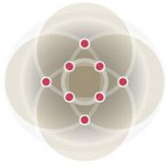
A very-short-text clustering method based on distributed representation to identify research capabilities of a Higher Education Institution

---

Jorge M. Carrasco
Jenny Marcela Sánchez
Fabio Augusto González

National University of Colombia, Bogotá



6th Annual
Global TechMining Conference
2016.09.13 • València, Spain

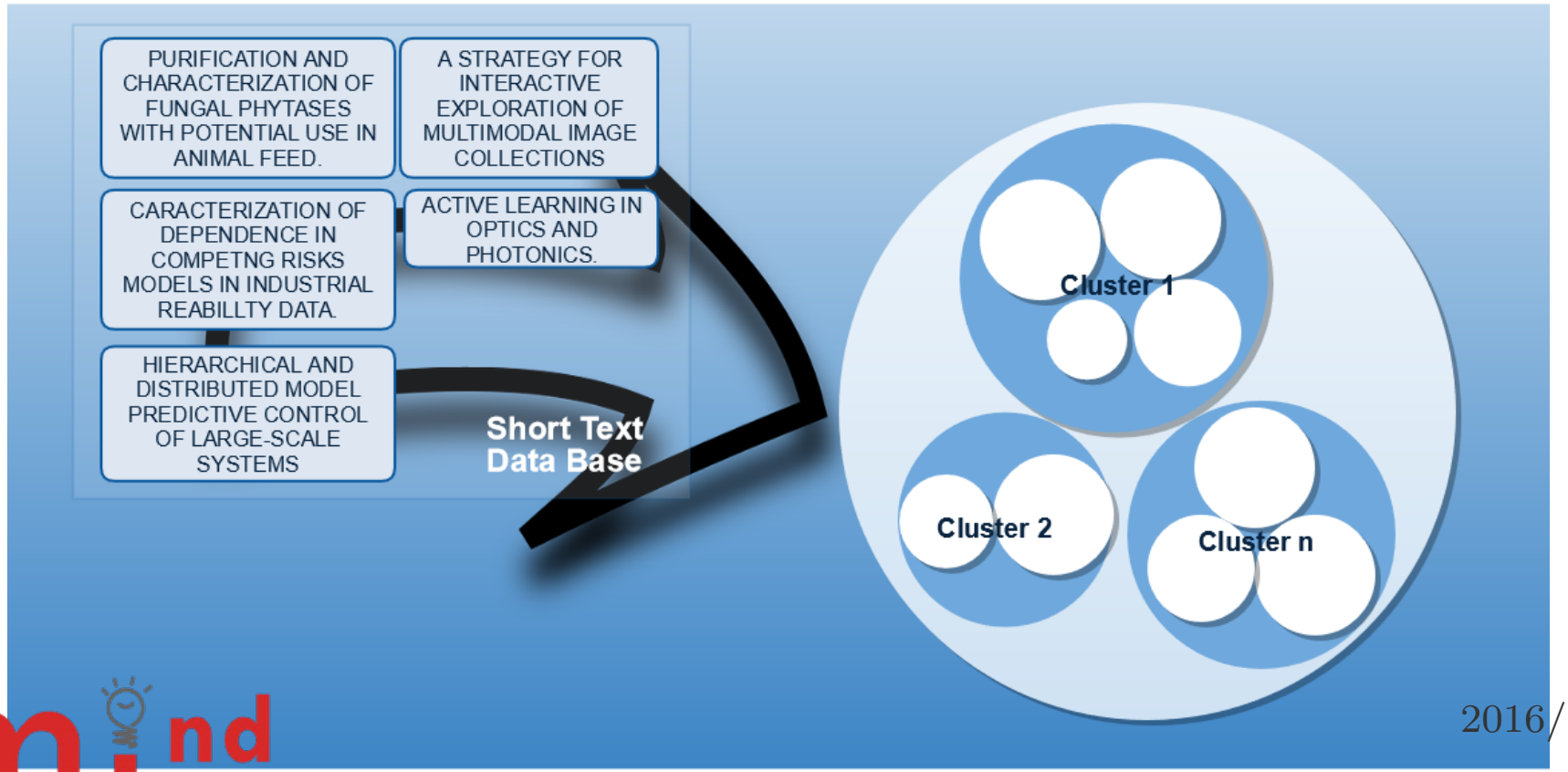
Overview

1. Motivation and Problem Formulation
2. Background
3. Proposed Methodology
4. Experimental Results

Motivation and Problem Formulation

Motivation

How to generate thematic groups of researches using the titles of the research projects of National University.



Problem Formulation

The low occurrence rate of a word across documents causes the small number of words in common among documents. An Example in a specific field:

1. (D_1) A review of theory and practice in scientometrics.

Term:	review	theory	practice	scientometrics
D_1	1	1	1	1

2. (D_2) Citation score normalized by cited references (CSNCR).

Term:	Citation	normalized	cited	references	CSNCR
D_2	1	1	1	1	1

3. (D_3) A review of the literature on citation impact indicators.

Term:	review	literature	citation	impact	indicators
D_3	1	1	1	1	1

Problem Formulation

The low occurrence rate of a word across documents causes the small number of words in common among documents. An Example in a specific field:

1. (D_1) A review of theory and practice in scientometrics.
2. (
3. (D_3) A review of the literature on citation impact indicators.

Term:	review	literature	citation	impact	indicators
D_3	1	1	1	1	1

The final BOW:

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}
D_1	1	1	1	1	0	0	0	0	0	0	0	0	0
D_2	0	0	0	0	1	1	1	1	1	1	0	0	0
D_3	1	0	0	0	1	0	0	0	0	0	0	1	1

Problem Formulation

The low occurrence rate of a word across documents causes the small number of words in common among documents. An Example in a specific field:

1. (D_1) A review of theory and practice in scientometrics.
2. (
3. (D_3) A review of the literature on citation impact indicators.

Term:	review	literature	citation	impact	indicators
D_3	1	1	1	1	1

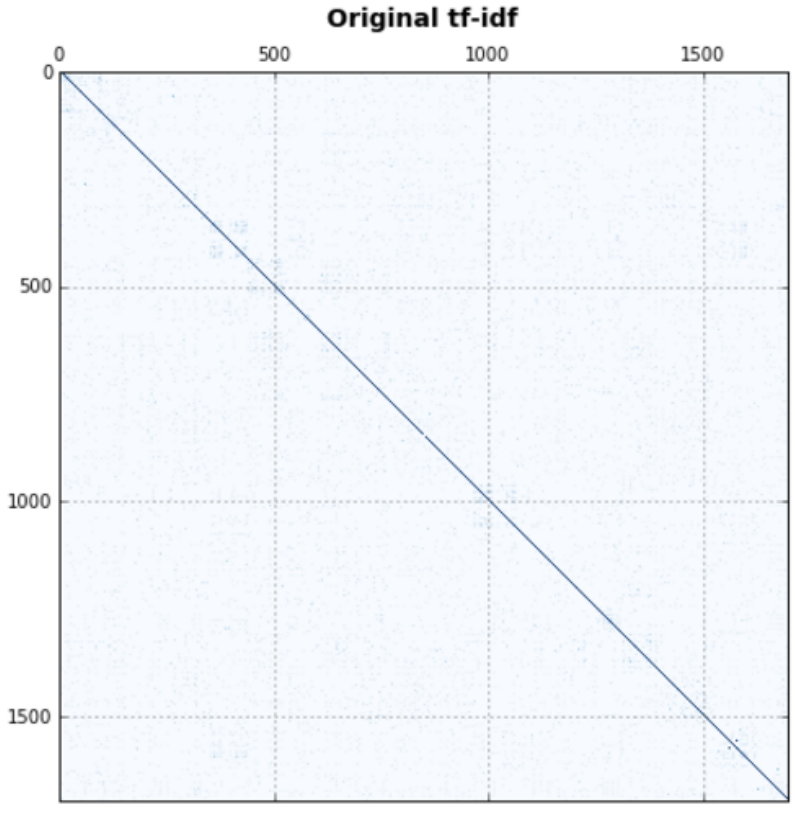
The final BOW:

represent a challenge for conventional text mining techniques.

												w_{12}	w_{13}
D_1													0
D_2													0
D_3	1	0	0	0	1	0	0	0	0	0	0	1	1

Problem Formulation

Small number of terms $\Longrightarrow$ poor statistical information to find any kind of relationships.

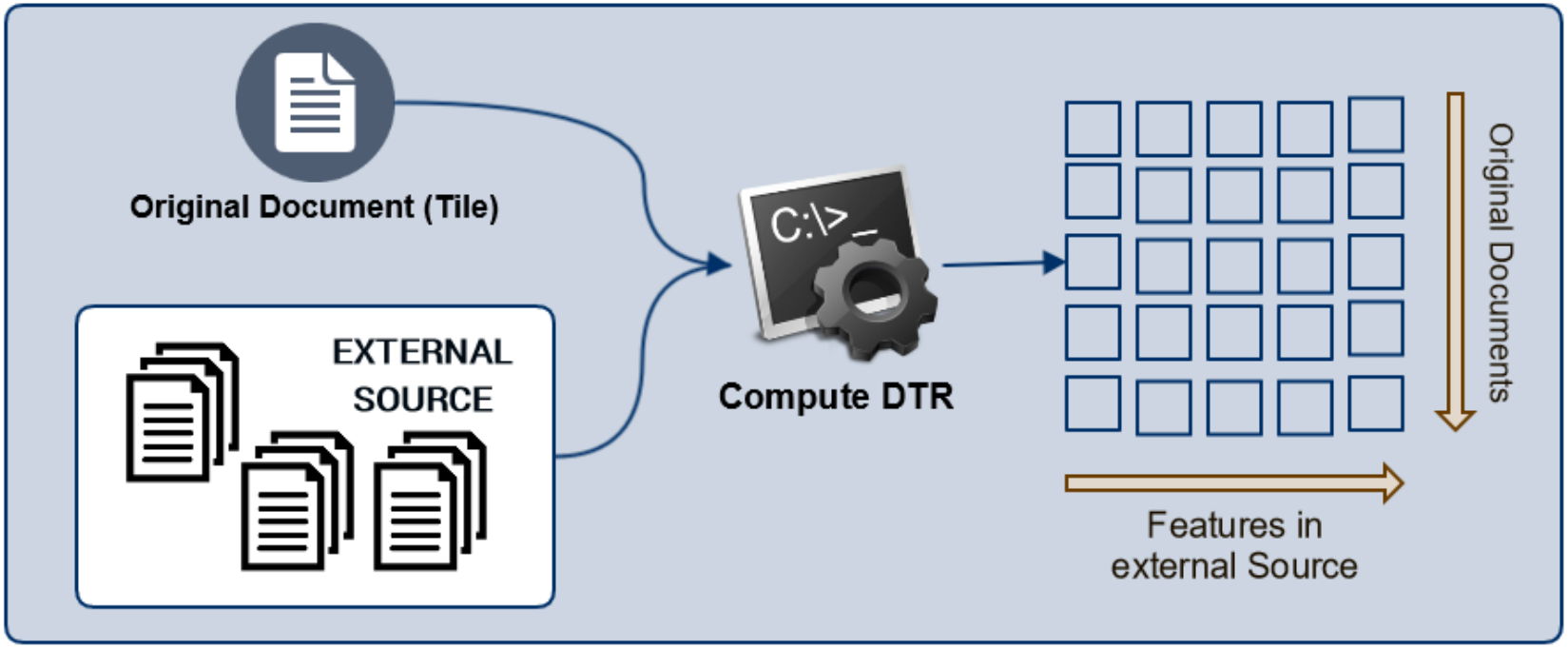


Cosine similarity ussing tf-idf representation of original text

Background

General Idea

Use external source to expand the original representation of short texts.



External Source problem

- Find a apropiate external source to expand the semantic meaning of the documents.

ScienceDirect

ELSEVIER

Google Scholar

Scopus

WordNet
A lexical database for English

WIKIPEDIA
The Free Encyclopedia

Distributional Term Representation (DTR)

A DTR is a way to expand semantic representation of terms, compute w_{jk} extracted of related text [Cabrera et al., 2013]. Let w_{jk} .The representation of a document d_i based on DTRs is:

$$d_i^{dtr} = \sum_{t_j \in d_i} \alpha_j w_{t_j}$$

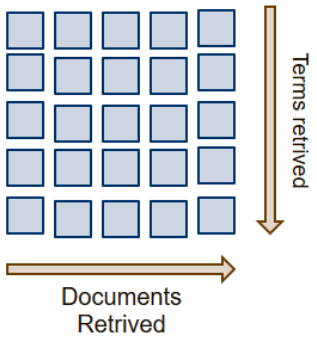
Where α_j is a scalar that weights the contribution of term $t_j \in d_i$ into the document representation. Many options are available for defining α_j.

DOR Representation

Let $w_{jk} \in [0,1]$ represents the contribution of k-th document to semantics representation of j-th text.

$$D = diag\left(\frac{|T|}{\pi(d_1)}, \ldots, \frac{|T|}{\pi(d_N)}\right)$$

$$DOR = \underbrace{(1 + log(A^T))}_{A'} D$$



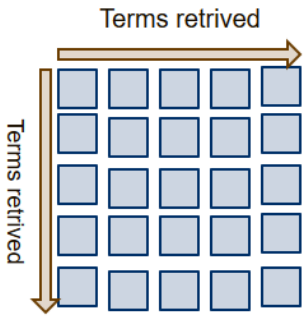
Where $A_{ij} = df(d_j, t_i)$, $\pi(d_k)$ is the number of diferent terms in the dictionary T, that appear in the document d_k, it's mean $\pi(d_k) = |\{t_i \mid t_i \in d_k \wedge t_i \in T\}|$.

TCOR Representation

We found $\vec{w}_j =< w_{j1}, \ldots, w_{j|T|} > \in R^{|T|}$, such that $t_j \in T$ and $w_{jk} \in [0, 1]$ that is the contribution of k-th term to semantics representation of j text.

$$D = diag\left(\frac{|T|}{\gamma(t_1)}, \ldots, \frac{|T|}{\gamma(t_{|T|})}\right)$$

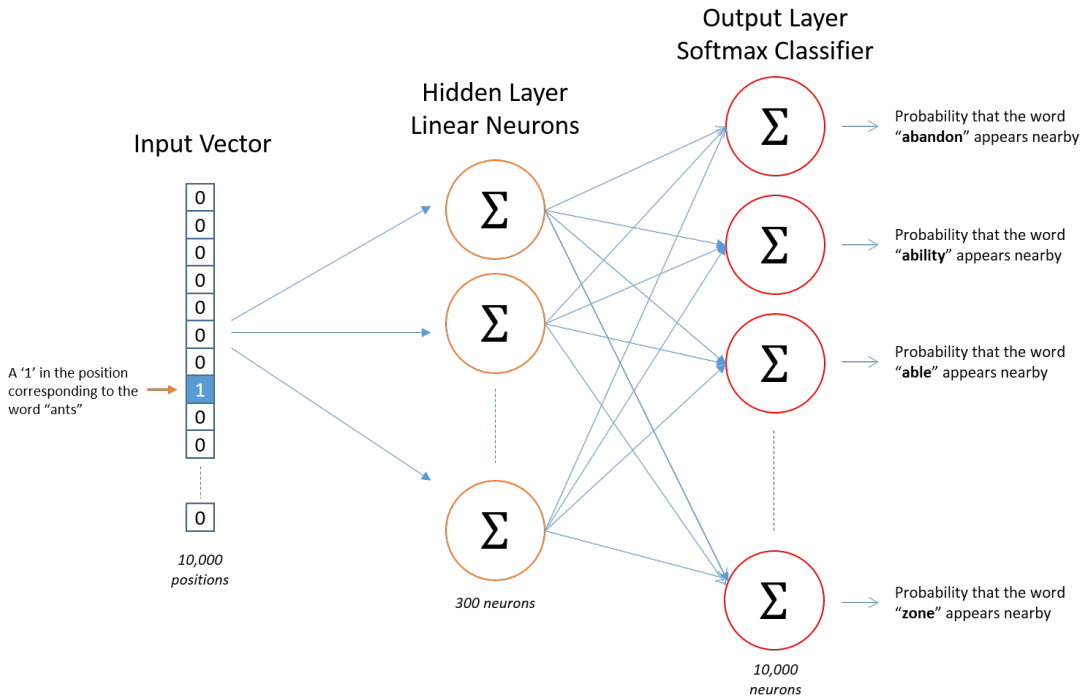
$$TCOR = DB' = D(1 + log(B^t * B))$$



Where $B_{ij} = \begin{cases} 1 & \text{si } t_j \in d_i \\ 0 & \text{e.o.c} \end{cases}$, $\gamma(t_j)$ represents the number of diferent terms in the dicctionary T that co-occurs with t_j in at least one document.

Word2Vec Representation

Maps words to continuos vector representations (i.e. point in an N-dimensional space), using the continuous skip-gram model [Mikolov et al., 2013].



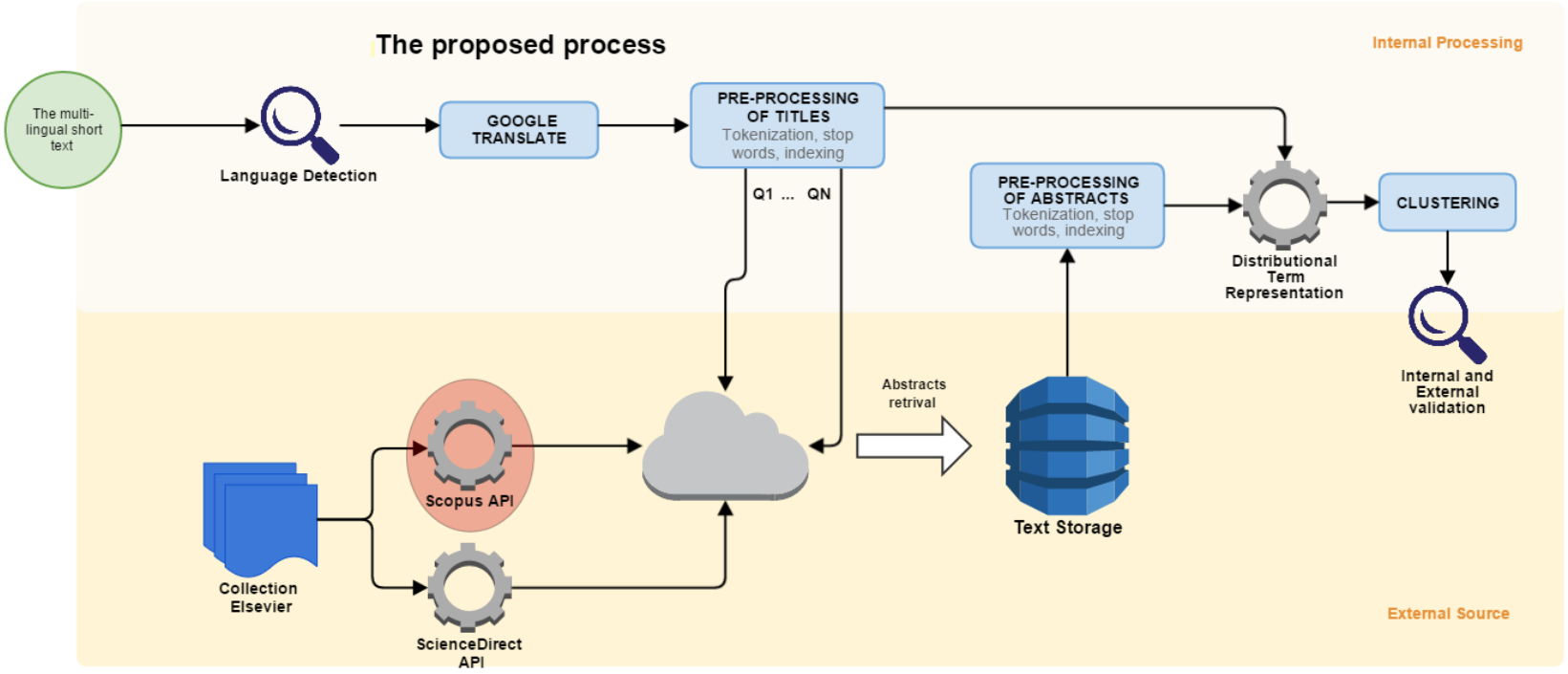
Word2Vec Representation

Post-training, associate every word $w \in W$ with a vector $\vec{w}_j$:

- $\vec{w}_j$ is the vector of synaptic strengthes connecting the input layer unit w_j to the hidden layer
- more meaningfully, $\vec{w}_j$ is the hidden-layer representation of the single-word context $C = w_j$.
- vectors are (artifically) normed to unit length (Euclidean norm), post-training

Proposed Methodology

General Process Map

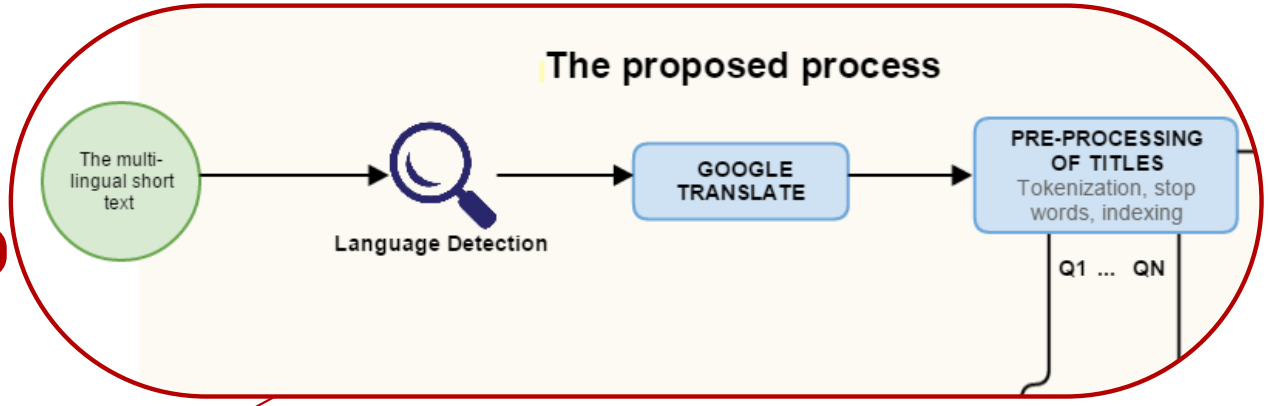


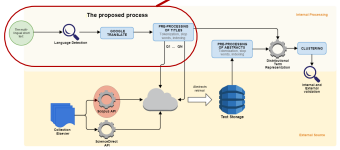
Short-text clustering method proposed

Preprocessing Step

Language Detection

- Constructing language classifier.
- Identify English/Spanish Documents.
- Google Translate API on Clud Google Service

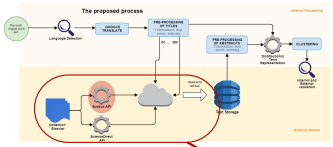




Pre-processing

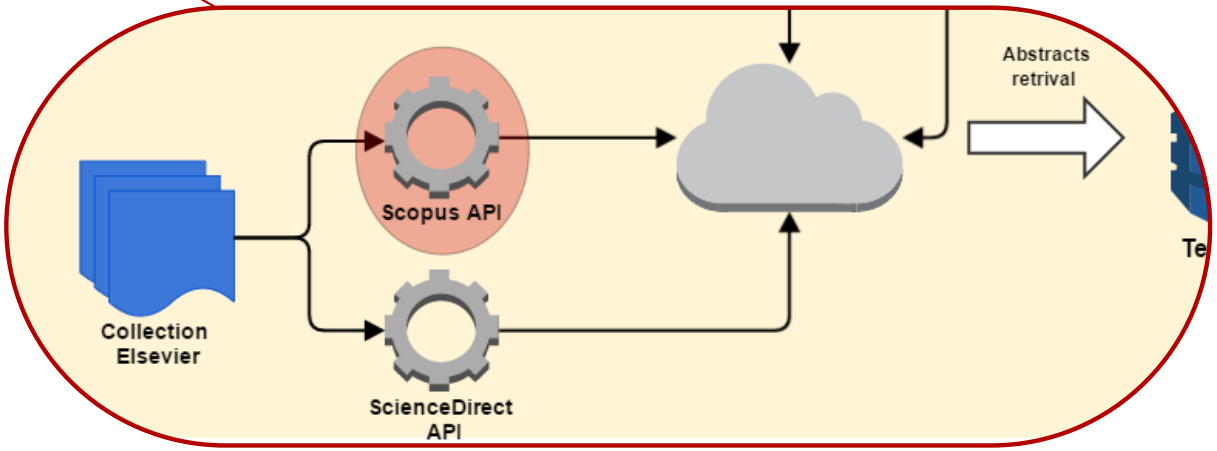
- Depurate original, remove recurrent words.
- Remove Stop Words (NO Steming).
- Tokenization and construct dictionary.

Query Process

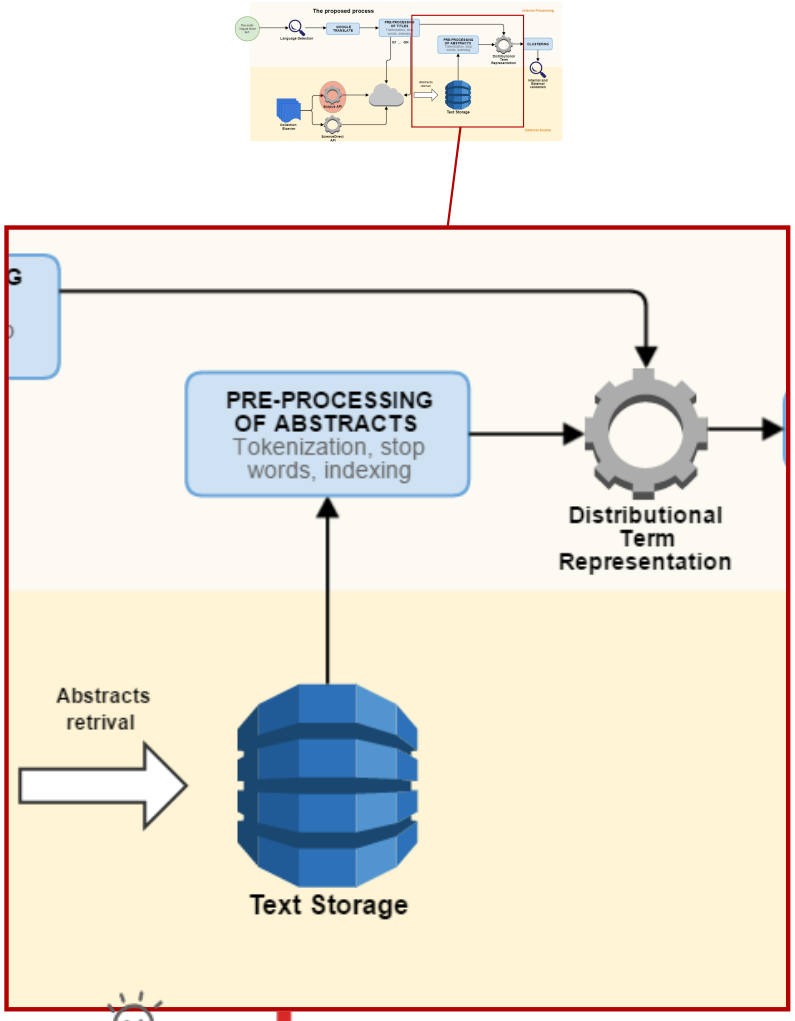


Query Process

- Use Scopus API and Scien Direc API
  - Maximum 100 articles retrived by query
  - Explore number of document
- Contruct a Python wrapper for Elsevier APIs



Distributional representation



Distributional Representation

1. Consolidate and depurate abstracs retrived.
2. Make co-ocurrence term matrix and
3. Build our own Word2Vec model based on Scopus data base.
4. Compute distributional term representation:
   - Using TCOR representation
   - Using DOR representation
   - Using Word2Vec models: (Google News, Wikipidea and own Word2Vec representation)

Clustering Algorithms

Methods

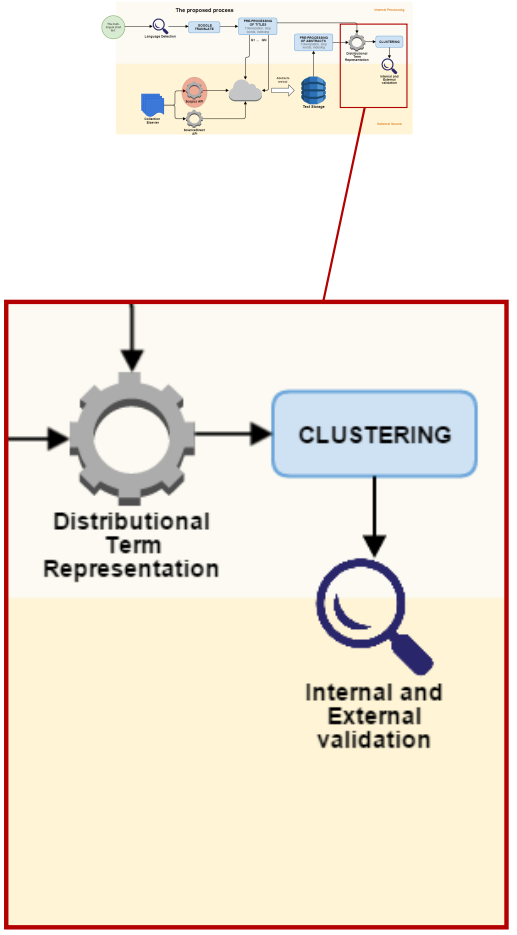
- Kernel K-Means (Cossine Kernel, Gaussian Kernel)
- Spectral Clustering
- Non-Negative matrix factorization
- Online Kernel Matrix Factorization (OKMF)

Internal Validation

- Davies - Bouldin
- QError
- Silhouette

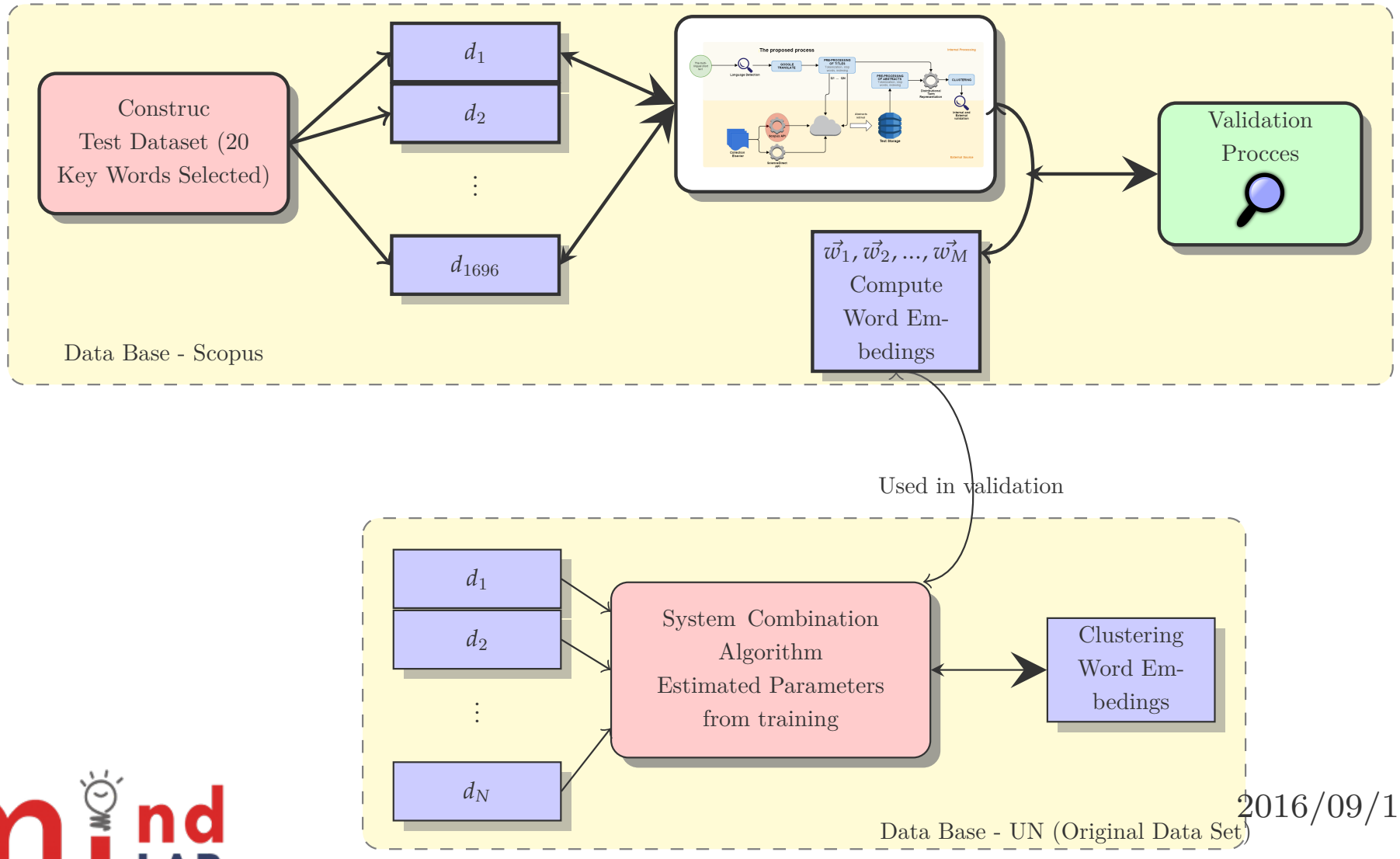
External Validation

- Purity
- Adjusted mutual information score
- V Measure score
- Adjusted rand score



Experimental Results

Experimental Setup



Data Sets

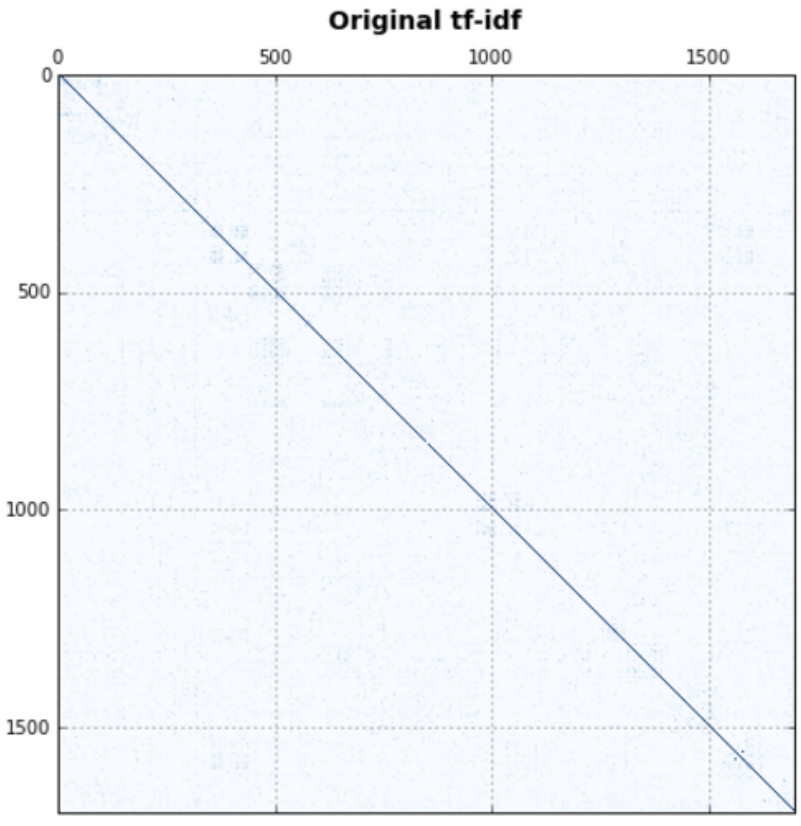
Scopus Data Set

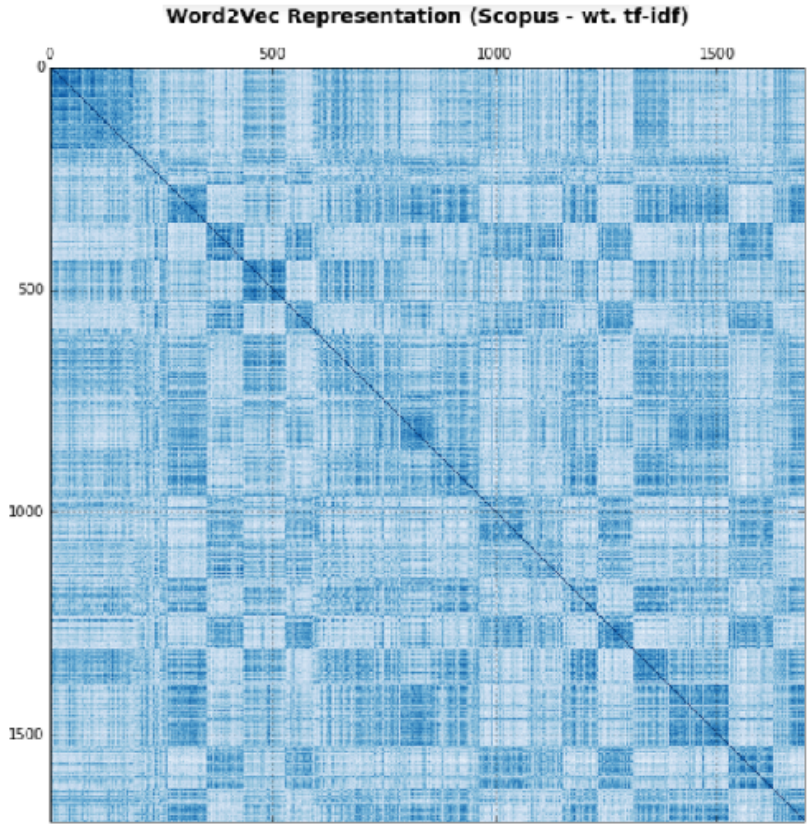
- We construct this data set selected 20 different key words.
- We make queries and consolidate a data set with 1696 titles (our gold standard).
- We have 76820 different terms and 22267 documents retrieved.

UN Data Set

- The initial data set has 3718.
- Processioning text removing some string using regular expression.
- In this data set we have 202792 different terms and 37069 documents retrieved.

Tf-idf vs Word2Vec-Representation





Cosine similarity using tf-idf representation of original text (left) and word2Vec expanded representation (right).

Comparison of clustering methods

	Davies-Bouldin	QError	Silhouette	adjusted rand score	homogeneity score	v measure score	adjusted mutual info score	Purity
Spectral(tf_idf)	0.974	7.513	0.010	0.046	0.207	0.223	0.174	0.253
Kernelkmeans(tf_idf)	0.979	9.704	0.007	0.064	0.176	0.177	0.142	0.238
Spectral(TCOR)	0.405	4.386	**-0.008**	0.205	0.389	0.398	0.364	**0.429**
Kernelkmeans(TCOR)	0.406	5.002	-0.041	0.197	0.383	0.393	0.358	0.391
Spectral(W2V_GOOGLE)	3.033	4.127	0.014	0.147	0.305	0.306	0.277	0.327
Kernelkmeans(W2V_GOOGLE)	3.006	4.025	-0.015	0.148	0.319	0.322	0.292	0.318
Kernelkmeans(W2V_SCOPUS)	**0.379**	**2.934**	0.041	**0.212**	**0.395**	**0.402**	**0.371**	0.407
Spectral(W2V_WIKIPEDIA)	12.336	3.573	0.001	0.167	0.341	0.342	0.314	0.380
Kernelkmeans(W2V_WIKIPEDIA)	12.265	3.646	0.003	0.169	0.352	0.354	0.326	0.364
Kernelkmeans(TCOR*W2V_SCOPUS)	0.379	2.934	0.041	0.212	0.395	0.402	0.371	0.407

Conclusions

- The performance of methods using external knowledge related to the document collection was better than the performance of traditional methods.
- Kernel Kmeans and Spectral Clustering showed better results that the other methods tested.
- (Future Work) We would be used a Spanish corpus to train word2Vec word embeding.

Thank you

¿Some Question?

References I

Cabrera, J. M., Escalante, H. J., and Montes-Y-Gómez, M. (2013).

Distributional term representations for short-text categorization.

In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 7817 LNCS, pages 335–346.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).

Distributed representations of words and phrases and their compositionality.

References II

In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, Advances in Neural Information Processing Systems 26, pages 3111–3119. Curran Associates, Inc.