222	8754	6183	1	83	186
janvier 2016	18146	52	78817	3135	23312	5164	20	149	233
février 2016	3297	356	76767	2738	13301	40245	1	207	510
mars 2016	5264	45	140035	3860	18741	6315	2	437	500
avril 2016	8214	60	209694	4815	23223	4478	9	700	901
mai 2016	13783	56	1949047	17319	34928	6419	5	563	703
juin 2016	12296	75	342273	7128	42452	5409	3	4397	732
juillet 2016	7180	307	142265	2624	50265	5460	37	285	1282
aout 2016	5305	336	242631	3025	37099	6517	35	146	1437
septembre 2016	2401	497	46471	2040	25452	2733	78	144	5218
octobre 2016	3494	81	59718	2709	40856	5916	139	258	325

III. Plongements

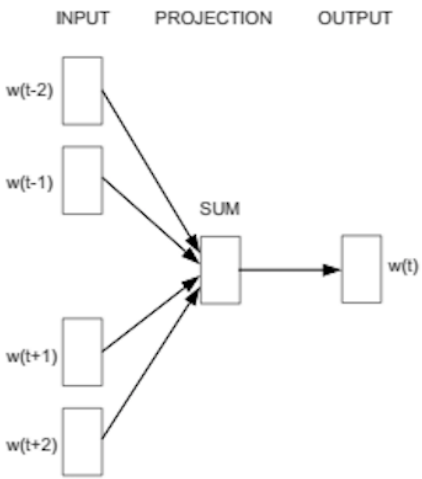
A) Temporels

Représentation graphique : ACP

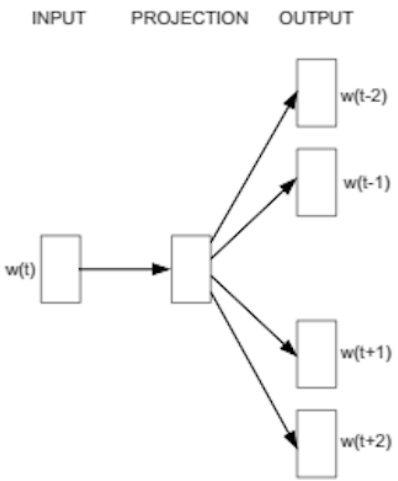


III. Plongements

B. Lexical (Word2Vec)



CBOW



Skip-gram

- **CBOW** recherche la représentation vectorielle qui permet de prédire un mot manquant dans context donné.
- **Skip** recherche le modèle qui permette de prédire le contexte pour un mot ciblé.

III. Plongements

B. Lexical (Word2Vec)

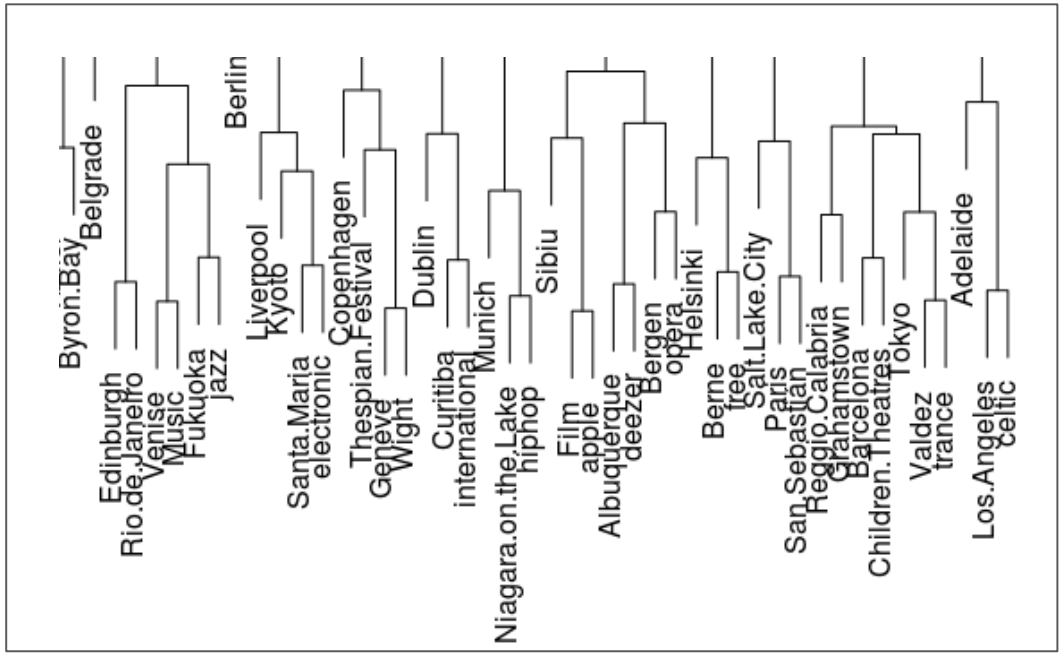
Table des distances cosinus : *1 – cos(u,v)*

	International.Theatre	Teatri	Garden	jardin	Music	musique	historic.music
Tokyo	0.7536	1.0200	0.8112	1.0622	0.7884	1.1086	0.9936
Berlin	0.7033	1.0219	0.7946	1.0450	0.8762	1.0840	0.9401
Vienna	0.7112	0.9683	0.7804	1.0291	0.8495	1.0236	0.8729
Copenhagen	0.7427	0.9980	0.8101	1.0137	0.8503	1.0500	0.9324
Munich	0.8327	1.0662	0.8608	0.9757	0.9105	1.0021	0.9498

III. Plongements

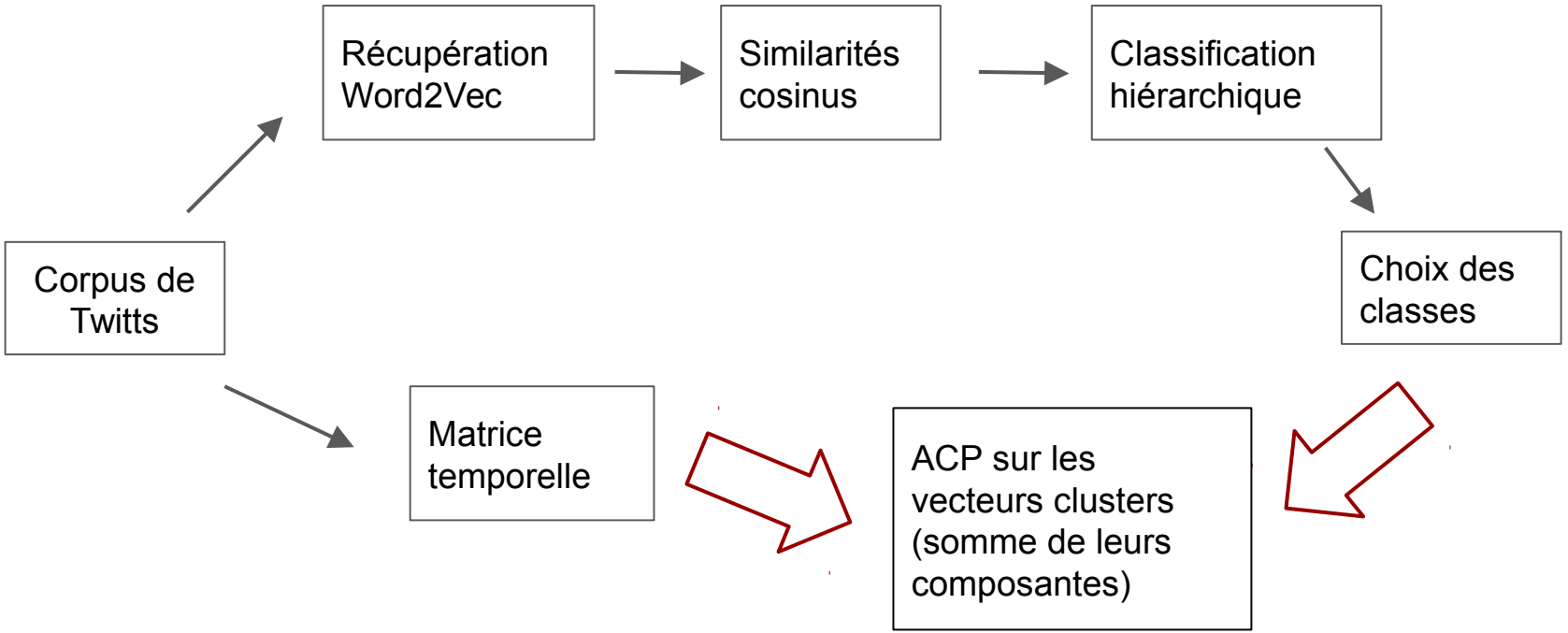
B. Lexical (Word2Vec)

Zoom sur la classification hiérarchique
réalisée à partir des similarités cosinus entre les vecteurs



III. Plongements

C. Hybride



III. Plongements

C. Hybride

Propriété d'additivité

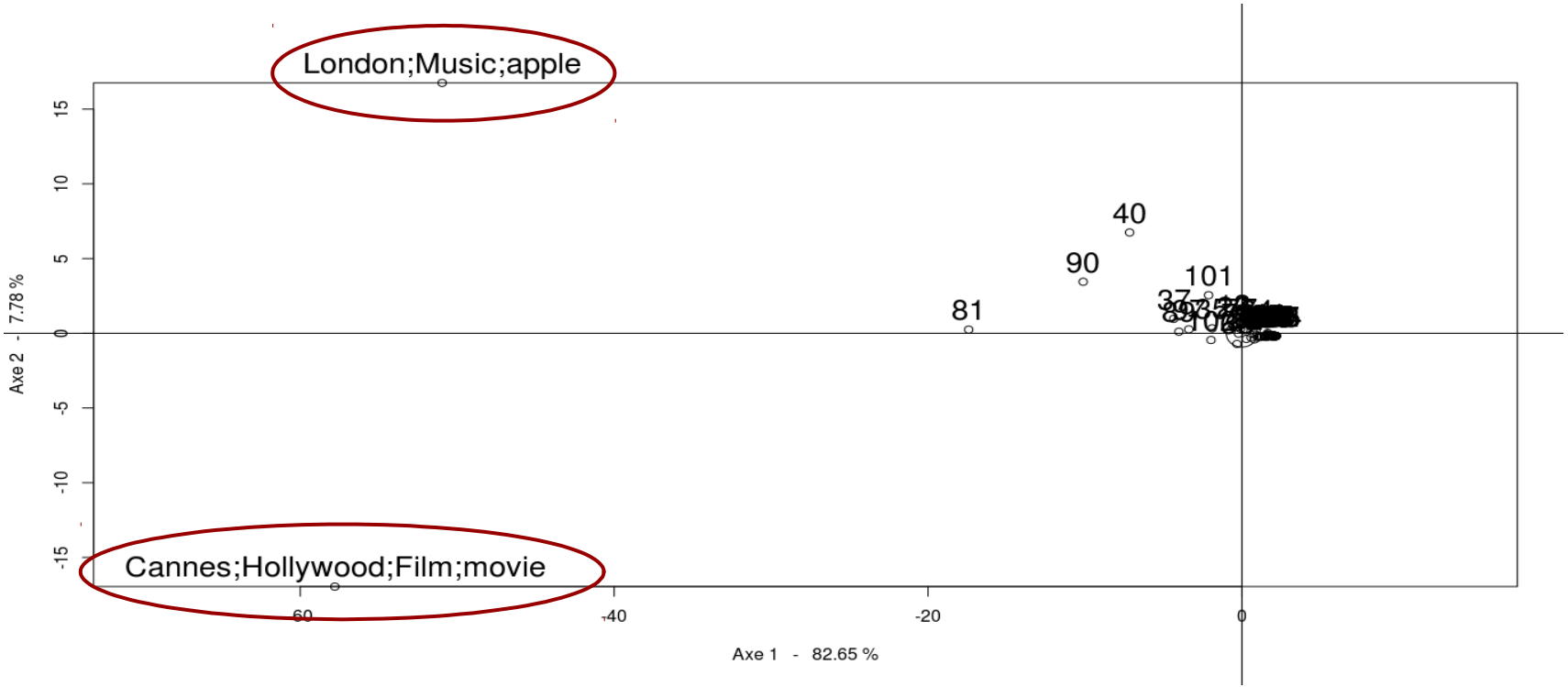
Mathématiquement, si on note RC la fonction de répartition chronologique d'un sous ensemble de twitts, pour deux mots A et B on a :
RC(A ET B) = RC(A) + RC(B) - RC(A INTER B)

La propriété d'additivité que j'ai voulu démontrer pour pouvoir appliquer la méthode hybride s'applique si A et B sont distincts, c'est à dire s'ils n'apparaissent pas dans les mêmes twitts, car dans ce cas :
RC(A ET B) = RC(A) + RC(B)

III. Plongements

C. Hybride

Résultats



Plongements

C. Hybride

Comparaison

Test de **Wilcoxon-Mann-Whitney**, test de comparaison de médianes qui représente l'équivalent non paramétrique du test de Student sur les moyennes : p-value < **2.2*10^-16**

Test de **Kendall**, test non paramétrique de comparaison d'ordonnancement qui s'apparente à un test de corrélation : p-value < **2.5*10^-6**

V. Conclusion

Réseaux construits par plongements

Hybridation pour détecter des événements diffus

Approches automatiques à partir d’un choix de lexique et d’un corpus

Analyses interactives et incrémentales

Références

- Thèse réalisée par Mohammed El Malki sur les données multidimensionnelles NoSQL. Technique permettant de gérer des mégadonnées : https://pdfs.semanticscholar.org/f277/c9f615f01ebe6cab7bf4327d1ebb9eab2aae.pdf?_ga=1.230207636.965685276.1492084013 https://hal.archives-ouvertes.fr/hal-01360873/document
- Cours sur les séries chronologiques et les lissages réalisé par Agnès Lagnoux, qui m'a permis une bonne réalisation des lissages sur mes données : http://www.math.univ-toulouse.fr/~lagnoux/Poly_SC.pdf
- Cours sur les similarités entre les mots (partie sur la similarité cosinus qui nous intéresse particulièrement) réalisé par Matthieu Constant : http://igm.univ-mlv.fr/ens/Master/M2/2007-2008/TAL/cours/mstal-1-3-m2.pdf
- Méthodes pour classification hybride et librairie factoextra : http://www.sthda.com/english/wiki/hybrid-hierarchical-k-means-clustering-for-optimizing-clustering-outputs-unsupervised-machine-learning
- Exposé réalisé par Terrence Szymanski sur le "Word Embeddings Over Time" : http://www-personal.umich.edu/~tdszyman/misc/InsightSIGNLP16.pdf
- Présentation de Youssef ESSTAFA sur l'estimation des modèles FARIMA avec un bruit non corrélé mais non indépendant : http://jds2017.sfds.asso.fr/program/Soumissions/subm242.pdf
- Présentation de Pascal MONESTIEZ sur la Modélisation statistique des données d'observation issues des Sciences Participatives : http://jds2017.sfds.asso.fr/program/Soumissions/subm160.pdf

Merci pour votre attention !

Cassandre Ollivier[1], Mathias Quillot[2], Eric SanJuan[1,2]

1. Inst. Univ. Tech. (IUT) Statistique Informatique Décisionnelle (STID)

2. Laboratoire d'Informatique (LIA)

Université d'Avignon