A text mining approach to investigate urban planning documents: study case of Reunion Island

Eve ETIENNE[1*], Jean-Philippe PRAENE[1], Divya LEDUCQ[2], Jean-Claude GATINA[1]

[1]Laboratory of Physics and Mathematical Engineering for Energy, Environment and Building (PIMENT), University of Reunion Island

[2]UMR CNRS 7324 Cités, Territoires, Environnement et Sociétés (CITERES), Polytechnic College of the University of Tours

***Corresponding author:**

Email: eve.etienne@univ-reunion.fr

Address: Eve ETIENNE – Laboratoire PIMENT
117, rue du Général Ailleret
97430 Le Tampon, LA REUNION, FRANCE

Abstract:

In France, land-use planning is governed by a set of planning documents. The latter are articulated among themselves at different spatial and hierarchical scales. At the municipality scale, the urban planning document used as reference is the Local Urban Plan (PLU). This one defines the own long-term orientations and rules of the planning of the place. Reunion Island is a French region, and since the last years, municipalities are replacing their earlier urban planning document, the Land Use Plan (POS), by a PLU. One of the documents composing the PLU is the Sustainable Planning and Development Project (PADD). It indicates the general politic orientations to contribute to the sustainable development of the locality, the general orientations for housing and amenities development and the concrete objectives concerning land-use and urban sprawl. However, the PADD seems to show similarities for one to another, while this document must be specific to its locality.

This paper aims to analyze the PADD of Reunion island municipalities with the text mining approach. The latest versions of the gathered PADD constitute the corpus. Frequencies and correlation analyses, document clustering and topic modeling methods have been used to understand the main topics of these documents better. Then, filtering methods have been applied to the corpus in order to extract strategies of these territories for their sustainable planning. After that, most common terms have been associated with most adapted main topics (chosen keywords) in order to traduce the information to a simplified model, based on the contribution of these keywords. Finally, these outcomes have been compared with results of the same methods applied on draft-copies of existing “Eco-PLU” in order to determinate the differences between them.

One of the issues discussed focus on the extent to which the localities characteristics are taken into account throw the urban planning. The second concerns the real adequacy of these PADD for the sustainable urban planning of their localities, and more particularly its capacity to deal with rapid changes of the territories’ facing to its long-term orientations.

Keywords: *Text mining, urban planning, sustainability, locality, Reunion Island*

1. Introduction

In France, land-use planning is regulated by a set of planning documents. They are interlocked at different spatial and hierarchical scales. Their objective is to define their own long-term orientations and rules of the planning of the place for their respective scale. Since 2000, the law on solidarity and urban renewal law (SRU – « *Solidarité et Renouvellement Urbains* ») imposes to replace precedent planning documents by news. Hence, the Scheme for Territorial Coherence (SCoT – « *Schéma de Cohérence Territoiriale* ») replaces the Director Scheme for the inter- municipality scale. With some exceptions, it is the highest territorial urban planning document. The Local Urban Plan (PLU – «*Plan Local d'Urbanisme*») replaces the Land Use Plan (POS – «*Plan d'Occupation des Sols* ») at the municipality scale and must be conforming to the SCoT. The SCoT and the PLU are each composed by a set of respective documents. Among these documents, the both have their own Sustainable Planning and Development Project (PADD - « *Plan d'Aménagement et de Développement Durable* »), which defines the general urban planning orientations and strategies to enhance the sustainable development of the territory.

In Reunion Island case, all municipalities have not ever defined their PLU yet: their creation is time-consuming and need money and administrative approbation. Concerning municipalities which have already defined theirs, the comparison of their PADD seem to show similarities from one to another, while this document must be specific to its locality. This might be explained by conformity and/or compatibility obligations of PLU to higher planning documents, all the more that in Reunion Island case, there is another planning document higher than the SCoT: it is the Regional Urban planning Scheme (SAR – « *Schéma d'Aménagement Régional* ») defined at regional scale (Bénard-Sora & Praene, 2018). The Figure 1 summarizes the hierarchical organization of the urban planning documents on the Reunion Island, and identifies the documents which compose the Local Urban Plan.

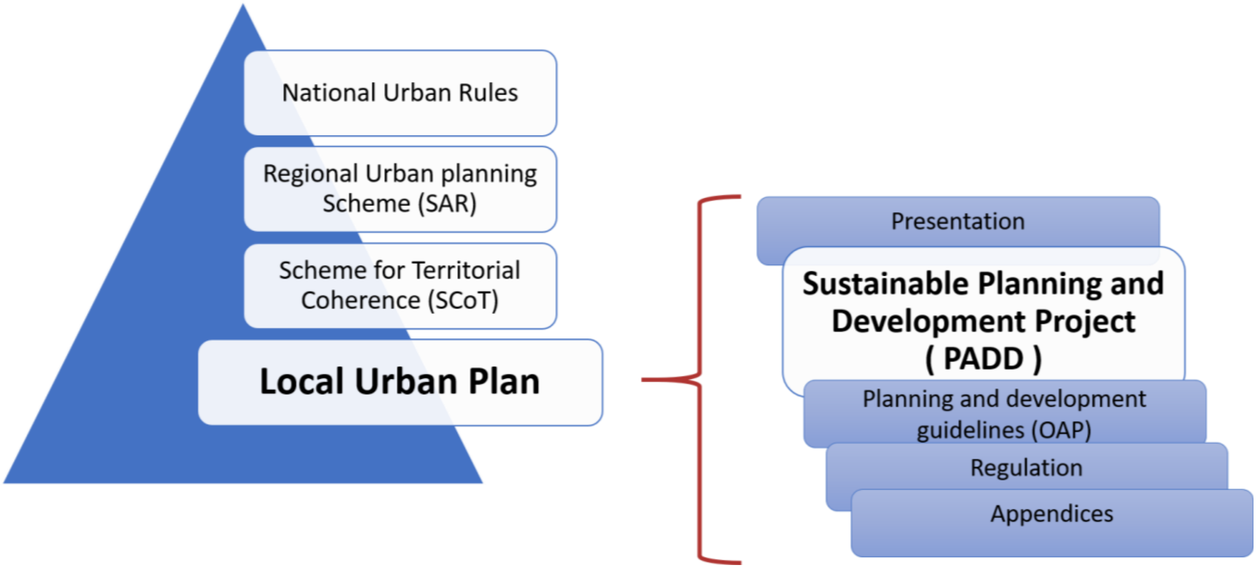
National Urban Rules
Regional Urban planning Scheme (SAR)
Scheme for Territorial Coherence (SCoT)
Local Urban Plan
Presentation
Sustainable Planning and Development Project (PADD)
Planning and development guidelines (OAP)
Regulation
Appendices

Figure 1: Hierarchical organization of urban planning documents in Reunion Island and composition of the Local Urban Plan (Eve ETIENNE, 2019)

The aim of this paper is to analyze the PADD of the Reunion Island's PLU to identify their strategies. A better comprehension of these PADD highlight their similarities and their characteristics. Furthermore, the reasons of these similarities are researched, to understand if they are due to the hierarchical constraints or not. It is also interesting to compare particularities reveled and localities characteristics to understand how the latter are taken into account in these planning documents. Furthermore, the adequacy of these PADD with territories and population needs is also questioned, because PLU are long-term orientations planning documents. So, an objective is to determinate in what extent these PADD are useful for urban planning face to the higher documents existing strategies.

This paper contributes to highlight strengths and weaknesses of actual system of urban planning in France, and focus on Reunion Island case. The utilization of text mining approach in urban studies is an emerging method: it presents a real advantage, because it allows comparing and analyzing several text data with a scientific and objective method.

2. Methods and tools

Text mining is part of big data family, where data sources used are textual documents (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Feldman & Sanger, 2007; Tan, 1999). One or several documents are tokenized (Hassler & Fliedl, 2006) (that means broken down) in chosen textual patterns, which can be term, group of terms, sentence, paragraph, section or entire document units. These defined patterns constitute the analyzed processing units. In this paper, data sources are the latest version of eleven municipalities' PADD available on-line, and the patterns are terms or group of terms following the different analyze methods.

RStudio is a free and open-source Integrated Development Environment (IDE) software for the R programming (Racine, 2012). This interface allows the installation of different packages following the user needs. This text mining study required the *tm* (Feinerer, Hornik, & Meyer, 2008), *tidytext* (Silge & Robinson, 2016), *cluster* (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2019) and *topicmodels* (Grün & Hornik, 2011) packages mainly.

The methodology is composed by the following steps presented in Figure 2.

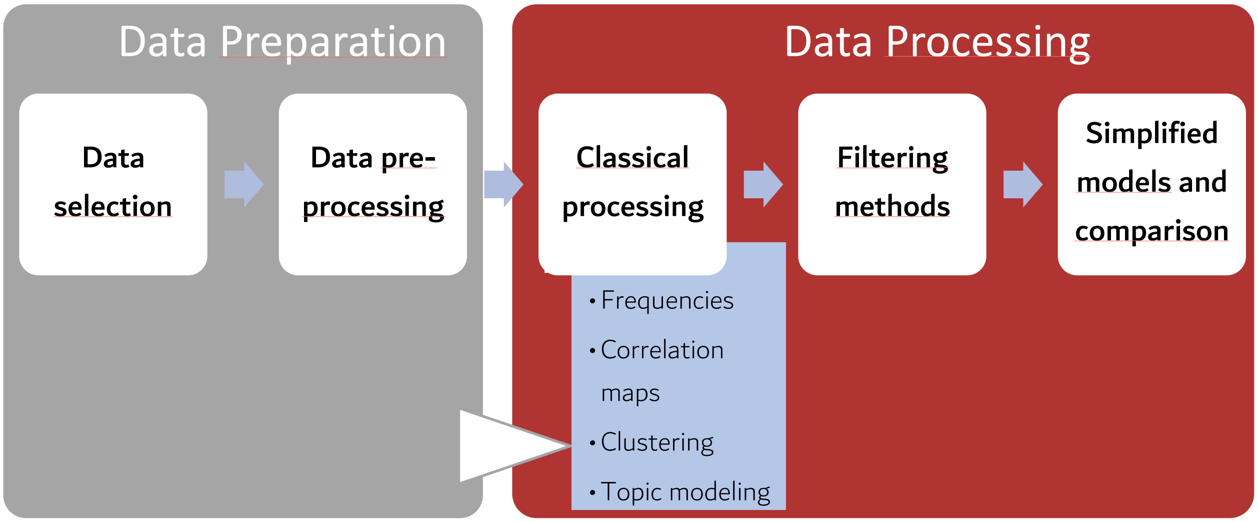

Figure 2: Methodology of the PADD's study (Eve ETIENNE, 2019)

2.1.Data selection and Preprocessing step

The corpus is constituted by PADDs. For each of the 24 island municipalities, their PLU has been searched and available documents have been downloaded. Eleven PADD have been distinguished and selected. They were produced or revised between 2004 and 2017. These PADDs are charged in a PDF format on RStudio interface.

The pre-processing is a necessary step in order to keep main information of the corpus and to reduce the dimensionality of the corpus. Therefore, documents are cleaned and transformed.

The cleaning of the documents allows removing several elements that can skew the occurrences measures. PDF format generally induced noise induced by the document formatting, as headers and footers for example. The latter are removed thanks to a program that detected same text strings at same position in the document. Then the documents are converted into a corpus object to be comprehensible by the computer (Allahyari et al., 2017; Jivani, 2013). Then, others elements are deleted as numbers, punctuation or common words (called stop-words).

The transforming step implies to manipulate the terms structure. Throughout this study, the stemming – which transforms the terms into their stem words – was used to reduce the dimensionality; and the tokenization into bigrams was explored to create groups of two terms into the whole corpus.

2.2.Classical processing

First processing steps are frequencies, calculated on the document-terms matrix. Frequencies studied are:

- Term frequency, to highlight the most frequent occurrences of the terms.
- Document frequency, which is a binary weighting to count the presence or not of the term in the documents of the corpus. Most frequent terms allow to identify the main topics in the whole corpus.
- Tf-idf frequency, which is a term frequency by the inverse document frequency weighting. Most frequents terms indicate terms which are frequent in a few documents, so particular cases (Aggarwal & Zhai, 2012).

Correlations between the terms and between the bigrams are also analyzed. They are calculated for the first five percent of the most frequent occurrences. These correlations permit to create correlation maps (semantic maps), that indicate relation strengths. The minimal correlation threshold is fixed at 0.5 for the plot. The just threshold is chosen following an iterative approach that reduce progressively the threshold value: the objective is to determinate the right balance between the best visualization in order to be able to analyze the results, and a relevant correlation threshold. The correlation map is drawn thanks to Gephi interface (Bastian, Heymann, & Jacomy, 2009).

A document clustering has been made on the corpus to identify PADDs which are most similar. The clustering is realized thanks to the euclidian distance calculed inyo the document terms matrix. The normalized tf-idf weighting of this matrix is used in order to discriminate as possible the documents. The clustering is generated following ward.D2 method (Murtagh & Legendre, 2014). The clusters' optimal number is determinate by the Dunn index (Pakhira, Bandyopadhyay, & Maulik, 2004). Same result is returned with K-means clustering algorithm (Bholowalia & Kumar, 2014; Jain, 2010). The corpus of each cluster is generated thanks to the documents that composed them in order to understand their characteristics.

Finally, a topic modeling has been realized to identify the main topics of the PADD of Reunion Island. It uses the Latent Dirichlet Allocution (LDA) model which generated topics from the corpus terms. Then it is possible to determinate the probability that a term has been generated by a topic (beta matrix) and the probability that a document has been generated by a topic (beta matrix).

The topics optimal number is estimated thanks to the Gibbs method. Several metrics are generated: the optimal number can correspond to "*CaoJuan2009*"and "*Arun2010*" minimum or to "*Griffiths2004*" maximum (Nidhi, 2017). The maximum of the harmonic mean metric determinate the optimal number too (Wallach, Murray, Salakhutdinov, & Mimno, 2009) .

2.3.Filtering methods

Following the first results which highlight the common vocabulary shared by the PADD, it is purposed to perform them with filtering methods for a better interpretation. This step could be regarded as a second pre-processing which allows to extract a new most relevant information.

First method was to remove the vocabulary linked to administrative form of the documents and to the urban planning. They are identified throw the most relevant document frequency occurrences. The second was to remove terms which appears in all the documents to get a less subjective approach. All the document frequency occurrences of the highest count were deleted.

Concerning the clustering results, the terms which appear in the more than one group are filtered, in order to highlight their specific composition.

The results of the topic modeling were performed with the selection of terms which present the highest beta average gaps against the others topics. Into these new relevant terms, the first thirty of each topic were kept, then terms which appears in more than one topic were filtered. That allows filtering terms with closed probabilities to belong to different topics and to identify particular terms of each topic.

Hence, these filtering methods allows to extract the new most relevant terms and to better understand the sustainable strategies presented by the PADDs.

2.4.Simplified models

Reasons why a document is affected to one or another cluster cannot be clearly determinate throw their most relevant occurrences. Concerning the realized topic modeling, it identifies several topics. Nonetheless, the determination of the topics through the terms which compose each of them is delicate. For a better comprehension, the alternative was to create a simplified model for the both cases. The processing is applied on the first ten percent of the corpora, because some authors consider that main information is contained in this part (Aggarwal, 2015) .

A list of sustainable planning keywords is selected into the most frequent terms: “*natur*”, “*econom*”, “*social*”, “*project*”, “*politic*” and “*culture*”. Then, each term of the topics is associated to the first keyword with which it is most correlated. As the correlation threshold is fixed at 0.5, if the term is not correlated with one of the terms of this list, it is associated with the notion “*autre*” (which means *other*). This processing allows to create a simplified model of the clusters and the topics, which are described by the contribution of the fixed keywords.

2.5.Comparison of PADD to the Eco-PLU

Dijon was the first city to purpose an Eco-PLU in 2010 (Boquet, 2010, 2014): it has replaced its POS by a PLU that includes a high environmental dimension, which became the first French Eco-PLU. Actually, there is any clear definition of the Eco-PLU, but French municipalities trend to purpose theirs, that traduces sustainability willingness of the policymakers.

Some municipalities of Reunion Island trend to edit their "Eco-PLU" too, which has to be an improved version of classical PLU. Actually, none have been approved yet, but draft versions indicate their objectives and strategies. Saint-Pierre is a municipality of Reunion island. Its latest Eco-PLU's PADD and the actual PADD are compared using their frequencies for different tokenization patterns. This final step allows to identify the existing orientations and strategies difference between these planning documents.

3. Results

3.1.Classical processing

3.1.1. Terms frequencies analyzes

The results of the term frequencies highlight the most frequent terms present in the corpus. The first ten are « *ville* », « *developpement* », « *commune* », « *amenagement* », « *projet* », « *espaces* », « *territoire* », « *urbain* », « *centre* » and « *logements* ». They might be assimilated to main planning objectives for Reunion island, but all of these terms are common of general urban planning vocabulary.

Furthermore, the document frequency analysis reveals that the first thirty terms are present in all the documents, and the first one hundred and sixty terms are present in more than 75 % of the corpus. It highlights that these terms are not enough discriminant to identify differences of strategies between the PADDs.

Concerning the tf-idf frequency, the highest frequency is lower than 0.010, that indicate that terms trend to appear in more than one document. It is interesting to note that the ten most frequent terms represent municipalities (« *saintlouis* », « *petiteile* », « *communale*» and «*commune*»), a PLU editor («*codra*») and a few other terms:

- « *changements* » for the new measures' notion,
- «*calendrier*» which induces the planning notion,
- «*contournante*» for road infrastructures planned to relieve congestion on the island city centers,
- «*menager*» that corresponds to the household,
- and «*grands*» in reference to the large-scale projects.

As shown for term frequency and document frequency, these most frequent terms are not enough discriminant: they indicate municipalities or common vocabulary and issues of the PADD.

3.1.2. Correlation maps

Correlation maps are drawn to understand relations between the chosen patterns. The Gephi interface allows formatting the map for a better comprehension. The more term is red and high and the more it is linked with the others. Inversely, the more it is blue and thin and the less it is linked with the others. Red links indicate correlations close to 1, while blue links correspond to correlations close to fixed threshold.

A first correlation map is drawn used the stem terms with a correlation threshold fixed at 0.8. It is presented by the following Figure 3.

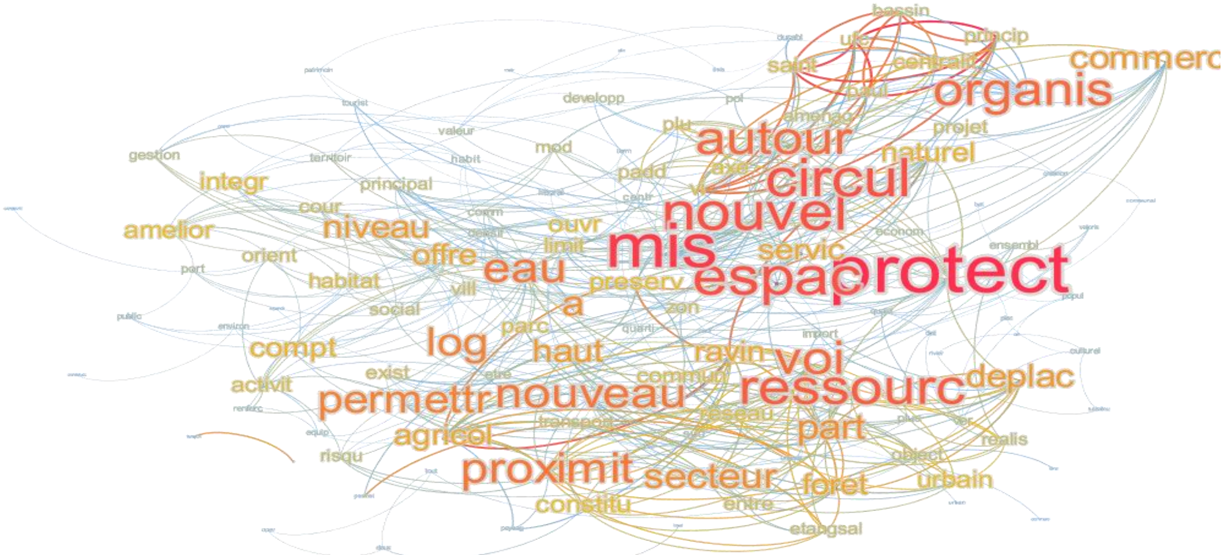

Figure 3: Correlation map of the stem terms (Eve ETIENNE, 2019)

This map highlights few red stems. The most linked stem is "*protect*" (protection), followed by stems as "*espac*" (space or area) and "*mis*" (to put, which can be relied on many action verbs in the French language). These stems presented many low links with the others, that traduces they appear in all the issues of the latter. This indicates the omnipresence of the protection policy in the PADD and urban plannings.

The most orange stems traduce the goals of the PADDs. The main group includes space organization notions with stems as "*circul*" (traffic), "*autour*" (around), "*organis*" (organization), "*voi*" (lane) or "*log*" (housing). Another concerns environmental issues as "*ressourc*" (ressources) and "*eau*" (water). Creations are also represented by stems as "*nouvel*" (new) or "*nouveau*" (new), and reasons justified in the document by the stem "*permettr*" (to permit). "*Proximit*" (proximity) is a key notion of the sustainable city and illustrates the municipalities' ambition.

The stem "*commerc*" (business) seems isolated from the cloud, but it presents many blue links with other stems, that indicates the economic aspect presence in the documents. Concerning the other terms, they are common vocabulary of urban planning, as the areas categories or the other presented above notions.

This correlation map illustrates the planning orientations of these PADDs. No particular measure different from the general objectives of sustainable planning is appears, except the stem "*ravin*" (gully). The very steep relief of the island implies indeed particular measures concerning the gullies' management, as the building approvals, the risk management or their biodiversity preservation.

The previous stemmed corpus is tokenized into bigrams (groups of two terms) and the second correlation map has been realized with them. The correlation threshold is fixed at 0.7. It is presented in the following Figure 4.

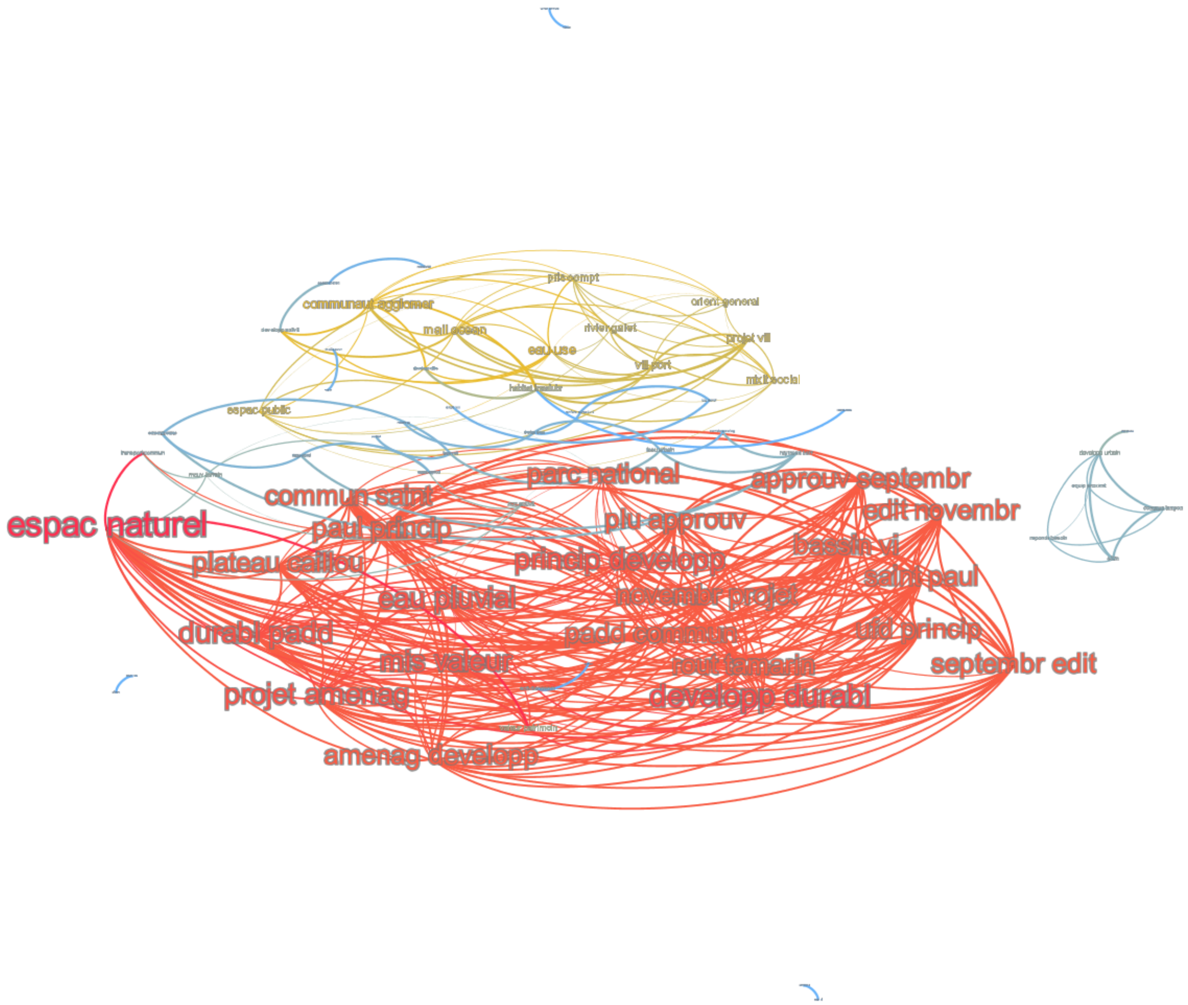

Figure 4: Bigrams correlation map (Eve ETIENNE, 2019)

This map highlights the red cloud dominance. The red bigrams indicate words which are the most linked with the others, and the red links indicate that the correlation is higher than 0.9. The most linked bigram is "*espac naturel*" (natural space), which indicate the high environmental dimension of the PADD, followed by "*developp durabl*" (sustainable development) and "*princip develop*" (development principle). Two main vocabulary groups appear: an environmental group; and a regulatory group, with the bigrams "*approuv septembr*" (approved in September), and "*edit novembr*" (edited in November). Others bigrams are linked with spatial notions (as names of places) or urban planning, with the bigram "*projet amenag*" (urban planning project). As a rule, the bigrams of the red cloud are strongly linked between themselves, that indicates strength ties between them.

The yellow cloud contains bigrams which less linked as previously, and the yellow links indicate correlations between 0.8 and 0.9. Its vocabulary group is most linked with urban planning strategies, as "*communaut agglomer*" (agglomeration community), "*mixit social*" (social diversity), "*eau use*" (waste water), "*orient general*" (general orientations), "*espac public*" (public spaces) and "*projet vill*" (city project). As previously, these bigrams are highly linked between them and are separated from red cloud: that traduces a real distinction between these two groups of terms. The blue bigrams represent bigrams which are few linked with others, and blue links represent correlations between 0.7 and 0.8. These bigrams link the red cloud and the yellow cloud. However, they are too much dissociated to analyze their relations.

This map could traduce the strategies that appears throw the PADDs. Main solutions concern natural spaces protection, the water management and environmental issue. Then the attention is axed on sustainable planning solutions. Hence, bigrams allow a better comprehension, but noise is induced by common urban planning vocabulary and administrative formatting of the documents.

3.1.3. Clustering

The Dunn index and the k-means method identify four optimal clusters. The Figure 5 below shows their distribution. However, the last two clusters are truly relevant and are analyzed in this study. The first cluster is composed by a single document ("*PADD St Joseph*") : it is isolated because it is a presentation formatting, so with fewer words than the others. The second cluster contains also one document ("*Plaine des Palmistes*") : it could not be process because it is composed by original PADD photocopies linked thanks to PDF format. The program recognizes these copies as images but it cannot identify textual strings. The two analyzed clusters are the third and the fourth.

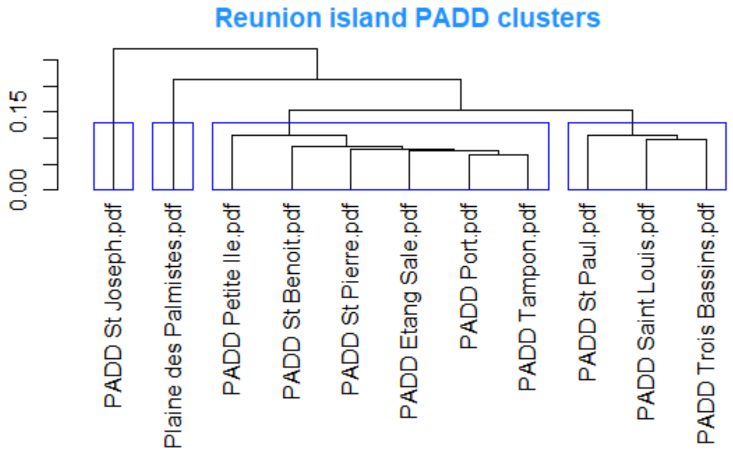

Figure 5: Reunion island PADDs clustering (Eve ETIENNE, 2019)

This clustering traduces the high existing similarity between the documents. The clusters partitioning cannot be explained by parameters as edition date, geographical area or intercommunalities. Indeed, municipalities have to edit their PLU – consequencly their PADD – following the SCoT of their intercommunality in order to comply with it. However, the documents' proximity cannot be explained by a SCoT influence, that means that the SCoTs' influences are not notable throw the PADDs.

Two corpora were generated from these two clusters documents to understand their composition. The most relevant stemmed terms of the clusters don't explain the partitioning too : same terms appear in the both clusters, with variable occurrences due to the size difference between the two corpora.

3.1.4. Topic modeling

The topic modeling of the PADDs needs to determinate the topics optimal number. The corpus with stemmed terms is used. The metrics' comparison returns different optimal numbers from one to another. It is chosen to keep the result which occurs in two metrics: "Griffiths2004" and the harmonic mean. It is also the lower result. Hence, nine topics are identified.

The most frequent stemmed terms of each topic are identified by means of the beta matrix. However, they trend to be the same from one to another, with different frequencies. They belong to common urban planning vocabulary. This fact complicates the characterization of the topics through its composition.

The PADDs' probabilities to be generated by a topic are determinated using the gamma matrix. They are presented by the following Figure 6. This result can be interpreted as the documents' composition by these topics.

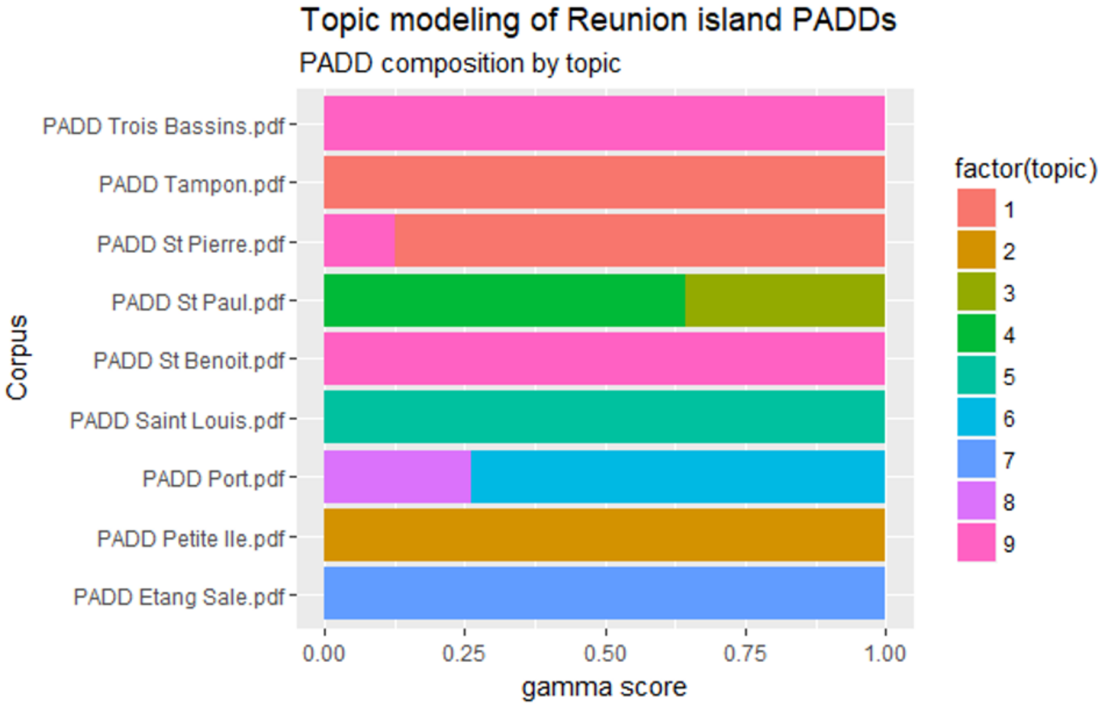

Figure 6: PADDs' composition by the topics (Eve ETIENNE, 2019)

Two thirds of the PADDs are entirely generated by a single specific topic, except for the PADDs of Trois-Bassins and St-Benoit, which are generated by the same topic. Saint-Pierre's PADD is mainly generated by the same topic as Tampon PADD. Each of the two other PADDs are generated by two specific topics. Hence, this result highlights that the majority of the PADDs has been generated by theirs own topics, that means the topics depend on each PADDs' composition mainly. This fact traduces the specificity of the PADDs to their territory.

These classical processing results reveal that the PADDs present high similarities from a municipality to another, due to their administrative form and their common subject: urban planning. However, Reunionese main orientations for sustainable planning are a protection policy and environmental sensibility. Economic aspect is also taken in account. Concerning strategies, they are also concentrated on environmental aspect and classical ways of sustainable planning. But the PADDs specificities are highlighted thanks to topic modeling, because each document is generated at least by one specific topic generally. Main inconvenient of this study part is the noise induced by the high occurrences of common vocabulary in the documents, that tf-idf weighting not amortize enough.

3.2.Filtering results

Faced to precedent results, different filtering methods have been applied on the precedent results to extract most relevant information.

The filtering of terms linked to administrative formatting and the urban planning vocabulary on the stemmed corpus and the tokenized corpus reveals poor results, as the filtering of all terms which occurs in all documents. Reasons they are not kept will be discussed. The presented results are focused on clusters and topics filtering.

3.2.1. Clusters filtering

Concerning the filtering of stemmed terms shared by two clusters, the first twenty specific terms highlight a different strategy between the clusters, as shown by the following Figure 7.

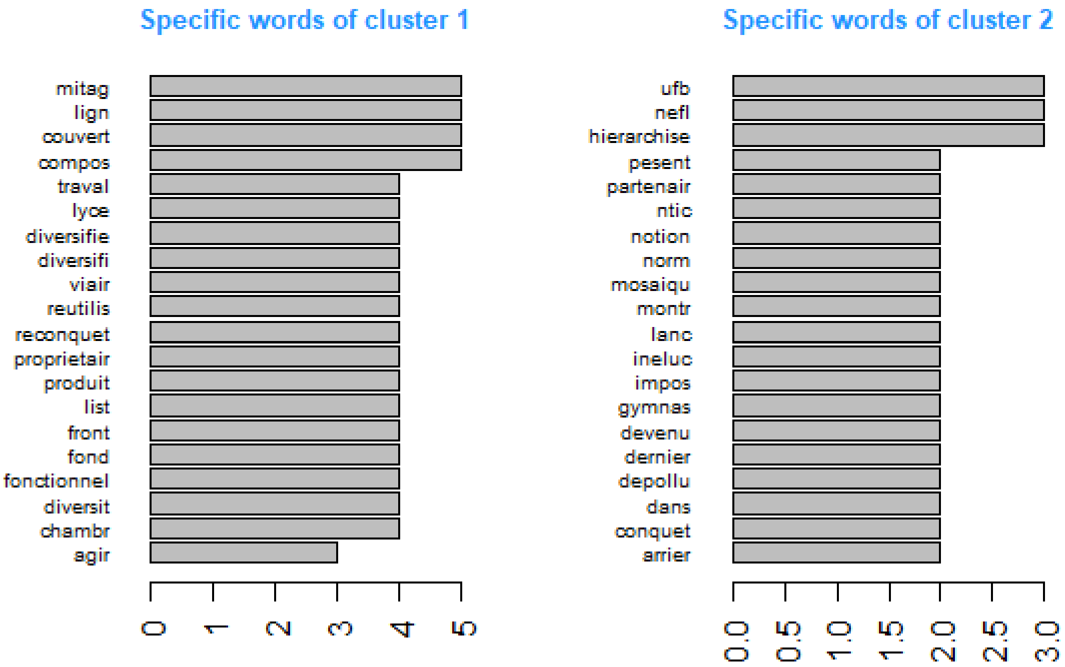

Figure 7: Specific terms of the clusters (Eve ETIENNE, 2019)

The documents of the first clusters present a most development strategy, with stemmed terms around space organization as "*diversifie*" and "*diversifi*" (diversification), "*reutilis*" (reutilization) and "*fonctionnel*" (functional). They concentrate on urban sprawl too – with "*mitag*" (sprawl) and "*reconquete*" (recapture) – which is a main issue of Reunion island. The traffic network is also considered with "*lign*" (lane) and "*viaire*" (related to road network). These action axes traduce long term strategies for a coherent planning.

The second cluster seems to present a most operational strategy. Stemmed terms concern the urban regulation vocabulary, as "*ufb*" term, which corresponds to an urban zoning, or "*hierarchise*" (hierarchical), "*notion*", "*norm*" (standard) or "*impos*" (to enforce). Others stemmed term concern particular point of interest, as "*ntic*" (NICT New Information and Communication Technologies) for example. Hence, these PADDs seem to adopt short-term and mid-term strategies, using a most operational strategy for urban planning.

3.2.2. Topics filtering

The topic filtering highlights particular terms for each of the topics that are presented by the following Figure 8. They trend to confirm the hypothesis that the topics are mainly specific to a PADD. For example, the first term of the topic 6 is the name of Le Port municipality, and the topic 3 shows "*bourg*" (bourough) and "*piton*" terms, which can be relied on Saint-Paul places (as Piton Maïdo or the Guillaume borough that is planning). However, these particular terms sample is too limiting to determinate clearly point of interests and strategies of the clusters.

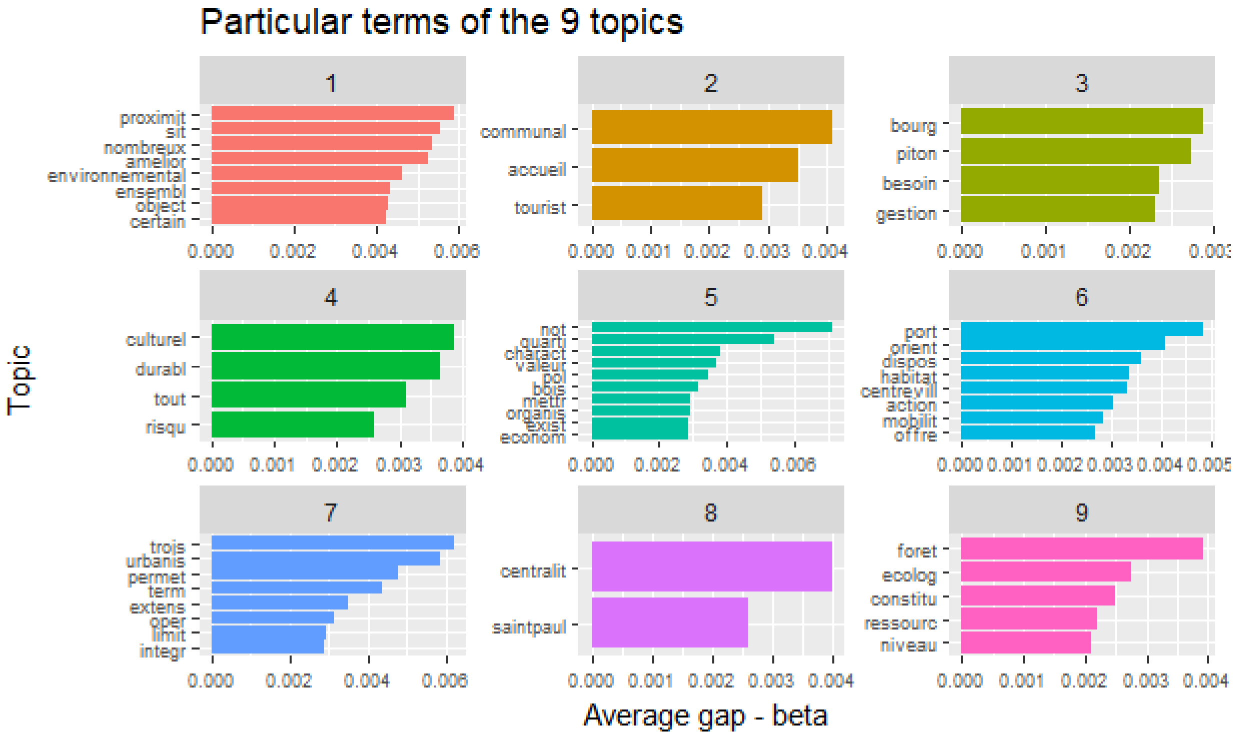

Figure 8: Particular terms of the topics (Eve ETIENNE, 2019)

These results show the filtering methods interest and potential of the in this study: they allow highlighting information that is masked by most by the predominance of common terms, without arbitrary removing.

3.3.Simplified models of the clusters and topics

The simplified models were applied on stemmed corpora of previous clusters and topics. The first ten percent were processed, for each case, in order to reduce the dimensionality.

3.3.1. Clusters simplified model

The clusters' simplified model explains the composition of the clusters by the keywords. The threshold is fixed at 0.8.

As shown by the following

Figure 9, the first cluster highlights a high importance around the project notion. That is followed by the social aspect, then the environmental and policies concerns. This cluster strategies could be traduced by a prospective vision, with policies measures. It confirms the previous achievements, where the development strategy has been highlighted. The sustainable development pillars taken in account are environment and society mainly.

The second cluster accords the main importance to the environmental aspect, that represents more a third of its composition. The other relevant interest is the cultural dimension, highest than in the first cluster. Previously, the operational strategy has been relieved: the environment predominance can be explained by the protection aspect and the urban zonings defined in these PADDs. The result can also traduce a politic ambition of cultural elements protection and valorization, as Saint-Paul for example, that is one of the first inhabited places of Reunion Island.

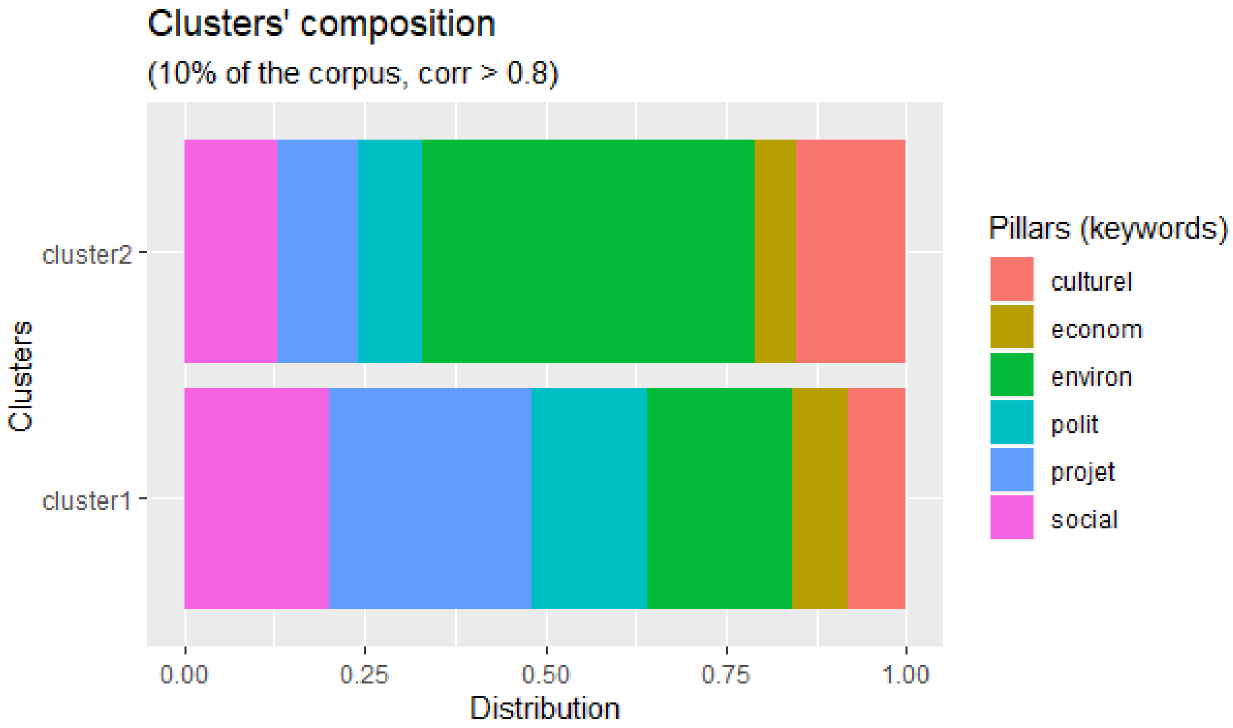

Figure 9: Clusters' composition by the keywords (Eve ETIENNE, 2019)

3.3.2. Topics simplified model

The topics simplified model generates the keywords presence rates on each topic. The correlation threshold is fixed at 0.5 to get more occurrences. They are presented by the following Figure 10.

The topics' composition by the chosen keywords varies, but the latter represent more than 40% of each topic. The keywords proportion trends to be the same from a topic to another, that confirms the previous results concerning the fact that PADDs trend to be represented by at least a specific topic. Furthermore, the high similarities between the keywords' distribution into the topics trend to confirm the hypothesis that the PADDs' composition seems to be the same. The main attention focuses on the project aspect, due to the planning objective of this document. Concerning the sustainable planning strategy, the attention focuses on environmental aspect, then the economic aspect. Depending on the topic – so the municipality – the point of interest can be cultural aspect for topic 4 (that is associated to Saint-Paul municipality), or political and social aspect for topic 6 (that is associated to Le Port municipality). The topic 1, which is composed by the keywords with more than 60 %, is associated to Le Tampon and Saint-Pierre municipality. It is possible to advance that these documents incorporate more sustainable planning terms than the others. They may better integrate the sustainable development vision into their planning strategy.

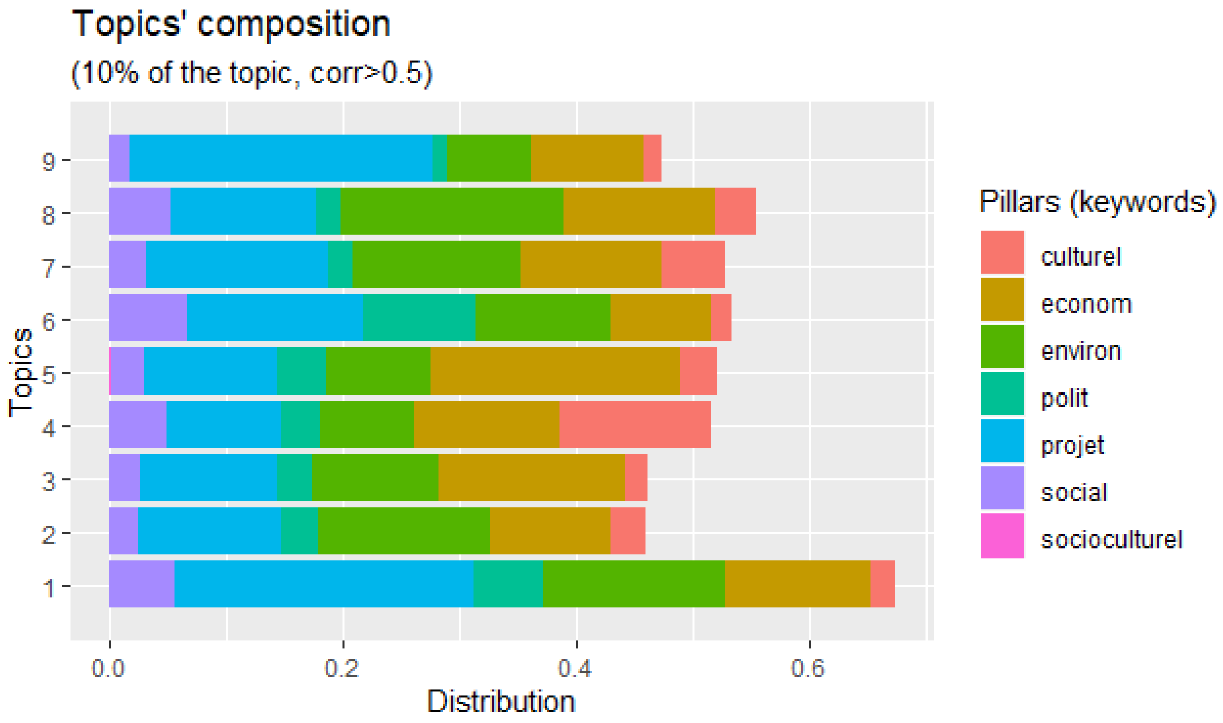

Figure 10: Topics composition by the keywords (Eve ETIENNE, 2019)

Hence, the simplified models permit a better comprehension of the clusters and topics. Furthermore, they trend to confirm previous analyzes.

3.4.Comparison of PADD to the Eco-PLU: Saint-Pierre's study case

In Reunion Island, few municipalities are beginning to working on their Eco-PLU. In this study, Saint-Pierre is chosen because it is the most advanced on this procedure. Its actual PADD and its latter version of the PADD of the Eco-PLU are compared, to understand the difference between these documents.

Each of the documents constitutes a corpus. The two corpora were compared using frequencies of different patterns for the tokenization: term, bigram and trigram. The most interesting result are the bigrams and the trigrams comparisons, presented in following Figure 11 and Figure 12.

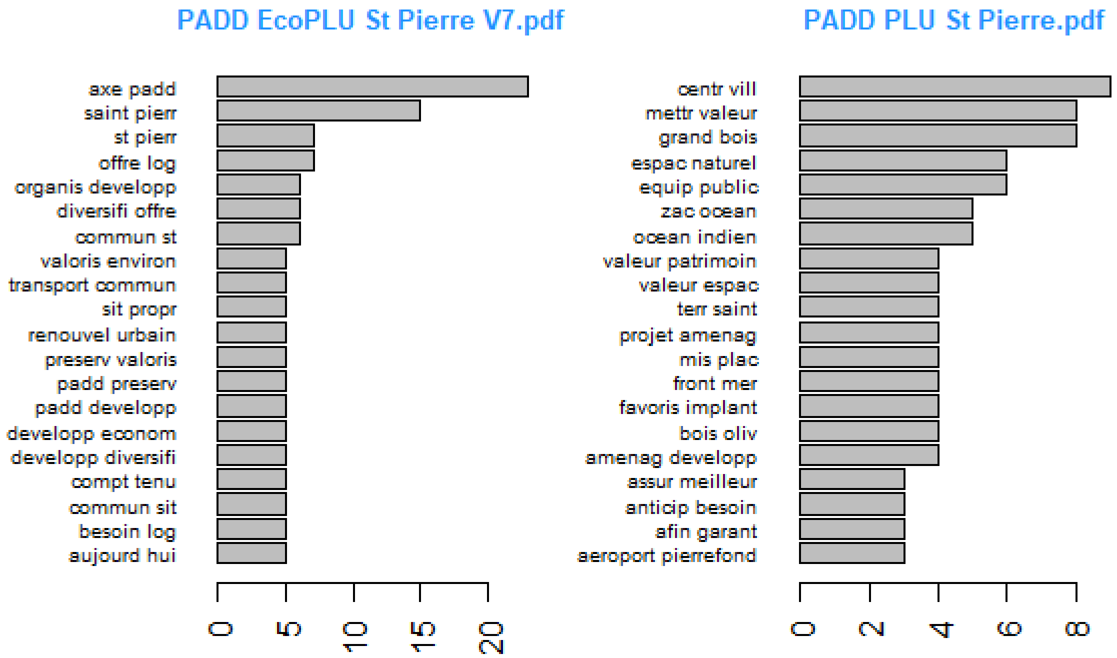

Figure 11: EcoPLU's PADD and PLU's PADD bigrams comparison (Eve ETIENNE, 2019)

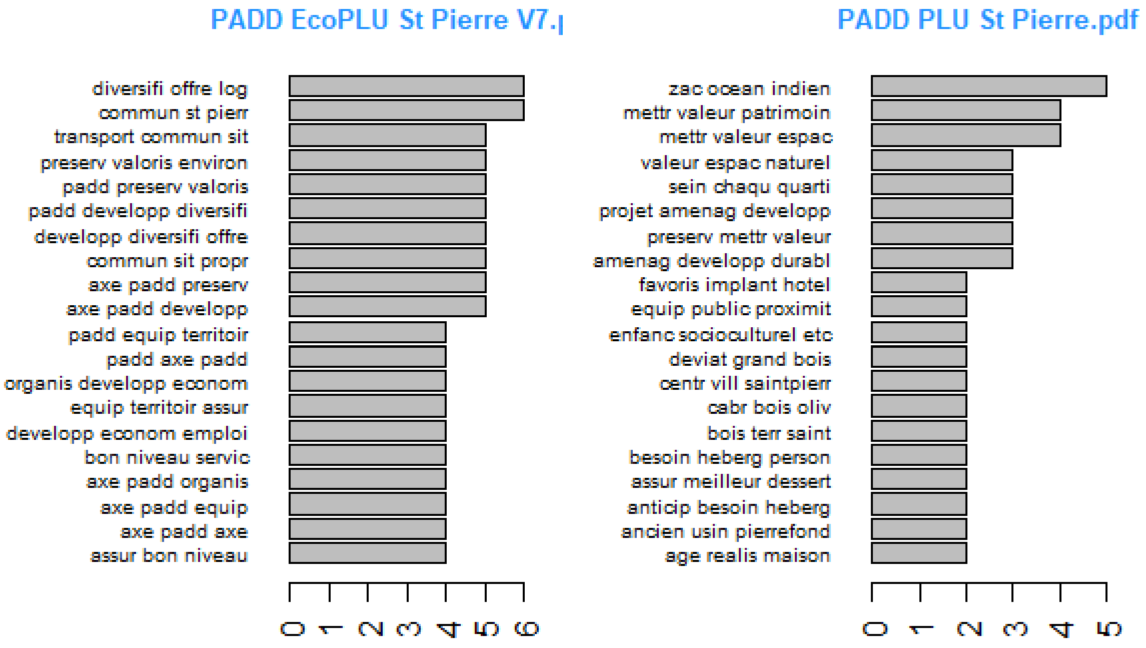

Figure 12: EcoPLU's PADD and PLU's PADD trigrams comparison (Eve ETIENNE, 2019)

The bigram comparison highlights in the actual PADD several Saint-Pierre's neighborhoods with associations as "*centr vill*" (city center), "*grand bois*" (Grand-Bois), "*terr saint*" (Terre-Sainte) and "*bois oliv*" (Bois d'Olive); or places with associations as "*zac ocean*" and "*ocean indien*" (ZAC Ocean Indien), "*front mer*" (seafront) and "*aeroport pierrefond*" (Pierrrefonds' airport). This observation induces a planning reflection axed on each neighborhood's needs. This hypothesis may be confirmed with the actual PADD's trigram ranking, with the associations "*sein chaqu quarti*" (within each neighborhood) and "*equip public proximit*" (local public facilities), and others that are:

- The previous neighborhoods or places, and others as "*cabr bois oliv*" (that corresponds to Ravine des Cabris, a surrounding neighborhood of Bois d'Olive) and "*ancien usin pierrefond*" (Pierrefonds' former factory)
- specific plannings, as "*deviat grand bois*" (Grand-Bois' detour road) and "*favoris implant hotel*" (promote a hotel installation).

The other bigrams and trigrams of the actual PADD are few general objectives and strategies, as the valorization and preservation of heritage and natural spaces, the housing demand forecast or the improvement of the service road.

Concerning the Eco-PLU's PADD, the main bigram associations are "*axe padd*" (PADD's strategic axis) and "*saint pierr*" (Saint-Pierre). These two points traduces a general strategic vision for the municipality. It is noted that any neighborhood appears in this ranking. The other association are information-poor, in sense they are general urban topics, as "*offr log*" (housing supply), "*renouvel urbain*" (urban renewal) and "*besoin log*" (housing needs). The trigram association provides a better understanding of these strategies, which are:

- Developing and diversifying the housing supply ("*diversifi offr log*" and "*developp diversifi offre*");
- working on the existing reserved-lane public transport systems ("*transport commun sit*" and "*commun site propr*");
- environment preservation and valorization ("*preserv valoris environ*");
- equipping the territory to ensure a good level of service ("*equip territoir assur*", "*assur bon niveau*" and "*bon niveau servic*");
- Organizing the economic development for employment "*organis developp econom*" and "*developp econom emploi*".

Hence, the comparison of the actual PADD and the further reveals two different approaches. On one hand, the first defines the actions for the different neighborhoods of the municipalities, with planning for the territories' needs. It favors their local development in line with the urban proximity. On the other hand, the further PADD seems to consider the global development of the municipality, with usual orientations and strategies to improve its sustainable development.

However, the lack of precise actions concerning the neighborhoods may be due to the non-definitive character of this latter document which highlights its axes, or to the municipality willingness to work first and foremost for a coherent development.

4. Discussion

4.1. Orientations and strategies trends of Reunion island urban planning

Reunion’s municipalities edit their PADD, following French urban regulation, in order to define their general urban planning orientations and strategies to enhance the sustainable development on their territory. The topic modeling analyzes highlight the specific character of these documents to their municipalities, but high similarities between them persist.

These similarities can be explained by the same objective of these documents: an administrative form and a same objective, that induce a common vocabulary utilization. But the tf-idf weighting cannot discriminate them enough, that means similarities extend beyond the language boundaries.

The correlation maps reveal that the municipalities orientations mainly focus on protection notion, concerning resources and areas. Others are common sustainable planning orientations. Concerning the strategies, they are concentrating on the development and the projects, then the usual ways for produce the sustainable city. However, the fact of producing the sustainable city on a small island must be considered with caution. The challenges of small islands’ territories are not same, on environmental, economic and social aspects (Douglas, 2006; Stratford, 2003).

It is interesting to point the model of the urban planning regulation on the island. As mentioned previously, three hierarchical levels of urban planning documents coexist on the territory, and each of them has to be conformed to the higher document level. Five inter-municipalities – hence five SCoTs, have to impose general orientations and strategies for their sustainable territory development. However, these intermediate documents’ influence is not perceptive throw the clustering of the PADDs, where neither of the two analyzed clusters highlight any proximity between municipalities belonging a same inter-municipality. This fact induces the questioning of the territory development planning coherence (Desjardins & Leroux, 2007; Leroux, 2012; Moscarelli, 2013).

The specific character of the PADD seems to indicate that municipalities define their orientations and strategies following their own territory and their political ambitions and capacities. Indeed, the clustering reveals that some municipalities adopt long-term development strategy while the others focus on the operational planning with short-term solutions. Furthermore, if all the PADDs integrate sustainable topics in their planning, the rate varies from one to another as shown by the simplified models. These two latter facts may also indicate the perception of sustainable planning – and higher urban planning documents – more than a constraint than a real planning tool. The issues of the hierarchical levels number and the application scale of the urban planning documents deserve to be examined.

Emerging of Eco-PLUs on the island traduces the policymakers’ willingness to enhance the sustainable development of their territories. These new versions of urban planning documents seem to be an opportunity to go further on the way to sustainable urban planning. They could be a useful tool to bridge identified gaps of the previous versions and to ensure a coherent planning on the municipalities, never forgetting the neighborhood’ scale.

4.2. The text mining: an interesting tool for urban studies to perform

The text mining analyzes trend to appear in urban studies. This method presents a real advantage because it permits treating high volume of documents following scientific approach. In this study case, text mining reveals its strengths and limits for similar documents processing.

4.2.1. A common urban planning vocabulary

The frequencies analyze allow identifying several vocabularies keywords of the documents and particularities. But they are difficult to exploit due to the high predominance of a common vocabulary. A filtering method was applied on the most relevant terms to remove urban planning common vocabulary. However, this processing induces a subjective approach, where the author has to choose concerning the pertinent character of each term. The second filtering method used removes all terms that are present in all the documents. However, this method removes all the terms without distinction between pertinent or not terms. These terms represent more of than 11% of the total information. Moreover, all of them belong to the top ten most frequent terms percent, and represent approximately 20% of this pertinent information. All of these reasons explain why the associated results of these methods are not kept.

This urban planning common vocabulary also appear throw the other classical processing, as the correlation maps, and the clustering and topic modeling composition. A careful watch is necessary to analyze and to understand the results. Indeed, this common vocabulary and knowledge concerning the studied territory are essential for the results' interpretation.

4.2.2. The balance between cleaning and pertinent information

The other filtering methods applied on the clustering and the topic modeling returns most comprehensible results, but the tricky issue of the balance between cleaning and the pertinent information remains. The filtering method applied on the clustering reveal specific terms, but these new corpora turned out to be too poor for generate correlation maps. Similarly, the particular terms' extraction of the topics returns variable terms numbers, that did not permit to clearly identify the topic.

The term transformation is useful to reduce the corpus dimensionality. It permits to unify the terms form for the stemming case. However, it implies to recognize the original terms forms. It is recommended to create a dictionary which kept the original term in order to facilitate the analyzes. Different terms can generate a same stemmed term, that skew the results. The lemmatization is a well alternative faced to French language complexity, but the occurrences between the classes of a same notion are not taken in account: a second processing could be necessary by identification and generating of lexical field of terms. Concerning the tokenization, it creates new and more comprehensible information. Nevertheless, some associations seem meaningless, due to the tokenization processing: some bigrams may be most comprehensible with a third term, while some terms are more meaning with their single form.

Removal of terms implies to take in account the possible lack of information. In this study, the choice is to keep the first ten percent of the corpus (Aggarwal, 2015), that represents more than 57% of the total information. Furthermore, the occurrences number is lower than 11 under this threshold. Concerning the correlation maps, they are calculated and drawn with the first five percent of the corpus, that represent approximately 43% of the total information. This choice is done in order to reduce the correlation matrix dimensionality, and to simplify the correlation maps. The thresholds' fixation also depends on the balance between pertinent information, visualization and computing capacity.

Preprocessing step and filtering methods are necessary for produce comprehensible results; however, the user has to fix these different parameters following the study needs, and to stay careful during the interpretations.

4.2.3. Simplified model

The simplified model was generated following the fixed keywords. The latter have been chosen because the aim of this paper was to identify sustainable orientations and strategies of the PADDs. It returns a sufficient result for the study needs.

However, the 0.5 fixed threshold for the attribution of a keyword can be considered as low. Furthermore, it is possible that other main notions exist in this document: consequently, the terms may present most important correlations with them than the fixed keywords. A well approach can be to determinate a second keywords level depending on the sustainable development pillars used. The comparison of these results with the actual clusters and topics composition may be an interesting approach to validate the previous result.

5. Conclusion

This article deals with urban planning sustainable orientations and strategies of Reunion Island municipalities. In order to identify them, the text mining approach is used to extract them from PADD of the municipalities. The developed method is progressive and exploratory, in order to extract the researched information from the corpus. It includes general processing, then the same steps on a filtered corpus and finally the comparison of the results with Eco-PLU, to determinate the evolution of these strategies.

All of these results highlight that the main sustainable planning orientations of the PADDs are closed of classical objective of the sustainable city, with a particular attention for protection questions, and environmental and cultural aspects. Concerning the strategies, the result is the same : they seem to adopt the same solutions that are proposed by European models. Nevertheless, two main approaches appears : the main group of municipalities promotes a development strategy, with planning project and long-term vision; while the second is focus on an operational urban planning, with a short-term strategy.

On the other hand, the island PADDs seem to be similar from one to another by their composition. On the other hand, these documents trend to be specific to their territory, as proved by the topic modeling analyzes. Hence, the coexistence of three levels of urban planning documents coexisting for this small territory is discussed, because the influence of the SCoTs' orientations and strategies are not revealed throw these results.

Emerging of the Eco-PLU may be an interesting tool to enhance the sustainable urban planning of the municipalities, but it has to remain coherent with the territories' needs. Further similar works on existing Eco-PLUs' PADDs may permit identifying their main orientations and strategies clearly, and their difference with actual PADDs.

The developed methods for this study permit providing first answers concerning the orientations and strategies on Reunion island strategies. Nevertheless, the results can by improved by using a hybrid method of tokenization, that mix ngrams and single terms. First tests with this approach highlight interesting first results, as shown by the following Figure 13.

Furthermore, considering the big data capacities, further studies could be realized with the same text mining approach on other PADDs of French municipalities, and compared with the SCoTs' PADDs of the country.

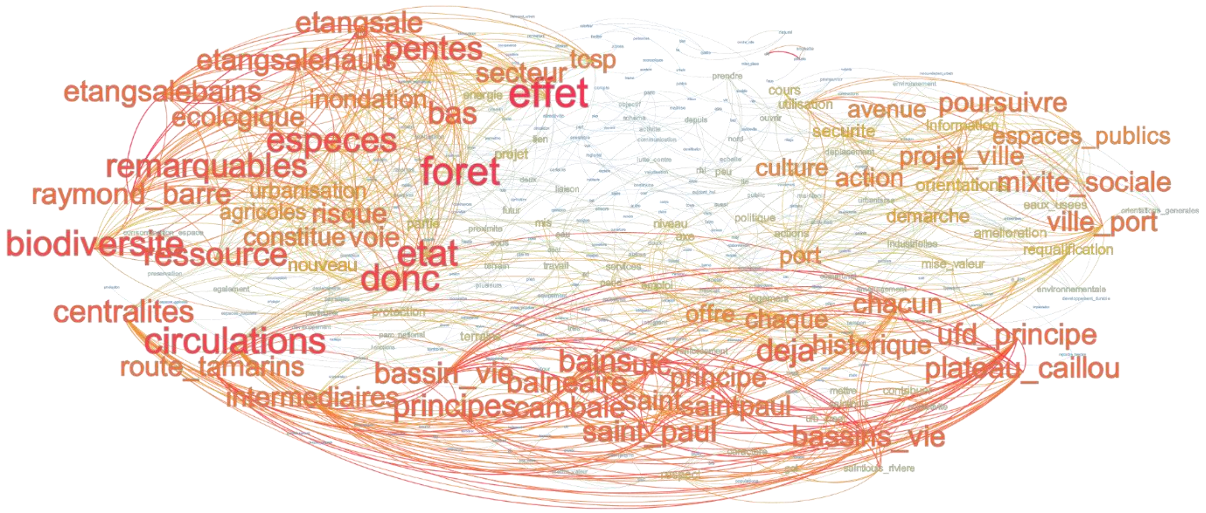

Figure 13:Correlation map of hybrid tokenization (Eve ETIENNE, 2019)

References

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer Publishing Company, Incorporated.

Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In *Mining text data* (p. 77–128). Springer.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *ArXiv:1707.02919 [Cs]*. Consulté à l'adresse http://arxiv.org/abs/1707.02919

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. *Third international AAAI conference on weblogs and social media*.

Bénard-Sora, F., & Praene, J. P. (2018). Sustainable urban planning for a successful energy transition on Reunion Island: From policy intentions to practical achievement. *Utilities Policy*, *55*, 1-13. https://doi.org/10.1016/j.jup.2018.08.007

Bholowalia, P., & Kumar, A. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, *105*(9), 17-24.

Boquet, Y. (2010). From PLU to Eco-PLU : Strategies for a sustainable city in Dijon, France. *The 1st International Conference on Sustainable Urbanization (ICSU)*, 1043-1051. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-00908992

Boquet, Y. (2014). Transportation and Sustainable Development in a Mid-Size French City, Dijon. *International Review for Spatial Planning and Sustainable Development*, *2*(2), 52-63. https://doi.org/10.14246/irspsd.2.2_52

Desjardins, X., & Leroux, B. (2007). Les schémas de cohérence territoriale : Des recettes du développement durable au bricolage territorial. *Flux*, *n° 69*(3), 6-20.

Douglas, C. H. (2006). Small island states and territories: Sustainable development issues and strategies – challenges for changing islands in a changing world. *Sustainable Development*, *14*(2), 75-80. https://doi.org/10.1002/sd.297

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Feinerer, I., Hornik, K., & Meyer, D. (2008). *Text Mining Infrastructure in R.* https://doi.org/10.18637/jss.v025.i05

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(1), 1-30. https://doi.org/10.18637/jss.v040.i13

Hassler, M., & Fliedl, G. (2006). Text preparation through extended tokenization. *Data Mining VII: Data, Text and Web Mining and Their Business Applications*, *1*, 13-21. https://doi.org/10.2495/DATA060021

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651-666. https://doi.org/10.1016/j.patrec.2009.09.011

Jivani, A. G. (2013). *A comparative study of text data mining algorithms and its applications* (PhD Thesis).

Leroux, E. (2012). Le SCOT : Un outil de Management public territorial au service du développement durable des territoires ? *Gestion et management public*, *Volume 1/n°1*(1), 38-52.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2019). *cluster: Cluster Analysis Basics and Extensions* [R package version 2.1.0].

Moscarelli, F. (2013). *Schéma de Cohérence Territoriale (SCOT) et développement durable en France : Enseignements à partir des cas grenoblois et montpelliérain*. Consulté à l'adresse https://tel.archives-ouvertes.fr/tel-00874429

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, *31*(3), 274-295. https://doi.org/10.1007/s00357-014-9161-z

Nidhi. (2017, mars 31). RPubs - Optimal Number of topics for LDA. Consulté 21 juin 2019, à l'adresse http://www.rpubs.com/MNidhi/NumberoftopicsLDA

Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition*, *37*(3), 487-501. https://doi.org/10.1016/j.patcog.2003.06.005

Racine, J. S. (2012). RStudio: A Platform-Independent IDE for R and Sweave. *Journal of Applied Econometrics*, *27*(1), 167-172. https://doi.org/10.1002/jae.1278

Silge, J., & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open Source Software*, *1*(3), 37. https://doi.org/10.21105/joss.00037

Stratford, E. (2003). Flows and boundaries: Small island discourses and the challenge of sustainability, community and local environments. *Local Environment*, *8*(5), 495-499. https://doi.org/10.1080/1354983032000143653

Tan, A.-H. (1999). Text mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, *8*, 65–70. sn.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation Methods for Topic Models. *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112. https://doi.org/10.1145/1553374.1553515