Master thesis on Interactive Intelligent Systems

Universitat Pompeu Fabra

Aiding decidim.barcelona: Clustering of Proposals

Esther González

Supervisor: Vicenç Gómez

September 2017

Master thesis on Interactive Intelligent Systems

Universitat Pompeu Fabra

Aiding decidim.barcelona: Clustering of Proposals

Esther González

Supervisor: Vicenç Gómez

September 2017

Contents

Abstract

Decidim Barcelona is an online participatory-democracy platform that boosts citizens participation to draw The Municipality Plan. A key step to build this plan is the selection of the proposals that will be included. This requires a manual selection and grouping of proposals that intend to tackle the same issue. The use of machine learning can speed up this process as well as improve user experience suggesting similar proposals, also decreasing the number of duplication. We have experimented several techniques to discover which one better solves the problem of automatically grouping similar proposals.

Keywords: e-democracy; machine learning; text clustering; decidim-barcelona.

Chapter 1

Introduction

The Barcelona City Council developed an online participatory-democracy platform in which citizens and government can create, discuss, comment and support proposals [3]. The goal of this platform is to elaborate the strategic plan of the municipality with the participation of the citizens, collecting proposals from different sources, giving citizens and neighborhoods the chance to speak up about improvements that can be carried out in the city and their neighborhoods.

Including citizens into the process of creation of The Municipality Plan through an e-participation platform is a practice that improve the relation among citizens, residents, and government. That is a common practice in smart cities, promoting cooperation, partnerships and participations [4]. The use of information and communications technologies in such platforms can upgrade the user experience, to allow a better interaction and as a consequence encourage their use.

For the case of Decidim Barcelona, avoid duplicated proposals, and automatically groups similar proposals would increase its effectiveness, user experience and would speed up the plan creation process. To achieve this we have carried severals experiments to discover to what extends proposals can be successfully clustered into semantically similar groups. The use of Machine Learning on similar platforms have been previously study to tackle similar problems in [4], building a recommender system for Decide Madrid platform.

Document clustering has been used widely for text categorization, search result grouping, information extraction. The process of clustering documents have these main steps: First, find a vector representation that correctly encode the semantic meaning of documents. Then, choose a metric to compute the distance between document's representation. Finally, select a clustering algorithm. to find a correct document representation is a critical factor when clustering documents, for the purpose of this study we chose three different representation: Tf-idf [21], LSA [6] and Word2Vec [15].

A problem we will face in this investigation is the length of the proposals. Several studies have used Word2Vec successfully, but the majority of them over documents represented by a hundreds of words [20]. We will use different documents representation that may overcome this limitation.

Motivation

To guide the decision-making of the developing team as well as identify whether a similar proposal already exists.

Challenges

The dataset is comprised of proposals and actuations written in two different languages, 94% in Catalan and 7% in Spanish. This feature makes the platform of decidim unique and, for the purpose of our thesis (clustering of proposals), requires additional steps to define a proper representation and similarity metric. We have analyzed possible ways to deal with this issue. The size of the data did not allow train a Neural Network model, that would probably have led to a better result than using a pre-trained model. Moreover, the length of proposals difficult the clustering process of semantically similar proposals.

To evaluate the quality of clusters, ideally, one should have a ground truth to compare the clustering result. There was not a trustworthy ground truth. The one used was the final result Id of approved proposals which group similar proposal. How-

ever, the groups of proposals found in several results are not always semantically similar, also semantically similar proposals might be across different results and, not all proposals appear only in one result. Consequently, we end up we a result that might give us some idea of correctness, but cannot be totally reliably.

Objectives

Discover to what extent decidim Barcelona proposals can be cluster obtaining good results.

Structure of the Report

Chapter 2 gives a brief explanation of the data used in this project as well as a theoretical introduction to methods and algorithm used, like different text representations, hierarchical clustering and its linkage methods, similarity measures and cluster quality evaluation.

Chapter 3 show and compare the results achieved after applying hierarchical clustering over decidim Barcelona dataset.

Finally, chapter 4 conclude this investigation summarizing what was discover when applying the proposed techniques.

Chapter 2

Background & Methods

The following chapter will give an introduction to some theoretical concepts related to the scope of this project. Also, a brief description of the libraries and tools used.

First, we will briefly describe the Decidim process and the components of its dataset used in this study. Later, the text transformation performed over the proposals text to obtain a cleaner data. Then, an explanation of the three different text representation implemented in this project. Next, a description of the clustering algorithm used and its variants. After, a definition of the similarity measures that were used to compute the similarity between documents. Finally, the metrics used to measure the quality of the final clustering.

Dedicim Barcelona

Decidim Barcelona aim to define The Municipal Plan with the active participation of government and citizens of the city. This plan will guide the actions to be developed during the four years the government last. The plan is composed by a set of proposed improvement to be carried out in the city, these proposals are divided into five categories: Plural Economy, Good Governance, Good Living, Ecological Transition and Global Justice.

The process to draw The Municipal Plan starts with an initial set of proposals

created by The Barcelona City Council. Then, new proposals are created by citizens either face to face or in a through the decidim website, these proposals along with the initial set are discussed, commented and supported. Later, the City Council study all the proposals, new proposals that count with a high level of support are added to the plan of the municipality, and similar proposals are grouped to draw up the final set of proposals that represents a collaborative version of the Municipal Plan. Finally, this version will be evaluated by the plenary of the Municipal Council which will approved proposals, The approved similar proposals are grouped into actuations (also named results indistinctly). The aim of this project is to automatically group similar proposals.

For this study, we will mainly use the text that correspond to the title, description, category and subcategory of each proposal. Results are used to know what would be a correct clustering output, each result has and Id and a list of proposals Ids, the result Id is used as the ground truth for its list of proposals.

Data Preprocessing and Translation

The original data was a mixture of Catalan and Spanish proposals and actuations.

Firstly, the text corresponding to the title and description of each proposa that was originally in either, Catalan or Spanish, was translated using Google Cloud Translate API [1], all proposals were translated to English using the auto detection source language of the API.

Secondly, stop words along with punctuation marks were removed to reduce noise and to avoid dealing with words that would not contribute to identifying similarity between words, moreover, words that appear only once in a proposal were also removed, this process was done using python NLTK implementation.

Lastly, to normalize words representation, words were stemmed in order to obtain their base form, the stemming algorithm used was Porter's Stemmer from NLTK package.

[1] `https://cloud.google.com/translate/`

Data Representation

Documents need to be represented in such a way that serves as an input to the target algorithm reflecting the document's content in a way that similar pair of documents can be identified. In this project we will focus on three representations, one simple model that works best to identify word overlapping, but that serve as input to the second method that is able to capture semantic similarity from a very simple structure, ultimately, a representation that has shown to capture semantic similarity successfully but that requires big corpus to achieve its best result.

The two most common and simple ways to represent a document is by building a Bag of Words (BOW)[19] or by their term frequency-inverse document frequency (TF-IDF) [21], both representations do not capture semantic information from documents. Latent Semantic Analysis (LSA) [6] tries to overcome this loss of information using singular value decomposition over the BOW or TF-IDF representation to arrive at a semantic feature space, where words with similar meanings are located close to each other, similarly, Word2Vec [15] capture semantical meaning from text, assuming that words with similar context might have similar meanings. Word2Vec produces word embedding that beats any previous model [15][11], however, unlike LSA, requires a corpus over the billions to achieve good results, and its quality is tied to the size of the dataset.

TF-IDF

Term frequency-inverse document frequency (TF-IDF) is a method that reflects the importance of a term to a document in a collection of documents. Term frequency (TF) assigns a weight to each term in a document equals to the number of times that a term t appears in document d. A drawback of TF is that considers all terms equally important, a term like *"the"* will occur in almost all document and, TF will assign a high weight. However, a term that occurs very often is usually not descriptive for a document [14]. Inverse document frequency (IDF) correct this penalizing those terms that occur too often across all documents of a collection.

TF-IDF built a $M \times N$ term-document matrix with the resulting TF-IDF value for all N terms in a collection of M documents. This result in a sparse matrix that can be used to compute similarity measures between each document.

Formally, term frequency-inverse document frequency of a term t in a document d is given by:

$$tfidf_{t,d} = tf_{t,d} \cdot idf_{t,D},$$

where $tf_{t,d}$ is the number of occurrences of the term t in document d and, $idf_{t,D}$ is the inverse document frequency. The value of $tfidf_{t,d}$ will be low when a term occurs in many documents or occurs a few times in a document, it will decrease when the quantity of documents it appears increase. When a term appears in several documents $tfidf_{t,d}$ will have a high value[14].

Inverse document frequency idf_t of a term t in a collection D is given by:

$$idf_{t,D} = \log \frac{N}{df_{t,D}},$$

where N is the number of documents in a collection, and $df_{t,D}$ is the number of documents in a collection D that contain term t.

This algorithm has many limitations. It cannot identify whether a word is present with small changes, if the word *play* and *playing* are present, they would have different entries in the resulting vector. TF-IDF is not able to capture the semantic relation of words, it relies on words overlapping to find similarities. Moreover, Tf-Idf outputs a very sparsely matrix, what produce documents vectors of a very high dimensionality. We will address this using a dimensionality reduction technique named LSA.

In this project, the term-document matrix was built for the collection of proposals, where each proposal was the text of its title concatenated with its description. Python library scikit-learn [18] was used for this propose.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method for dimensionality reduction that takes as input the term-document matrix built using TF-IDF [6]. LSA finds a projection with the maximal variance of data along axes. The idea behind LSA is that terms that appear together in a collection of documents will have a similar projection on the new vector space. Documents that were not similar by TF-IDF might be similar in a LSA representation. LSA and Principal Component Analysis (PCA) are both methods for dimensionality reduction. Their differ in that PCA uses SVD over a covariance matrix, while LSA uses SVD over a term-frequency matrix.

LSA uses Singular Value Decomposition (SVD) to decompose a term-document matrix M into three matrices $M = U\Sigma_k V^\top$, where U and V are column-orthogonal matrices and Σ_k is a diagonal $k \times k$ matrix that contains k singular values of M, such that the singular values are in descending order. Finally, a $k' \ll k$ is chosen to obtain a reconstructed M' matrix from multiplying $U\Sigma'_k V^T$ [20].

The implementation of LSA over the set of documents (proposals) of decidim.barcelona was carried out using Python library scikit-learn [18]. Then, cosine distance was used to build a distance matrix that served as input to the hierarchical clustering process.

Word2Vec

Word2Vec is an extensively used word embedding algorithm that convert words into vectors that capture semantical meaning [5]. Two neural network models (Continuous Bag of Words (CBOW) and Skip-gram) that outperformed all previous architecture were introduced in [15]. The main assumption of these models is that words with similar context have a similar meaning. Word2Vec defines a window of size k, also called context, and scans the corpus training the model for each word keeping k words around the target word. CBOW and Skip-gram differ in how this context is used.

CBOW predicts w_i given the window of surrounding words. It is composed of three layers (see Figure 1): the input layer corresponds to the context of the target word,

the hidden layer projects each word from the input layer into the weight matrix, this matrix has one row for every word in the vocabulary of the training corpus and as much columns as features (300 features for the model used in this study). Finally, the third layer is the output layer, that for this model will output the target word for the given context.

Skip-gram predicts the surrounding words given the center word w_i. It is also composed of three layers (see Figure 1), where the input layer is the target word w_i while the output layer is its context.

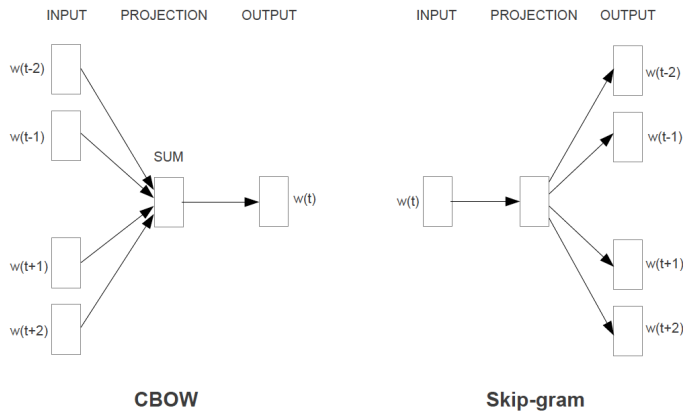


Figure 1: The CBOW model predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word [15].

Word2Vec will produce a vector representation for words present in its training set. To extend this representation to a document or sentence scope one can take the element-wise sum or mean of the word embeddings across all the words in the document, this will produce a vector that will encode the semantical meaning of the document. This approach has been used successfully in [16] [12].

Word2Vec typically requires a large training dataset to achieve good results (over billion words in[15]). Since the dataset for this project is too limited in that sense, we used a pre-trained word2vec model [13] [11].

In this project, we used the pre-trained model from Google, which has 300 feature word vectors for 3 million words and it was trained on 100 billion words Google

News dataset [2].

Clustering Algorithm

Clustering is a form of unsupervised learning that divides a dataset into different groups or clusters. It aims to build a distinct cluster where members of each cluster are similar between them and consequently, members of distinct clusters are different. Clustering algorithms don't have labeled categories where to fit the data, as classification algorithms which are a form of supervised learning. A key input in any clustering algorithm is the distance measure selection. This measure will determine which element belongs to each cluster, different distance measures will produce different clusters[14].

There are two main categories of clustering methods: partitional clustering and hierarchical clustering, the last was the one chosen for the propose of this study.

Hierarchical Clustering

Hierarchical clustering builds a tree-like hierarchical structure called dendrogram (see Figure 2) that offers a visualization of the obtain clustering at different scales.

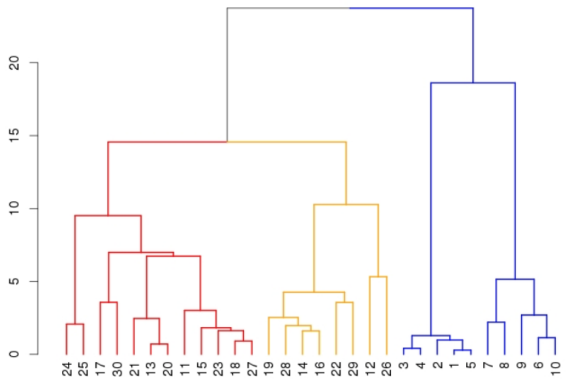


Figure 2: Example of a dendrogram

Hierarchical clustering algorithms can be split into two categories: Agglomerative, where all leaves start being a cluster and they are merged from bottom to top,

[2] `https://code.google.com/archive/p/word2vec/`

and divisive, where all leaves belong to a single cluster, and the clustering is created splitting recursively from top to bottom[1]. In this project, we will focus on agglomerative hierarchical clustering, where leaves represent proposal.

Agglomerative clustering on decidim's proposals initiate merging the two closest single proposals forming a new node that contains two proposals, this node is then merged to its closest single proposal or node, this process is repeated across all single proposals and nodes until there is one single node that contains all previously created nodes [17]. In agglomerative clustering, the fundamental step is the selection of two clusters that are merged to create a new cluster.

There are different methods to compute the similarity or distance between nodes; called linkage. These methods depend on a similarity measure 2.5, by default in Python Scipy library [10] *dist* is defined as the Euclidean distance 2.5.1. Below are the linkage methods used in this study:

Complete: The similarity of two clusters u and v is defined as the similarity of the two most dissimilar members of each cluster[17].

$$d(u, v) = \max(dist(u[i], v[j])),$$

for all members i in cluster v and j in cluster u.

Average: The similarity of two clusters is defined as the average similarity between all pairs of members [17].

$$d(u, v) = \sum_{i,j} \frac{dist(u[i], v[j])}{|u| \cdot |v|},$$

for all members i in cluster v with cardinality $|v|$ and points j in cluster u with cardinality $|u|$.

Weighted: The similarity of two clusters is defined as the average of the similarity between all pairs of members from different clusters and it is weighted

based on the quantity of members in each cluster[17].

$$d(u, v) = \frac{dist(s, v) + dist(t, v)}{2},$$

where u is a cluster formed with cluster s and t, and v is a remaining cluster. v.

The hierarchical clustering algorithm does not require a pre-defined number of clusters. Instead, the obtained dendrogram can be used to analyze and decide at which height the tree should be cut to generate the final clustering. In Figure 2 the x axis represents the data points and the y axis the distance at which each node was merged. The cutting threshold can be decided visually, or alternatively using a measure computed from the clustering at each level, such as silhouette coefficient [7].

Similarity Measures

Document clustering group a collection of documents into clusters, ideally placing similar documents in the same cluster, and different documents far apart in a different cluster. The goal of clustering is to achieve high intra-clustering similarity, documents in the same cluster are similar; and low inter-cluster similarity, documents from different cluster are dissimilar[14]. An appropriate definition of documents similarity or distance depends on the problem properties[8]. Different measures would lead to completely different clustering. In this project, we focus on two: Euclidean Distance and Cosine Similarity.

Euclidean Distance

Euclidean distance is widely used to determine the level of similarity between vectors, specially in the case of clustering. It is the length of a straight line between two points, by default multiple clustering algorithms use this distance measure. Given two documents d_a and d_b, represented by vectors $\vec{V_a}$ and $\vec{V_b}$, their euclidean distance

is defined as [8]:

$$D_E(\vec{V_a}, \vec{V_b}) = (|V_a - V_b|^2)^{1/2}$$

Cosine Similarity

Cosine similarity is one of the most used measures to determine documents similarity. It measures the angle between the vector representation of two documents. Given two documents d_a and d_b, represented by vectors $\vec{V_a}$ and $\vec{V_b}$, their cosine similarity is defined as [8]:

$$SIM_C(\vec{V_a}, \vec{V_b}) = \cos(\vec{V_a}, \vec{V_b}) = \frac{\vec{V_a}.\vec{V_b}}{|\vec{V_a}| \cdot |\vec{V_b}|}$$

The cosine distance will be used as a metric to compute the distance between documents, it is defined as:

$$D_C(\vec{V_a}, \vec{V_b}) = 1 - SIM_C(\vec{V_a}, \vec{V_b})$$

Cluster Evaluation

Evaluate the quality of a clustering solution is a critical and difficult task. Decide the goodness of clusters is a key step to choose a solution that will group the data successfully. Cluster quality measures can be categorized into two classes, external criterion of quality and internal criterion of quality [14]. They difference is that once required external information; a ground truth, and the other only rely on the data itself.

For the propose of this study we decided to use silhouette coefficient; an internal criterion, and purity and normalized mutual information; both external criterion. One of the main challenges we had was the absence of an accurate ground truth, we had the results (actuations) which group multiple proposals in a single result which tackle similar problems. However, this grouping is far from being accurate, similar proposals are not in an unique group, and also group not related proposals into the same result. However, as this was the closest we had to the expected output, we

used it along with an internal measure.

Silhouette Coefficient

When the ground truth is not available, we have to use an internal criterion to know the clustering quality. Internal criterion evaluates clustering by examining how compact each cluster is, and how separated different clusters are from each one[7].

To measure the silhouette coefficient for a document $d \in D$ we need to measure the minimum average distance from d to all different clusters to which d does not belong to and, the average distance between d and all other documents in the same cluster where d belongs to. [7].

The Silhouette coefficient for a document d in a dataset D of n documents partitioned into k clusters $\{w_1, ..., w_k\}$ is defines by,

$$silhouette(d) = \frac{b(d) - a(d)}{\max\{a(d), b(d)\}}, \tag{2.1}$$

where:

$$b(d) = min_{w_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{d' \in w_j} dist(d, d')}{|w_j|} \right\}, \tag{2.2}$$

and,

$$a(d) = \frac{\sum_{d' \in w_i, d \neq d'} dist(d, d')}{|w_i| - 1} \tag{2.3}$$

The value of the silhouette coefficient (2.1) goes from -1 to 1. The value of $a(d)$ (2.3) represents the compactness of the cluster, a small $a(d)$ represent a compact cluster. While the value of $b(d)$ (2.2) capture how separated a document d is to another clusters, the larger is $b(d)$ the more distant d is to another clusters. A silhouette coefficient close to 1 means that the cluster containing d is compact and d is far from other clusters. When the silhouette coefficient is negative ($b(d) < a(d)$) means that d is closer to a documents in a different cluster than to documents in the same cluster. The desired result is a silhouette coefficient close to 1 [7].

To measure the cluster quality, first, the silhouette coefficient is computed for all documents in the dataset. Then, the silhouette coefficient of a clustering result is the average of the silhouette coefficient of all documents in the dataset.

Purity

The purity is a simple external criterion that rely on the ground truth to measure the goodness of the clustering result. The purity measure if a cluster contains documents of a single class, the majority of documents determine the correct class of a cluster. The purity is defined by:

$$purity(\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |w_k \cap c_j|$$

, where $\omega = \{w_1, ..., w_k\}$ is the set of resulting cluster and $C = c_1, ..., c_j$ is the set of classes [14].

The purity value goes from 0 to 1. A bad clustering according to the ground truth would have a value near to 0, while a good clustering would have a purity near to 1 [14]. However, good purity is easy to achieve when the number of clusters is large, for example if each cluster has only one element the purity would be 1. For this reason, purity by itself is not a fair measure of the quality of a cluster, but it can give some idea of the quality of a solution when used along other metrics.

Normalized Mutual Information

Normalized Mutual Information (NMI) is an external criterion that has it origin in Information Theory and is based in the concept of entropy. The Entropy of k clusters in Ω is defined by [14]:

$$H(\Omega) = -\sum_{k} P(w_k) log(P(w_k)),$$

where $P(w_k)$ is the probability of a document of belonging to cluster w_k.

The entropy of a cluster Ω measure the level of uncertainty that can be reduced when a random document w is taken from Ω. The concept of entropy can be extended to Mutual Information, which measures how much information can be gained from the actual class of an element when its cluster is known. Mutual Information is defined by [14]:

$$I(\Omega, C) = \sum_k \sum_j P(w_k \cap c_j) log \frac{P(w_k \cap c_j)}{P(w_k)P(c_j)},$$

where $P(w_k), P(c_j)$, and $P(w_k \cap c_j)$ are the probabilities of a document belonging to cluster w_k, class c_j and to the intersection of w_k and c_j. The value of $I(\Omega, C)$ is 0 if knowing the cluster of a document give no information about its actual class. The highest values for $I(\Omega, C)$ is 1, that is the case when Ω define a perfect cluster, that match the ground truth, but, a maximum mutual information can be achieved when the number of clusters is equal to the number of documents. Mutual information presents the same problem that purity. Normalized mutual information penalize high cardinalities, and is defined by [14]:

$$NMI(\Omega, C) = \frac{I(\Omega, C)}{[H(\Omega) + H(C)]/2},$$

However, this penalty does not avoid NMI to favor clustering with a high number of clusters [2].

Adjusted Rand Index

Adjusted Rand Index (ARI) is another external criterion that measures the congruence between a clustering result and a ground truth. ARI is an improvement to Rand Index, which lead to high values even when there is a high level of disagreement between clusters and actual classes [9]. ARI is defined by:

$$\frac{\sum_{i,j} \binom{n_{i,j}}{2} - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right] / \binom{S}{2}}{1/2 \left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right] / \binom{S}{2}},$$

where $n_{i,j}$ is the number of document of class c_i present in a cluster w_j, n_i is the number of documents in class c_i and, n_j is the number of document in the cluster w_j.ARI values go from -1 to 1, being 1 the best score and -1 being worst than a random result [20].

Chapter 3

Results

In this chapter, we will show the results obtained from several tests using hierarchical clustering over Decidim Barcelona proposals data.

First, we describe Decidim Barcelona proposals to understand the challenge and limitation for clustering this data. Then, we present two sets of results: a positive result using proposals from very different categories and a different ground truth, and a negative result using proposals from one category. We decided to start with a controlled set of clearly different proposals. Then, we carried the same experiments over a set of proposals from one category that could represent the whole dataset more properly.

Dataset Details

In our experiments, we used proposals represented by their title and description in English to perform several tests using hierarchical clustering with different linkage methods and distance measures. Only those proposals that were accepted and as a consequence are grouped in a result with more than one proposal were selected for these experiments.

Table 1 shows how the proposals are distributed in the different categories, together with the initial number of proposals, the number of accepted proposals and the

quantity of proposals in Catalan and Spanish.

Dataset Categories	Total	#Accepted	#Accepted	
			Catalan	Spanish
Global Justice	62	45	44	1
Ecological Transition	3675	2710	2474	228
Plural Economy	1368	1140	1110	25
Good Living	4867	4010	3823	178
Good Governance	888	712	695	17

Table 1: Dataset description.

An advantage of using hierarchical clustering is that clustering at different scales can be analyzed. The output dendrogram illustrates the clustering process and allows to visually decide at which height cut it to obtain a sensible clustering. However, this feature can be feasible only for small datasets, since more than a hundred of samples will make a dendrogram unreadable. In our case, where we have more than a thousands of proposals, the output from a clustering in the whole dataset was not useful for the case of our study (see Figure 3). To overcome this limitation and study the proposed methods in decidim.barcelona we decided to split the dataset in smalls groups, and run our experiments over those groups. The results presented in this chapter will be limited to two groups, one where the clustering performs poorly for the given ground truth, and other where results were good but using a more reliable ground truth.

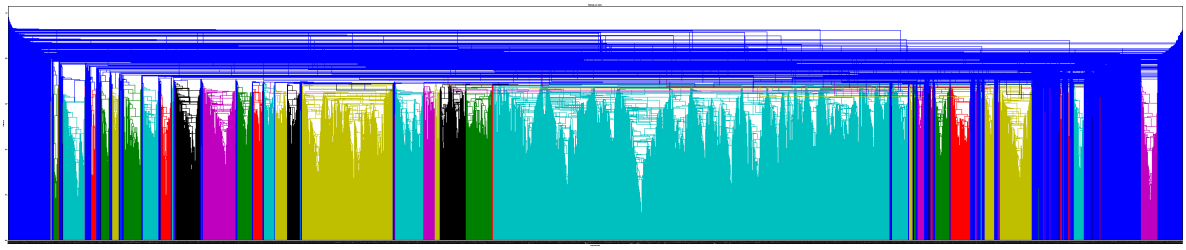
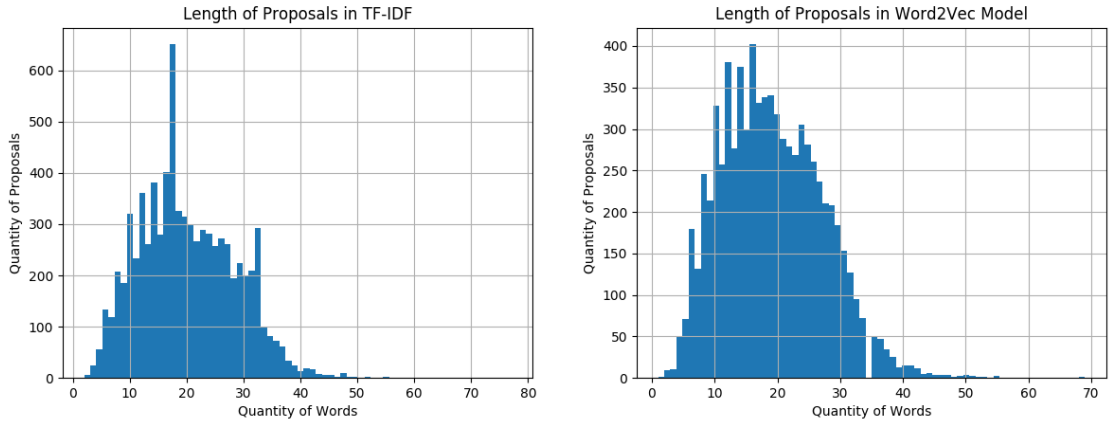


Figure 3: Dendrogram of all proposals.

An important feature of our dataset was that besides being small, the text was short. As Figure 4 shows, the majority of proposals have a length between 10 to 20 words per proposal (title + description), which is a clear limitation for our analysis. Only few proposals have more than 30 words, this adds a challenge to the process of clustering similar proposals as it is more difficult to find similarities in short text

as the co-occurrence of words decreased significantly when comparing to texts of hundreds of words.



(a) (b)

Figure 4: Distribution of the number of words per proposal for two different representations. Most of the proposals have around 20 words, which is a limitation for our study.

Positive Results

To achieve a result where external criterion output good values we had to change our approach. We decided to use a set of very different proposals and to use subcategories as our ground truth. We chose some proposals from very different subcategories to see which document representation gave the best result. The subcategories and distribution of proposals can be seen in table 2. The content of these proposals is shown in table 4.

Subcategories	Quantity
Culture	7
Health	15
Host City	7
Sports	12
Transparent government	13

Table 2: Dataset description.

In table 5 we can see the best quality measures obtained from the clustering of the set of proposals from different subcategories using three different data representation; Tf-Idf, LSA and Word2Vec. We used different linkage methods, but only those

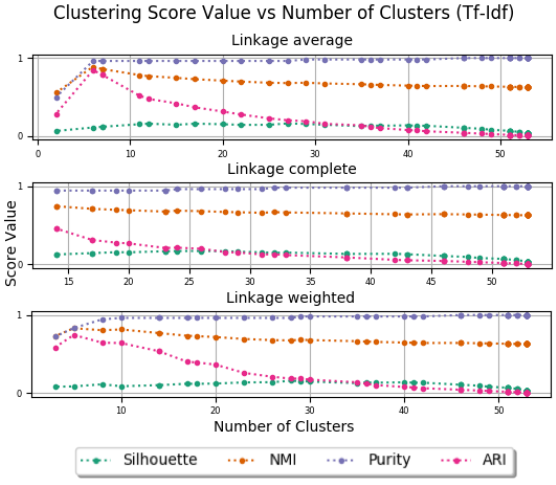
that showed best results for at least one quality measure are shown in this section. Similarly, even when Euclidean distance was used for the experiments of this study, its results were in general worse than those obtained when using cosine distance. Therefore, we will limit the results of this chapter to those using cosine distance.

In table 3 we can notice that in general the three document representation used achieve good results. The main difference between this dataset and the global justice's dataset is that the classes in this one are highly different from one class to another making easier to divide proposals correctly.

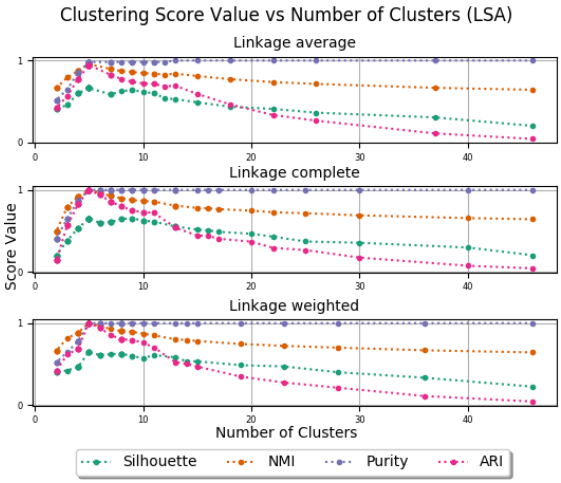
Representation	Distance	Linkage	Cut Threshold	#Clusters	NMI	Purity	ARI	Silhouette
Tf-Idf	Cosine	Complete	0.0200	53	0.6279	**1.0000**	0.0055	0.0370
		Average	0.9400	6	**0.8893**	0.9630	**0.8454**	0.1113
		Complete	0.8000	26	0.6808	0.9630	0.1992	**0.1711**
LSA	Cosine	Complete	0.0200	46	0.6426	**1.0000**	0.0431	0.2026
		Complete	0.7000	5	**1.0000**	1.0000	**1.0000**	0.6492
		Average	0.3600	5	0.9574	0.9815	0.9438	**0.6638**
Word2Vec	Cosine	Average	0.4400	9	**0.8972**	1.0000	0.8105	0.1927
		Complete	0.0200	53	0.6279	**1.0000**	0.0055	0.0370
		Average	0.4400	9	0.8972	1.0000	**0.8105**	0.1927
		Average	0.5600	2	0.1029	0.2963	0.0086	**0.2131**

Table 3: Quality measures for subset of proposals from different categories.

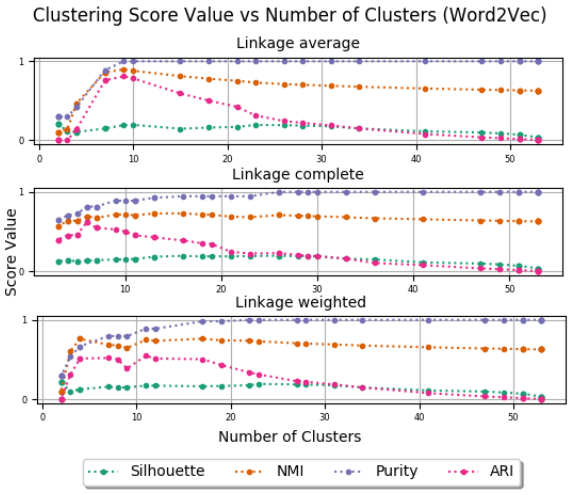
Figure 5 shows the performance measured using purity, NMI, ARI and silhouette coefficient accross different number of clusters. For the three document representations is notable that at certain point (around 10 clusters) the highest score is achieved. In table 3 and figure 5 can be observed that NMI and ARI are maximized at the same number of cluster (actually implies the same cut threshold). This led us to think that combining more than one quality score can be more informative and helpful when deciding which is the best solution.



(a) Tf-Idf.



(b) LSA.

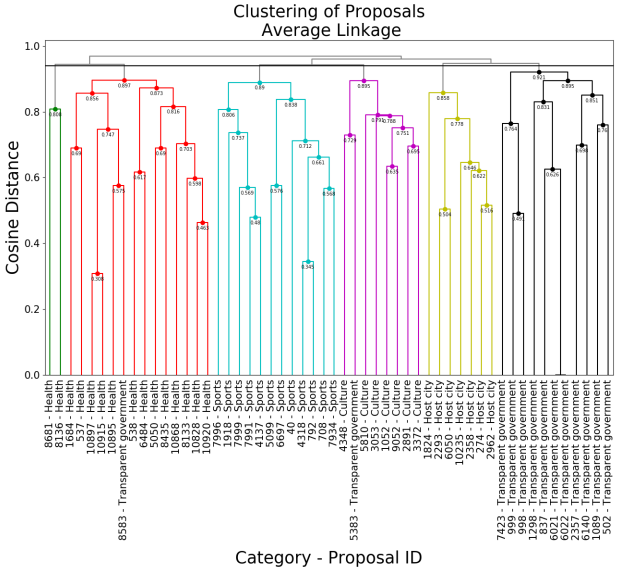


(c) Word2Vec.

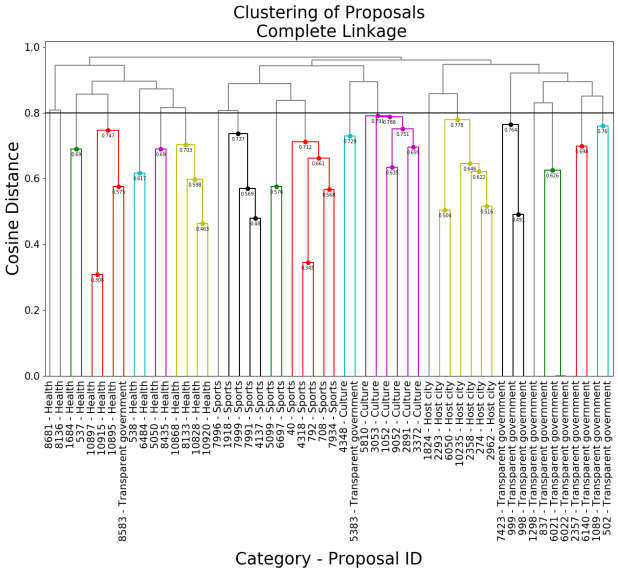
Figure 5: Quality measure results for Tf-Idf, LSA and Word2Vec of subset of proposals.

Tf-Idf Results

In figure 6 we can see the dendrograms corresponding to the highest ARI and silhouette coefficient when using Tf-Idf over the proposals from different subcategories 2. In these dendrograms we can observe that Tf-Idf success grouping proposals from same subcategories, this mainly product of the clearly different vocabulary used in each of these subcategories (see table 4).



(a) Clustering with highest ARI.

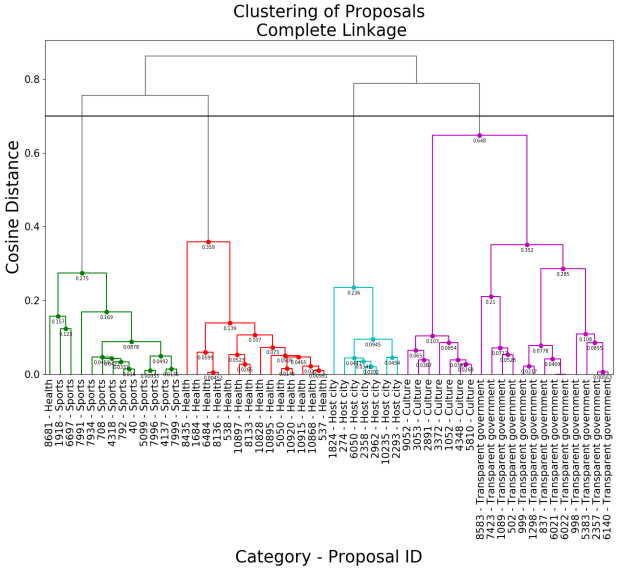


(b) Clustering with highest Silhouette.

Figure 6: Dendrogram with highest quality score of proposals from different categories (Tf-Idf).

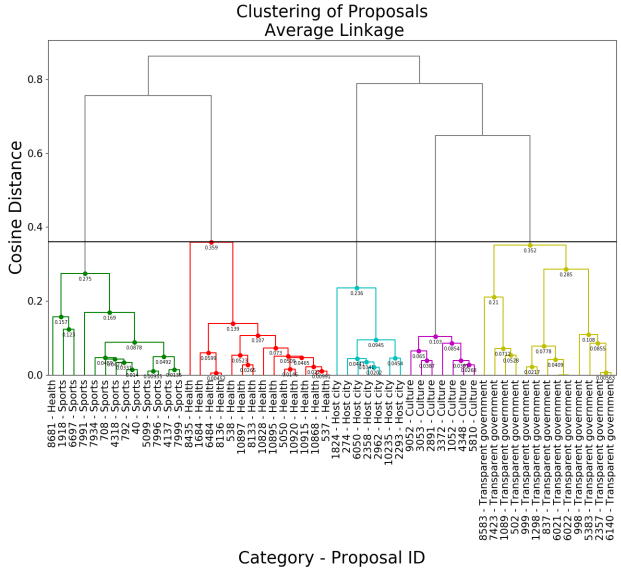
LSA Results

LSA gave of the best result from all the experiments carried over this dataset. We justify this by the difference of vocabulary between proposals of different subcategories; a property successfully caught by Tf-Idf. LSA improved Tf-Idf results by keeping more valuable features, removing all the noise captured by Tf-Idf. This can be clearly observed in the two dendrograms in figure 7.



(a) Clustering with highest ARI.

In figure 7a we see that the clusters obtained using LSA have a low intra-cluster distance and a high inter-cluster distance, what is a main goal when clustering data. This shows that LSA successfully groups this set of proposals.

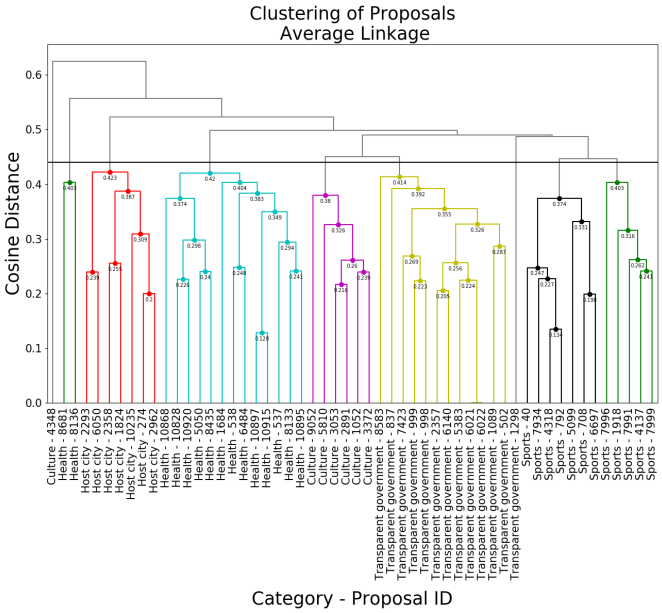


(b) Clustering with highest Silhouette.

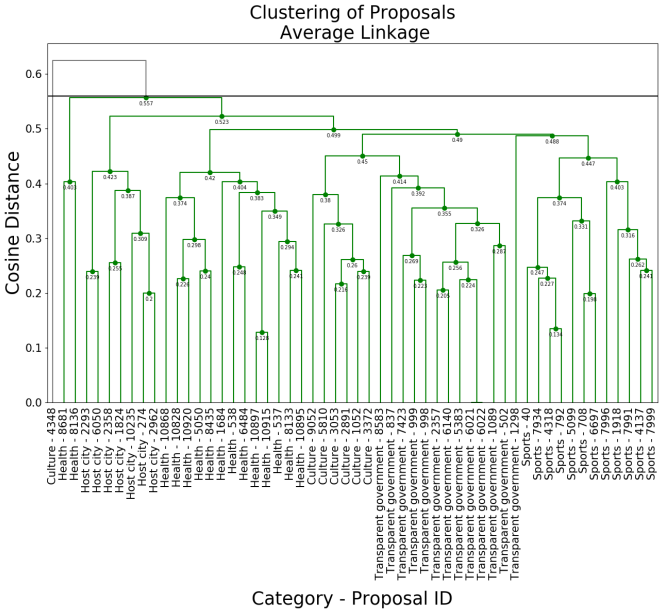
Figure 7: Dendrogram with highest quality score of proposals from different categories (LSA).

Word2Vec Results

Given the particularity of this dataset, already shown by Tf-Idf, it would have been surprisingly to find an extremely bad result in Word2Vec. We can see in figure 8 that Word2Vec is able to split the proposals into their subcategories, with only a few mistakes. Similarly to Tf-Idf, the clusters defined by Word2Vec are not compact, and the inter-cluster distance is relatively small, especially when compared to LSA results. This is caught by the best result of the silhouette coefficient, that given the high intra-cluster distance and the low inter-cluster distance found the best cut at 0.56 where the highest gap inter-clusters is reached.



(a) Clustering with highest ARI.



(b) Clustering with highest Silhouette.

Figure 8: Dendrograms with highest quality score of proposals from different categories (Word2Vec).

Title	Description	Proposal	Category
Visibilize the cultural activities of the district in the city circuit	Promote proposals and cultural productions that take place in Nou Barris and propose one of these projects in the city to enter the cultural map of Barcelona.	1052	Culture
Open the exhibition spaces of the district to the entities and the neighborhood of the territory	Generate competition bases to open the four exhibition halls of the city to all kinds of proposals, to turn them into spaces to promote popular and cultural exhibits, as well as personal artistic proposals.	3053	Culture
Promote the participation of the families of students in cultural activities	Promote the participation of the families of students in cultural activities.	9052	Culture
Commercial axis as a vehicle for cultural dissemination	Commercial axis: entities and stores as vehicles of cultural dissemination The dissemination can be done through the shops	4348	Culture
Coordination of the networking of large cultural facilities and those with entities, communities and collectives, both in the cultural and academic worlds	Encourage networking and new collaborative projects between libraries, civic centers, cultural entities, large museums or research and creation spaces to ensure the maximum circulation of contents and artistic practices.	2891	Culture
Coordination of the city's culture	Create a brand that coordinates the entire cultural offer in Barcelona, both from public and private entities.	3372	Culture
The libraries of the District as cultural agents of knowledge of the territory.	Activities organized by the libraries and other entities to make known the Heritage and the History of the District to the users.	5810	Culture
Day We create health to reduce inequalities: Go to the causes of causes.	To give importance to the diagnoses to go to the causes of the causes.	10897	Health
Mental health improvement projects	Encourage projects of entities in the sector to improve the mental health of the population.	538	Health
Strengthen the network of Mental Health for children and adolescents	Strengthen the network of Mental Health for children and adolescents to support situations of crisis and consultation in schools and caregivers.To attend more quickly and to prioritize infant at high risk.	6484	Health
Mental Health Plan: Support programs for professionals.	Training and specific resources for professionals in childhood and adolescence.Areas of coordination and work with young and homeless spaces.	10828	Health
Speeches of talks by neighborhoods of health professionals	Make a talk series of health professionals with entities, Neighborhood Associations etc.To improve the information.	8133	Health
Day We create health to reduce inequalities: symbiotic interventions.	Symbiotic interventions where different populations and needs benefit mutually from their strengths, being voluntary or new sources of employment.	10915	Health
Health in Barcelona: Review the patient care model.	To train professionals from the empowerment model.Go from the assistance model to an empowering model.Do not generate employee dependence.	10868	Health
Day We create health to reduce inequalities: More resources are needed so that professionals can do more work and community actions.	For example, more home care, street work, community mediation, etc ...).	10895	Health
Increase medical assistance - Horta Guinardó.	It is proposed to increase the number of doctors to offer a better medical assistance service, especially in general medicine.	8681	Health
Reopen / create the services previously referred to as 'family planning'	The life cycle of women has sexual characteristics specific to the potential and / or reproductive activity: menstruation, pregnancy, menopause are phases, not diseases, which should be able to monitor proximity to all neighborhoods.Women's sexual health centers would be very useful.	1684	Health

Title	Description	Proposal	Category
Training in health workers in functional diversity	Training health workers as gynecologists in functional diversity so that they can break a little with the taboos of this group and be able to accompany them in a more integral way	5050	Health
Ensure the existence of quality public services and specific services for young people, incorporating a gender perspective and respect for affectivosexual diversity, both in primary care centers and in educational centers	Promotion of training from the ASPB to professionals in the field of medicine, especially those that work around the ASSiR in LGBTI issues, gender perspective, sexual diversity Maintain and expand the specialized care services of the young people taking into account gender perspective, gender and sexual diversity within the CAPs.	8435	Health
Training in interculturality for health professionals.	Training in interculturality for health professionals.	10920	Health
Comprehensive community health plan, women, the elderly and the immigrant population	Create a Health Table equipped with your own resources	537	Health
Guarantee the stability of budgets and recovery of health services and outsourced services:	Recover all those sanitary services that have been eliminated due to the cuts and all the services that have been derived to private entities.https://decidim.barcelona/proposals/reduccio-de-les-externalizaciones-i-les-privatizaciones-de-serveis-sanitaris	8136	Health
Incorporate LGBT perspective in the refugee support plan	Incorporate the LGTBI perspective in the refugee support plan and take into account its rights as LGBT people.	10235	Host city
Initiatives with refugees.	creation of shelters, reception.	2293	Host city
Attention to refugees.	Barcelona City Refuge, neighborhood committed to the refugees, platform creation or support group	2358	Host city
Start the "Refuge" plan	The District of the Eixample undertakes to develop the "Refuge" plan and to promote the involvement of entities and groups in the different neighborhoods in this plan.	274	Host city
Start the plan "Barcelona, a city refuge"	Start up an operational and strategic plan for the reception of the city.To create a stable and permanent support structure for refugees and asylums that is complementary to state programs, with their own criteria and in collaboration with entities that work on the subject.Pay special attention to asylum seekers who are left out of the state social support program.Create a city-to-city cooperation plan for the high-density municipalities receiving refugees and migrants.	2962	Host city
Participation in the Social Area Barcelona City Refuge	Participation in the Social Area Barcelona City Refuge	1824	Host city
Create a bank of resources for refugees at the level of the District connected to the Generalitat in two levels (reception and awareness)	The reception of refugees in the city must be planned, work in conjunction with the Generalitat and establish measures to: - The reception (accommodation, support, language, etc.) - Awareness (educational centers and citizenship in general)	6050	Host city
Installation of gymnastics devices in the public parks of La Marina - Zona Franca.	Promote young people's sports practice.	4318	Sports
Comprehensive reform of the Municipal Narcís Hall Camp	Continue with the process of integral reform of the Narcí Sala Municipal Camp and execute phases 2, 3 and 4 during the years 2016, 2017 and 2018 respectively.This process will always count with the participation and supervision of the managing entity, which will also be revamped with provisional facilities while the works affect parts of the stadium.	1918	Sports
Take advantage of the dimensions of the old soccer field in Santander street for sports that require large sizes	and avoid dividing it into small clues, since the opportunity that this great space is giving away is wasted	5099	Sports

Title	Description	Proposal	Category
Coherence in municipal sports facilities with the environmental policies promoted by the City Council.	Be consistent in sports facilities with the policies promoted by the City Council, especially with the selective collection of waste.	7991	Sports
Encourage sports programs for children and young people during the holiday season	Promote sports practice during holiday periods, in order to guarantee the rights to the access and the sporting practice of young people and children, regardless of their conditions and the territory in which they live, and to promote the reconciliation of family life.	792	Sports
Creation of sports and play areas in the public space	Generate new areas of sports and play areas in the public space in order to promote a healthy, active and community life.	708	Sports
Create a sports area for young people in the Free Trade Zone	In the face-to-face event that took place at the IES Montjuïc, the 4th ESO students proposed the creation of a sports area for young people in the Free Trade Zone	7934	Sports
Adapt the maintenance of sports spaces to the volume of users.	In order to guarantee the proper maintenance of the equipment, adapt it to the volume of use of the equipment.Track in a specific and adapted way to each facility.Especially in the summer season.	7996	Sports
Improve and transfer sports spaces	It is valued to promote health from these entities through the availability of these spaces.They put the example of the chromium of the youth house of the courts, which should be improved and give way to the football field	6697	Sports
Children's Sports Festival	Promote the Children's Sports Festival as a showcase for sports activities and the different sports groups of Gràcia.Promote the participation of schools and school sports.	40	Sports
Open more municipal sports facilities.	Open more municipal sports facilities.In the case of Poble Nou, they can only access Can Felipa, who perceive that it is saturated and has limited and inadequate facilities.	4137	Sports
Reduction and adaptation of the quotas for municipal sports facilities.	Reduce access prices to municipal sports facilities.Adapt the quotas to the different groups, make specific quotas for families, large families, retirees, and other groups.Set the hour division into two strips.Universalise the price of the equipment.Flexible quotas to include a minimum service of activities.	7999	Sports
Encourage and increase municipal information	Support the radio of Sants.To make campaigns for the dissemination and awareness of participation.Increase political transparency.More understandable information for citizens, nearby.	5383	Transparent government
To achieve the full transparency and retention of accounts of public management	Recognize and facilitate the work of popular control initiatives, such as citizen observatories against corruption, transparency and good governance and promote a citizen advice that audits the proceedings of the district.	6021	Transparent government
Prepare the District Transparency Plan	Design and launch a plan with information that is neat and easy to consult through the Internet, which allows you to know in detail the budgets, the use made of the public resources and the actions carried out by the Administration and its institutional representatives.	1089	Transparent government
To achieve the full transparency and retention of accounts of public management	Recognize and facilitate the work of popular control initiatives, such as citizen observatories against corruption, transparency and good governance and promote a citizen advice that audits the proceedings of the district.	6022	Transparent government
Change how the City Council operates internally on some fronts linked to the common good: strengthen the existing procommun communities, respecting their autonomy	As a general principle, support communities of the existing procommunication and / or technologies, instead of replacing them with the administration, promoting the consolidation or creation of communities that can be self-managed and allow them to be autonomous, They respect ethical questions and work for the common good.	7423	Transparent government
Improvement in communication channels with the City Council.	Parity work sessions, e-Government, Open Data, Technology and politics, e-voting, information on right access to information, web where citizens can propose projects.	2357	Transparent government
Develop an ethical code for elected officials and management personnel	Develop the code that aims to establish the principles and ethical values and the rules of conduct that must be respected by the elected officials and municipal management personnel in the exercise of their functions, with the main purpose of guaranteeing efficient, complete management and transparent from the municipal administration.	999	Transparent government

Title	Description	Proposal	Category
Enable an ethical mailbox	Enabling an ethical mailbox conceived as a tool for participation so that citizens can safely communicate to the City Council facts and behaviors that are contrary to an ethical management of the Administration.	998	Transparent government
Portal of transparency	Publish the councilor's agenda, as well as its declaration of assets and interests;The declarations of goods and interests of District Counselors will also be published, and the District's estimates and investments will be made public.	837	Transparent government
District communication plan	Write and execute the Communications Plan of the District of the Cortes in order to improve the tools that give logic, coherence, purpose and effectiveness to internal and external communication.	502	Transparent government
Improve the information and accessibility of the websites of the City Council	Facilitate the knowledge of the activities, actions of the city council.Improve transparency.	6140	Transparent government
Improve general communication and transparency on ongoing works and on the street	I still see works every day without knowing what they are doing (eg, a type of orchard street Mare de Deus de las Neus, we have a look at it from home, we do not know after 2 years that it is neither for whom). They still continue to do the works clear poster that announces that something will be done soon, for whom, the final result, the financing.	8583	Transparent government
waiting room for saints	The Citizen Attention Office dl district of Sants has a ludic waiting room for the large number of citizens who are going to take steps.There is no maximum capacity or limit.There are also windows and therefore not natural light.There is air conditioning but when there are so many people in the summer it is not noticeable.It's small and claustrophobic.You need a larger room	1298	Transparent government

Table 4: Set of correctly clustered proposals.

Negative Results

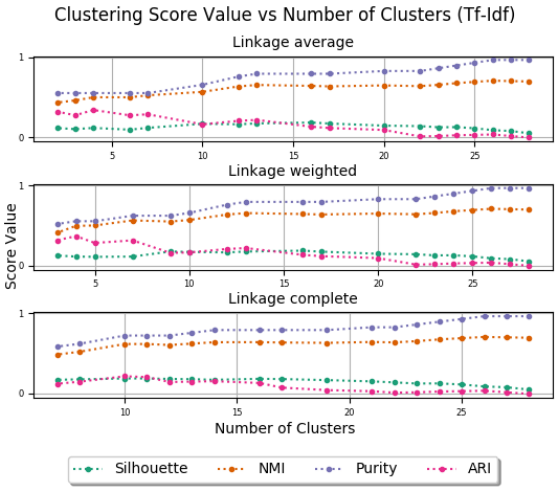
The results presented in this section correspond to Global Justice's proposals. This category has a total of 45 accepted proposals 1. Actuations (also called results) are composed by one to many proposals. From the 45 global justice's proposals we will use only the 29 proposals that are within an actuation of more than one proposal. For details of the content of these proposals look at table 6.

In table 5 we can see the best quality measures obtained from the clustering of global justice's proposals using Tf-Idf, LSA and Word2Vec.

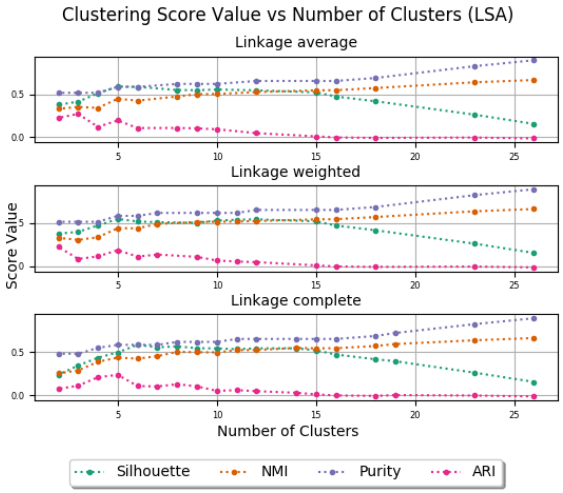
In figure 9 we can see the behavior of the four scores used to measure the quality of our output. We can see how purity and NMI tend to favor those result with a higher number of clusters. For this subset of proposals we found more reliable to use ARI and Silhouette coefficient to decide which would be the best clustering results, as both of them do not favor solutions with high cardinality.

Representation	Distance	Linkage	Cut Threshold	#Clusters	NMI	Purity	ARI	Silhouette
Tf-Idf	Cosine	Complete	0.4606	26	**0.7068**	0.9655	0.0344	0.0880
		Complete	0.2206	28	0.6965	**0.9655**	-0.0049	0.0489
		Average	0.8806	4	0.4991	0.5517	**0.3399**	0.1163
		Complete	0.9006	10	0.6183	0.7241	0.2152	**0.1879**
LSA	Cosine	Complete	0.0371	26	**0.6660**	0.8966	-0.0144	0.1544
		Complete	0.0371	26	0.6660	**0.8966**	-0.0144	0.1544
		Average	0.7171	3	0.3521	0.5172	**0.2711**	0.4137
		Average	0.4371	5	0.4476	0.5862	0.1949	**0.5899**
Word2Vec	Cosine	Complete	0.1577	27	**0.7016**	0.9655	0.0149	0.0498
		Complete	0.0977	28	0.6965	**0.9655**	-0.0049	0.0337
		Average	0.5177	4	0.4367	0.5517	**0.1652**	0.2890
		Average	0.5977	2	0.1123	0.4138	0.0033	**0.3491**

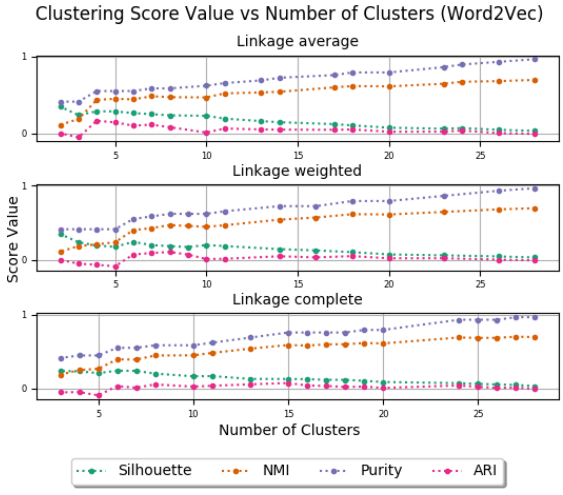
Table 5: Quality measures for Global Justice's Proposals.



(a) Tf-Idf.



(b) LSA.



(c) Word2Vec.

Figure 9: Quality measure results for Tf-Idf, LSA and Word2Vec of Global Justice's Proposals.

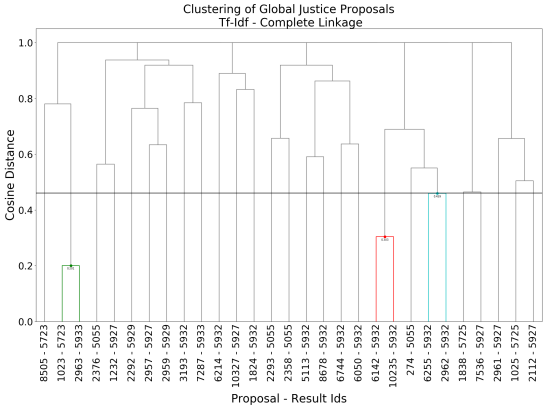
The original number of classes of global justice's proposals is seven. In table 5 we can see that none of the best scores results in this number of clusters. The closest clustering result was the obtained by using LSA + Average Linkage. However, the obtained clustering in this case does not agree with our ground truth, as ARI shows. Nevertheless, the resulting clustering shows a high silhouette coefficient, and giving the fact that our ground truth was not totally reliable, we tend to trust in results with high silhouette coefficient, this shows that the proposals in each cluster are very similar between then and clearly different from proposals in other clusters.

Tf-Idf Results

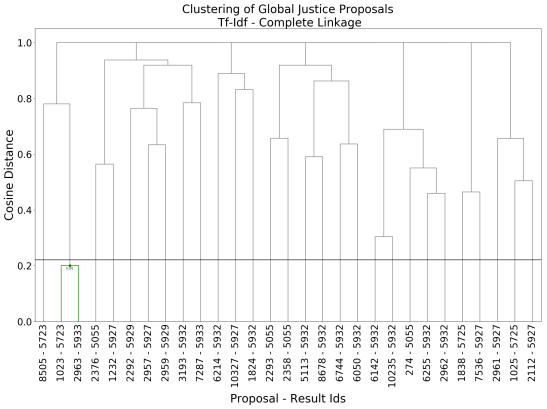
Looking at table 5 we will automatically discards the results where the dendrogram is cut at 0.4606 and 0.2206 (see Figure 10a and 10b respectively), there can be seen that this result show not useful clusters. The solutions with the closest number of clusters to the actual number of classes are the one from using average linkage and cutting the dendrogram at 0.9006. In 10c can be seen 4 clusters, which any of then perfectly matching the ground truth, however, looking at the content of the proposals for each of these clusters it is evident that at least 3 of these 4 clusters have a clear tendency to group proposals that tackle the main topic. From 10c) we can see these topics:

1. Green cluster: Global justice, international network.
2. Red cluster: International cooperation, exchange of practices across cities.
3. Blue cluster: Refugees.
4. Purple cluster: This is the least clear, it mainly talks about specific cities in the Mideast, humanitarian crisis and refugees.

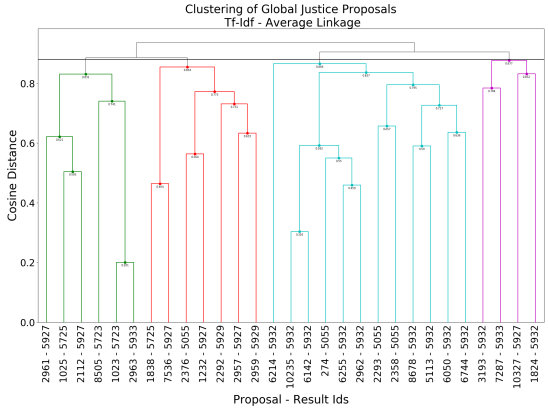
From the above results, we can conclude that, even when the clustering was not perfect, specially compared to the ground truth, in general proposals that have similar words were in the same cluster. This result is what we expected from Tf-Idf, which is good at finding proposals that have word overlapping.



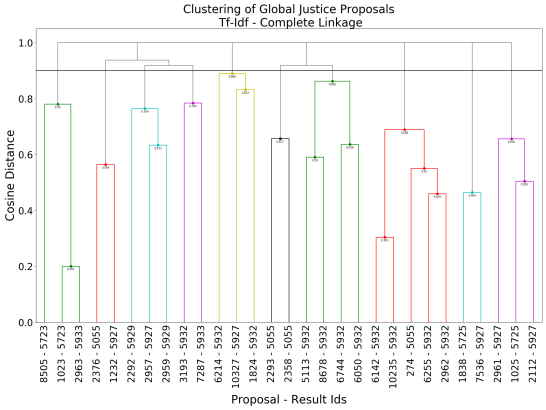
(a) Clustering with highest NMI.



(b) Clustering with highest Purity.



(c) Clustering with highest ARI.

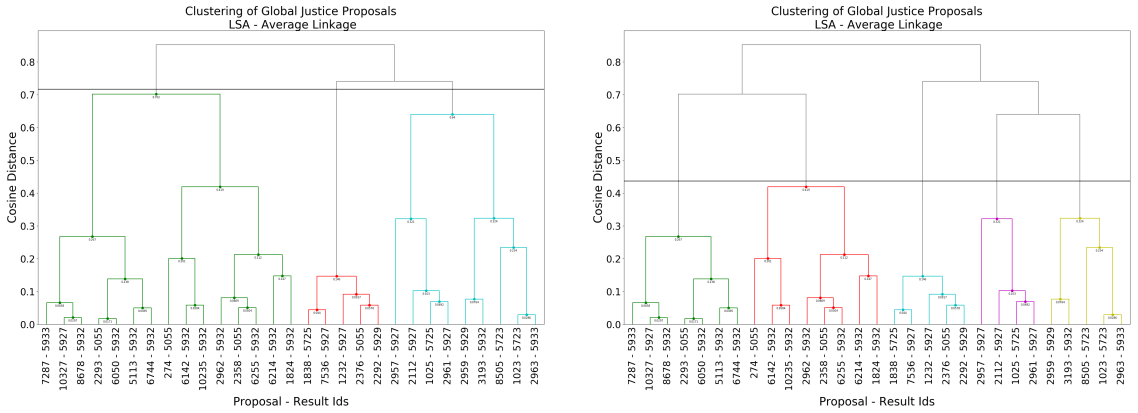


(d) Clustering with highest Silhouette C.

Figure 10: Dendrogram with highest quality score of Global Justice Proposals for Tf-Idf.

LSA Results

In figure 11 we can see the dendrograms for the best scores when using LSA. Dendrograms for the best NMI and purity have been ignored given the little information they provide. In these dendrograms we can see why LSA reached the highest silhouette coefficient in table (see table 5) as this produce very well defined clusters with elements within the same cluster very close between them and far from elements in different clusters . LSA outputs very compact clusters, specially when compare to the output of Tf-Idf. The dimensionality reduction of the proposals' vectors removes noise from the output of Tf-Idf, which produces a sparse matrix.



(a) Clustering with highest ARI. (b) Clustering with highest Silhouette C.

Figure 11: Dendrogram with highest quality score of Global Justice Proposals for LSA.

The highest ARI and Silhouette score was produced by the average linkage. Their differ in the cut height. In figure 11b five clusters were created which is the closest result to the actual number of clusters. From the clusters in 11b we can extract the following main topics per cluster:

1. Green cluster: Refugees.
2. Red cluster: Refugees.
3. Blue cluster: International cooperation, exchange of practices across cities.
4. Purple cluster: Global justice.
5. Yellow cluster: International network, human rights (the word Barcelona is present in all proposals of this cluster).

When looking at the clusters in 11b we could see that the proposals in each cluster have a high degree of relation between them. If we split the dataset content into two main topics it would be very clear to define a group of proposals that speak about refugees and another group about international cooperations. This division would be perfectly recreated by LSA if we cut the dendrogram at 0.8.

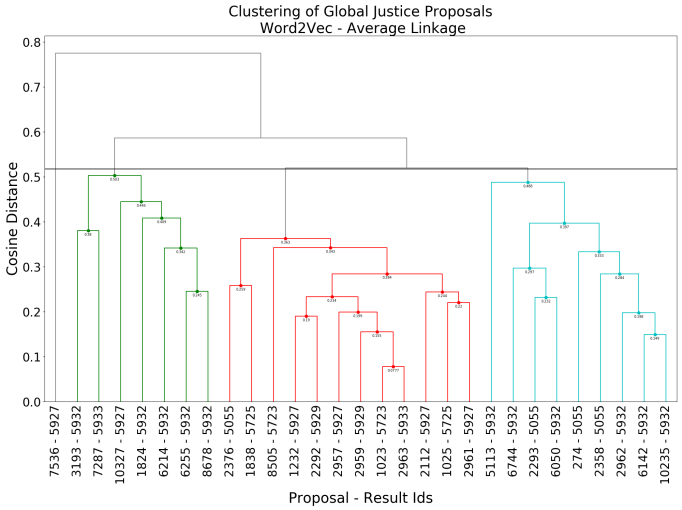
Word2Vec Results

In figure 12 we can see the resulting dendrogram from Word2Vec + Average linkage. In this case the silhouette coefficient is maximized when only two clusters are defined (see figure 12b). It can be noticed that silhouette coefficient tends to cut the dendrogram at the point where the highest inter clustering distance is found.

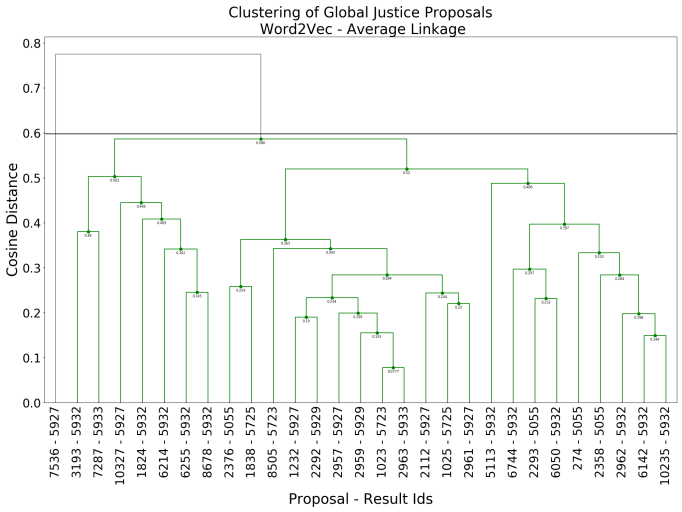
Looking at figure 12a three clusters were defined:

1. Green cluster: International cooperation, exchange of practices across cities, refugees.
2. Red cluster: Global justice, refugee.
3. Blue cluster: International cooperation, human rights, global justice.

We can see that Word2Vec failed to split proposals into groups of related topics. This result was unexpected, given that since its publication in 2013 many researches have used Word2Vec to improve state-of-the-art algorithms in fields like natural language [5]. Certainly, there is a lack of investigation on how effective is Word2Vec capturing the semantical meaning from short sentences [5].



(a) Clustering with highest ARI.



(b) Clustering with highest Silhouette C.

Figure 12: Dendrogram with highest quality score of Global Justice Proposals for Word2Vec.

Title	Description	Proposal	Result
Actions on cooperation.	Promote shared projects with other reference cities to: share experience, visibilize the world that it is possible to do it in a different way with example around the world.	2376	5055, 5927
Start the "Refuge" plan	The District of the Eixample undertakes to develop the "Refuge" plan and to promote the involvement of entities and groups in the different neighborhoods in this plan.	274	5055
Initiatives with refugees.	creation of shelters, reception.	2293	5055, 5932
Attention to refugees.	Barcelona City Refuge, neighborhood committed to the refugees, platform creation or support group	2358	5055, 5932
Participate in international networks to support the rights of migrants	It strives to coordinate and to be part of the international initiatives in defense of the rights of the migrants, as much at institutional level as citizen.	8505	5723
Commit Barcelona to the global fight for human rights	Increase efforts to make public debate, support organizations that work for regional justice and initiate processes of international impact to contribute to the achievement of a system of global governance that protects the human rights of people and peoples in the Euro-Mediterranean region.	1023	5723
Educate for global justice	Progressively increase the weight of education for global justice.We must move towards an identified citizenry with a solidarity Barcelona and committed to global justice, and involved in the different transnational networks that work in this regard from civil society.	1025	5725
Report international cooperation projects	disseminate good practices, horizontally and vertically, to share and improve the actions that are carried out	1838	5725
Insubmission to TTP, TTIP and TISA.	Declare Barcelona as a city insubmised in the treaties of the TTP, TTIP and TISA, to make a campaign for information and dissemination for the commerce of the proximity of this treaty and its involvement in trade.	10327	5927
Renew the spaces and processes of participation and sensitization of the city's agents in matters of global justice	Taking into account the changes suffered during the last years in the set of actors related to global justice, it is advisable to review and search the most effective ways of co-participation between citizens, the sector, public institutions and the City Council.	2961	5927
Exchange of good practices between European cities	Export and import activities and good practices that are working	7536	5927
Develop a global action plan and justice 2016-2020	The outer action plan will outline the guidelines of the city's foreign policy in all its dimensions (governance, relations with other international cities, international networks, economic relations, culture, cooperation, etc.).It is an instrument of definition and planning of the public policy of global justice, which is especially vigilant for the coherence of the external action of the municipal Government.	2957	5927
Reinforce the policies of relations and international cooperation of the city council	Strengthen the relationship and cooperation of the city council with other cities around the world and its external projection in order to learn from the best practices and promote the development and inclusion experiences driven locally.Work in horizontal networks of cities and deepen development cooperation projects.	1232	5927
Transition towards a model of cooperation aimed at global justice	Work on sensitization and solidarity actions.	2112	5927
Measures on cooperation.	Promote a municipal cooperation network to act and move forward with the major issues that concern the State and Europe: the economic model, distribution of wealth, consumer culture, ecology, etc.	2292	5929
Coordinate and strengthen the international networks present in the city	Develop a work strategy that strengthens the principles and values that define the municipal cooperation, based on the fact that in Barcelona there are headquarters and general secretariats of the main international municipal networks, and it is the city that participates in a greater number of international networks, both generalist and thematic (in key areas such as the environment, the fight against climate change, local governance, education, etc.).Guarantee also that this know how to make use of the economic fabric of the city.	2959	5929
Increase support for all refugees	Mitjans school plan, family plan ...	6142	5932
Incorporate LGBT perspective in the refugee support plan	Incorporate the LGTBI perspective in the refugee support plan and take into account its rights as LGBT people.	10235	5932

Title	Description	Proposal	Result
Motion for refugee reception	Join the motion approved by the District Council of the District of Sants-Montjuic by the Senate Sectoral Council.	5113	5932
Create reception spaces for immigrant and refugee population	Often the population that arrives in Barcelona fleeing poverty or conflicts in their countries is relocated.You need to create reception and education spaces where you can meet your first needs.	6744	5932
Help Syrian people	Barcelona should give an example and collaborate with economic funds with pro-active open arms every day where its volunteers play their lives saving people who flee from the war in Syria.It is unfortunate as a cosmopolitan and open city Enacara has not acted in this great humanitarian crisis with concrete facts and easy to carry out	3193	5932
Barcelona city refuge in Sarrià - Sant Gervasi	Support the proposal "start the Barcelona city refugee plan" and let it be set up in this district.	6255	5932
Start the plan "Barcelona, a city refuge"	Start up an operational and strategic plan for the reception of the city.To create a stable and permanent support structure for refugees and asylums that is complementary to state programs, with their own criteria and in collaboration with entities that work on the subject.Pay special attention to asylum seekers who are left out of the state social support program.Create a city-to-city cooperation plan for the high-density municipalities receiving refugees and migrants.	2962	5932
Withdrawal of the EU flag to any official entity in Barcelona	Withdrawal of the EU flag to any official entity of Barcelona as a sign of rejection for the lamentable action that is being carried out with the refugees	6214	5932
Leadership of Barcelona host city.	The city of Barcelona, with the city council at the head, leads the reception of refugees.	8678	5932
Create a bank of resources for refugees at the level of the District connected to the Generalitat in two levels (reception and awareness)	The reception of refugees in the city must be planned, work in conjunction with the Generalitat and establish measures to: - The reception (accommodation, support, language, etc.) - Awareness (educational centers and citizenship in general)	6050	5932
Participation in the Social Area Barcelona City Refuge	Participation in the Social Area Barcelona City Refuge	1824	5932
Engage in the reconstruction of the city of Kobane, in the Kurdish region of Syria.	The city of Kobane was a battle scene of the Kurdish militia against Daesh (Islamic state).The battle was won but the city was destroyed by 70%.Now you have to rebuild it.Thus emancipating policies in the Middle East are supported and conditions are created so that the population can return instead of fleeing to Europe.	7287	5933
Boost Barcelona as the capital for peace in the Mediterranean	Promote cooperation agreements and transnational twinning with cities considered as the arrival points of refugees and migrants from the Mediterranean in coordination with existing networks and the Refugee Cities Network.Increase the efforts to make a public debate, support organizations that work for regional justice and initiate processes of international impact to contribute to the achievement of a global governance system that protects the human rights of people and the towns in the region.	2963	5933

Table 6: Global Justice proposals.

Chapter 4

Conclusions

In this study, we have applied several techniques to a small corpus composed of short documents collected from an e-governance application. These two properties (the size of the corpus and the length of the proposals) represented a challenge when applying standard techniques, like the widely used Word2Vec. We applied different linkage methods to find the most suitable for our dataset. We found that when the clusters are compact, the difference between linkages is small. However, as every dataset has its own particular features, we suggest testing with several linkage methods.

To evaluate the clustering of proposals, we have used four different metrics; Purity, Normalized Mutual Information, Adjust Rand Index and Silhouette Coefficient. We have shown, that none of these measures is able to accurately evaluate the final clustering, and we found useful to combine more than one metric to select a final result as the best. Also, we suggest using an internal evaluation criterion, mainly due to the possibility of having a not accurate ground truth.

Word2Vec did not show a superior solution for any of the subset of proposals used in our experiments. We believe this is due to the lack of a model training with our data, and due to the length of proposals that affect the final found similarities. LSA

worked very well removing noise and keeping features that allow finding similarities more accurately. We cannot conclude if any of these methods can capture semantics accurately, as we found a high word co-ocurrence between proposals cluster in the same groups. Tf-Idf did perform as expected and was very useful to remove words that did not add any meaningful information to sentences.

We propose to train a Word2Vec model after collecting more proposals and define a more accurate ground truth. A promising direction would be to use the text from the proposals to build the corpus. In the mean-while, we believe LSA might lead to significant and informative clustering that can be used by Decidim Barcelona.

List of Figures

List of Tables

Bibliography

[1] C. Aggarwal and C. Zhai. A survey of text clustering algorithms. 08 2012.

[2] A. Amelio and C. Pizzuti. Is normalized mutual information a fair measure for comparing community detection methods? 08 2015.

[3] P. Aragón, A. Kaltenbrunner, A. Calleja-López, A. Pereira, A. Monterde, X. E. Barandiaran, and V. Gómez. Deliberative platform design: The case study of the online discussions in decidim barcelona. In Social Informatics: 9th International Conference (SocInfo) (vol II), pages 277–287. Springer, 2017.

[4] I. Cantador, A. Bellogín, M. Cortés-Cediel, and O. Gil. Personalized recommendations in e-participation: Offline experiments for the 'decide madrid' platform. 08 2017.

[5] C. De Boom, S. Van Canneyt, T. Demeester, and B. Dhoedt. Representation learning for very short texts using weighted word embedding aggregation. 80, 06 2016.

[6] S. Deerwester, S. T. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. 41:391–407, 09 1990.

[7] J. Han, M. Kamber, and J. Pei. Data mining concepts and techniques, third edition. Morgan Kaufmann Publishers, 2012.

[8] A. Huang. Similarity measures for text document clustering. 01 2008.

[9] L. Hubert and P. Arabie. Comparing partitions. 2:193–218, 02 1985.

[10] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed <today>].

[11] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. pages 957–966, 01 2015.

[12] M. Kågebäck, O. Mogren, N. Tahmasebi, and D. Dubhashi. Extractive summarization using continuous vector space models. 04 2014.

[13] Y.-Y. Lee, H. Ke, H.-H. Huang, and H.-H. Chen. Combining word embedding and lexical database for semantic relatedness measurement. pages 73–74, 04 2016.

[14] C. D. Manning, P. Raghavan, and H. Schütze. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK, 2008.

[15] T. Mikolov, G. Corrado, K. Chen, and J. Dean. Efficient estimation of word representations in vector space. pages 1–12, 01 2013.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. 26, 10 2013.

[17] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. 09 2011.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[19] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. 18:613–, 11 1975.

[20] P. Shrestha, C. Jacquin, and B. Daille. Clustering short text and its evaluation. pages 169–180, 03 2012.

[21] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. 08 2018.