NATIXIS
BEYOND BANKING

Document Clustering with United Nation Corpus and Natixis Economic Research

Delong LI
01.09.2020

GROUPE BPCE

CONTENTS

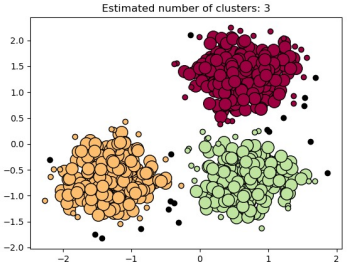
OBJECTIVE OF PROJECT

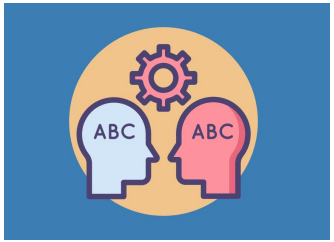
What do we want to study

1 Objective of Project

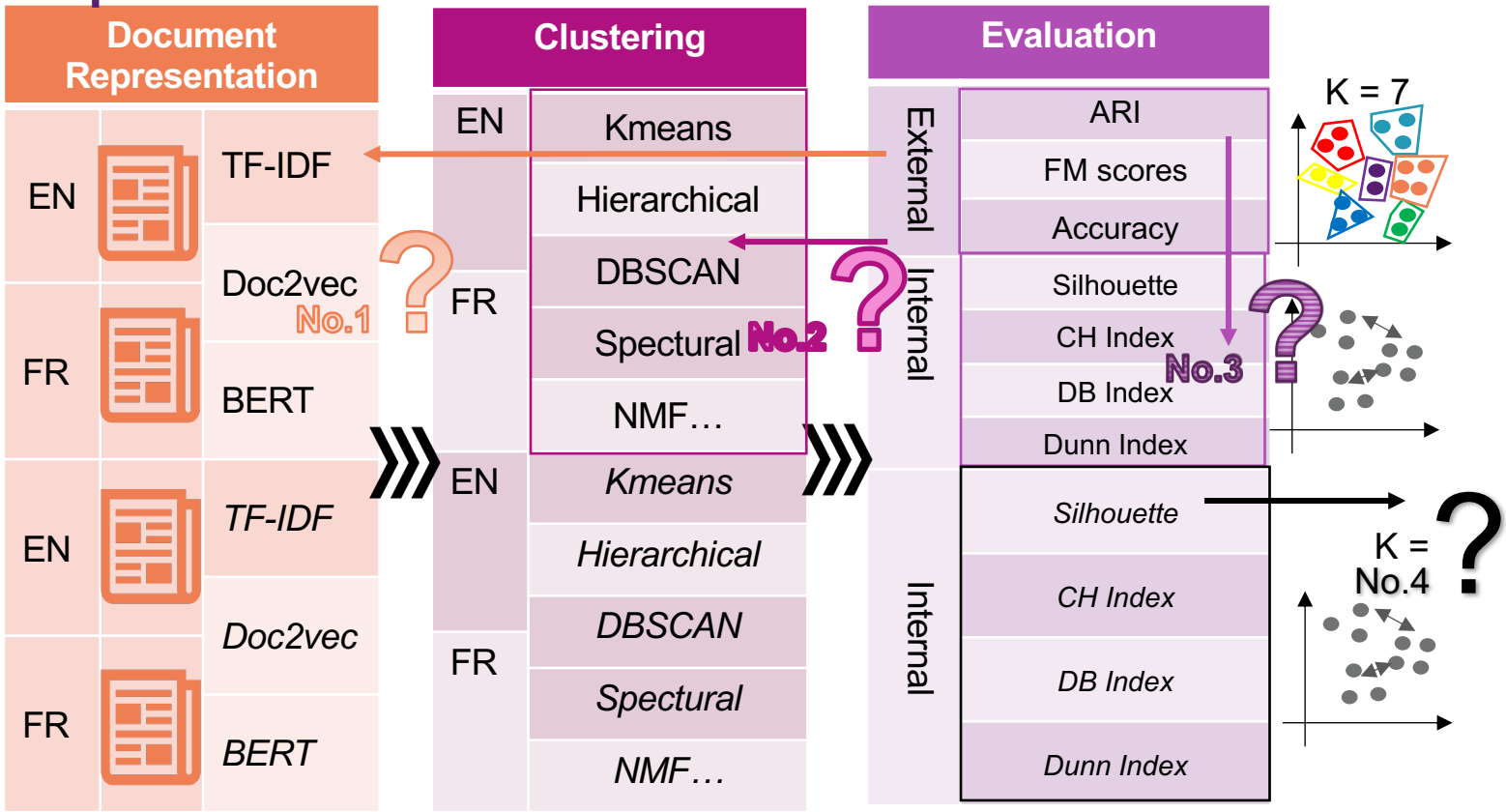
Smart Search…

- Find the effective document embeddings and clustering models
- Find appropriate internal evaluation metrics by reference of classified text corpus
- Exam the compatibility of preprocessing steps for both French and English





Route Map


Document Representation
Clustering
Evaluation
EN
FR
TF-IDF
Doc2vec
No.1
BERT
Kmeans
Hierarchical
DBSCAN
Spectural
No.2
NMF…
External
Internal
ARI
FM scores
Accuracy
Silhouette
CH Index
DB Index
Dunn Index
No.3
K = 7
K = No.4

DESCRIPTION OF DATA

Which dataset are we working on

1 Description of Data

37 000+
DOCUMENTS
EACH LANGUAGE

1990-2014
25 YEARS OF OFFICIAL DOCUMENT

XML FORMAT
MANNUALLY TRANSLATED BY EXPERTS

Metadata:Symbol, Date, sub-directory **(label)**, Keywords…

United Nations Parallel Corpus

Introduction

The United Nations Parallel Corpus v1.0 is composed of official records and other parliamentary documents of the United Nations that are in the public domain. These documents are mostly available in the six official languages of the United Nations. The current version of the corpus contains content that was produced and manually translated between 1990 and 2014, including sentence-level alignments.

The corpus was created as part of the United Nations commitment to multilingualism and as a reaction to the growing importance of statistical machine translation (SMT) within the Department for General Assembly and Conference Management (DGACM) translation services and the United Nations SMT system, Tapta4UN.

The purpose of the corpus is to allow access to multilingual language resources and facilitate research and progress in various natural language processing tasks, including machine translation. For convenience, the corpus is also available pre-packaged as language-specific bi-texts and as a six-language parallel corpus subset.

When using the United Nations Parallel Corpus, the user must acknowledge the United Nations as the source of the information. When making reference to the United Nations Parallel Corpus, please cite this reference: Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B., (2016), The United Nations Parallel Corpus, Language Resources and Evaluation (LREC'16), Portorož, Slovenia, May 2016.

In our study, 7 refined sub-directory documents were selected, in total 3000+ documents.

2 Description of Data

4585
DOCUMENTS IN BOTH EN AND FR

2016-2020
NEARLY 5 YEARS OF PUBLICATIONS

TXT FORMAT
MANNUALLY TRANSLATED BY EXPERTS

Metadata:
Text, Filename

Placée sous la responsabilité de Patrick Artus, la Recherche Économique dispose d'une **expertise** forte et diversifiée. Elle propose des **prestations évolutives** et adaptées aux **besoins des clients** -domestiques et internationaux- de la banque et des réseaux actionnaires de BPCE.

Pour mieux répondre aux besoins des clients, la gamme des publications est structurée par grands thèmes.

Ces publications couvrent des domaines tels que l'analyse macro-économiques, les différents marchés financiers (taux, change, actions, commodities), l'analyse technique et des indicateurs.

PREPROCESSING

NLP treatment on original corpus data

1 Preprocessing

Text Cleaning

III. La première session ordinaire de 2010 du Conseil d'administration de l'UNICEF aura lieu du mardi 12 au jeudi 14 janvier 2010 dans la salle de conférence 3 (TNLB) aux États-Unis. On peut consulter version 5.0 de toute la documentation de la session sur le site Web suivant : <www.unicef.org/about/execboard/index_51167.html>

1st FILTER

- Match, delete address and path
- Match, delete text inside parenthesis ()
- Only keep characters and numbers A-Za-z0-9ÀÉÈÇâàçéèëêîïôüùû

III. La première session ordinaire de 2010 ddd du Conseil d administration de l-UNICEF aura lieu du mardi 12 nomjour ddd au jeudi 14 nomjour ddd janvier 2010 nommois ddd dans la salle de conférence 3 ddd aux États-Unis On peut consulter une version 5.0 fff de toute la documentation de la session sur le site Web suivant

2nd REPLACE

- Replace day and month with common token
- Replace number with special tokens (ddd for integer; fff for float)

1 Preprocessing

~~iii~~ ~~la~~ première session ordinaire ~~de~~ ddd ~~du~~ conseil d administration ~~de~~ ~~l~~ unicef aura lieu ~~du~~ nomjour ddd ~~au~~ nomjour ddd nommois ddd dans ~~la~~ salle ~~de~~ conférence ddd états unis ~~on~~ peut consulter ~~une~~ version fff ~~de~~ toute ~~la~~ documentation ~~de la~~ session sur ~~le~~ site web suivant

3rd
TOKENIZATION (ENTITY)

- Recognise entities that are compound nouns and tokenize (country, person) correspondingly.

4th
FILTER

- Remove stopwords and words that are less than 2 characters
- Lemmatize all the remainning words in lowercase

premier, session, ordinaire, ddd, conseil d administration, unicef, lieu, nomjour, ddd, nomjour, ddd, nommois, ddd, salle, conférence, ddd, états unis, pouvoir, consulter, version, fff, documentation, session, site, web, suivant

DOCUMENT REPRESENTATION

Represent text in numeric form

Document R

Document ID	Textual description
1	“Data science is fun”
2	”Artificial intelligence is the future”
3	“Business and artificial intelligence combination is the key”

TF-IDF

TF

Word / Document	“data”	“science”	“fun”	”artificial”	“intelligence”	”future”	“business”	“combination”	“key”
1	1	1	1	0	0	0	0	0	0
2	0	0	0	1	1	1	0	0	0
3	0	0	0	1	1	0	1	1	1

X

IDF

Word / Document	“data”	“science”	“fun”	”artificial”	“intelligence”	”future”	“business”	“combination”	“key”
1	0,48	0,48	0,48	0,18	0,18	0,48	0,48	0,48	0,48
2	0,48	0,48	0,48	0,18	0,18	0,48	0,48	0,48	0,48
3	0,48	0,48	0,48	0,18	0,18	0,48	0,48	0,48	0,48

=

TF-IDF

Word / Document	“data”	“science”	“fun”	”artificial”	“intelligence”	”future”	“business”	“combination”	“key”
1	0,48	0,48	0,48	0,00	0,00	0,00	0,00	0,00	0,00
2	0,00	0,00	0,00	0,18	0,18	0,48	0,00	0,00	0,00
3	0,00	0,00	0,00	0,18	0,18	0,00	0,48	0,48	0,48

document

ion the

frequently

ur rarely

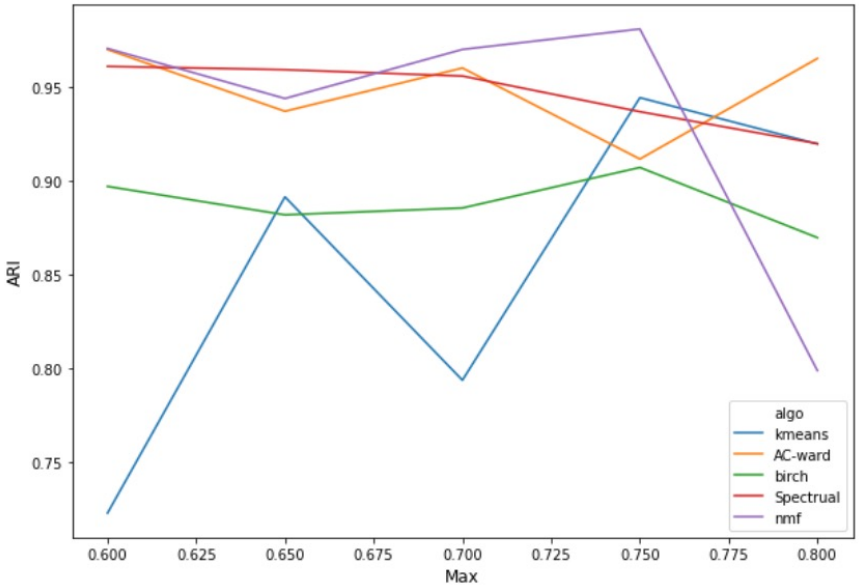
t in D, t is a

TF-IDF

- It is necessary and reasonable to ignore the words that have a document frequency strictly higher or lower than the given threshold

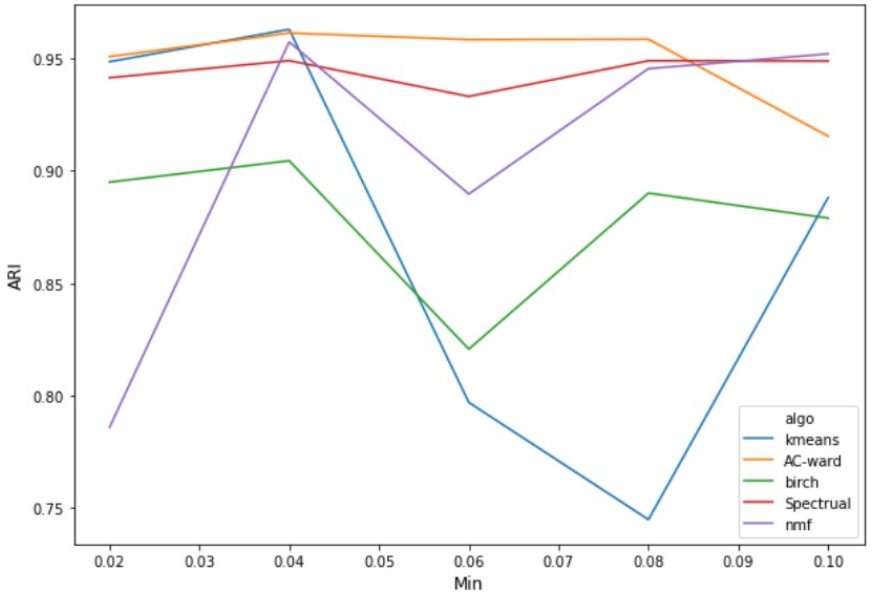
df_max=0.6,0.65,0.7,0.75,0.8

df_min = 0.05



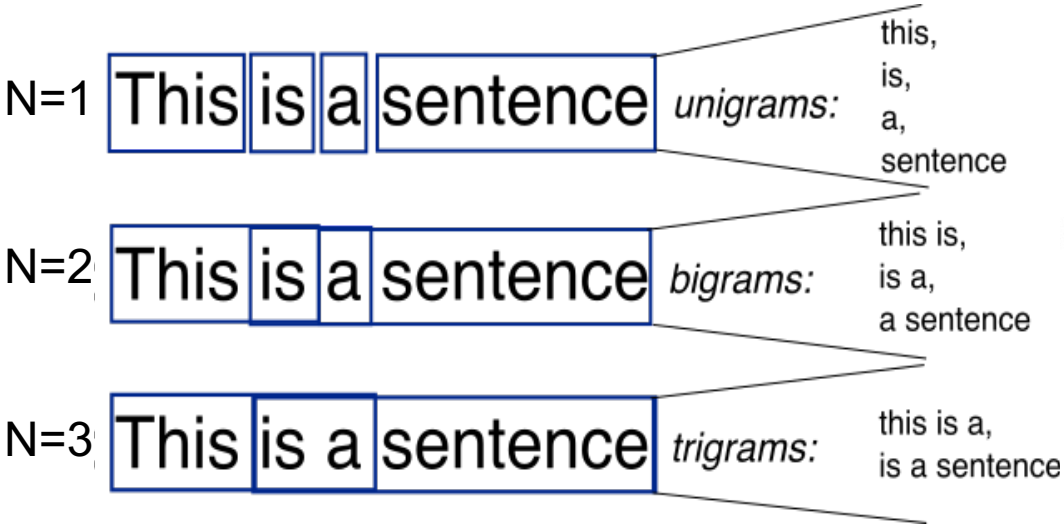
df_max = 0.75

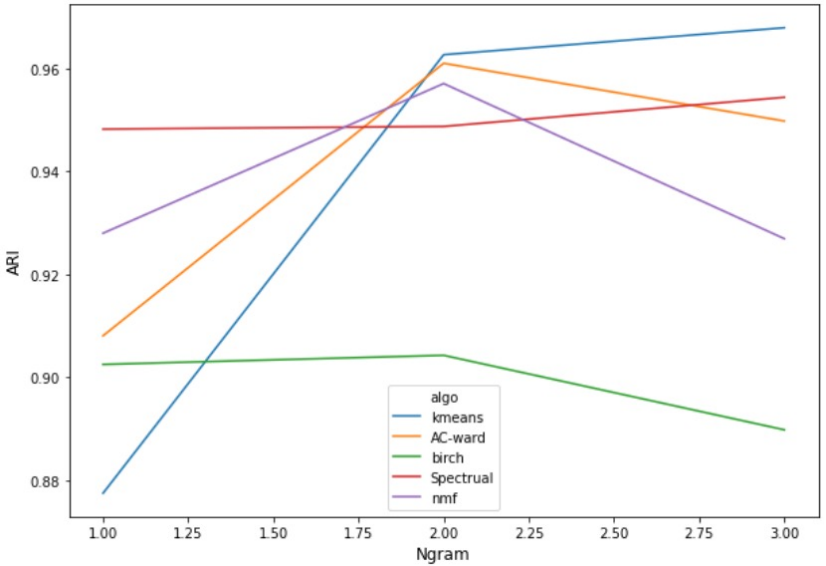
df_min = 0.02,0.04,0.06,0.08,0.1



TF-IDF

An n-gram is a sequence of N words, which helps to find meaningful group of words. It is extensively used in NLP tasks.

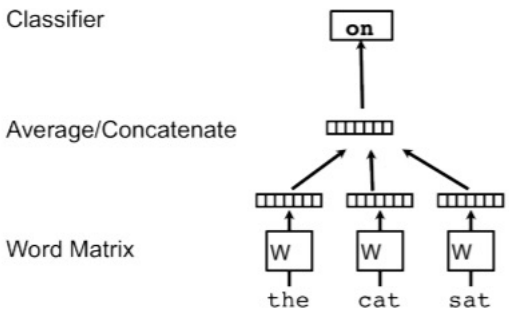


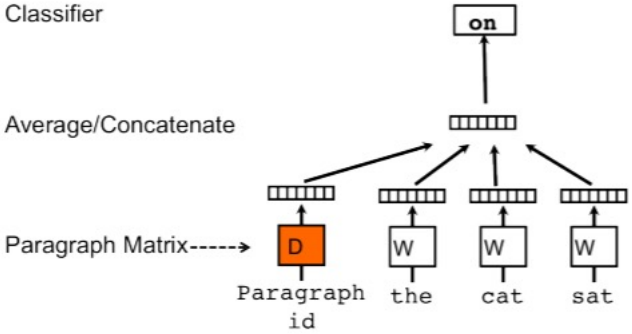


Document Representation

Doc2 vec

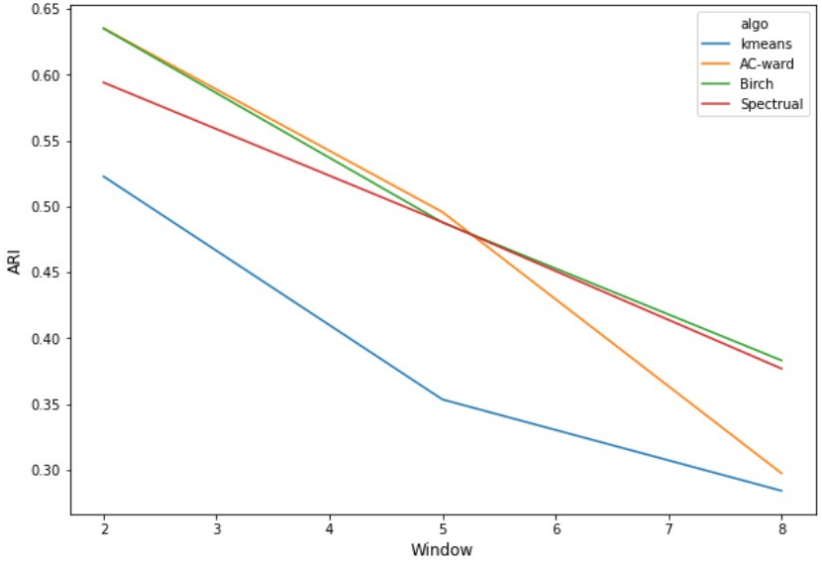
- Word2Vec is to predict a word given the other words in a context with simple neural network with a single hidden layer. The goal is to learn the weights of the hidden layer, which are actually the “word vectors”.
- Doc2vec model act as if a document has another floating word-like vector, which contributes to all training predictions, and is updated like other word-vectors, but we will call it a doc-vector.



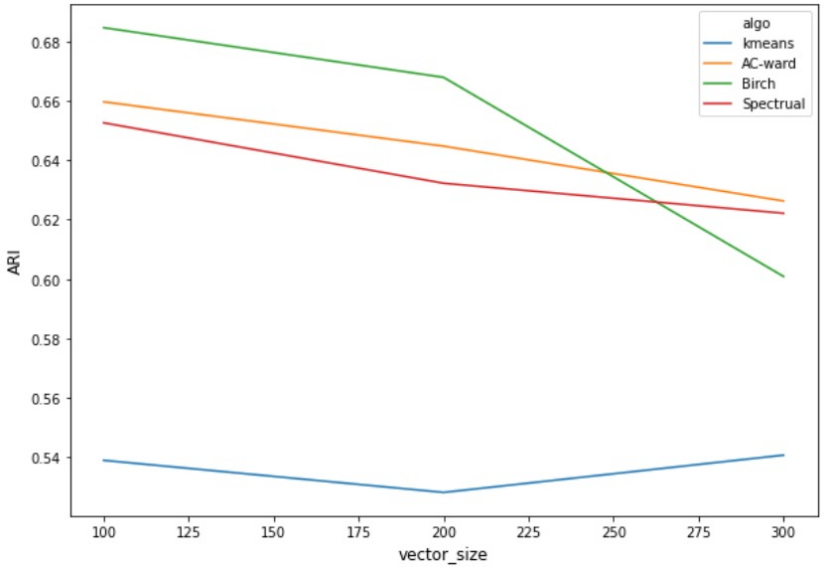


2 Doc2vec

WINDOW = 2,5,8
VECTOR_SIZE = 200



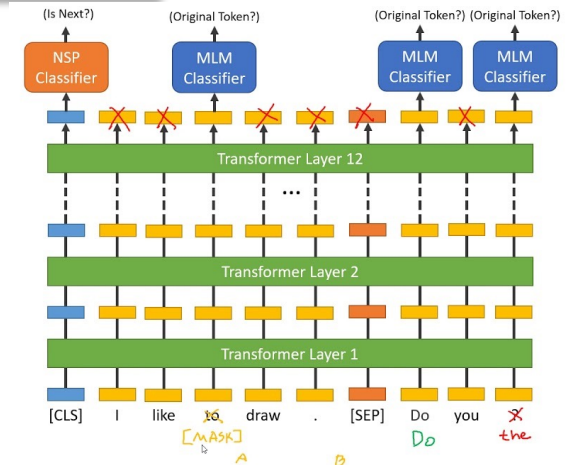
VECTOR_SIZE = 100, 200, 300
WINDOW = 2

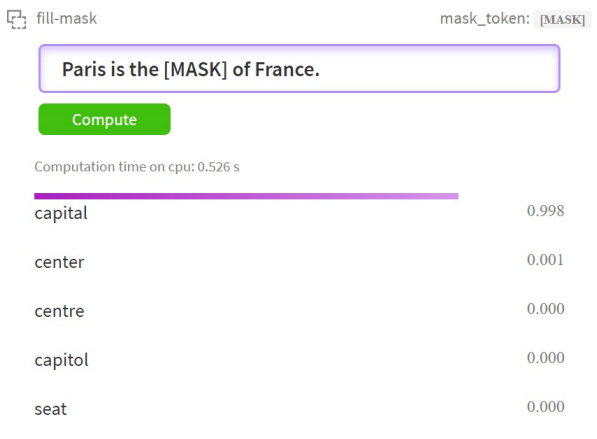


Document Representation

BERT

- A batch of models that pre-trained deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context; models can be finetuned with just one additional output layer for various tasks.
- we pre-train new representation based on our own corpus and extract sentence embeddings for clustering.

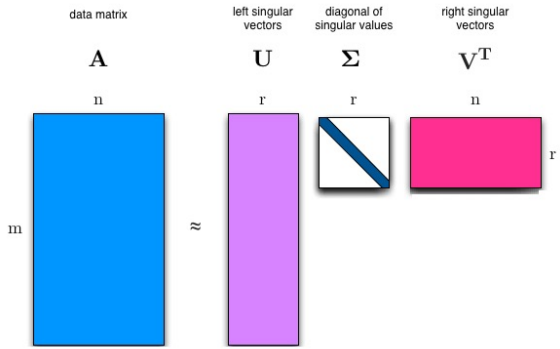




Dimension Reduction

LSI/SVD

- **LSI** is the application of **SVD** **on** term-by-document matrices (term count/tf-idf) .

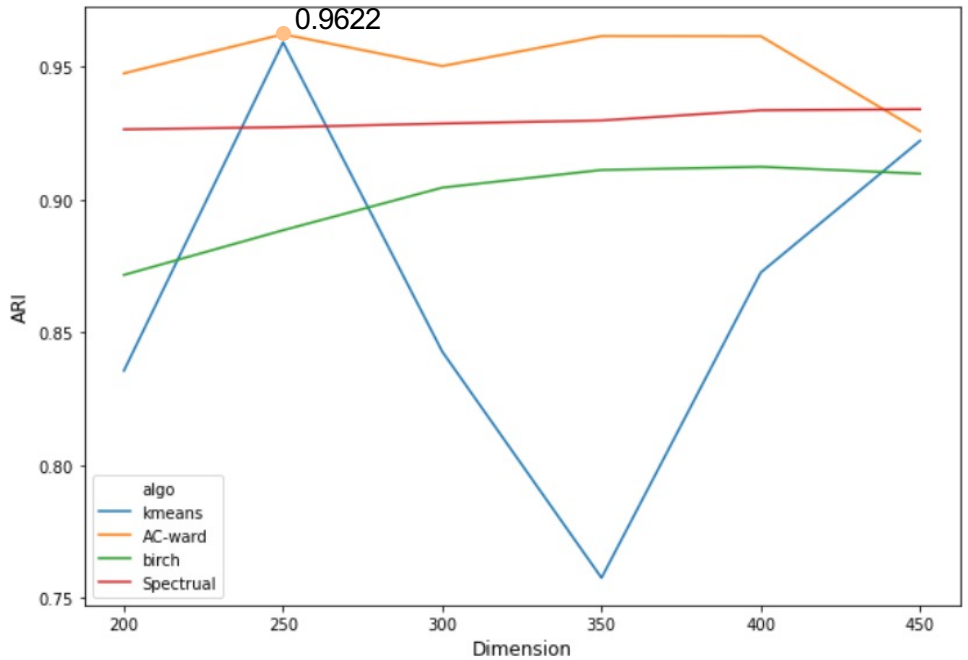


- Here A is the term-document matrix
- The k largest singular values and the associated vectors are kept

NMF

- A feature transformation method which is particularly suited to clustering
- Find two non-negative matrices (W, H) whose product approximates the non-negative matrix X (term-by-document matrices) .
- The basis in the new space are not necessarily orthonormal
- The cluster membership for a document can be determined by examining the largest component

Dimension Reduction



0.9626
0.9610
0.9487

0.9043

Normal Dimension

Dimensionality reduction is not significantly effective in our case

CLUSTERING ALGORITHME

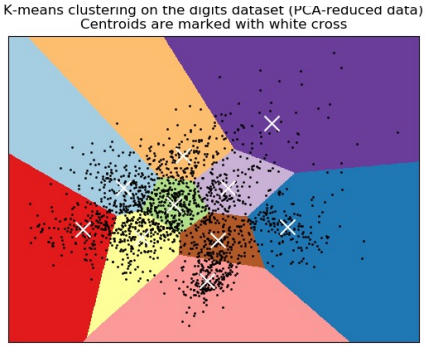
Brief introduction of clustering methods

Clustering Algorithm

Kmeans

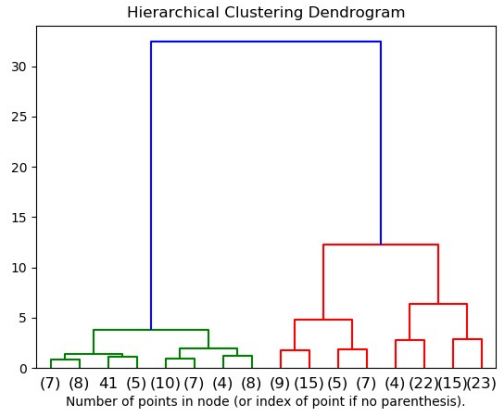
Cluster data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the within-cluster sum-of-squares

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$



Hierarchical Clustering

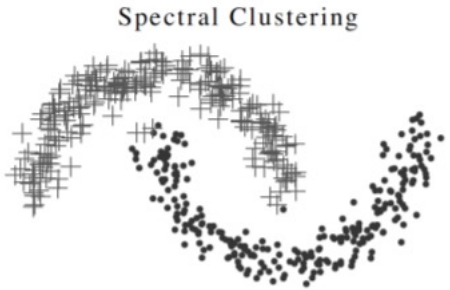
Successively merge data into clusters based on their linkage distance (ward, average, etc,)with one another. This hierarchy of clusters is represented as a tree, with root gathering all samples and leaves representing only one sample.



Clustering Algorithm

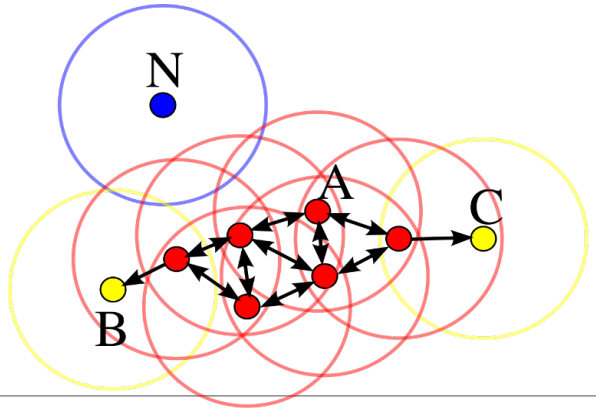
Spectral Clustering

Dataset is represented with G=(V,E) with object as the vertex and the similarity among objects as the weighted edge: W is an adjacency or affinity (e.g., K-nearest neighbour) of G.Then data are embedded to a lower space using eigenvectors of Laplacian matrix of W. Labels are assigned by basic clustering model, e.g., K-Means.



(H)DBSCAN

DBSCAN algorithm views clusters as areas of high density (core & reachable points) separated by sparser areas of noise by setting min_ samples, radius parameters. This method is deterministic and no number of clusters has to be fixed



FIND BEST MODEL

Clustering results of different embeddings and clustering algorithms

	TF-IDF			Doc2vec			BERT		
	ARI	FMI	Silhouette	ARI	FMI	Silhouette	ARI	FMI	Silhouette
K-means	0.962	0.969	**0.153**	0.539	0.652	**0.109**	0.943	0.952	0.148
Hierarchical	**0.966**	**0.971**	0.142	**0.659**	**0.741**	0.088	**0.957**	**0.961**	**0.152**
NMF	0.957	0.965	0.145	/	/	/	/	/	/
Spectral	0.949	0.958	0.144	0.595	0.673	0.086	0.936	0.949	**0.152**

- Hierarchical clusering and TF-IDF representation seem to have stable and excellent reuslts

MODEL EVALUATION

NLP treatment on original corpus data

Evaluation metrics

External

Comparing clustering results with reference – ground truth

- Adjusted Rand Index
- Fowlkes-Mallows scores
- Accuracy

Internal

evaluate structure of found clusters without reference

- Silhouette Coefficient
- Calinski-Harabasz Index
- Davies-Bouldin Index
- Dunn index

1 Adjusted Rand Index

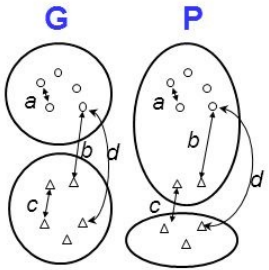
- RI is like accuracy in supervised learning, however RI computing is based on pairs

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- **Adjusted RI** establishes a baseline by using the expected similarity of a random model to guarantee random assignment will have a bad score
- Bounded range [-1, 1]

Rand and Adjusted Rand index

[Rand, 1971] [Hubert and Arabie, 1985]



Agreement: a, d
Disagreement: b, c

$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

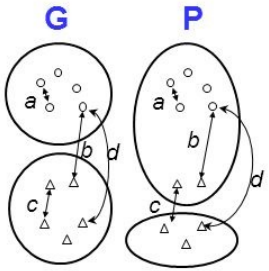
$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

Fowlkes and Mallows Index

- Based on the same pairs classification, Fowlkes and Mallows Index reform Adjusted RI as to give a simpler and deterministic evaluation, in the range of [0,1]

Fowlkes and Mallows Index

[Fowlkes, Edward B., and Colin L. Mallows, 1985]



Agreement: a, d
Disagreement: b, c

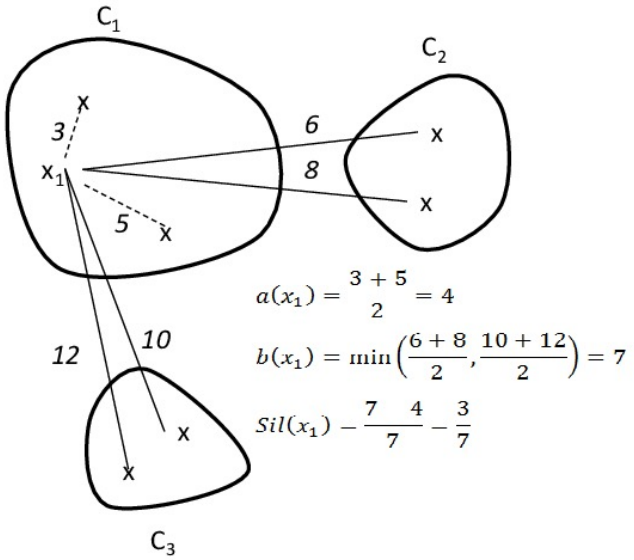
$$FMI = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

3 Silhouette Coefficient

- a: The mean distance between a sample and all other points in the same class.
- b: The mean distance between a sample and all other points in the next nearest cluster.
- The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

- Bounded range [-1, 1]



4 Davies-Bouldin Index

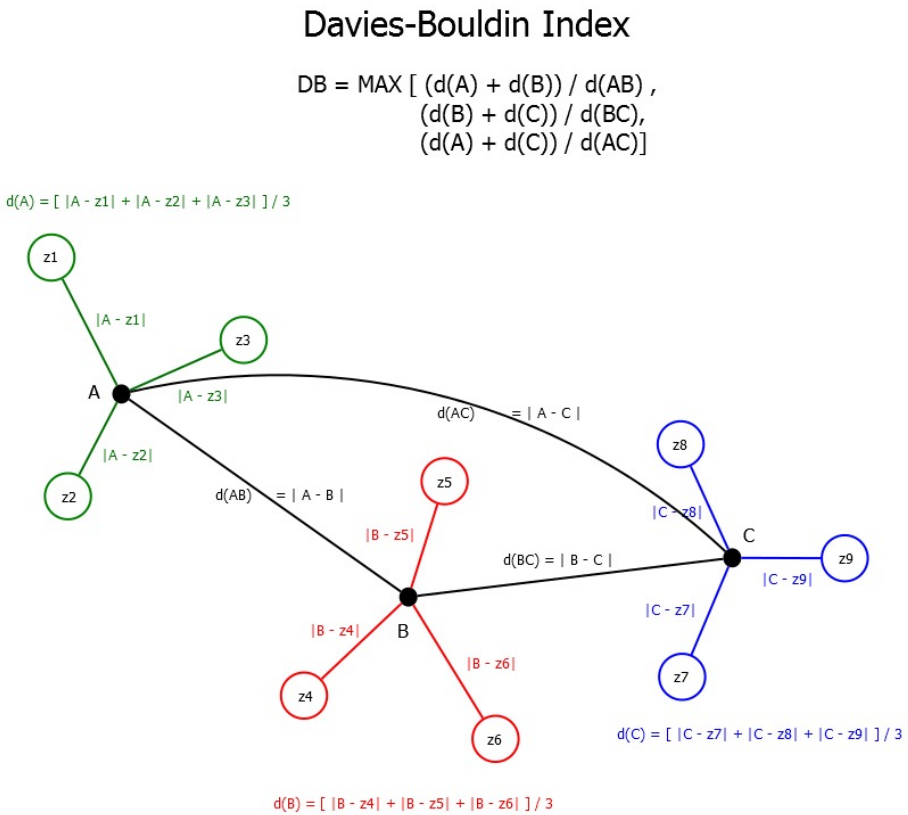
- It is constituted by two kinds of distance measurement:

  s_i: the average distance between each point of cluster i and the centroid of that cluster – also know as cluster diameter.

  d_{ij}: the distance between cluster centroids i and j.

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} \frac{s_i + s_j}{d_{ij}}$$

- Principle: the lower the better, zero indicates perfect clustering

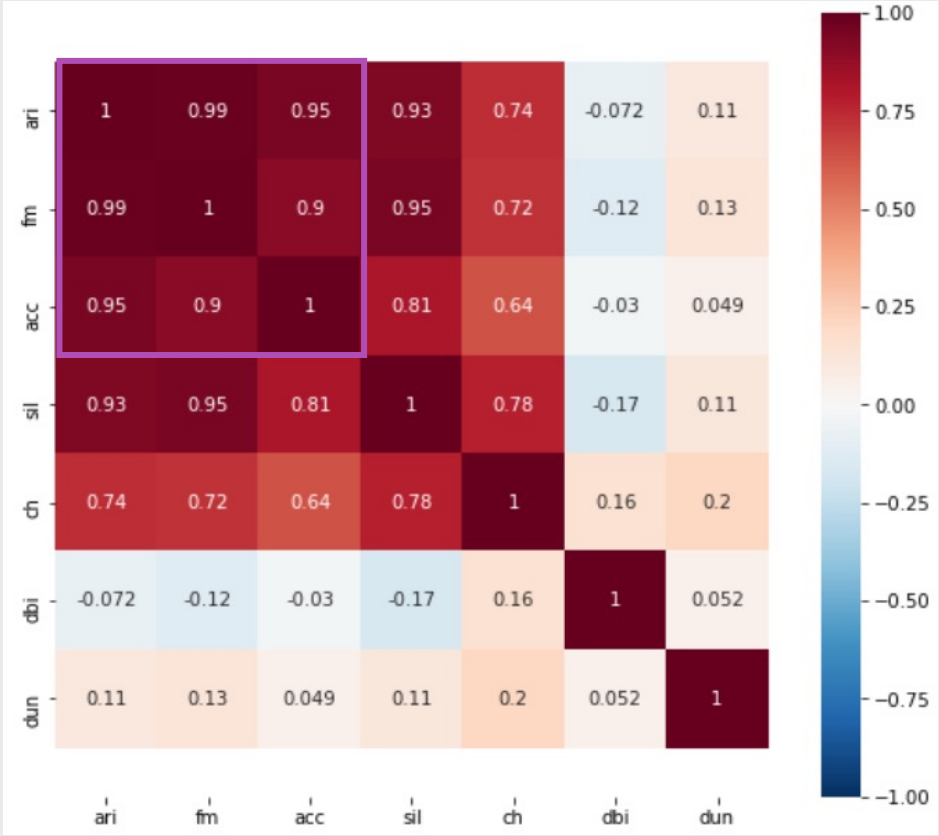


FIND PROPER METRIC

- *This plot is based on comprehensive clustering results with various algorithms as well as parameters, (which including clusters that are very good and also very bad) to test which internal evaluation is most consistent with external evaluation.*

- Silhouette coefficient (sil) seems to be the most consistent with ARI,FM and Accuracy

Correlation Coefficients of evaluation metrics

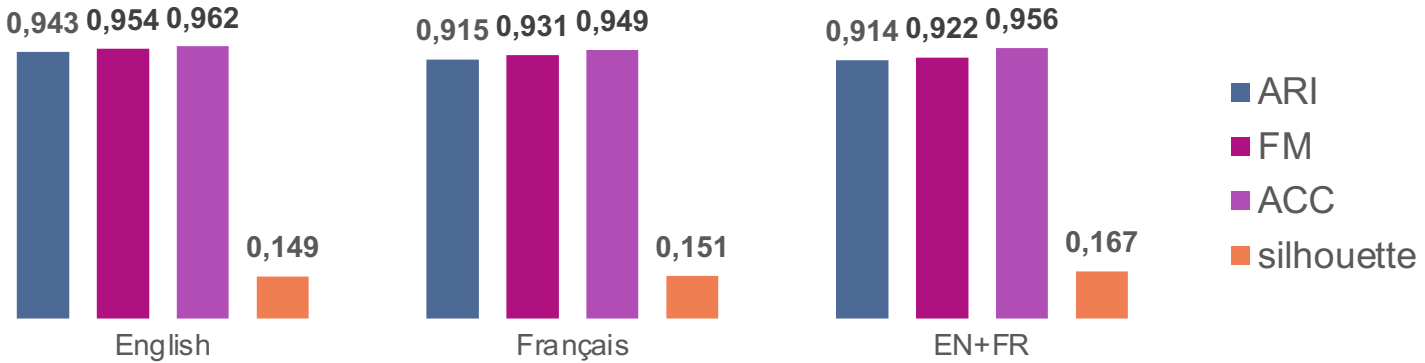


EXPERIMENT RESULT

NLP treatment on original corpus data

Test on Language

UN Corpus

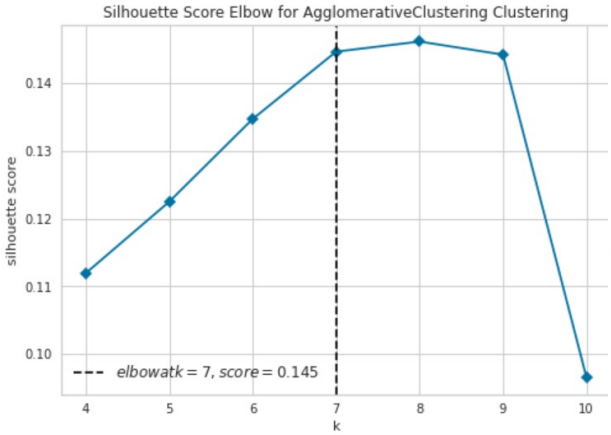


* *Experiments of Hierarchical Clustering with various tfidf parameters on 3 UN datasets : English(k=7), Français(k=7), Union set of both English and Français(k=14)*

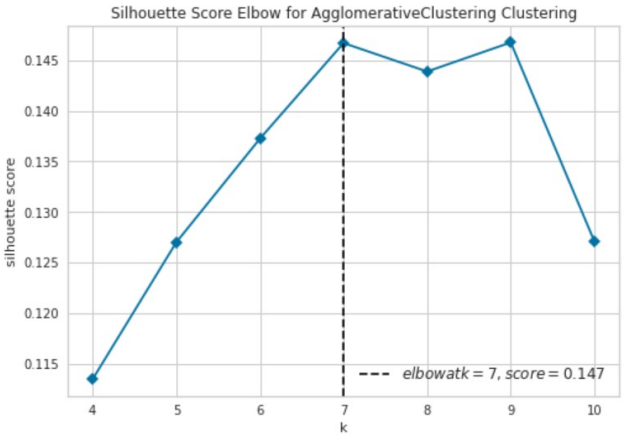
Note de bas de page

Test on Number of Cluster

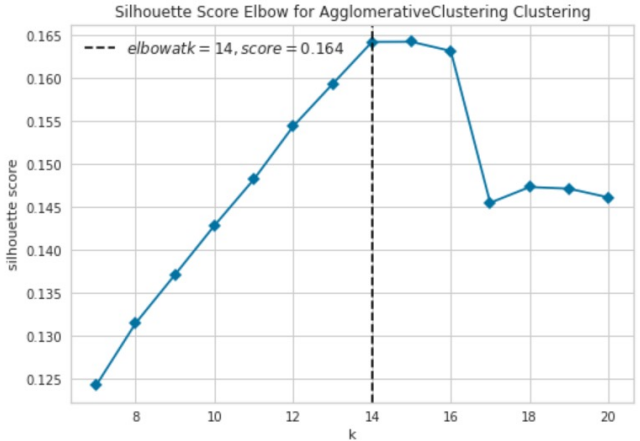
UN Corpus



English



Français

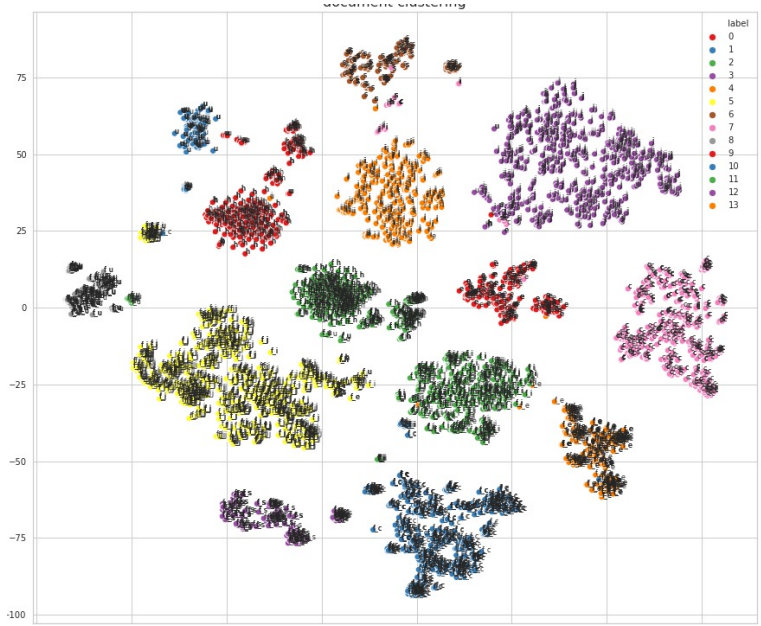


En + Fr

** Experiments of Hierarchical Clustering with the same tfidf parameters but various n_cluster (k) on 3 UN datasets : English, Français, Union set of both English and Français*

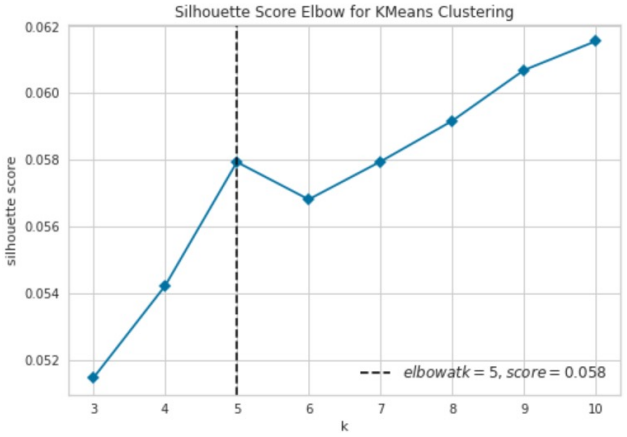
Visualization of clustering result

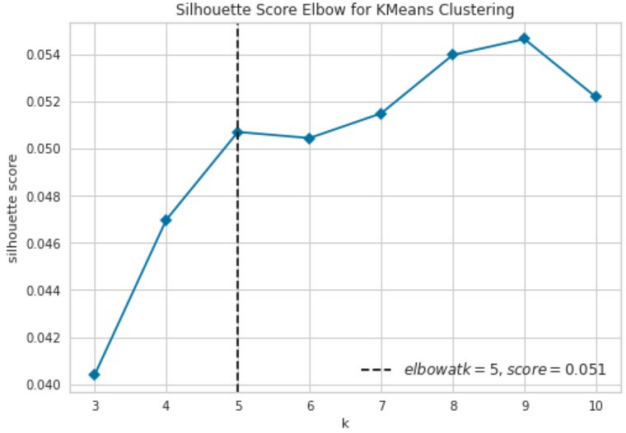
UN Corpus



	0	1	2	3	4	5	6	7	8	9	10	11	12	13
e_ccw	0	0	595	0	0	0	0	0	11	0	0	0	0	0
e_energy	0	0	0	0	0	0	270	0	0	0	0	0	0	0
e_hri	0	372	1	0	0	0	6	0	0	0	6	0	0	0
e_iccd	0	0	0	11	0	0	10	9	503	0	0	0	0	0
e_journal	0	0	0	893	0	0	0	0	0	0	0	0	0	0
e_splos	0	0	3	0	0	0	0	0	0	4	222	0	0	0
e_unodc	0	1	22	0	0	0	19	0	6	0	0	0	133	0
f_ccw	599	0	0	0	0	0	0	7	0	0	0	0	0	0
f_energy	1	0	0	0	0	0	0	7	0	0	0	0	0	262
f_hri	0	0	0	0	379	0	0	6	0	0	0	0	0	0
f_iccd	0	0	0	0	0	3	0	528	2	0	0	0	0	0
f_journal	0	0	0	0	0	893	0	0	0	0	0	0	0	0
f_splos	3	0	0	0	0	1	0	3	0	222	0	0	0	0
f_unodc	0	0	0	0	1	2	0	47	0	0	0	131	0	0

Natixis Economic Research Corpus





Elbow method on silhouette score indicates that 5 is the best cut.

EN\FR	0	1	2	3	4
0	3	2	1	71	987
1	0	5	321	1	0
2	73	32	5	1358	28
3	955	5	1	99	45
4	12	529	0	43	9

Natixis Economic Research Corpus

Topics and keywords

Sum of TFIDF score

Cluster	Top words
0	france, wage, productivity, labour, employment
1	fff, fff fff, around fff, around, resistance
2	euro zone, country, the united states, growth, china
3	interest, interest rate, rate, policy, euro zone
4	date, date date, fff, person, person person

Cluster	Top words
0	intérêt, taux intérêt, zone euro, inflation, public
1	date, fff, personne, personne personne, mois
2	fff, fff fff, support, vers, résistance
3	zone euro, pays, croissance, etats unis, extérieur
4	emploi, france, entreprise, productivité, chômage

LDA topic modeling

Topic	Top words
0	growth, law, rate, france, increase, wage
1	rate, date, inflation, interest, price
2	fff. date, person, around, will, towards
3	country, china, external, global, the united states
4	rate, euro zone, interest,policy, debt, fiscal

Topic	Top words
0	croissance, emploi, entreprise, taux, pays
1	taux, zone euro, intérêt, pays, public
2	personne, taux, date, fff, hausse, marché
3	fff, vers, support, croissance,rebond, résistance
4	personne,, fff, date, tddd, gouvernement, mesure

Merci

INTERLOCUTEURS

M. Delong LI
delong.li-ext@natixis.com

ADRESSE

NATIXIS
30, avenue Pierre Mendès France
75013 Paris - France
www.natixis.com