UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona

Heart Failure factors

A database approach

A Degree Thesis Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona
Universitat Politècnica de Catalunya

by

Gerard Ion Gállego Olsina

In partial fulfilment of the requirements for the

Bachelor's degree in Telecommunications Technologies and Services Engineering
Major in Audiovisual Systems

Co-Directors:
Marta Ruiz Costa-jussà, PhD
Carlos Escolano, MSc
Jordi Cortadella, PhD

Barcelona, March 2019

"The journey is the reward"

Steve Jobs

Abstract

The first cause of death in our country is cardiovascular disease, which risk factors are widely-known, being stress an important one. In today's world, people access information almost instantly, and all these stimuli may trigger emotional reactions and behavior. Could this be considered a stress factor? Is there a way to relate those stimuli with heart attacks?

This project aims to find relationships between psychological stress factors and heart attacks that took place in Catalunya between 2010 and 2016. We have measured these factors through the news that were published in *La Vanguardia*'s Twitter account, processed by Machine Learning techniques, such as Word Embeddings or Clustering.

We have found three groups of words related to psychosocial stress factors. *Football related words* represent an association of +0.9 % in the mean number of heart attacks, while *Spanish and Catalan politics related words* represent a stronger relationship of +1.51 %. we found a negative association in the *International news related words* with a -0.8 %. We have also studied other relationships based on population subgroups.

Although these results do not prove causality, interesting associations have been obtained and represent a first approach in this type of study.

Resumen

La primera causa de muerte en nuestro país son las enfermedades cardiovasculares, cuyos factores de riesgo son ampliamente conocidos, siendo el estrés uno de los importantes. Hoy en día, el acceso a la información es prácticamente instantáneo, y todos estos estímulos pueden desencadenar reacciones emocionales y de comportamiento en la población. Podría considerarse esto un factor de estrés? Hay alguna forma de relacionar estos estímulos con los infartos?

El objetivo de este proyecto es encontrar las relaciones entre factores de estrés psicosocial y ataques al corazón que se produjeron en Cataluña entre 2010 y 2016. Estos factores se miden a través de las noticias publicadas en la cuenta de Twitter de *La Vanguardia*, procesadas con técnicas de Machine Learning, tales como Word Embeddings y Clustering.

Principalmente, se han encontrado tres grupos de palabras relacionadas con factores de estrés psicosocial. El *grupo de palabras relacionadas con el fútbol* representa una asociación del +0.9 % sobre la media diaria de ataques al corazón, mientras que el *grupo de palabras relacionadas con la política catalana y española* representa una relación mayor, del +1.51 %. Sorprendentemente, se ha encontrado una asociación negativa en el *grupo de palabras relacionadas con noticias internacionales* con un -0.8 %. Otras relaciones basadas en subgrupos de población también han sido estudiadas.

Aunque estos resultados no prueben causalidad, se han encontrado relaciones interesantes y representan un primer acercamiento a este tipo de estudio.

Resum

La primera causa de mort al nostre país són les malalties cardiovasculars, de les quals els seus factors de risc són àmpliament coneguts, sent l'estrès un dels més importants. Avui en dia, l'accés a la informació és pràcticament instantani, i tots aquests estímuls poden desencadenar reaccions emocionals i de comportament a la població. Es podria considerar això un factor d'estrès? Hi ha alguna manera de relacionar aquests estímuls amb els infarts?

L'objectiu d'aquest projecte és trobar les relacions entre factors d'estrès psicosocial i els atacs de cor que es van produir a Catalunya entre 2010 i 2016. Aquests factors es mesuren a través de les notícies publicades al compte de Twitter de *La Vanguardia*, processades amb tècnques de Machine Learning, com Word Embeddings i Clustering.

Principalment, s'han trobat tres grups de paraules relacionades amb factors d'estrès psicosocial. El *grup de paraules relacionades amb el futbol* representa una associació del +0.9 % sobre la mitjana diària d'atacs de cor, mentre que el *grup de paraules relacionades amb la política catalana i espanyola* representa una relació major, del +1.51 %. Sorprenentment, s'ha trobat una associació negativa en el *grup de paraules relacionades amb notícies internacionals* amb un -0.8 %. Altres relacions basades en subgrups de població també han sigut estudiades.

Encara que aquests resultats no provin causalitat, s'han trobat relacions interessants i representen un primer pas en aquest tipus d'estudi.

I dedicate my work to my parents, for their unconditional support, and to my grandmother, who always desired to be present in my graduation.
I also dedicate it to my girlfriend, Laura, for her patience during the development of this project. I love you.
Finally, I would like to give a special thanks to Arreplegats de la Zona Universitària, that have made my years in college the best of my life so far, and to Tura Gimeno, who has helped me countless times since the first day in college.

Acknowledgements

I am very thankful to the supervisors of the project: Marta Ruiz (UPC), Carlos Escolano (UPC) and Jordi Cortadella (UPC), whose guidance throughout the whole project have been excellent.

This project could not have been possible without the collaboration of Dr. Barrabés and Dr. Bañeras from Hospital Vall d'Hebron, who proposed to find relationships between heart attacks and the level of stress in population and provided data from Codi Infart Catalunya.

Revision history and approval record

Revision	Date	Purpose
0	15/02/2019	Document creation
1	28/02/2019	Document revision
2	10/03/2019	Document revision
3	14/03/2019	Document revision
4	15/03/2019	Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Gerard Gállego	gerard.gaol@gmail.com
Marta Ruiz	martaruizcostajussa@gmail.com
Carlos Escolano	carlos.escolano@tsc.upc.edu
Jordi Cortadella	jordi.cortadella@upc.edu

Written by:		Reviewed and approved by:	
Date	15/03/2019	**Date**	15/03/2019
Name	Gerard Gállego	**Name**	Marta Ruiz
Position	Project Author	**Position**	Supervisor

Table of contents

List of Figures

List of Tables

Nomenclature

AI	Artificial Intelligence
AMI	Acute Myocardial Infarction (Heart attack)
CIC	Codi Infart de Catalunya
CSV	Comma-Separated Values
DL	Deep Learning
EM	Expectation-Maximization
GMM	Gaussian Mixture Model
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
URL	Uniform Resource Locator (Web address)

Chapter 1

Introduction

In this chapter, I detailed the general outline of the project. First, a statement of purpose is defined, where I describe the primary motivation of the project and its overview. Then, requirements and specifications are defined, as well as the methods used to accomplish them. Finally, I present the work plan followed and the most significant incidences that we have found.

1.1. Statement of purpose

In a globalized world, where information arrives everywhere almost instantly, people are increasingly sensitive to psychosocial stress factors and, therefore, their health might be affected. The first cause of death in our country is cardiovascular diseases (Fig. 1.1.1), which causes are widely-known: congenital heart defects, coronary artery disease, high blood pressure, smoking and also stress. Is there a way to measure it? Is there any possibility to find associations between any psychosocial stress factor and cardiovascular diseases?

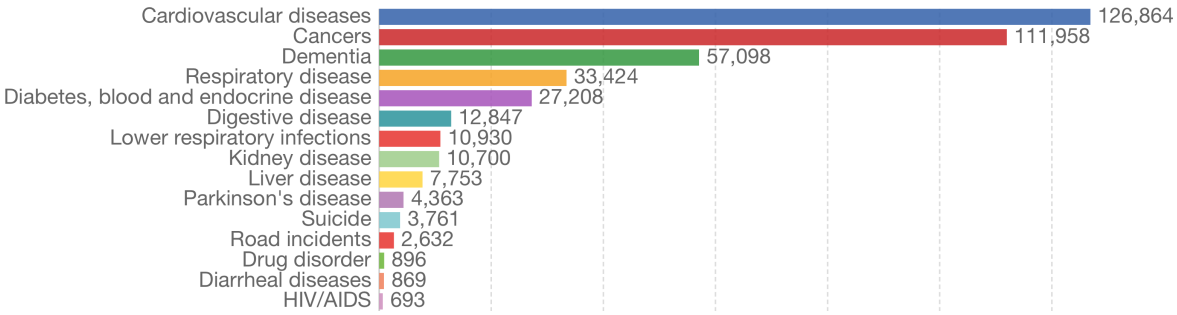


Figure 1.1.1: Annual number of deaths by cause (Top 15), Spain, 2016.[1]

[1]Source: Institute of Health Metrics and Evaluation (IHME); Global Terrorism Database (GTD); Amnesty International. OurWorldInData.org/causes-of-death/

Moreover, recent advances in data science, machine learning and deep learning are helping researchers to process more information in a more powerful way, which is helping them to understand better the world where we live. Furthermore, thanks to Natural Language Processing (Section 3.1) and social media information extraction, a wide range of possibilities is triggered.

Although other studies focused into relating a particular psychosocial stress factor to heart diseases or its mortality, none of them has tried to find associations from a more generalist approach, trying to see, globally, which are the factors that may be more or less related, positively or negatively (Fig. 1.1.2). In this project, taking advantage of recent machine learning techniques, mass media Twitter data and thanks to the medical data Dr. Barrabés and Dr. Bañeras provided us[2], we have tried to find which are these relationships.

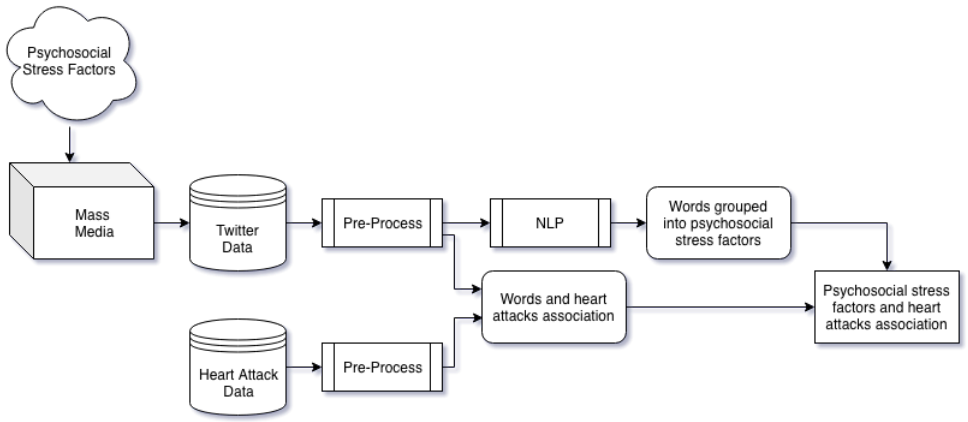


Figure 1.1.2: General schema of the project.

The project has been carried out at the Signal Theory and Communications (TSC) of Universitat Politècnica de Catalunya (UPC), in collaboration with Dr. Barrabés and Dr. Bañeras from Hospital Vall d'Hebron, during the academic year 2018/19.

1.2. Requirements and specifications

The requirements of this project are:

- Create a model of the language used in a mass media Twitter account (Word Embedding, Section 3.1.2).
- Perform a clustering of the language model, to obtain groups of words related to psychosocial stress factors (politics, sports, economy...) (Section 3.2).
- Find associations between the number of Acute Myocardial Infarction (AMI) and clusters.

[2]From *Codi Infart de Catalunya*

Moreover, the specifications of this project are the following ones:

- Some clusters obtained have to be identifiable as psychosocial stress factors
- The relationships found must be significant.
- Population sub-groups (gender, age ranges...) have to be studied too.

1.3. Methods and procedures

This project aims to find a relationship between the average number of AMI per day (in Catalunya, between 2010 and 2016), and the words (grouped into psychosocial stress factors) that appear in *La Vanguardia*'s Twitter account.

First, medical data has been preprocessed to clean it up and correctly format. Twitter data has been acquired from *La Vanguardia*'s account and has also been preprocessed, tokenizing and lemmatizing the tweets. Once data has been ready, we have computed relationships between words and the number of AMI per day.

Then, a Word Embedding model has been created, whose results have been clustered, trying to group words related to a psychosocial stress factor together. Finally, we have computed relationships between these clusters and the number of AMI per day too.

A wide variety of specific programming tools have been used to perform all the described procedures.

On the one hand, medical data has been preprocessed using *Bash* scripts with *csvkit*[3] and *Python 3* with *Pandas* library, to aggregate the columns per day.

On the other hand, Twitter data has been acquired using the *Get Old Tweets Programatically* project[4]. It has been preprocessed using *Bash* scripts and *Python 3* with the libraries: *Pandas* for the CSV management, *NLTK* for the tokenization of the tweets and *Pattern*[5] for the lemmatization (Section 3.1.1).

For the rest of the project a *Jupyter Notebook* running a *Python 3* kernel has been used, with the libraries: *Numpy* to be able to work with arrays, *Pandas* for the CSV and table management, *Gensim's Word2Vec* for the word embedding model (Section 3.1.2), *Sci-Kit Learn* for the clustering (Section 3.2), *UMAP* for the dimensionality reduction (Section 3.3), *Plotly* for the result graphs and *Scipy* for the P-Value calculation.

[3] Available in GitHub: https://github.com/wireservice/csvkit
[4] Available in GitHub: https://github.com/Jefferson-Henrique/GetOldTweets-python
[5] By: Computational Linguistics & Psycholinguistics Research Center (CLiPS)

1.4. Work plan

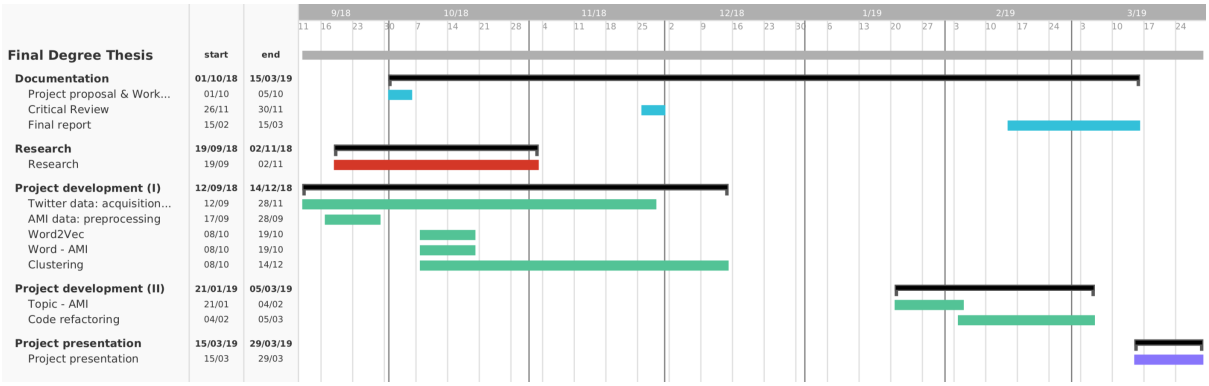


Figure 1.4.1: Gantt diagram. Created with TeamGantt.

1.5. Incidences

The main modification is that I did a break in the project since the 15th of December until the 21st of January, due to academic work overload.

When I came back to work in the project, I decided to work in a *Jupyter notebook* environment. I also had to improve how the system stored the results. For these reasons, the task "Code refactoring" has been added.

Chapter 2

Literature Review

There are lots of researchers who have been interested in relating psychosocial stressors with cardiovascular disease and mortality. Some studies have concluded that these factors "can be both a cause and a consequence of cardiovascular disease events" [Figueredo, 2009], and a wide range of psychosocial factors have been studied. In this chapter, I expose a review of some of the most important articles in this field.

Many studies have covered different sports, which is the most researched factor: football [Kirkup and Merrick, 2003], hockey [Gebhard et al., 2018] or American football. Some of them have focused in a concrete event, like a Super Bowl Championship [Kloner et al., 2009], a Football World Cup [Wilbert-Lampen et al., 2008][Berthier and Boulay, 2003] or a European Football Championship [Witte et al., 2000]. Most of them have concluded that there is a relationship, but sometimes this has been found only in man population or in young people. However, other studies have concluded that the association was non-existent [Niederseer et al., 2013] or insignificant [Barone-Adesi et al., 2010].

Some studies have also found a relationship between cardiovascular diseases and factors like meteorology [Bai et al., 2018], pollution [Bañeras et al., 2018] or volatility of the stock market [Ma et al., 2011]. Some of them conclude that the association is stronger in the old population [Dilaveris et al., 2006][Vanasse et al., 2017]. Other researchers have also taken into account factors like an earthquake disaster [Aoki et al., 2012] or a terrorist attack [Chi et al., 2003], the last one without finding any association.

However, all these studies only have analyzed one stress factor each one, that differs from the globalized objective of the present study, and none of them has used the social media to measure the degree of psychosocial stress. The most similar research that we have found has used "language expressed on Twitter to characterize community-level psychological correlates of age-adjusted mortality from Atherosclerotic Heart Disease (AHD) [Eichstaedt et al., 2015].

Chapter 3

Theoretical Background

The system that we have developed takes advantage of some techniques which are part of the vast family of machine learning, composed of many different algorithms, specialties and applications. In this chapter, I describe the basic concepts of the techniques we have used.

The first appearance of the concept "machine learning" was in the article *Some Studies in Machine Learning Using the Game of Checkers [Samuel, 1959].* There, the author said that "a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program". Furthermore than learning to do this "in a remarkably short period of time" and "when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters". This article concludes that those techniques could be also applied to real-life problems. Moreover, this author provided the very first definition of machine learning:

> *"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed."*
>
> Arthur L. Samuel, 1959

During the 1990s, researchers started to change their primary objective. Instead of focusing on artificial intelligence, they began to be interested in statistical and pattern-recognition approaches [Langley, 2011], which is more related in which, nowadays, is understood as machine learning. While AI is a program that can sense, reason, act and adapt; ML is a subset of AI (Fig. 3.0.1) that refers to algorithms whose performance improve as they are exposed to more data, as was defined in 1997 by Tom Mitchell [Mitchell, 1997]:

> *"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."*

Machine learning can deal with problems that cannot be solved by traditional methods because it can understand complex models and relationships in data.

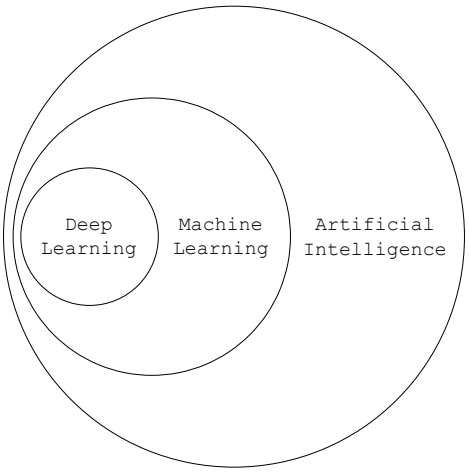


Figure 3.0.1: Relationship between AI, ML and DL.

Moreover, deep learning has been popular since the 2000s, it consists of adding hidden layers to a neural network layer [Goodfellow et al., 2016], and it is considered a subset of machine learning techniques (Fig. 3.0.1). However, we have not used deep learning techniques in this project.

Some examples of the uses of machine learning are the estimation of the market value of a house, automatic speech recognition or email filtering.

Machine learning techniques can be distinguished in two main subgroups, based on the way machines or algorithms learn from data: supervised and unsupervised learning.

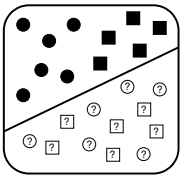


Figure 3.0.2: Supervised learning algorithms schema.

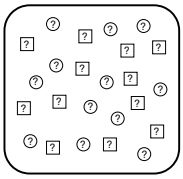


Figure 3.0.3: Unsupervised learning algorithms schema.

Supervised learning, which is the most common, consists of training the model with a training set, which is a subset of input data whose correct output values are known (labeled data). The algorithm iterates over training data, improving the accuracy of its predictions, correcting itself when the outputs are wrong, and trying to obtain the optimal mathematical model which can predict the output for new inputs, that were not present in training data (Fig. 3.0.2). Supervised learning includes classification problems (outputs restricted to a limited set of values) and regression problems (outputs are numerical values within a range) (Tab.

3.0.1). The most used algorithms are logistic regression, support vector machines and neural networks.

In unsupervised learning, data is not labeled, so the algorithms learn from data without knowing the "correct" outputs beforehand (Fig. 3.0.3). The program finds structures and similarities present in data; it recognizes similar objects, and they are grouped (the machine design the labels). Other unsupervised processes also use mathematical reduction of redundancy. The typical problems to solve are clustering and dimensionality reduction (Tab. 3.0.1), and the most used algorithms are k-means clustering, hierarchical clustering, principal component analysis, T-distributed stochastic neighbor embedding...

	Supervised Learning	**Unsupervised Learning**
Discrete	Classification	Clustering
Continuous	Regression	Dimensionality Reduction

Table 3.0.1: Supervised learning and unsupervised learning comparison.

3.1. Natural Language Processing

Computers are used to work with standardized and structured data, and they can work faster than humans do, but they have some problems if data is unstructured, like human languages. This issue is in what natural language processing is focused, which is a subfield of study between computer science, artificial engineering and information engineering. The *Encyclopedia of Library and Information Science* [Liddy, 2001] gived a commonly accepted definition of this field:

> *"Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications."*

Natural language processing started back in the 1950s, when Georgetown & IBM began working in automatic translation [Reynolds, 1954]. The following decades the scientists also studied chatterbots, but the revolution did not start until the 1980s, when they introduced machine learning algorithms for language processing. In the 2010s, the use of neural networks has allowed performing tasks like word embeddings or text summarization [Goldberg, 2016].

Nowadays, some of the most common uses of NLP are machine translation, natural language generation, sentiment analysis, speech recognition... However, also others, that are part of significant machine learning systems, like word embeddings, part-of-speech tagging or lemmatization.

The methods of natural language processing are based on the "levels of language" approach, which divides the language into seven layers [Liddy, 2001]:

- Phonology: Focuses on speech sounds within and across words.
- Morphology: Studies the morphemes, the smallest units of meaning.
- Lexical: Interpretation of the meaning of individual words.
- Syntactic: Analyses the grammatical structure of a sentence.
- Semantic: Studies the interactions of words in a sentence to obtain the global sentence meaning.
- Discourse: Works with units of text longer than a sentence and its objective is to understand the properties of the whole text.
- Pragmatic: Applies context and world knowledge to understand the extra meaning

One of the main difficulties of NLP is The Ambiguity Problem, that appears when an NLP system finds a fragment of text with multiple interpretations and has to decide the more appropriated [Màrquez, 2000]. Some examples are word selection in speech recognition, semantic ambiguity in polysemic words or structural ambiguity in parsing.

3.1.1. Lemmatization

Lemmatization is the process to group the different forms of a word into their lemma. It is used in natural language processing, text mining and other linguistics fields [Plisson et al., 2004], because the simplification of the words is very useful to reduce the vocabulary in a text.

It is similar than stemming, that consists in removing the last part of a word, following some rules; but the first one is more complex because it takes into account the part-of-speech information and context information [Müller et al., 2015].

For instance, if the words like "worse" or "worst" were stemmed the result would probably be "wors", while a lemmatizer would obtain its correct lemma "bad". In other cases, the lemma is different depending on the context and the part-of-speech, where the stemmer also fails, but lemmatizer does not [Manning et al., 2009]

While there are good stemmers for English words like Porter's stemmer [Porter, 1980], these don't work correctly for very inflected languages [Mladenic, 2002].

3.1.2. Word Embeddings

Word embeddings are a group of unsupervised learning techniques (sometimes considered self-supervised) which consist in representing words by multi-dimensional feature vectors, in such way that representations of words with similar meaning are close and word relations are equivalent to vectors (Fig. 3.1.1).

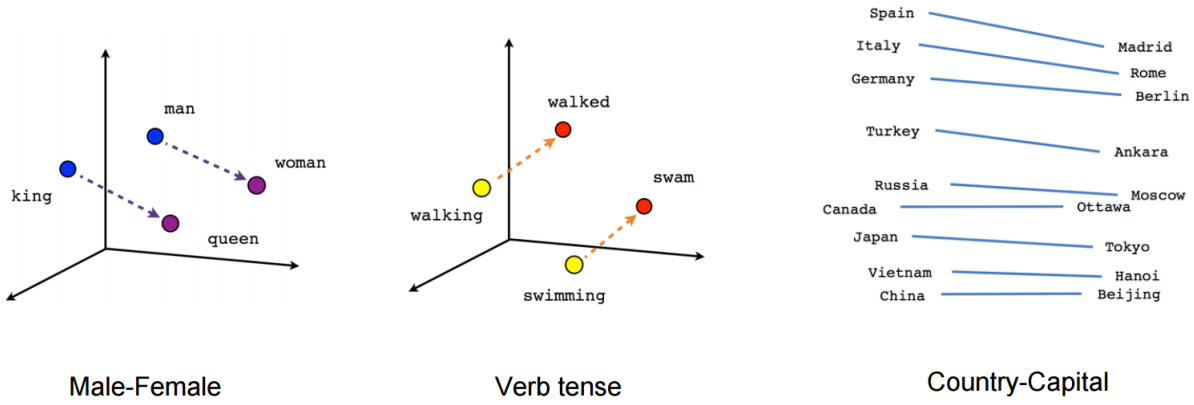


Figure 3.1.1: Word relationships and their equivalence to vectors.[1]

Without these techniques, words are typically represented with the bag-of-words model (one-hot encoding), whose vector representations are too sparse (much more dimensions than word embeddings), and they don't include word semantics information [Bengio et al., 2003].

Although the utility of pre-trained word embeddings was shown five years before [Collobert and Weston, 2008], the popularization of word embeddings is attributed to the creator of Word2Vec [Mikolov et al., 2013a][Mikolov et al., 2013b], a toolkit that allows an easy training and usage of pre-trained word embeddings.

Other toolkits like GloVe were released later [Pennington et al., 2014], but, in this project, we have used Word2Vec, as the tutors recommended.

Word2Vec

Word2Vec is probably the most important word embedding algorithm of our times. It consists of training a neural network that learns a language model, calculating conditional probabilities of words within sentences. The neural network only has one hidden layer, so it cannot be considered a deep learning technique. One of the main benefits over previous approaches [Bengio et al., 2003] is that it is computationally less expensive.

[1]Source: Vector Representations of Words tutorial (Tensorflow).

There are two architectures proposed: continuous bag-of-words (CBOW) and Skip-gram [Mikolov et al., 2013a].

On the one hand, CBOW consists of guessing the middle word of groups of surrounding words. At each step, the model receives a window of words (the size is defined) and the network is adjusted to predict the central word. In the end, the weights between the hidden layer and the output layer are taken as the word vectors (Fig.3.1.2). Its main advantages are *"several times faster to train than the skip-gram, slightly better accuracy for the frequent words"* [Mikolov, 2013].

On the other hand, Skip-Gram is the opposite of CBOW. It tries to guess the surrounding words from the middle words. At each step, the model receives a central word and the network is adjusted to predict the context words. Finally, the weights between the input layer and the hidden layer are taken as the word vectors (Fig.3.1.3). Its main advantages are that *"works well with small amount of the training data, represents well even rare words or phrases"* [Mikolov, 2013].

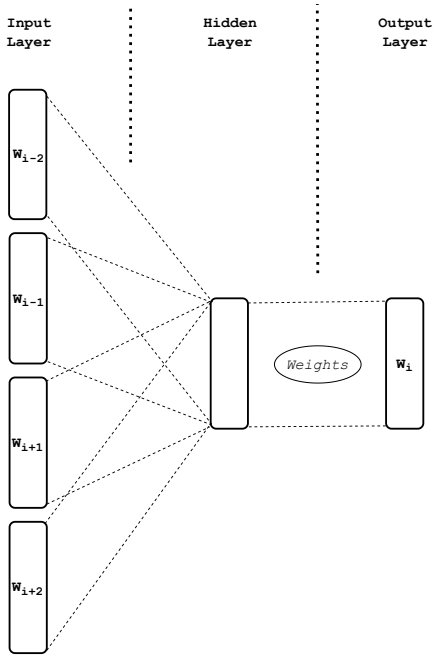


Figure 3.1.2: Word2Vec's CBOW architecture.

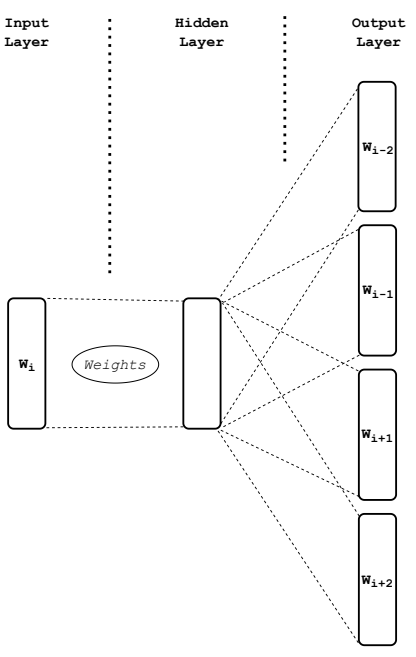


Figure 3.1.3: Word2Vec's Skip-Gram architecture.

3.2. Clustering

Clustering is a method of unsupervised learning in charge of grouping data points with similar features into the same cluster and points with different features into different clusters. Once the algorithm has partitioned into different groups, then they are labeled. Clustering is a widely used technique in lots of machine learning fields, included natural language processing Baker and McCallum [1998].

There are different clustering algorithms approaches, but the appropriate one and its parameters depend on the data set where it is applied.

3.2.1. Centroid-based clustering

In this approach, the cluster is represented by a central vector. The most extended algorithm is K-Means [MacQueen, 1967], which is fast and simple because it only has to calculate distances between points and centroids.

K-Means algorithm has three main steps:

1. Randomly chooses initial centroids.
2. Classifies each data point into the nearest centroid cluster.
3. Recalculates the centroid taking the mean value of all points in the cluster.

The algorithm repeats the last two steps until the centroids do not move significantly.

The main disadvantages of K-Means are that the number of clusters must be known in advance and that the random initialization may yield to the results not being repeatable.

3.2.2. Connectivity-based clustering

In this type of clustering, the clusters are understood like groups of "connected" data points, and it is also known as Hierarchical Clustering because it is a method which tries to build a hierarchy. There are two ways of working: bottom-up or top-down.

On the one hand, bottom-up algorithms start assigning a cluster to each data point and, then, begin agglomerating pairs of clusters recursively, with the smallest distance, until there is only one cluster.

On the other hand, top-down algorithms start assigning all data points to the same cluster and, then, start splitting the biggest clusters recursively in pairs until no more divisions are possible.

The representation of the hierarchy is a dendrogram, where the root is the global cluster and the leaves are the clusters with only one data point.

The distance metric (euclidian, Manhattan...) and linkage criteria (complete-linkage, single-linkage...) are the parameters of the algorithm. The number of clusters does not need to be specified beforehand and can be chosen once the dendrogram is built.

3.2.3. Distribution-based clustering

In this approach, the clustering model is related to statistics, based on distribution models. The clusters are defined as groups of data points that belong to the same distribution. These distributions are represented as Gaussian Mixture Models, which is a generalization of the K-Means algorithm, that fails in cases where clusters are not circular. In this case, the assumption is that data points are Gaussian distributed, which is less restrictive than the circular distribution of the K-Means (based on computing radial distances to the centroid)

The mean and standard deviation are the parameters of the clusters, which are found with the Expectation-Maximization optimization algorithm [Dempster et al., 1977]. They are randomly initialized and recomputed iteratively to maximize the probabilities of data points to belong to the distributions. The algorithm will converge to a local optimum, so it may obtain different results in different runs of the algorithm.

The main advantages of using GMM clustering are that clusters can take an ellipse shape and that overlapping clusters are allowed (it supports mixed membership). The most significant disadvantage is that, as every technique based in distances to the center (K-Means included), standard deviations are the same in all dimensions. Another important drawback is that it suffers overfitting.

3.2.4. Density-based clustering

In this type of methods, clusters are higher density areas than the rest of the data set. The most used algorithm is DBSCAN [Ester et al., 1996] [Kriegel et al., 2011], that is based in connecting a minimum number of points, which distance is below a threshold. These are the needed parameters of the algorithm, the distance threshold and the minimum number of data points.

The algorithm iterates all data points and joins them into clusters that meet the conditions established by the parameters. All those data points which the process has not assigned to any cluster are considered noise.

The main advantages of DBSCAN are: it does not require the number of clusters beforehand, it identifies outliers as noise, it finds clusters of any shape. The main disadvantage is that it does not perform well with clusters of different density.

3.3. Dimensionality Reduction

It is the process of reducing the number of dimensions of a data set. Machine learning engineers widely use it in their projects as a previous step before plotting multidimensional data into a 2-dimensional graph. Another important use is to avoid the "Curse of dimensionality", that is a group of phenomena that appear with high-dimensional data, due to the sparsity of the points [Hughes, 1968][Trunk, 1979][Beyer et al., 1999].

There are lots of techniques to reduce dimensionality, one of the most used is t-Distributed Stochastic Neighbor Embedding (t-SNE) [van der Maaten and Hinton, 2008], but in this project a brand new method called Uniform Manifold Approximation and Projection (UMAP) has been used, which is "constructed from a theoretical framework based in Riemannian geometry and algebraic topology. The result is a practical scalable algorithm that applies to real world data. The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance. Furthermore, UMAP has no computational restrictions on embedding dimension, making it viable as a general purpose dimension reduction technique for machine learning" [McInnes et al., 2018].

The most critical parameters for UMAP are: the number of neighbors, that controls how much it preserves the overall structure (high number of neighbors) or the local structure (low number of neighbors); and the minimum distance, that controls how tightly can the points be packed (low minimum distances result in finding small connected points).

Chapter 4

Datasets

The purpose of this chapter is to define how the necessary data for this project is acquired and pre-processed, clearly describing how we treated the two main datasets.

4.1. Heart attack data pre-processing

The dataset provided includes all the Acute Myocardial Infarction cases in Catalunya between 2010 and 2016 (more than 17,000 AMI cases). Furthermore, each row had more than a hundred variables, so we had to make some efforts before we start working with this dataset.

The first pre-process that we carried out consisted of getting rid of all the unnecessary information. We followed different criteria:

- Patient's identification data.
- Data which was considered irrelevant (e.g., the nationality of the patient).
- Columns with too much variability (e.g., the medical center where the patient went).
- Columns that contain information about the treatment.
- Duplicated columns[1].

Once we finished the basic cleanup, other arrangements were necessary. The main changes we did are:

- Binary columns were converted to the standard format to work easier with them later.
- Date format was adapted to the standard.
- Some columns with high variability, were grouped into ranges[2].

[1] e.g. Two columns were referring to the gender, the first one had the values "man" or "woman", and the other had the values "0" or "1".

[2] e.g. Information about the moment of initial pain was grouped into three ranges (morning, night and afternoon).

Finally, summary tables were extracted performing an aggregation by day (Figure 4.1.1). They were stored into different files with similar information, to facilitate work later on. The aggregations are grouped by:

- Total number of heart attacks
- Gender
- Mortality
- Age range (20 years)
- Medical history
- Moment of initial pain
- Diagnostic

Date	Number of IAM	20-39a	40-59a	60-79a	80-99a	100-119a
2010-01-01	1	0	0	1	0	0
2010-01-02	6	0	0	5	1	0
2010-01-03	3	0	1	2	0	0
2010-01-04	6	1	1	4	0	0
2010-01-05	8	1	2	5	0	0

Figure 4.1.1: Extract of the AMI summary table by age range.

4.2. Twitter data acquisition

Once we had the heart attack data ready, we needed to obtain a database where psychosocial stressors were reflected, for what we chose Twitter data from mass media accounts. The premise was that these tweets would reflect if anything happened in the world that could affect society.

To take benefit of the seven years of heart attack data available, obtaining Twitter data of the same time was also necessary. This task was difficult because acquiring Twitter data using the official API required a Developer account, which was obtained but, even so, acquiring all the necessary data was not possible. Instead of that, as described in Section 1.3, we used the *Get Old Tweets Programatically* project, that bypasses limitations of the official API.

The process of acquiring Twitter data took longer than expected (sometimes a whole night with the computer working), because the script downloads packets of a hundred tweets, with a short stop between them.

Although we acquired data from different sources: *La Vanguardia* (259,858 tweets), *Diari ARA* (77,757 tweets), *324* (127,953 tweets) and *Agencia Catalana de Notícies* (46,026 tweets); finally, we only used the first dataset, due to the language differences between them and some difficulties to work in Catalan in the lemmatization process mentioned in Section 4.3.

4.3. Twitter data pre-processing

Similarly to what was done before, preprocessing the Twitter data was necessary. We divided this task into the data arrangement and the content of the tweets pre-processing.

On the one hand, the first one consisted in:

- Properly formatting the CSV. The main problem was that, sometimes, the text content of the tweet had semicolons, the symbol used to separate columns in the CSV format, which caused bad data splitting.
- Removing unnecessary fields.
- Removing URLs from the content of the tweet.
- Formatting dates into the standard format and removing time information.

On the other hand, the tweet content pre-processing consisted in:

- Text tokenization of the content of the tweet.
- Removing tokens which are not words (e.g., numbers or symbols).
- Filtering "Stop words" (i.e., the most common words in a language).
- Lemmatization of the tokens (Section 3.1.1).

As detailed in Section 1.3; *Pandas*, *NLTK* and *Pattern* were used in this procedure.

username	date	retweets	favorites	text	geo	men
LaVanguardia	2016-12-03 19:01:00	2	4	"El resultado no era justo; estoy contento con el gol" http:// dlvr.it/MpBB78 pic.twitter.com/KNZhlHpSRa		

Figure 4.3.1: Extract of the original Twitter data.

date	hour	username	text	retweets	favorites
2016-12-03	19	LaVanguardia	[resultado, ser, justo, contentar, gol]	2	4

Figure 4.3.2: Extract of the pre-processed Twitter data.

Chapter 5

Methodology

This chapter describes how we obtained the results from the proposed objectives, clearly explaining which are the procedures that we have followed until the project completion.

5.1. Heart attacks associated with words appearances

Once we had preprocessed all the necessary information, the next step was to relate the number of AMI (at each aggregation; i.e., gender, age range...) and the words that appeared. We performed this procedure calculating the mean number of heart attacks for each word; however, we followed two different approaches:

- Mean value of all the days the word appeared at least one time.
- Weighted mean value depending on the number of appearances of the word each day.

Although the first one seems to fit better the primary assumption of this project (if a word appears one day, it could represent a psychosocial stress factor for the population), the second one is interesting because it adds information about how important was the word each day it appeared.

The computation was done iterating all the tweets and storing the information as the words were found, the number of days or the number of times they appeared was also saved in the exported CSV files (Fig. 5.1.1 & 5.1.2).

word	Days	Medical history	Previous IAM	Diabetes	Previous Angioplasty	Previous Coronary surgery	RT	FAV
aguirre	279	3,237…	0,728…	1,448…	0,566…	0,079…	9,207…	2,857…
decir	1.538	3,174…	0,715…	1,424…	0,564…	0,092…	5,734…	2,292…
zapatero	376	2,346…	0,702…	1,316…	0,524…	0,064…	2,079…	0,393…
llamar	639	3,567…	0,725…	1,410…	0,592…	0,108…	7,743…	3,312…
estrangulador	1	2,000…	1,000…	2,000…	0,000…	0,000…	0,000…	0,000…

Figure 5.1.1: Extract of the Word2AMI table by medical history following the basic mean approach.

word	Appearances	Medical history	Previous IAM	Diabetes	Previous Angioplasty	Previous Coronary surgery	RT	FAV
aguirre	391	3,353…	0,688…	1,437…	0,578…	0,066…	9,207…	2,857…
decir	3.135	3,138…	0,721…	1,422…	0,563…	0,101…	5,734…	2,292…
zapatero	868	2,282…	0,691…	1,368…	0,515…	0,065…	2,079…	0,393…
llamar	848	3,618…	0,717…	1,407…	0,587…	0,098…	7,743…	3,312…
estrangulador	1	2,000…	1,000…	2,000…	0,000…	0,000…	0,000…	0,000…

Figure 5.1.2: Extract of the Word2AMI table by medical history following the weighted mean approach.

5.2. Words grouped into clusters

The goal of this process is to find automatically possible psychosocial stress factors in words, which has been done using an NLP technique called Word Embedding (Section 3.1.2), and a machine learning kind of algorithm called Clustering (Section 3.2).

These are unsupervised learning algorithms (Table 3.0.1), so objectively evaluating their performance is difficult because data is not labeled. Although this could have been a drawback to adjust algorithm parameters, this was done using some references, theoretical information and a bit of trial and error. Finally, results were validated by subjective evaluation, checking that some of the groups of words represented a stress factor.

5.2.1. Word Embedding

The first step to group words is to capture context information of every word, to have an overview of which words are used in similar ways, what is done automatically by the Word Embedding technique.

A model is created using this technique in text data, where every word is represented by a multi-dimensional vector, which contains, as mentioned before, its context information. This method is very interesting for two reasons:

- It allows working in fewer dimensions than other techniques.
- Words with similar contexts tend to be close in the multi-dimensional space.

The Word Embedding algorithm used in this project is called Word2Vec[Mikolov et al., 2013a][Mikolov et al., 2013b] which is divided into two architectures: Skip-gram and CBOW. We chose the second one because it is better for frequent words, as said in section 3.1.2, what fits in this project because only the most frequent words were taken into account, as will be explained later on.

The number of dimensions of vectors that represent words must be specified beforehand. Intuitively, using more dimensions is better because more features of the context information can be represented. However, this means a higher computational effort and also possible dimensionality problems, as explained in section 3.3. For this reason, a number of dimensions sufficiently high has to be chosen, but not too high. In this case, we chose 128.

The context information is extracted, by the algorithm, defining a window size, which specifies the number of surrounding words that have to be considered. Often a size of 5 is used for the CBOW architecture, but in this case, we chose a higher value (7), because while little values extract local surroundings information, higher ones obtain a better representation of the whole context.

The dataset we have used contains 69,931 different words, but 61,158 of them appear less than 20 times in the seven years studied. For this reason, we reduced the number of words in the Word2Vec model and only the 1,500 more frequent words were represented (the minimum number of appearances is 212) because words with fewer appearances are not considered interesting in this study. Furthermore, the number of words has also been chosen taking into account data visualization.

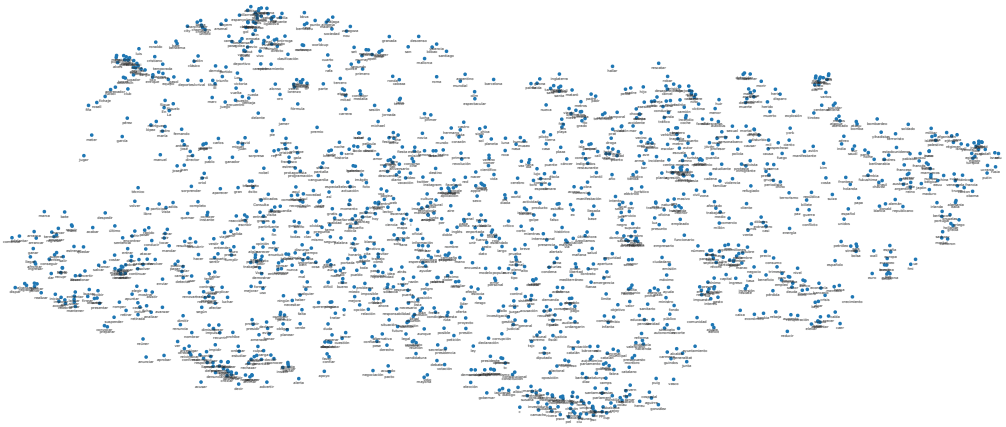


Figure 5.2.1: Word2AMI scatter plot. Mean number of AMI per day.

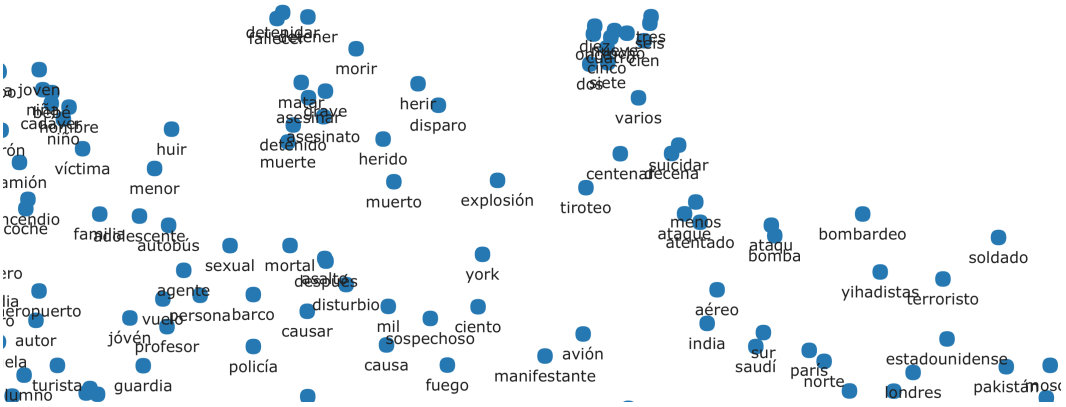


Figure 5.2.2: Zoom in of the Word2Vec model, reduced to 2 dimensions.[0]

5.2.2. Clustering

As seen in the previous section, the word data points were placed in a multi-dimensional space, where similar words were closer than others. Since the primary goal was to group all related words, applying a clustering algorithm was needed.

Although there are four types of clustering (Section 3.2), the code has been prepared to be able to work with any of them. However, the kind of data that is needed to cluster and its application is a critical point in choosing the best clustering algorithm. The existence of data points in undefined regions of the space, where no groups or stress factors can be delimited, penalizes centroid-based, connectivity-based and distribution-based clustering, that assign all data points to a cluster. In this case, grouping these points to any cluster is not needed, so we chose density-based clustering (DBSCAN algorithm) because it only clusters those data points with a minimum density.

The most important parameters of this algorithm that had to be adjusted are:

- Maximum distance between two data points to be considered in the same neighborhood.
- Minimum number of data points in a neighborhood to be considered as a cluster.

The first one is the one in charge to define the size of the existing clusters. If its value is too high, the algorithm will put inside the cluster some data points that should not be there and, in contrast, if it is too small, some words will not be classified in any cluster when they should. In this case, the value used was 0.28.

The minimum number of data points parameter is in charge to regulate the number of clusters that are created. If its value is too small, the algorithm creates lots of unnecessary tiny clusters and, in contrast, if it is too big, only a few big clusters are obtained. For this project, the value that was used is 12.

Before clustering the data points, a dimensionality reduction was performed to convert the space to 10 dimensions and avoid the "Curse of dimensionality" (Section 3.3), as explained in section 5.2.3.

5.2.3. Dimensionality reduction

As mentioned before, a dimensionality reduction process was necessary for two situations:

- To visualize data points in a bidimensional space.

[0]Data points of 128 dimensions had to be dimensionality reduced to be able to plot them, as explained in section 5.2.3. Coordinates of the points do not represent anything; just the word surroundings are interesting from these plots.

- To reduce dimensionality before clustering.

We used an algorithm called UMAP (Section 3.3). The most important parameters of this method are:

- Number of neighbors
- Minimum distance

The number of neighbors is in charge of control what is more preserved: the local structure of the multi-dimensional data (low values) or its global structure (high values). We used a large value (200) because, for this project, is more necessary to preserve the overall structure before the clustering.

The minimum distance refers to how close the data points are permitted to be glued together when the dimensionality is reduced. We used very little value (0.002) because there is no problem in grouping similar points very close.

5.3. Heart attacks associated with clustered words appearances

On the one hand, we found the relationship between words and heart attacks. On the other hand, we grouped the words into clusters. The final step was to find the association between clusters and heart attacks (at each aggregation; i.e., gender, age range...).

This process was done calculating the mean value of AMI from all words grouped in a cluster, weighted with the number of appearances or days of each word. Moreover, we calculated the variation ratio with the global mean.

We also needed P-Value[1] to be sure which results were statistically significant.

[1]Statistical measure, typical accepted results are smaller than 0.05.

Chapter 6

Results

In this chapter, I present a selection and a discussion of the most relevant results; the rest of them can be found in the appendices.

Once we had calculated the association of heart attacks and words(Section 5.1), we combined it with the Word Embedding model (Section 5.2.1). Resulting scatter plots (Appendix A), named "Word2AMI", consist in a cloud of word points, where the color of each one represents the mean number of AMI per day associated with that word (Fig. 6.0.1).

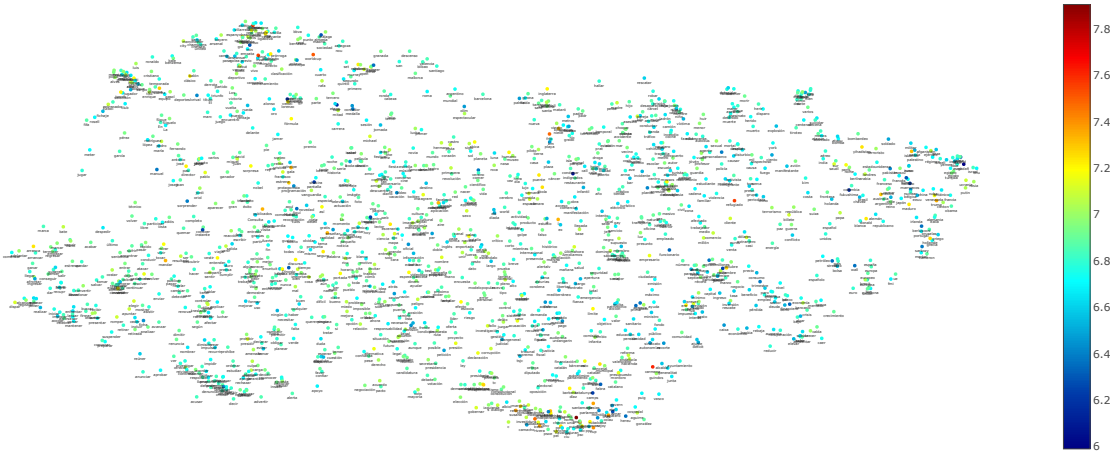


Figure 6.0.1: Word2AMI scatter plot. Mean number of AMI per day.

These scatter plots do not allow to extract general conclusions as clearly as desired. For this reason, we performed a clustering over the Word Embedding model (Section 5.2.2). Resulting scatter plots (Appendix C), named "Cluster2AMI", consist in a cloud of word points, where

the color of each one represents the mean number of AMI per day associated with the cluster correspondent to that word (Fig. 6.0.2 & 6.0.3).

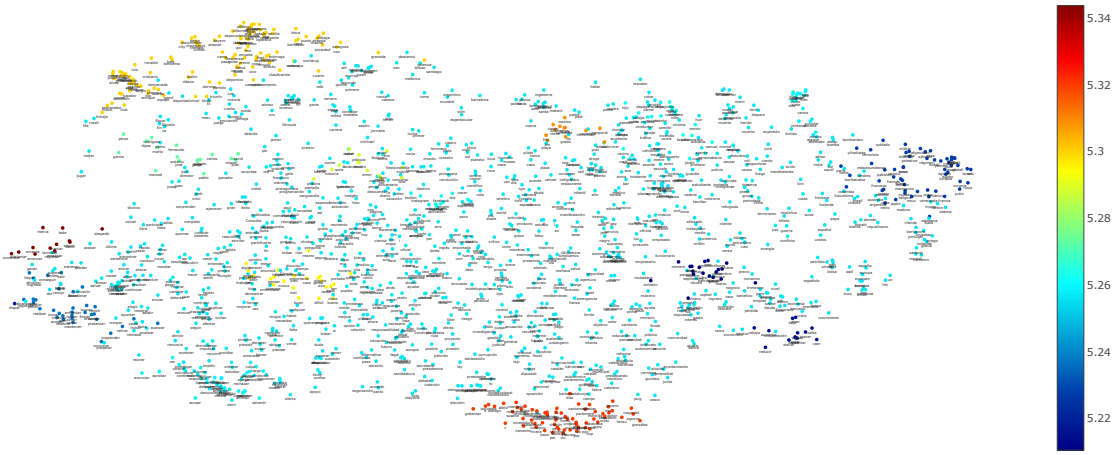


Figure 6.0.2: Cluster2AMI scatter plot. Mean number of AMI per day of men population.

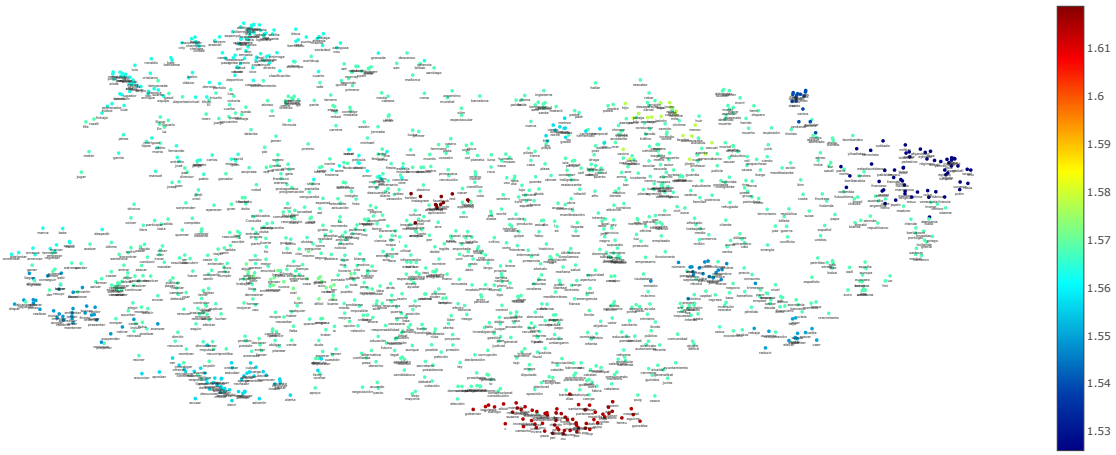


Figure 6.0.3: Cluster2AMI scatter plot. Mean number of AMI per day of women population.

Clearer conclusions can be extracted from "Cluster2AMI" scatter plots. For instance, in figures 6.0.2 and 6.0.3 can be appreciated differences in the location of higher number of AMI averages, depending on the gender of population, specially in the football-related words cluster (#1)[1].

[1]Cluster contents are described in Appendix B.

Before extracting conclusions, we computed the P-Value of the experiments to be sure that they were statistically significant (P-Value < 0.05). In the previous example (Fig. 6.0.3), the P-Value confirms that the variation rate for cluster #1 in women population is not statistically significant (Table 6.0.1), so this result is not considered in the discussion.

Cluster	Mean	Var. rate	P-Value
0	1.5667	-0.121 %	0.663
1	1.56761	-0.063 %	0.473
2	1.56354	-0.322 %	0.156
3	1.60193	2.125 %	0.001
4	1.55649	-0.772 %	0.121
5	1.5377	-1.97 %	0.001
6	1.57258	0.254 %	0.372
7	1.55863	-0.635 %	0.234
8	1.54349	-1.6 %	0.353
9	1.56982	0.078 %	0.822
10	1.55142	-1.095 %	0.039
11	1.56762	-0.062 %	0.855
12	1.55895	-0.615 %	0.737
13	1.54024	-1.808 %	0.706
14	1.55806	-0.672 %	0.708
15	1.60325	2.209 %	0.1

Table 6.0.1: Cluster2AMI table. Mean number of AMI per day of women population.

We extracted three main groups of words from the clustering; these clusters can be recognized as possible psychosocial stress factors, while the others not. However, they cannot be considered as solid representations of those factors, since there are some words of the same field that are outside those clusters. This issue may have happened as a result of a too restrictive configuration of the parameters. The three main clusters that we consider in the results are:

- Cluster #1: Football related words
- Cluster #3: Spanish and Catalan politics related words
- Cluster #5: International news related words

Finally, we extracted table 6.0.2 to sum all the statistically significant results, from which we reach the following conclusions.

In the case of the *Football related words* (Cluster #1), we found a positive association with AMI (+0.9 %). We noticed a stronger relationship in men population (+1.17 %), in AMI with initial pain between 2 p.m. and 10 p.m. (+2.11 %) and between 10 p.m. and 6 a.m (+2 %); but also in those AMI diagnosed as lateral Q-wave (+2.9 %) and in population between 40 and 59 years old (+2.92 %). The strongest associations we found are in people with a previous angioplasty (+3.46 %) and with a previous AMI (+3.8 %).

We found a stronger positive association in the *Spanish and Catalan politics related words* (Cluster #3) (+1.51 %). This relation is lower in the case of men population (+1.24 %) but higher in women population (+2.44 %). We noticed a stronger association in AMI with initial pain between 2 p.m. and 10 p.m., and between 10 p.m. and 6 a.m (+2.04 %), in people with diabetes (+2.8 %), in those AMI diagnosed as lateral Q-wave (+2.93 %) and in population between 80 and 99 years old (+4.39 %). The strongest relationship we found is in people with a medical history (+7.92 %).

Surprisingly, we found a negative association in the *International news related words* (Cluster #5) (-0.8 %), that is stronger in women population (-3.59 %), in people with medical history (-3.53 %), with a previous AMI (-3.58 %) and with a previous angioplasty (-4.28%). We discovered an exception in the case of mortal AMI, where there is a positive relationship (+1.74 %).

	Cluster #1	*Cluster #3*	*Cluster #5*
Total number of AMI	+0.9 %	+1.51 %	-0.8 %
Age range 40-59	+2.92 %	-	-
Age range 60-79	-	+1.42 %	-
Age range 80-99	-	+4.39 %	-
Men	+1.17 %	+1.24 %	-
Women	-	+2.44 %	-3.59 %
Mortal AMI	-	-	+1.74 %
Diag. Anterior Q-wave	-	-	-
Diag. Inferior Q-wave	+0.9 %	+1.76 %	-
Diag. Lateral Q-wave	+2.9 %	+2.93 %	-2.46 %
Init. Pain: Morning	-	-	-
Init. Pain: Afternoon	+2.11 %	+2.04 %	-
Init. Pain: Night	+2 %	+2.04 %	-
Diabetes	-0.62 %	+2.8 %	-
Medical History	-	+7.92 %	-3.53 %
Previous AMI	+3.8 %	-	-3.58 %
Previous Angioplasty	+3.46 %	-	-4.28 %

Table 6.0.2: Summary table of variation rates of the weighted mean number of AMI per day, for different population groups, depending on the cluster.

Chapter 7

Budget

The software engineering effort and computing resources dominate the cost of the project. That is why the main costs come from labor, but we also have to take into account the computer where we developed the software. No software licenses have been needed.

The total duration of the project has been 24 weeks, without taking into account the break (Section 1.4).

Considering my position as an intern engineer and the three supervisors as senior engineers, who have worked 1h each one every week, the labor costs are described in Table 7.0.1.

	Dedication	Wage / Hour	Total Hours	Total Cost
Intern Engineer	20 h/week	8 EUR	480 h	3,840 EUR
Senior Engineer	1 h/week	50 EUR	24 h	1,200 EUR
Senior Engineer	1 h/week	50 EUR	24 h	1,200 EUR
Senior Engineer	1 h/week	50 EUR	24 h	1,200 EUR
TOTAL				**7,440 EUR**

Table 7.0.1: Labor costs

As said before, the only equipment needed is a computer, so an approximation of its cost has been made in Table 7.0.2.

Concept	**Acquisition Cost**	**Scrap Value**	**Useful Life**	**Yearly Deprecation**	**Weekly Cost**	**Total Cost (24 weeks)**
Computer	800 EUR	266.67 EUR	5 years	106.67 EUR	2.05 EUR	49.20 EUR
TOTAL						**49.20 EUR**

Table 7.0.2: Equipment costs

The total cost of the project is **7,489.20 EUR**.

Chapter 8

Conclusions and future development

We have created a system whose objective is to find relationships between heart attacks and word clusters. This system allows to change parameter values and/or data inputs easily. Furthermore, it automatically generates plots and tables and stores them in a well-structured folder system.

The clusters are expected to contain groups of words related to psychosocial stress factors. Despite only finding three clusters which meet that condition, the results obtained for these clusters are impressive, because they demonstrate a relationship between groups of words and an increase or decrease of the number of heart attacks (Table 6.0.2).

However, those clusters are too restrictive, and the system has not grouped some words of the same fiend into the correspondent cluster. For this reason, we cannot draw global conclusions, but we can say that *a relationship exists for specific groups of words, that seem to be related to psychosocial stress factors.*

Although we have found many relationships between AMI and word clusters, no causality has been proven. Nevertheless, these results may be considered as a good starting point for further research, specially because this project is the first approach to a new way to investigate these type of relationships, since all previous studies were focused in a particular factor (or circumstance), as seen in Literature Review (Chapter 2).

Despite the great results obtained, some improvements, which I describe below, could be made in the future to improve them.

First, if the system had more input data, from other mass media accounts, a better Word Embedding would be found. Furthermore, the current version of the system is only considering the number of AMI the day a word appeared and, if it took into account the following days too, deeper conclusions could be extracted.

Finally, multi-word tokenization and proper noun detection should be implemented to achieve disambiguation of words. Moreover, a way to deal with polysemy of words could improve the whole system quality too.

Bibliography

Tatsuo Aoki, Yoshihiro Fukumoto, Satoshi Yasuda, Yasuhiko Sakata, Kenta Ito, Jun Takahashi, Satoshi Miyata, Ichiro Tsuji, and Hiroaki Shimokawa. The Great East Japan Earthquake Disaster and cardiovascular diseases. *European Heart Journal*, 33(22):2796–2803, 2012. ISSN 0195668X. doi: 10.1093/eurheartj/ehs288.

Li Bai, Qiongsi Li, Jun Wang, Eric Lavigne, Antonio Gasparrini, Ray Copes, Abderrahmane Yagouti, Richard T. Burnett, Mark S. Goldberg, Sabit Cakmak, and Hong Chen. Increased coronary heart disease and stroke hospitalisations from ambient temperatures in Ontario. *Heart*, 104(8):673–679, 2018. ISSN 1468201X. doi: 10.1136/heartjnl-2017-311821.

L Douglas Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, pages 96–103, New York, New York, USA, 1998. ACM Press. ISBN 1581130155. doi: 10.1145/290941.290970. URL `http://portal.acm.org/citation.cfm?doid=290941.290970`.

Jordi Bañeras, Ignacio Ferreira-González, Josep Ramon Marsal, José A Barrabés, Aida Ribera, Rosa Maria Lidón, Enric Domingo, Gerard Martí, and David García-Dorado. Short-term exposure to air pollutants increases the risk of ST elevation myocardial infarction and of infarct-related ventricular arrhythmias and mortality. *International Journal of Cardiology*, 250:35–42, 2018. ISSN 18741754. doi: 10.1016/j.ijcard.2017.10.004. URL `https://doi.org/10.1016/j.ijcard.2017.10.004`.

Francesco Barone-Adesi, Loredana Vizzini, Franco Merletti, and Lorenzo Richiardi. It is just a game: Lack of association between watching football matches and the risk of acute cardiovascular events. *International Journal of Epidemiology*, 39(4):1006–1013, 2010. ISSN 03005771. doi: 10.1093/ije/dyq007.

Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

F. Berthier and F. Boulay. Lower myocardial infarction mortality in French men the day

France won the 1998 World Cup of football. *Heart*, 89(5):555–556, 2003. ISSN 00070769. doi: 10.1136/heart.89.5.555.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When Is "Nearest Neighbor" Meaningful? In *Int. Conf. on Database Theory*, pages 217–235, 1999. doi: 10.1007/3-540-49257-7{_}15. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1422http://link.springer.com/10.1007/3-540-49257-7_15.

Jason S. Chi, W. Kenneth Poole, Sarah C. Kandefer, and Robert A. Kloner. Cardiovascular mortality in New York City after September 11, 2001. *American Journal of Cardiology*, 92 (7):857–861, 2003. ISSN 00029149. doi: 10.1016/S0002-9149(03)00901-9.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 160–167, New York, New York, USA, 2008. ACM Press. ISBN 9781605582054. doi: 10.1145/1390156.1390177. URL http://portal.acm.org/citation.cfm?doid=1390156.1390177.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 9 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL http://doi.wiley.com/10.1111/j.2517-6161.1977.tb01600.x.

P. Dilaveris, A. Synetos, G. Giannopoulos, E. Gialafos, A. Pantazis, and C. Stefanadis. CLimate impacts on Myocardial infarction deaths in the Athens TErritory: The CLIMATE study. *Heart*, 92(12):1747–1751, 2006. ISSN 13556037. doi: 10.1136/hrt.2006.091884.

Johannes C. Eichstaedt, Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, Megha Agrawal, Lukasz A. Dziurzynski, Maarten Sap, Christopher Weeg, Emily E. Larson, Lyle H. Ungar, and Martin E.P. Seligman. Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science*, 26(2):159–169, 2015. ISSN 14679280. doi: 10.1177/0956797614557867.

M. Ester, H.P. Kriegel, J. Sander, and Xu Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, 1996. URL http://doi.wiley.com/10.1002/widm.30.

Vincent M. Figueredo. The Time Has Come for Physicians to Take Notice: The Impact of Psychosocial Stressors on the Heart. *American Journal of Medicine*, 122(8):704–712, 2009. ISSN 00029343. doi: 10.1016/j.amjmed.2009.05.001. URL http://dx.doi.org/10.1016/j.amjmed.2009.05.001.

Caroline E. Gebhard, Catherine Gebhard, Foued Maafi, Marie Jeanne Bertrand, Barbara E. Stähli, Karin Wildi, Zurine Galvan, Aurel Toma, Zheng W. Zhang, David Smith, and Hung Q. Ly. Hockey Games and the Incidence of ST-Elevation Myocardial Infarction. *Canadian Journal of Cardiology*, 34(6):744–751, 2018. ISSN 0828282X. doi: 10.1016/j.cjca.2017.12.028.

Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016. ISSN 10769757.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1 1968. ISSN 0018-9448. doi: 10.1109/TIT.1968.1054102. URL `http://ieeexplore.ieee.org/document/1054102/`.

W. Kirkup and D. W. Merrick. A matter of life and death: Population mortality and football results. *Journal of Epidemiology and Community Health*, 57(6):429–432, 2003. ISSN 0143005X. doi: 10.1136/jech.57.6.429.

Robert A. Kloner, Scott McDonald, Justin Leeka, and W. Kenneth Poole. Comparison of Total and Cardiovascular Death Rates in the Same City During a Losing Versus Winning Super Bowl Championship. *American Journal of Cardiology*, 103(12):1647–1650, 2009. ISSN 00029149. doi: 10.1016/j.amjcard.2009.02.012. URL `http://dx.doi.org/10.1016/j.amjcard.2009.02.012`.

Hans Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011. doi: 10.1002/widm.30.

Pat Langley. The changing science of machine learning. *Machine Learning*, 82(3):275–279, 2011. ISSN 0885-6125. doi: 10.1007/s10994-011-5242-y.

Elizabeth D Liddy. Natural Language Processing, 2001.

Wenjuan Ma, Honglei Chen, Lili Jiang, Guixiang Song, and Haidong Kan. Stock volatility as a risk factor for coronary heart disease death. *European Heart Journal*, 32(8):1006–1011, 2011. ISSN 0195668X. doi: 10.1093/eurheartj/ehq495.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkeley, Calif., 1967. University of California Press.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction*

to Information Retrieval. Cambridge University Press, Cambridge, 2009. ISBN 9780511809071. doi: 10.1017/CBO9780511809071. URL http://ebooks.cambridge.org/ref/id/CBO9780511809071.

Lluís Màrquez. Machine Learning and Natural Language Processing. Technical report, Centre de recerca TALP, Universitat Politècnica de Catalunya, Barcelona, 2000.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2 2018. ISSN 2475-9066. doi: 10.21105/joss.00861. URL http://arxiv.org/abs/1802.03426.

Tomas Mikolov. Google Groups: de-obfuscated Python + question, 2013. URL https://groups.google.com/d/msg/word2vec-toolkit/NLvYXU99cAM/E5ld8LcDxlAJ.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. 1 2013a. URL http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *IEEE transactions on neural networks*, 14(6):1569–72, 10 2013b. URL http://arxiv.org/abs/1310.4546.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. ISBN 978-0-07-042807-2.

Dunja Mladenic. Automatic word lemmatization. *Proceedings of the 5th International Multi-Conference Information Society (IS-2002)*, pages 153–159, 2002.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. Joint Lemmatization and Morphological Tagging with Lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, number September, pages 2268–2274, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1272. URL http://aclweb.org/anthology/D15-1272.

David Niederseer, Christoph W. Thaler, Andreas Egger, Michaela C. Niederseer, Martin Plöderl, and Josef Niebauer. Watching soccer is not associated with an increase in cardiac events. *International Journal of Cardiology*, 170(2):189–194, 2013. ISSN 01675273. doi: 10.1016/j.ijcard.2013.10.066. URL http://dx.doi.org/10.1016/j.ijcard.2013.10.066.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume s5-IV, pages 1532–1543, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics. ISBN 9781937284961. doi: 10.3115/v1/D14-1162. URL http://aclweb.org/anthology/D14-1162.

Joël Plisson, Nada Lavrac, and Dr. Dunja Mladenić. A rule based approach to word lemmatization. *Proceedings of the 7th International Multiconference Information Society (IS'04)*, (November):83–86, 2004. URL http://eprints.pascal-network.org/archive/00000715/.

M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 3 1980. ISSN 0033-0337. doi: 10.1108/eb046814. URL http://www.emeraldinsight.com/doi/10.1108/eb046814.

A. Craig Reynolds. The conference on mechanical translation. *Mechanical Translation*, 1(3): 47–55, 1954.

A. L. Samuel. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3):210–229, 7 1959. ISSN 0018-8646. doi: 10.1147/rd.33.0210. URL http://ieeexplore.ieee.org/document/5392560/.

G. V. Trunk. A Problem of Dimensionality: A Simple Example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(3):306–307, 7 1979. ISSN 0162-8828. doi: 10.1109/TPAMI.1979.4766926. URL http://ieeexplore.ieee.org/document/4766926/.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. ISSN 02624079. doi: 10.1007/s10479-011-0841-3.

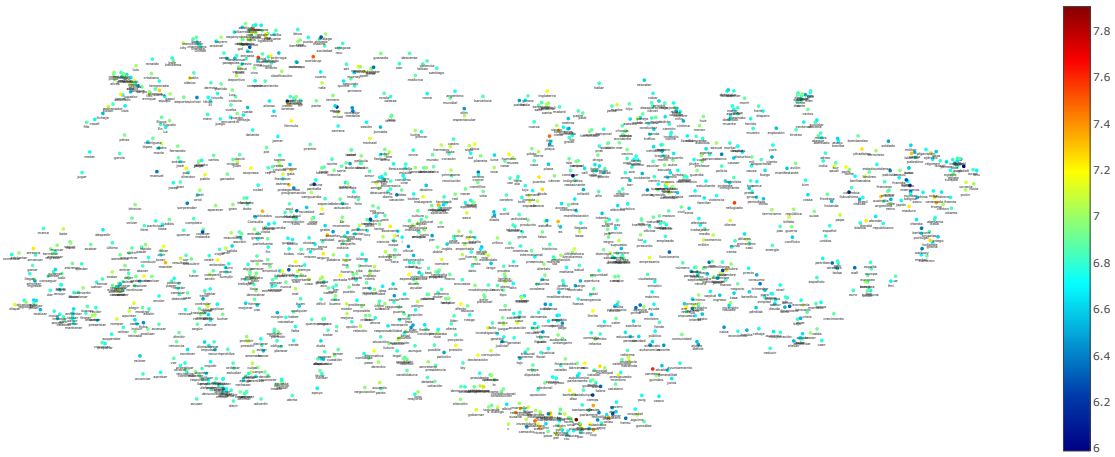
Alain Vanasse, Denis Talbot, Fateh Chebana, Diane Bélanger, Claudia Blais, Philippe Gamache, Jean Xavier Giroux, Roxanne Dault, and Pierre Gosselin. Effects of climate and fine particulate matter on hospitalizations and deaths for heart failure in elderly: A population-based cohort study. *Environment International*, 106(July):257–266, 2017. ISSN 18736750. doi: 10.1016/j.envint.2017.06.001. URL http://dx.doi.org/10.1016/j.envint.2017.06.001.

Ute Wilbert-Lampen, David Leistner, Sonja Greven, Tilmann Pohl, Sebastian Sper, Christoph Völker, Denise Güthlin, Andrea Plasse, Andreas Knez, Helmut Küchenhoff, and Gerhard Steinbeck. Cardiovascular Events during World Cup Soccer. *New England Journal of Medicine*, 358(5):475–483, 2008. ISSN 0028-4793. doi: 10.1056/NEJMoa0707427. URL http://www.nejm.org/doi/abs/10.1056/NEJMoa0707427.

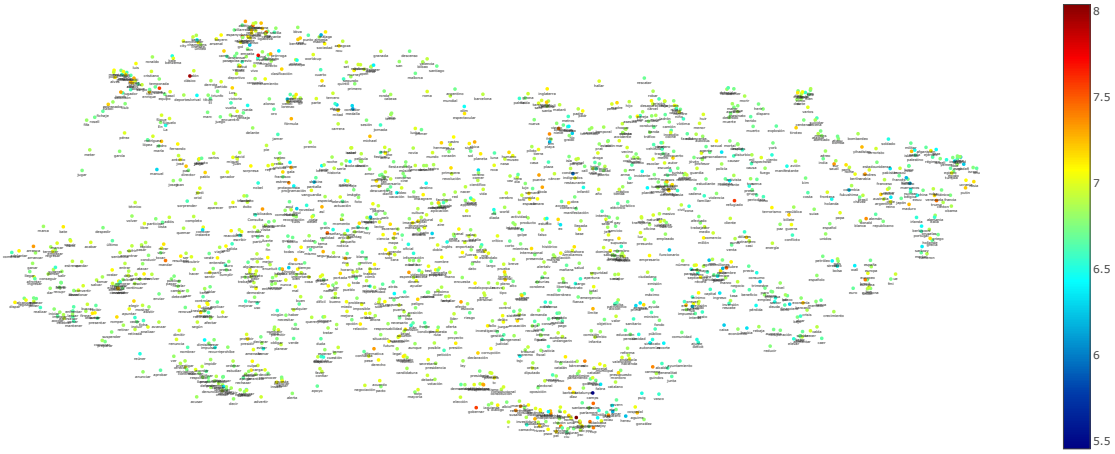
D. R Witte, M. L Bots, A. W Hoes, and D. E Grobbee. Cardiovascular mortality in Dutch men during 1996 European football championship: longitudinal population study. *Bmj*, 321(7276):1552–1554, 2000. ISSN 0959-8138. doi: 10.1136/bmj.321.7276.1552. URL http://www.bmj.com/cgi/doi/10.1136/bmj.321.7276.1552.

Appendix A

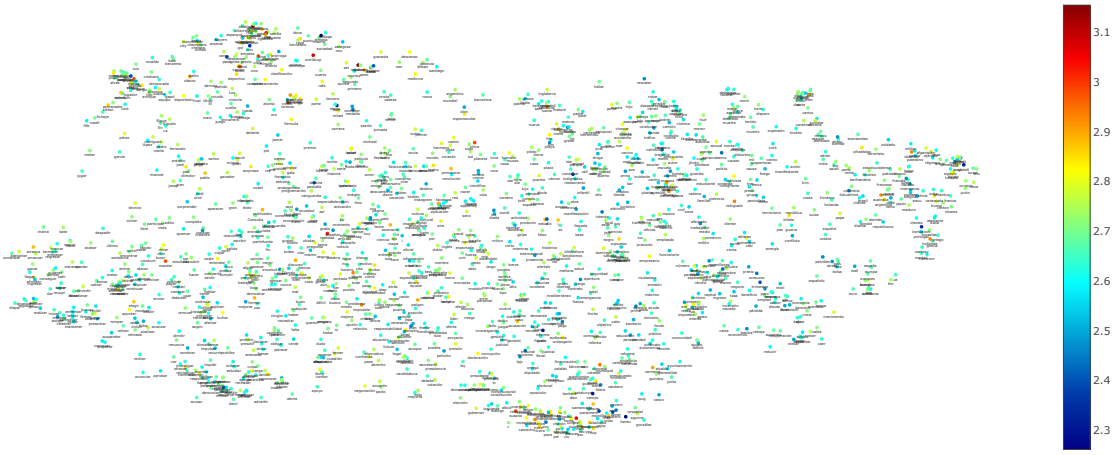
Heart attacks associated to words



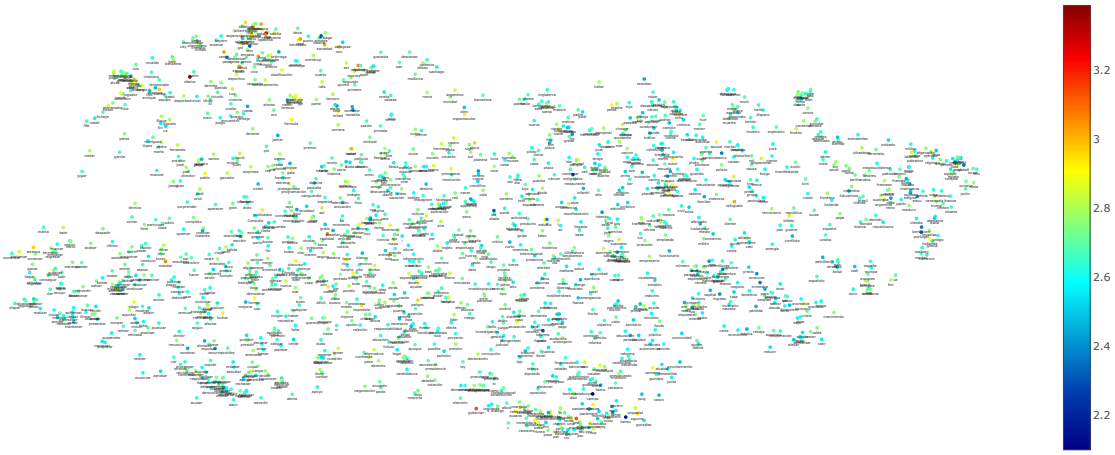
Word2AMI scatter plot. Mean number of AMI per day.



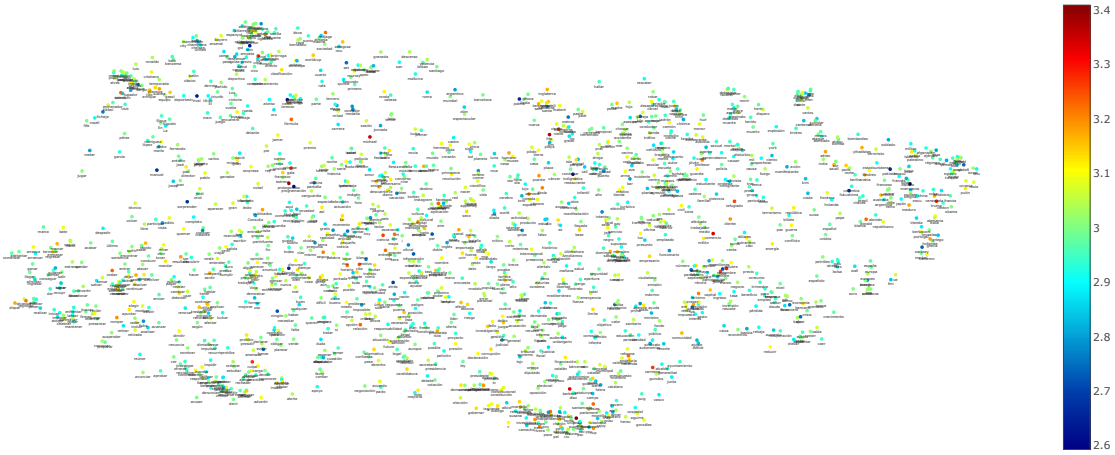
Word2AMI scatter plot. Weighted mean number of AMI per day.



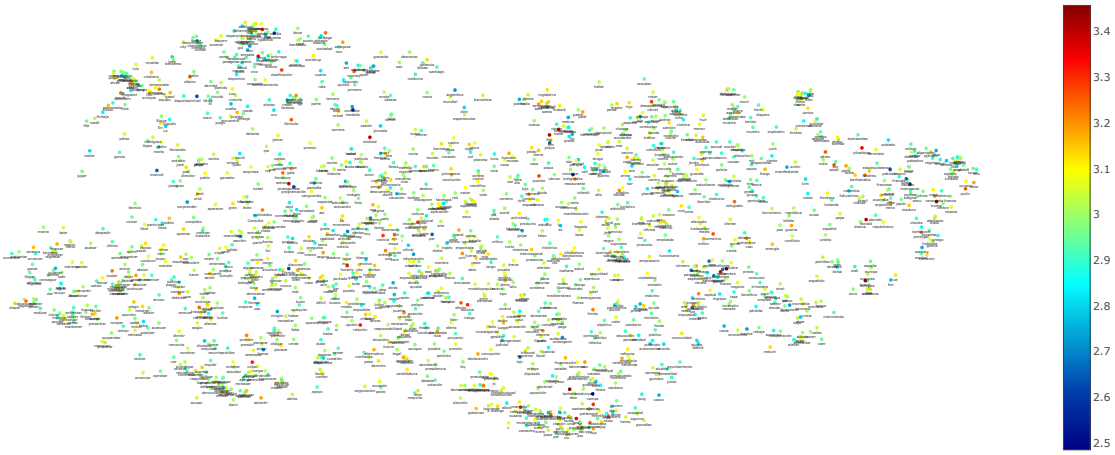
Word2AMI scatter plot. Mean number of AMI per day of population between 40 and 59 years old.



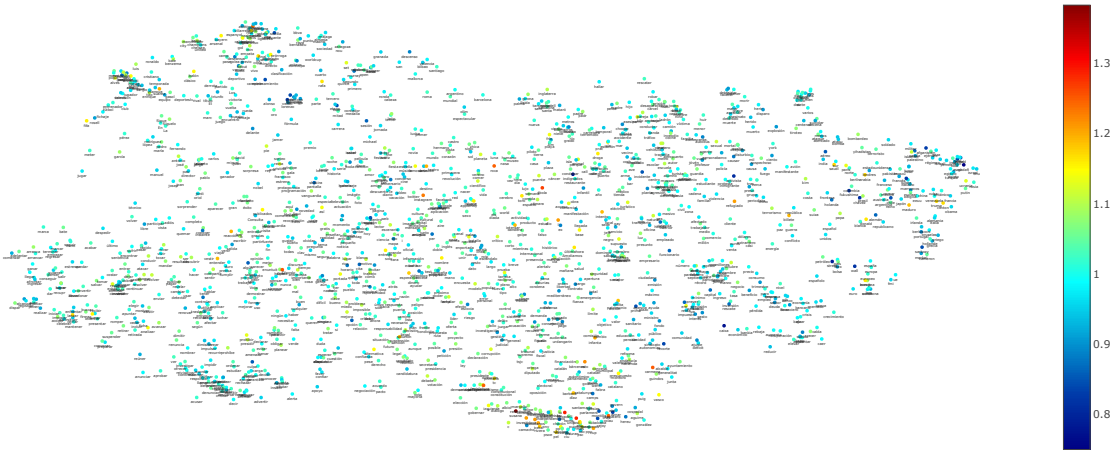
Word2AMI scatter plot. Weighted mean number of AMI per day of population between 40 and 59 years old.



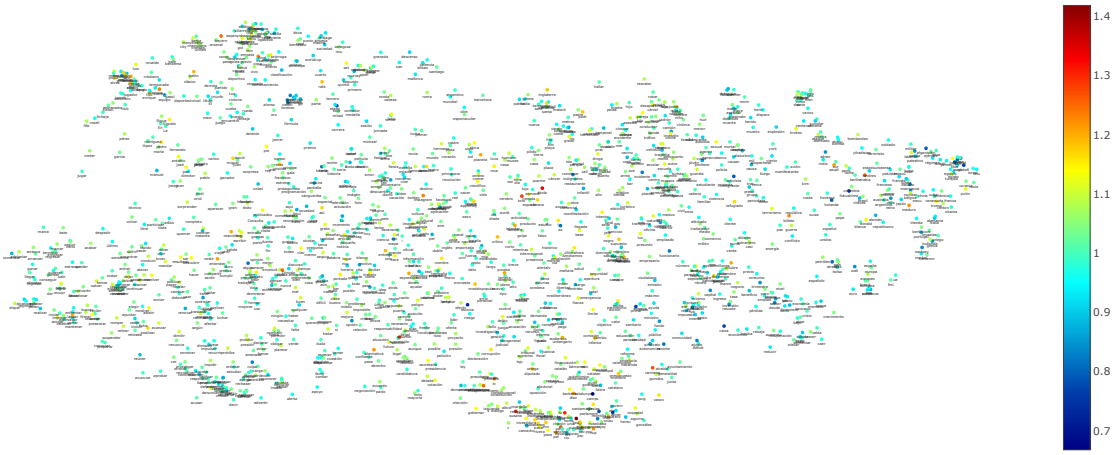
Word2AMI scatter plot. Mean number of AMI per day of population between 60 and 79 years old.



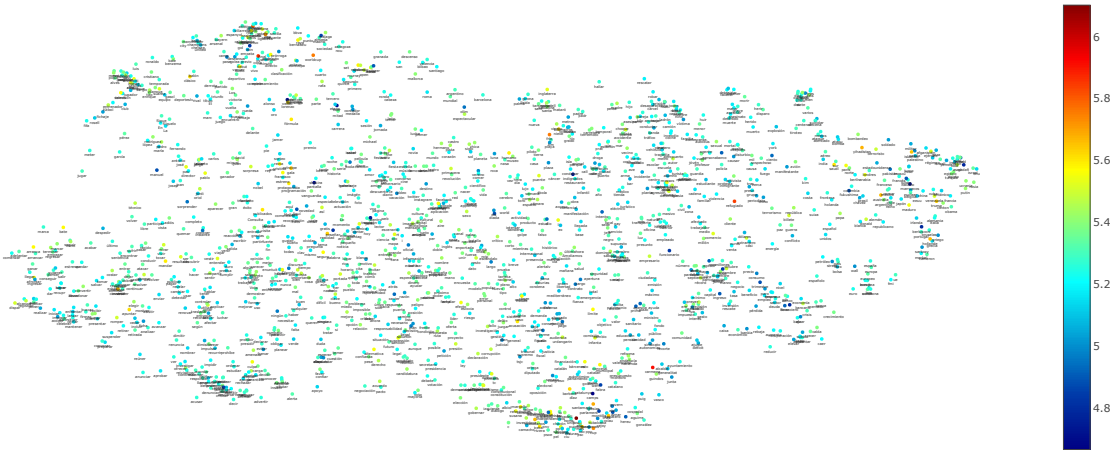
Word2AMI scatter plot. Weighted mean number of AMI per day of population between 60 and 79 years old.



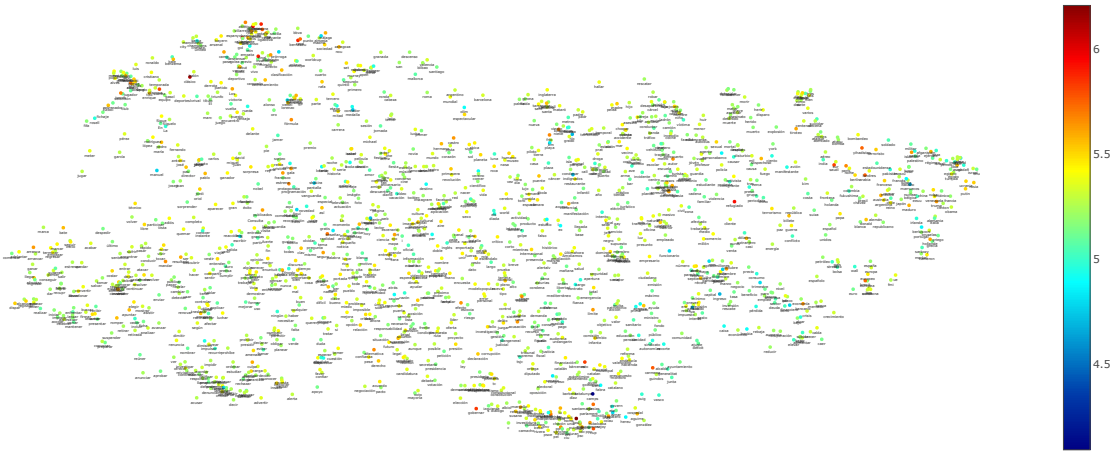
Word2AMI scatter plot. Mean number of AMI per day of population between 80 and 99 years old.



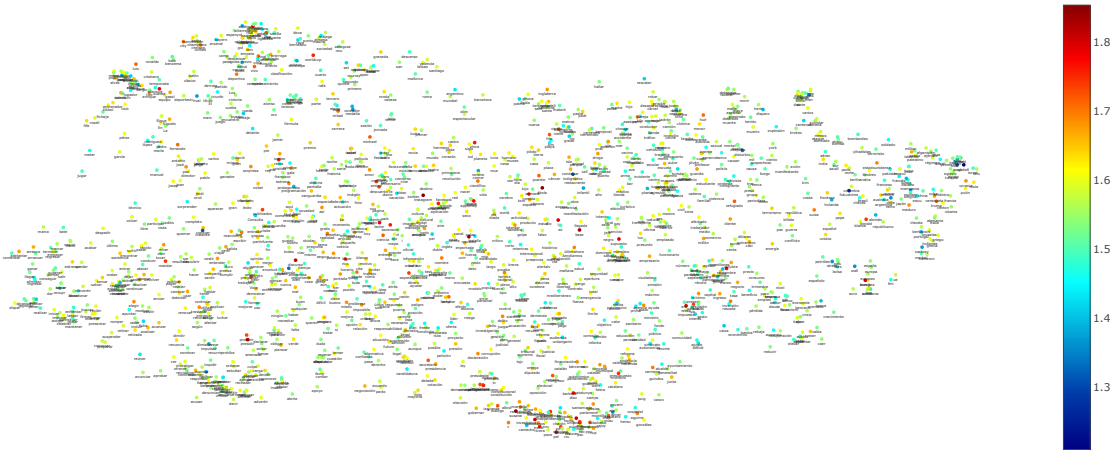
Word2AMI scatter plot. Weighted mean number of AMI per day of population between 80 and 99 years old.



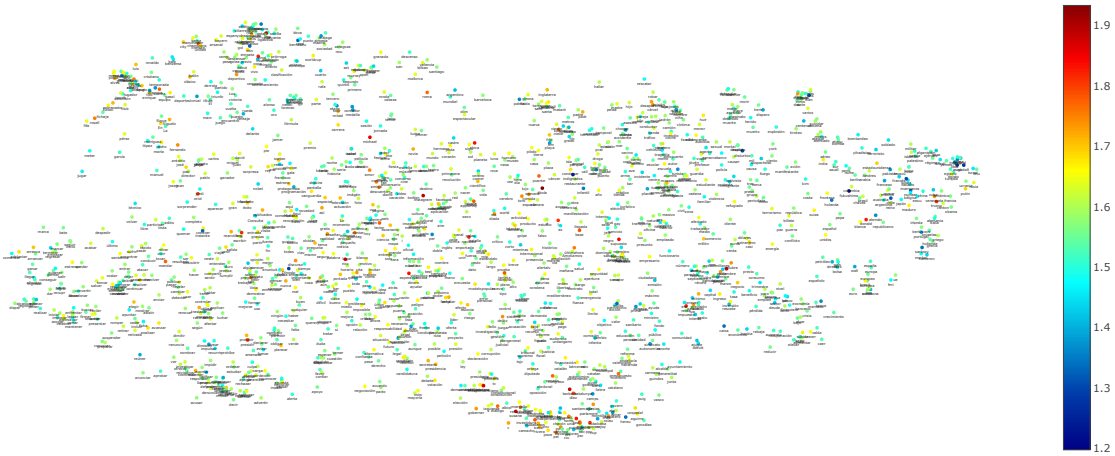
Word2AMI scatter plot. Mean number of AMI per day of men population.



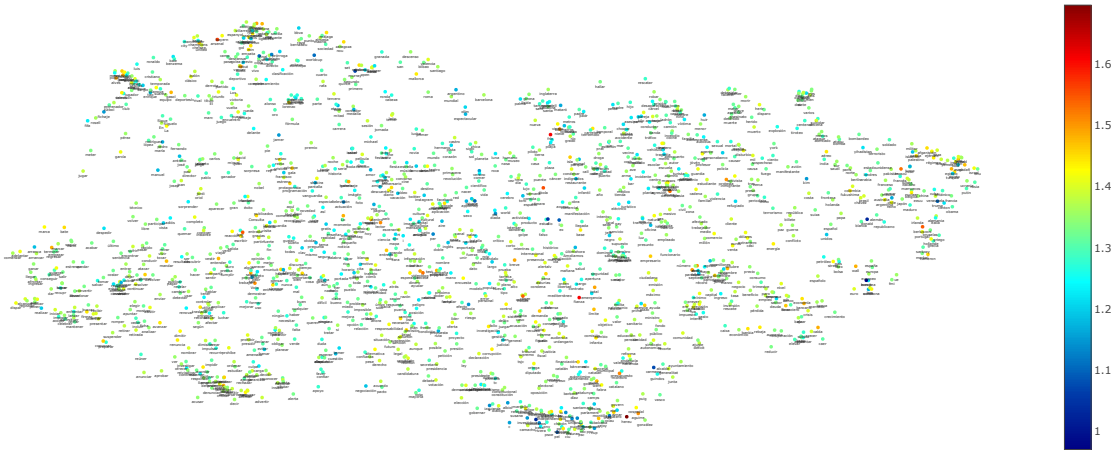
Word2AMI scatter plot. Weighted mean number of AMI per day of men population.



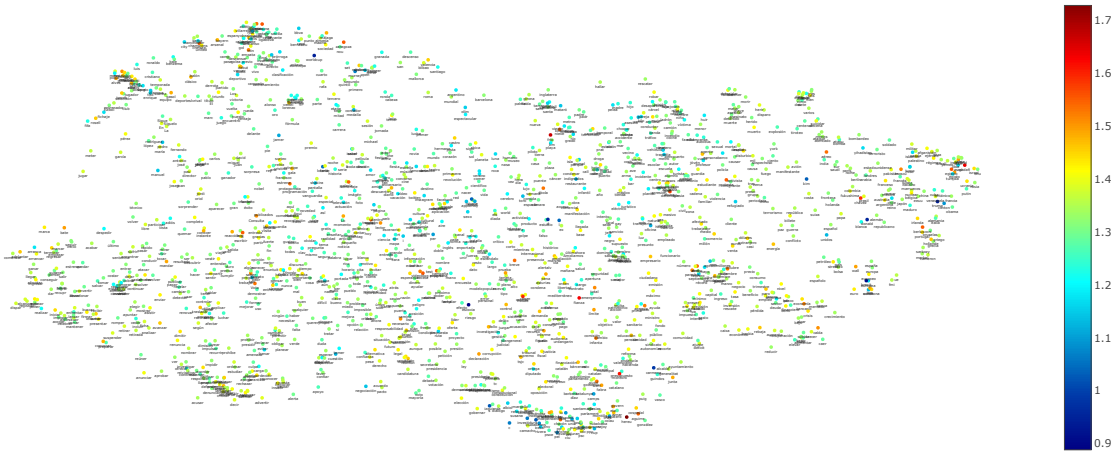
Word2AMI scatter plot. Mean number of AMI per day of women population.



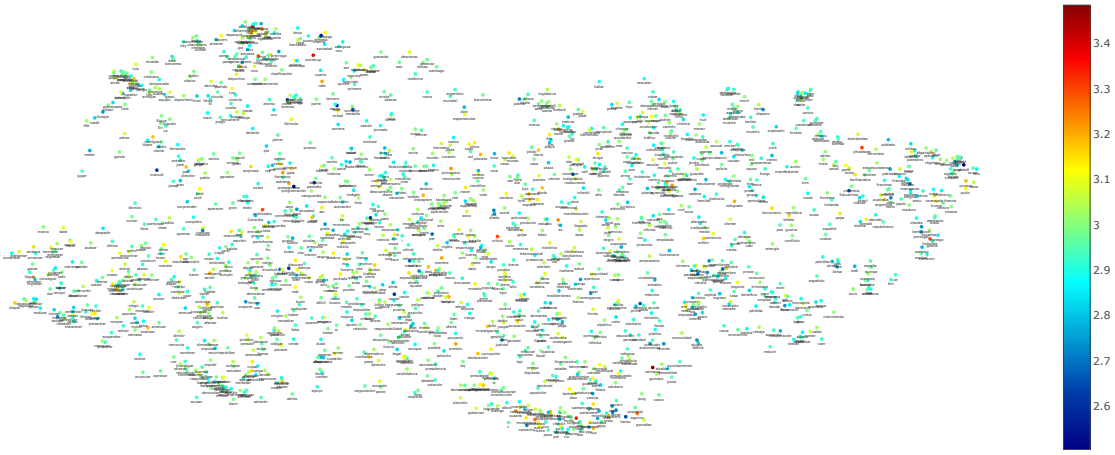
Word2AMI scatter plot. Weighted mean number of AMI per day of women population.



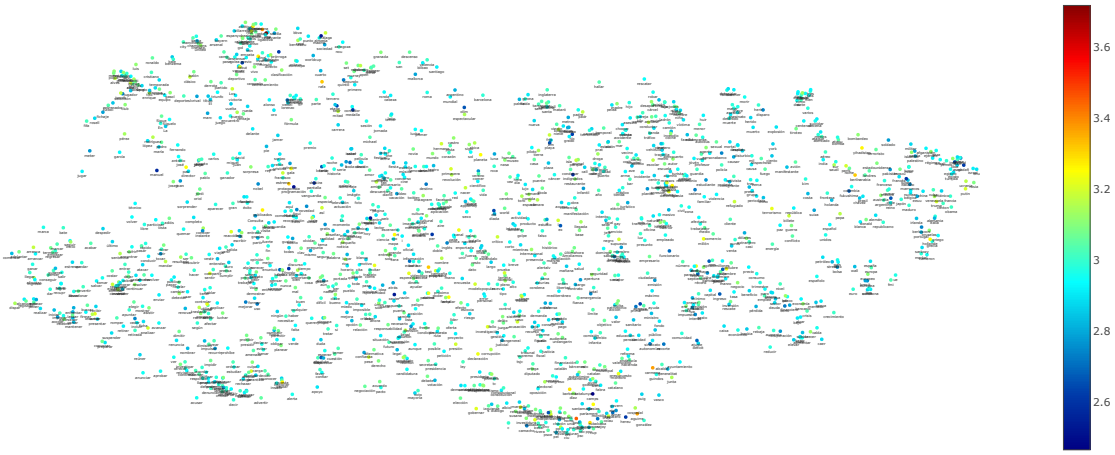
Word2AMI scatter plot. Mean number of mortal AMI per day.



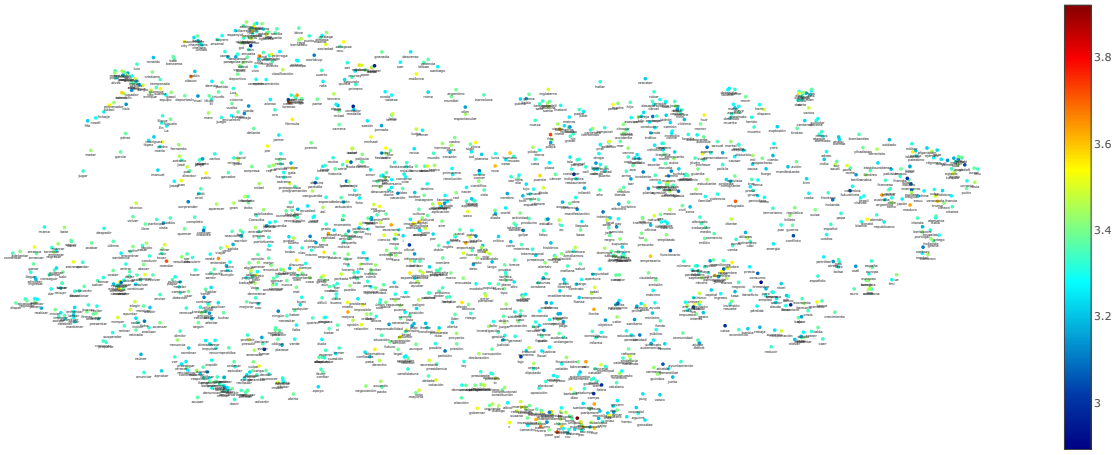
Word2AMI scatter plot. Weighted mean number of mortal AMI per day.



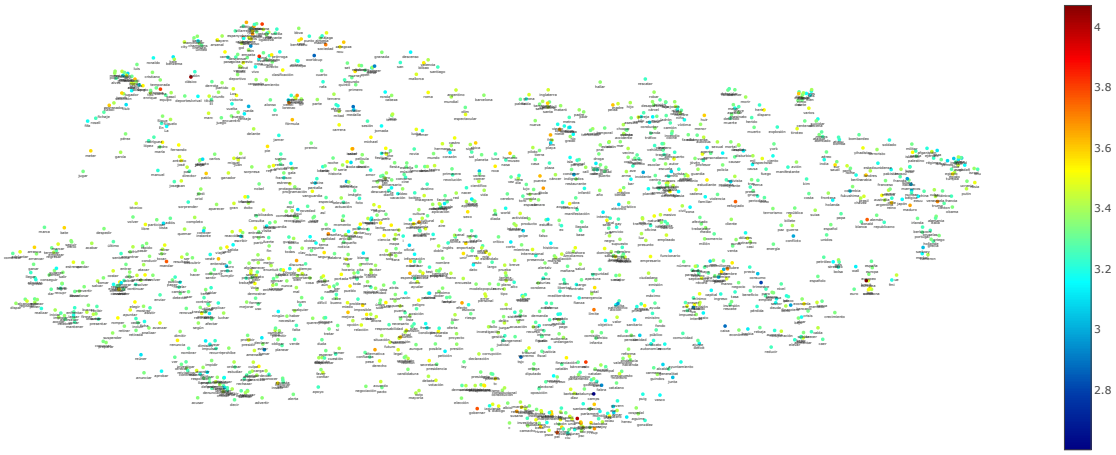
Word2AMI scatter plot. Mean number of AMI, diagnosed as anterior Q-wave, per day.



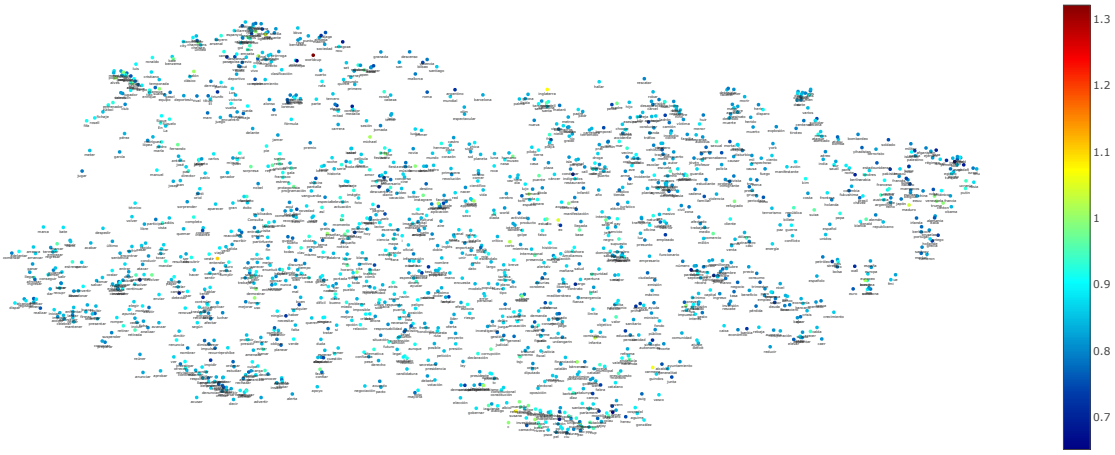
Word2AMI scatter plot. Weighted mean number of AMI, diagnosed as anterior Q-wave, per day.



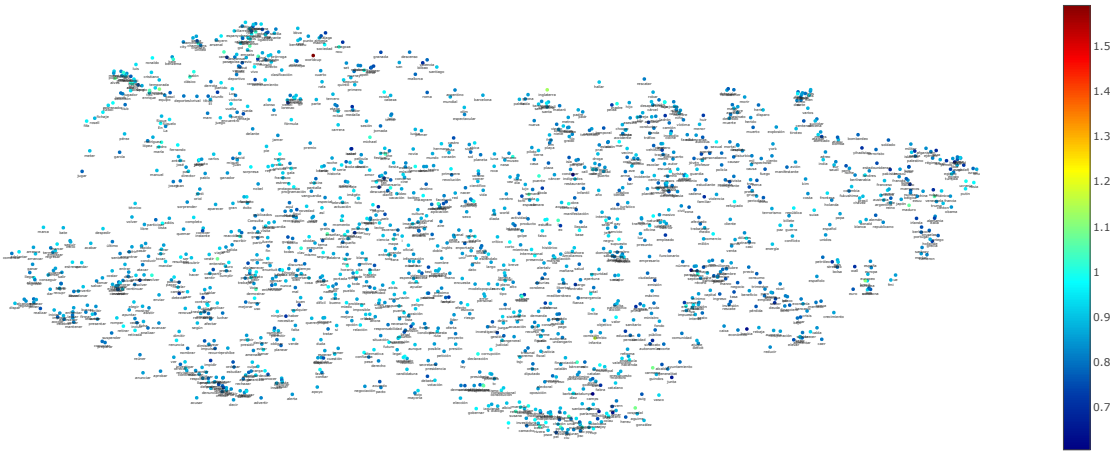
Word2AMI scatter plot. Mean number of AMI, diagnosed as inferior Q-wave, per day.



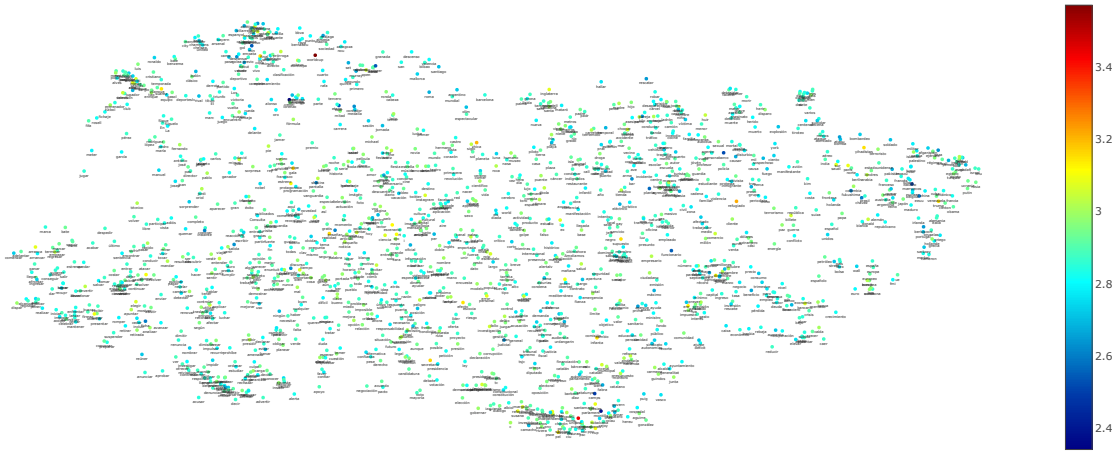
Word2AMI scatter plot. Weighted mean number of AMI, diagnosed as inferior Q-wave, per day.



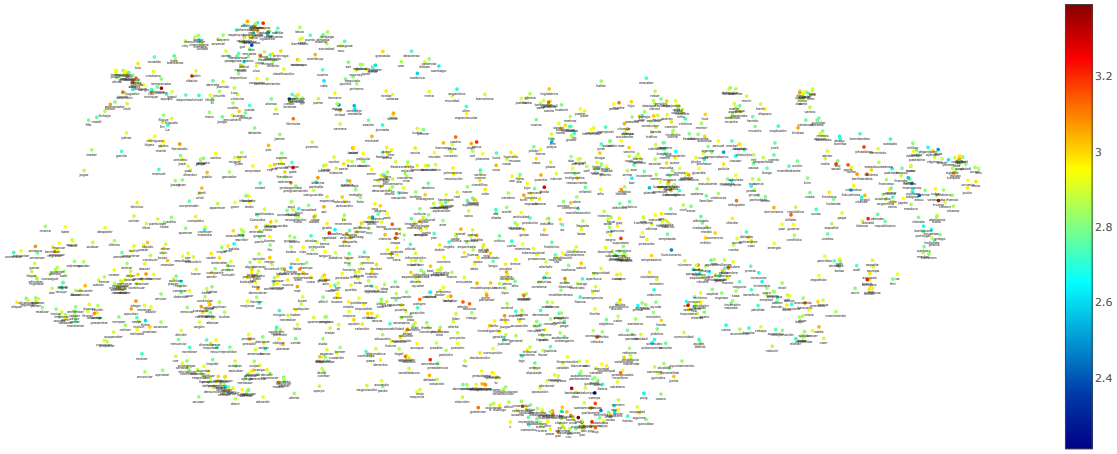
Word2AMI scatter plot. Mean number of AMI, diagnosed as lateral Q-wave, per day.



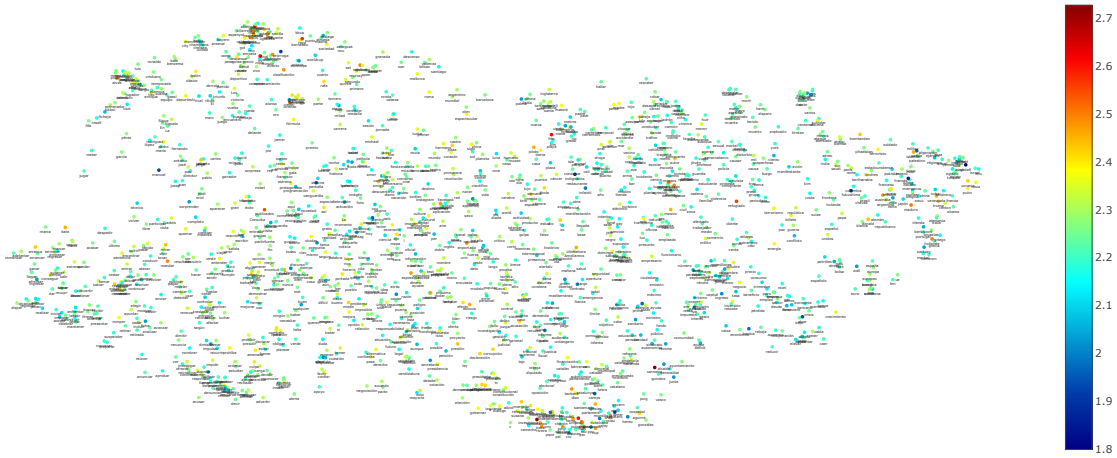
Word2AMI scatter plot. Weighted mean number of AMI, diagnosed as lateral Q-wave, per day.



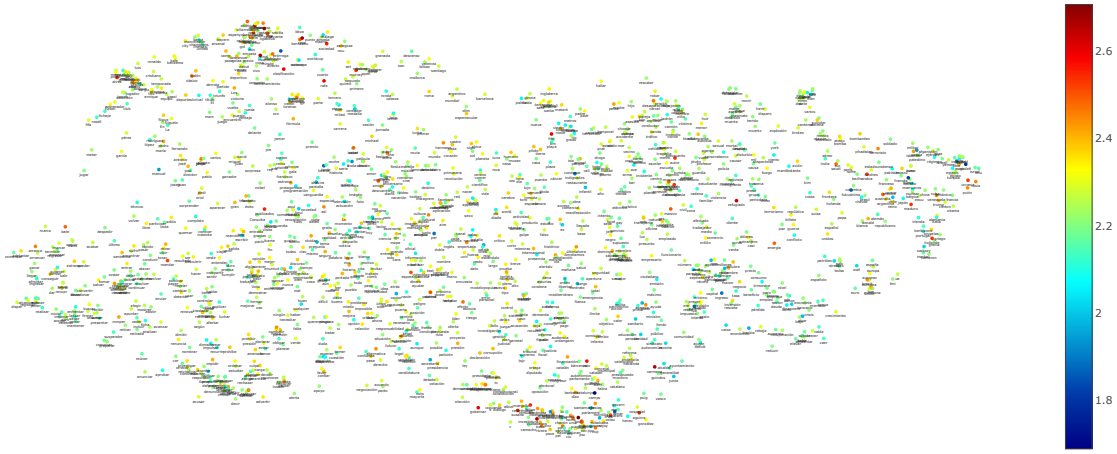
Word2AMI scatter plot. Mean number of AMI per day, with initial pain between 6 a.m. and 2 p.m.



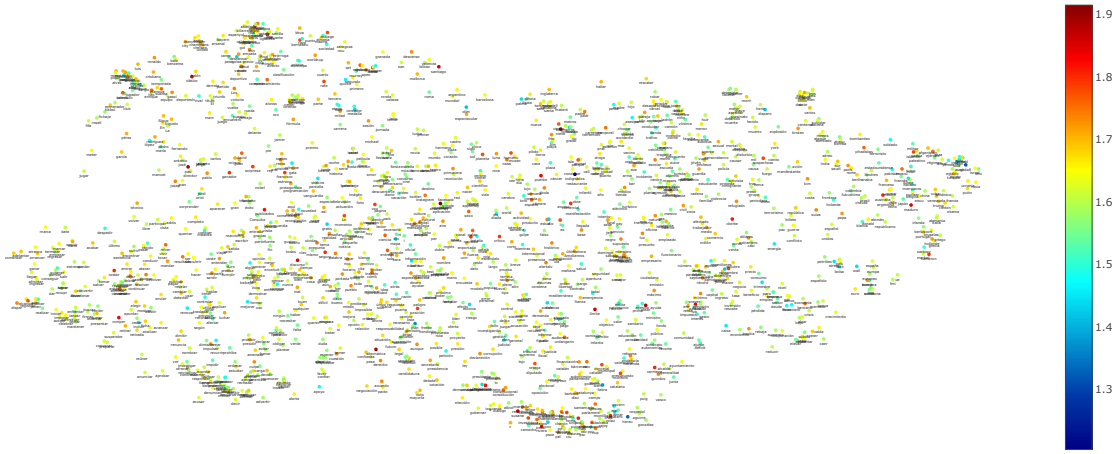
Word2AMI scatter plot. Weighted mean number of AMI per day, with initial pain between 6 a.m. and 2 p.m.



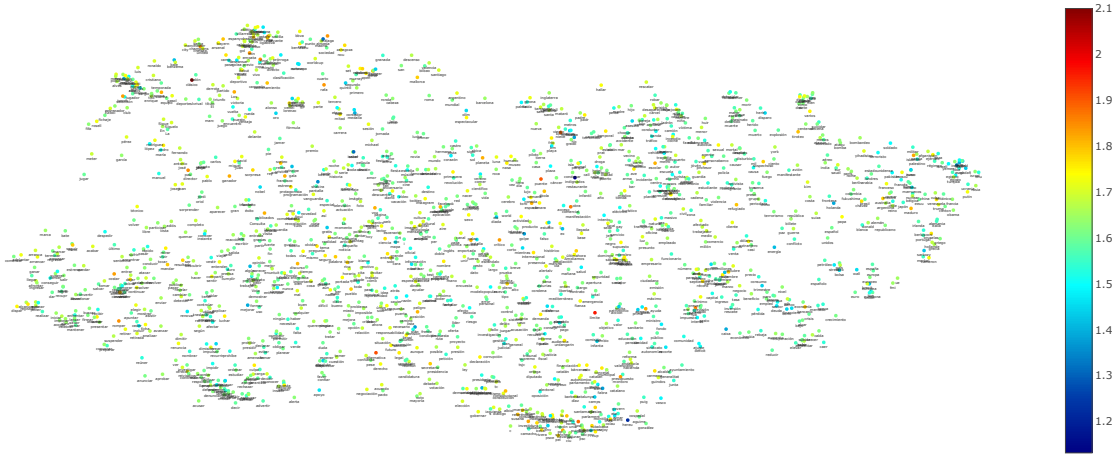
Word2AMI scatter plot. Mean number of AMI per day, with initial pain between 2 p.m. and 10 p.m.



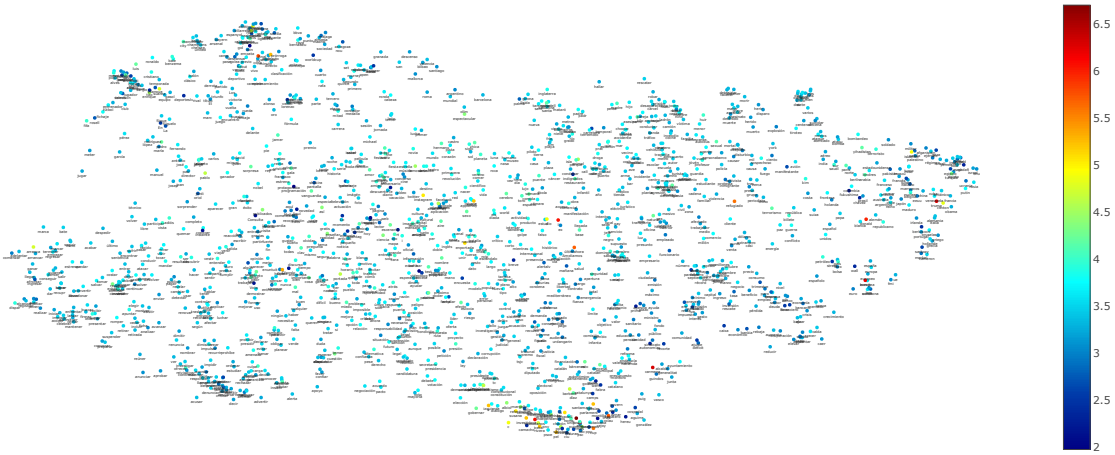
Word2AMI scatter plot. Weighted mean number of AMI per day, with initial pain between 2 p.m. and 10 p.m.



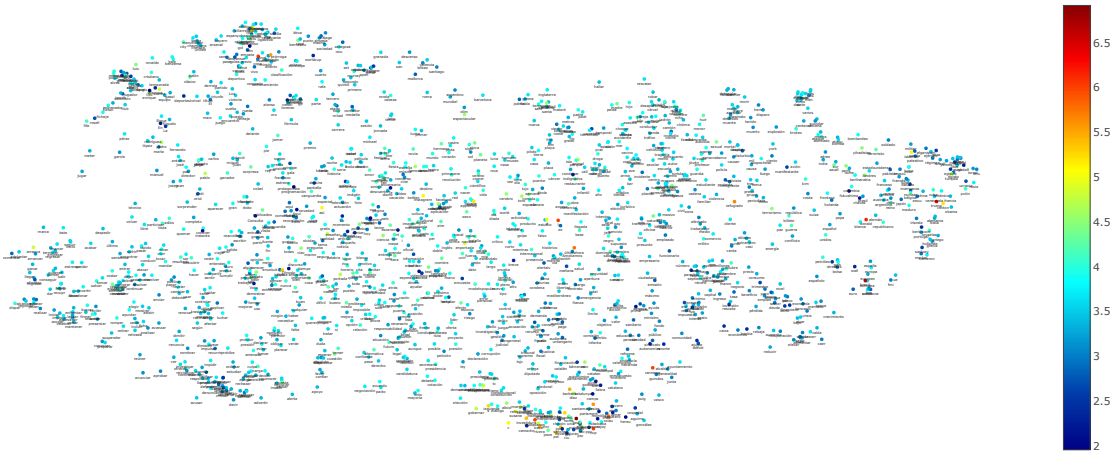
Word2AMI scatter plot. Mean number of AMI per day, with initial pain between 10 p.m. and 6 a.m.



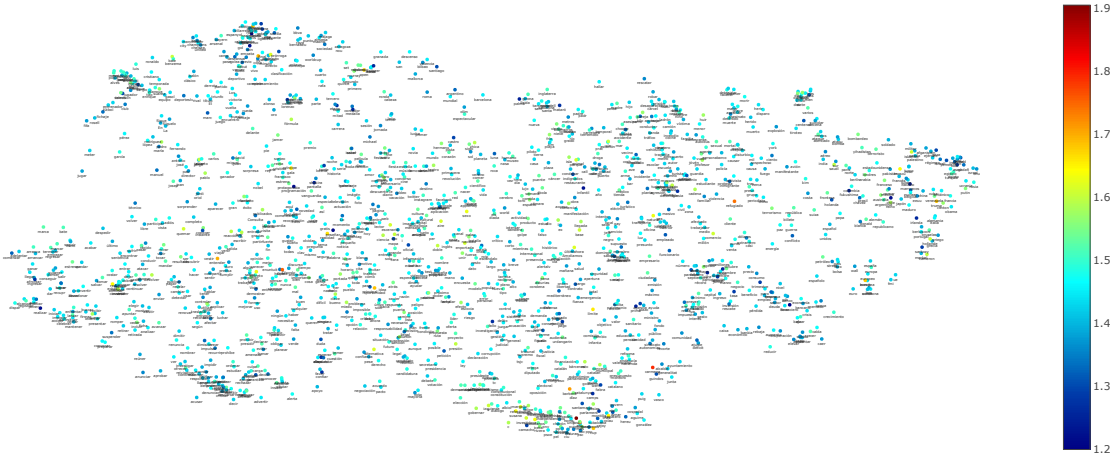
Word2AMI scatter plot. Weighted mean number of AMI per day, with initial pain between 10 p.m. and 6 a.m.



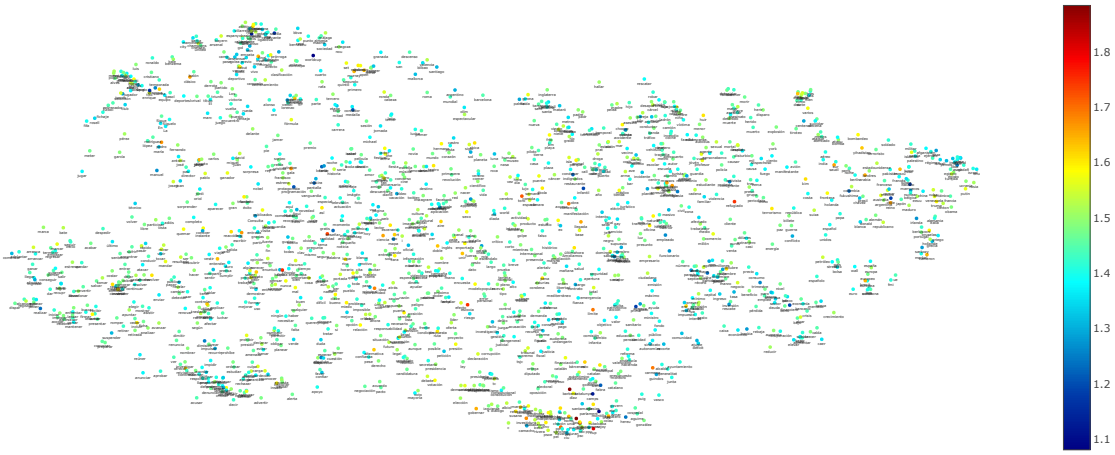
Word2AMI scatter plot. Mean number of AMI per day, from people with medical history.



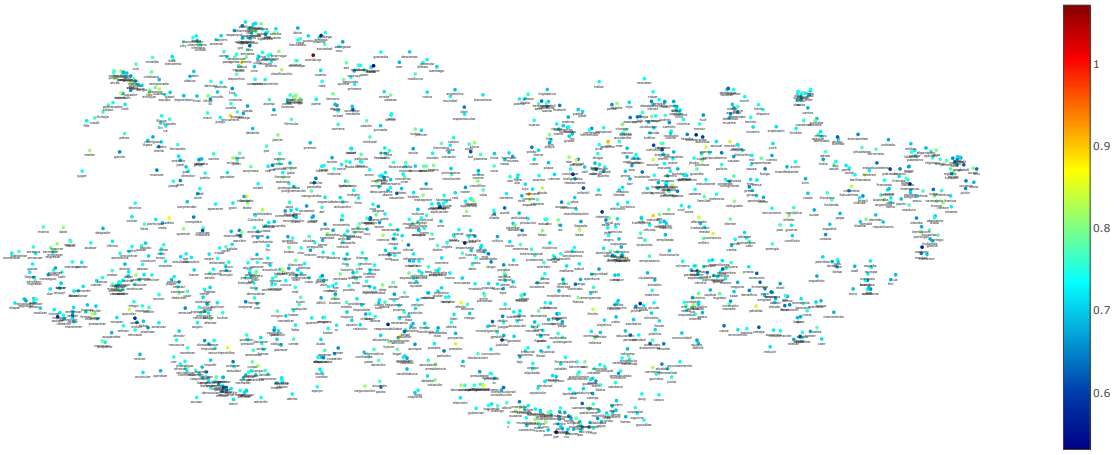
Word2AMI scatter plot. Weighted mean number of AMI per day, from people with medical history.



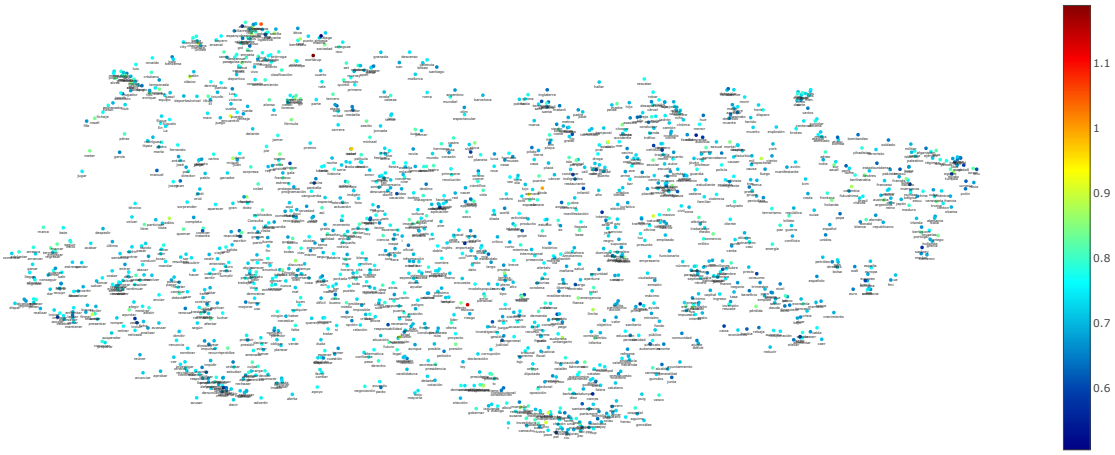
Word2AMI scatter plot. Mean number of AMI per day, from people with diabetes.



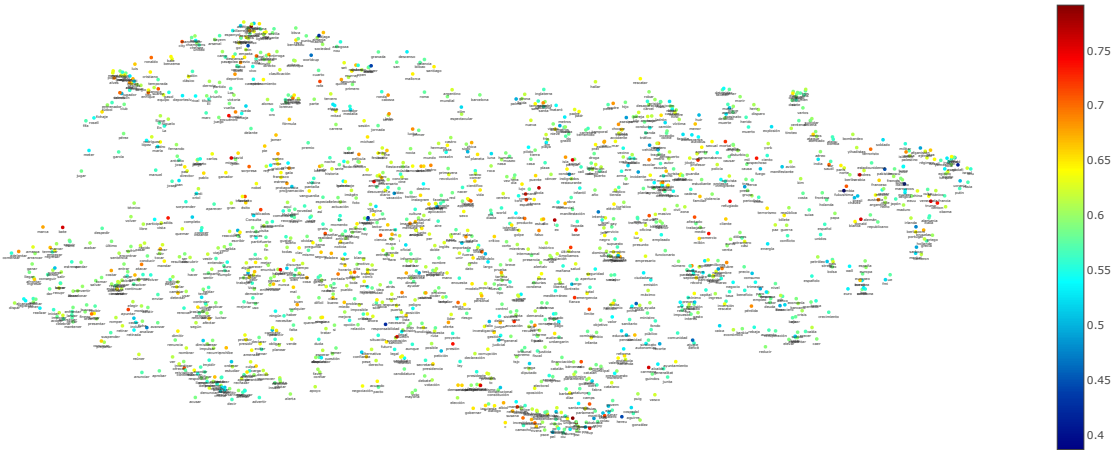
Word2AMI scatter plot. Weighted mean number of AMI per day, from people with diabetes.



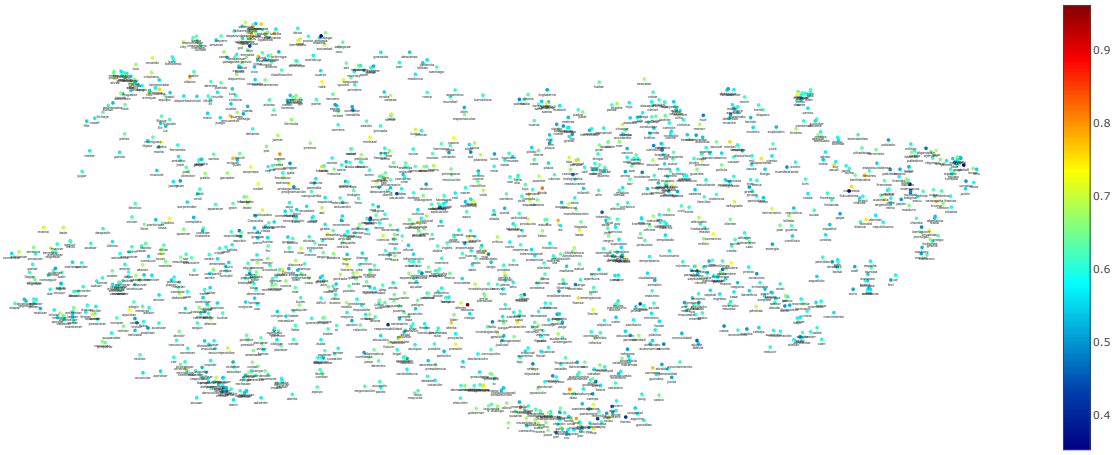
Word2AMI scatter plot. Mean number of AMI per day, from people with a previous AMI.



Word2AMI scatter plot. Weighted mean number of AMI per day, from people with a previous AMI.



Word2AMI scatter plot. Mean number of AMI per day, from people with a previous angioplasty.



Word2AMI scatter plot. Weighted mean number of AMI per day, from people with a previous angioplasty.

Appendix B

Clustering

Cluster #1 (105)
madrid, fútbol, bale, málaga, entrenador, atlético, villarreal, manchester, city, barça, sevilla, golea, pase, getafe, minuto, equipo, guardiola, partido, atleti, bernabéu, empate, espanyol, punto, directo, cuarto, gol, messi, ronaldo, xavi, título, mourinho, bayern, liderato, final, luis, real, selección, cristiano, campeón, valencia, zaragoza, deportivo, penalti, chelsea, torres, osasuna, balón, levante, almería, athletic, alves, piqué, derrota, sociedad, club, copa, debut, calderón, camp, nou, previo, prórroga, octavo, liga, granada, betis, clasificación, clásico, iniesta, vencer, bosque, arsenal, cesc, united, bbva, jugador, fichaje, rayo, vivo, semifinal, sergio, duelo, ramos, champions, diego, suárez, rival, league, enrique, descansar, empata, neymar, vilanova, descanso, benzema, alba, casillas, ancelotti, celta, vs, tito, min, realmadrid, ligabbva, endirecto

Cluster #2 (50)
decir, creer, criticar, afirmar, admitir, considr, anunciar, ver, defender, abogar, pedir, advertir, culpa, lamentar, asegurar, plantear, proponer, prometer, reivindicar, aprobar, alerta, acusar, insta, exigir, apuesta, apoyar, ofrecer, reclamar, denunciar, negociar, reconocer, garantizar, descartar, destacar, decidir, rechazar, confirmar, estudiar, pactar, negar, carga, acordar, aceptar, insistir, avisar, favor, calificar, dispuestar, ordenar, impulsar

Cluster #3 (62)
aguirre, zapatero, pp, rajoy, montilla, felipe, gonzález, ciu, psoe, fernández, referéndum, cospedal, rubalcaba, mas, psc, independentisto, icv, izquierda, duran, govern, gobernar, eta, iu, diálogo, pnv, erc, albiol, ppc, camacho, aznar, hereu, rivera, cdc, sánchez, díaz, iglesias, navarro, parlament, investidura, independencia, chacón, santamaría, iceta, trias, c, consulta, soberanista, homs, s, pel, cup, susana, unió, ciudadanos, bildu, puigdemont, junqueras, colau, margallo, anc, forcadell, jxsí

Cluster #4 (15)
seis, siete, menos, dos, diez, cuatro, tres, centenar, ocho, varios, decena, cinco, nueve, cien, once

Cluster #5 (55)
china, venezuela, cuba, pakistán, afganistán, rusia, italiano, militar, británico, qaeda, iraq, marruecos, alemania, nuclear, soldado, estadounidense, ruso, francia, franceso, méxico, ejército, moscú, terrorista, israel, colombia, francés, berlín, israelí, reino, onu, parís, japón, corear, yemen, eeuu, bélgica, rebeld, egipcio, otan, gaza, turco, egipto, palestino, islámico, régimen, turquía, putin, gadafi, túnez, siria, sirio, libia, ucrania, asad, yihadistas

Cluster #6 (20)
jóvén, hombre, persona, menor, niño, hospital, sexual, pareja, hijo, familia, mujer, joven, anciano, bebé, hija, perro, niña, ladrón, vecino, adolescente

Cluster #7 (14)
bajo, caer, dispar, subir, elevar, caída, aumentar, reducir, bajar, crecer, subida, aumento, alzar, rebaja

Cluster #8 (19)
trimestre, prever, diciembre, septiembre, participación, emisión, noviembre, número, octubre, junio, agosto, mes, enero, abril, julio, febrero, mayo, marzo, cifra

Cluster #9 (16)
david, pérez, pablo, jorge, josé, fernando, garcía, carlos, antonio, maría, manuel, lópez, juan, javier, miguel, alberto, rodríguez

Cluster #10 (34)
buscar, presentar, llegar, cerrar, llevar, recibir, alcanzar, salir, realizar, abrir, poner, mantener, afrontar, iniciar, celebrar, supr, conseguir, lanzar, acoger, avanzar, convocar, lograr, convertir, recupr, colocar, situar, imponer, suspender, retirar, sumar, ceder, enfrentar, levantar, firmar

Cluster #11 (23)
ir, peor, saber, bueno, siempre, nunca, buen, tan, nadie, hecho, her, malo, difícil, gente, aún, bien, mal, pensar, cosa, mismo, alguna, quién, algún

Cluster #12 (12)
adelantar, seguir, comenzar, marca, inicio, marcar, despedir, terminar, empezar, arrancar, bate, arranca

Cluster #13 (12)
tormenta, temporal, lluvia, metros, frío, viento, km, terremoto, grado, nieve, kilómetros, ola

Cluster #14 (13)
canción, historia, cine, the, película, fan, disco, moda, serie, amor, música, oscar, shakira

Cluster #15 (12)

red, facebook, internet, twitter, móvil, google, aplicación, tecnología, iphone, usuario, instagram, whatsapp

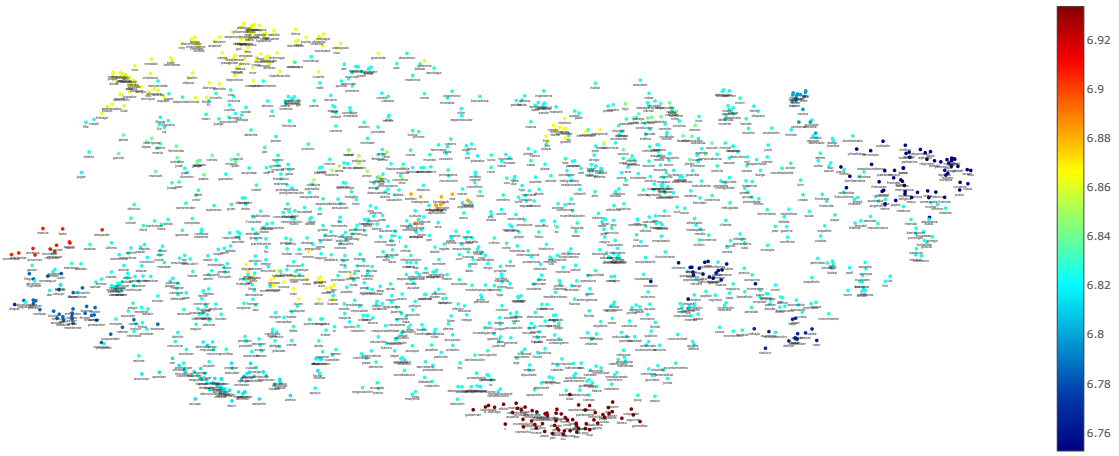
Cluster #0 (1027, "other")

llamar, invertir, nuevo, presidente, poder, dinero, clav, escándalo, har, déficit, elección, català, responder, fórmula, pregunta, respuesta, tener, sólo, ser, si, demostrar, interés, español, marc, márquez, papa, gasol, presunto, asesino, robar, banco, hacer, año, querer, semana, nasa, éxito, acto, sanitario, mundo, según, obama, nueva, york, visita, barcelona, rescatar, vettel, favorito, vida, tras, perder, victoria, detener, conducir, jugar, mejor, temporada, roma, acabar, palau, ronda, wall, street, regresar, nivel, tomar, leer, dejar, aire, matar, carretera, bomba, unidos, incidente, gran, contrato, aeropuerto, conflicto, europa, ruta, santiago, haber, alternativa, proyecto, propio, papel, crear, marcha, político, actuación, españa, caso, avión, pasajero, centro, muerto, ciudadano, hallar, pleno, parte, plan, seguridad, morir, ex, costa, grupo, catalano, medio, deber, opinión, mayoría, víctima, terrorismo, obra, coch, objetivo, economía, registrar, crecimiento, tercer, pib, sant, joan, retirada, problema, corazón, comisión, europea, modelo, negocio, digital, explosión, demanda, último, vuelo, después, michael, venta, congreso, eliminar, gas, luz, roca, trabajo, segundo, primero, libre, preguntar, jamer, londres, gustar, miembro, san, herido, ataque, suicidar, noticia, abandonar, posible, delito, libertad, gobierno, premio, nacional, claro, jornada, consumo, multa, euros, abuso, romper, trabajar, preparar, fiscal, colegio, vivir, país, rico, día, televisión, noche, nombrar, ere, afectar, trabajador, dar, mayor, electoral, legislatura, cambio, primo, riesgo, bono, volver, test, incendio, casa, sur, escuela, recuperación, fallecer, accidente, aéreo, constitución, experto, amigo, conocer, ganar, caja, sabadell, cameron, quedar, policía, alonso, turismo, ingreso, josep, puig, i, ayudar, estación, pueblo, constitucional, toro, precio, ue, detenidar, tratar, fiscalía, cambiar, madre, planear, lorenzo, tiempo, vuelta, oro, duda, comercio, lleida, amenaza, cárcel, ningún, generalitat, servicio, mínimo, detenido, més, empleo, tasa, paro, vídeo, revelar, estrella, festival, cadena, atentado, martes, pasado, reacción, falta, periodista, juicio, activista, voto, prisión, error, badalona, educación, paso, hacia, unión, vez, tierra, renuncia, derecho, tráfico, transporte, fin, huelga, absoluto, oriol, pujol, nombre, millones, infantil, meter, dentro, entrenamiento, causa, norte, través, alto, agresión, reúner, información, intento, iglesia, fmi, acuerdo, internacional, histórico, rojo, mostrar, hoy, tipo, sorpresa, herir, tercero, sesión, fiesta, ingresar, cerca, protesta, reforma, pensión, central, mejora, nacer, entrar, pole, asalto, hora, policial, impidir, manifestante, casi, manifestación, forma, salida, crisis, recuperar, valor, carrera, pasar, autor, sorprender, delante, cara, jefe, pelea, ferrer, primer, español, actor, responsabilidad, abogado, parecer, berlusconi, girona, b, vasco, rueda, autobús, camión, justicia, arabia, saudí, frente, argentina, brasil, imágén, técnico, mirar, rápido, juego, olím, unir, calle, grecia, mil, diputado, salvar, planta, catalan, encuentro, condena, ley, catalán, fondo, especial, financiación, negociación, ciento, participar, matrimonio, sueño, realidad, extranjero, empresa, dólares, próximo, base, defensa, libro, cultura, alemán, cualquier, palabra, india, puerta, razón, apoyo, evitar, reina, isabel, ii, página, propuesta, orden, violencia, gp, hotel, mundial, social, echar, declaración, catalunya, viaje, mano, municipal, descenso, republicano, impuesto, hamilton, triunfo, plazo, cabeza, griego, económico, descubrir, importante, construir, público, líder, espectacular, ataqu, lunes, negro, total, pau, momento, verano, olvidar, mitad, concurso, vacación, vender, acción, vehículo, pese, huir, debate, activo, situación, mensaje, contar, campaña, presencia, asistir, recordar, navidad, investigar, robo, primavera, tribunal, ilegal, blanco, lucha, pequeño, siglo, beneficio, caixa, general, relación, preso, kilo, tarragona, controlar, vista, pacto, hablar, fuerte, padr, recurrir, supremo, nadal, operación, mallorca, ave, sede, formar, resultado, condenar, petición, grande, encuesta, cargo, reunión, firma, luchar, cáncer, esperar, metro, bolsa, chile, agente, pena, viajar, fallo, consejo, positivo, contador, ampliar, anuncio, ampliación, junta, argentino, fifa, guerra, parados, universidad, autonómico, muerte, oposición, plán, barco, refugiado, frontera, disturbio, batalla, laboral, foto, destino, provocar, desaparecer, toda, gratis, parlamento, entrada, concierto, uso, votar, ahora, italia, novedad, ganador, tour, obligar, motivo, inversión, mapa, web, rey, permitir, programa, mossos, plaza, aniversario, actuar, turístico, corto, todas, ciudad, cámara, vestir, edad, sistema, dato, financiero, opción, sentencia, padre, aunque, zona, asesinar, bruselas, duro, europeo, gestión, gala, entrega, alcalde, efecto, declarar, récord, límite, tsjc, salud, playa, apuntar, solución, mientras, detención, producto, pantalla, viernes, sol, vivienda, frenar, afectado, pagar, agua, fuego, dimitir, directivo, ayuda, presupuesto, santa, guardia, príncipe, humano, atacar, miedo, cumplir, enviar, gasto, sueldo, ministro, animal, comercial, mar, coche, eléctrico, cómo, cada, mañana, informe, candidato, inmigrante, único, compra, tarjeta, comida, recurso, cliente, estudiante, call, droga, falso, científico, atrás, ibex, cierre, chica, control, jueves, aparecer, cadáver, clinton, confiar, paz, piso, acceso, galería, prueba, barrio, polémica, director, entrevista, investigación, resolver, adiós, analizar, canal, imagen, tocar, explicar, presión, juez, ayuntamiento, sábado, domingo, cobrar, lotería, arte, necesario, oficial, museo, devolver, esperanza, millón, encontrar, mediterráneo, sentir, deuda, futuro, escribir, restaurante, local, clave, intentar, mortal, línea, necesitar, voz, grave, personal, arma, crítica, familiar, proceso, sindicato, fuerza, patrimonio, máximo, audiencia, interior, chino, inglaterra, usar, vía, legal, tienda, portugal, solo, tesoro, cuenta, asesinato, gastar, sanción, mandar, incluir, desnudo, cita, quinto, pérdida, banca, popular, estrenar, artur, independentismo, prohibir, puente, luna, mejorar, judicial, oportunidad, entrevistar, doble, boda, escolar, juzgar, discurso, tensión, posición, irlanda, rescate, apple, recorte, elegir, eurozona, disparo, bombardeo, cuestión, confianza, puerto, causar, sarkozy, dimisión, federer, diario, comunidad, continuar, publicar, novia, supuesto, golpe, medida, pista, cumbre, andalucía, puesto, homenaje, condición, comentario, estudio, sacar, campo, mercado, invitar, junto, merkel, hermano, castro, teatro, bce, tren, camps, valenciano, presidir, euro, jordi, edificio, espacio, salario, sufrir, pie, tema, lista, nobel, principal, peligro, sanidad, sector, decisión, portada, médico, funcionario, seguro, candidatura, piloto, completo, kim, carta, conductor, ficha, bilbao, rato, acusación, apertura, cerebro, oficina, enseñar, pedro, superar, repetir, pago, cultural, prensa,

medalla, famoso, detectar, república, vanguardia, oferta, ciencia, fecha, camino, ninguna, ajuste, choque, lugar, ventaja, idea, bar, bancario, sorteo, alumno, suspensión, xavier, gay, chávez, dónde, secreto, emergencia, llegada, empleado, democracia, civil, prat, horario, sospechoso, descuento, partir, ahorrar, corrupción, empresario, francisco, ojo, crítico, energía, caza, universitario, combatir, secretario, planeta, fianza, antiguo, utilizar, gracias, largo, tarde, agencia, capital, masters, entender, comer, bandera, turista, profesor, holanda, detall, miércoles, els, murray, presidencia, así, protagonista, cuerpo, parque, recopilación, iva, djokovic, actriz, verde, prohíbe, león, asturias, isla, temer, autonomías, tc, fraude, quemar, alarma, rafa, billete, natural, ana, sexo, comprar, votación, banda, per, río, blanca, renovar, rosell, resto, triunfar, dirección, hacienda, aquí, tiroteo, unidad, venir, world, masivo, tragedia, receta, primario, etapa, cristina, senado, costar, lector, ciclista, suiza, palma, artículo, previsión, etiqueta, escocia, eurocopa, barberá, australia, inglés, infanta, trasladar, escenario, petróleo, perfil, d, actividad, renta, lujo, programación, revolución, mataró, fabra, compartir, concejal, estreno, ortega, sección, La, motogp, open, set, montoro, breve, imposible, El, lavanguardia, instante, indignados, diada, Los, bankia, fukushima, En, hashtag, rt, bárcenas, tuits, Aquí, Te, trump, Hoy, Dínoslo, Ampliamos, ébola, maduro, Sigue, Mañana, donald, Lo, hollande, Consulta, l, Opina, Síguelo, Vota, deporteslv, videoblog, urdangarin, guindos, wert, alertalv, enuntuit, publicados, worldcup, últimahora, carmena, enportada, brexit

Appendix C

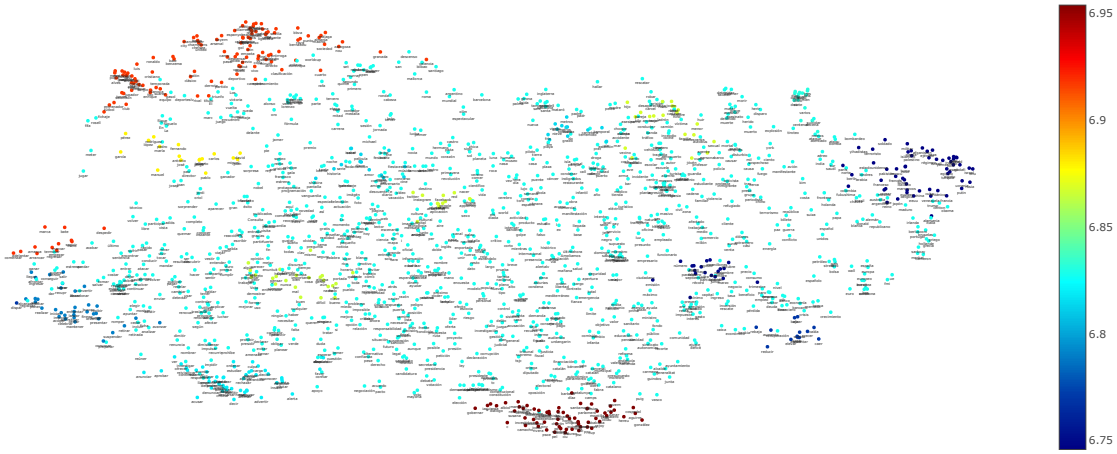
Heart attacks associated to clusters



Cluster2AMI scatter plot. Mean number of AMI per day.

Cluster	Mean	Var. rate	P-Value
0	6.82621	-0.053 %	0.312
1	6.85743	0.404 %	0.129
2	6.81325	-0.243 %	0.366
3	6.90796	1.144 %	0.013
4	6.79957	-0.443 %	0.28
5	6.78881	-0.6 %	0.028
6	6.83312	0.048 %	0.707
7	6.78511	-0.655 %	0.175
8	6.75342	-1.119 %	0.244
9	6.84283	0.191 %	0.863
10	6.79013	-0.581 %	0.043
11	6.86115	0.459 %	0.146
12	6.86332	0.49 %	0.122
13	6.82547	-0.064 %	0.674
14	6.86578	0.527 %	0.484
15	6.846	0.237 %	0.399

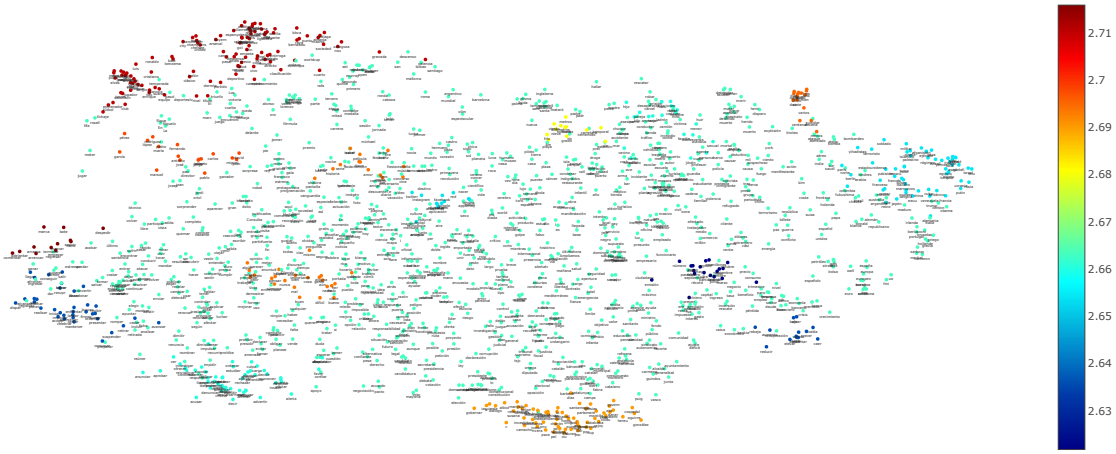
Cluster2AMI table. Mean number of AMI per day.



Cluster2AMI scatter plot. Weighted mean number of AMI per day.

Cluster	Mean	Var. rate	P-Value
0	6.83676	-0.059 %	0.045
1	6.90244	0.901 %	0.008
2	6.79004	-0.742 %	0.291
3	6.94442	1.514 %	0.011
4	6.81761	-0.339 %	0.779
5	6.78632	-0.797 %	0.011
6	6.84409	0.048 %	0.373
7	6.76303	-1.137 %	0.105
8	6.74504	-1.4 %	0.246
9	6.89043	0.725 %	0.482
10	6.78184	-0.862 %	0.04
11	6.84911	0.121 %	0.404
12	6.85451	0.2 %	0.195
13	6.78009	-0.888 %	0.757
14	6.83469	-0.09 %	0.245
15	6.84225	0.021 %	0.664

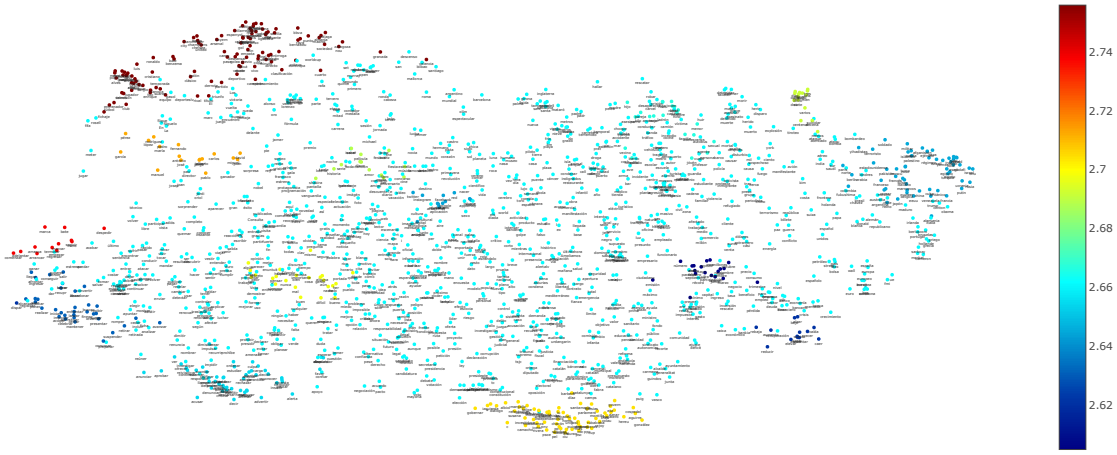
Cluster2AMI table. Weighted mean number of AMI per day.



Cluster2AMI scatter plot. Mean number of AMI per day of population between 40 and 59 years old.

Cluster	Mean	Var. rate	P-Value
0	2.66537	0.075 %	0.934
1	2.70876	1.704 %	0.0
2	2.66222	-0.043 %	0.762
3	2.68175	0.69 %	0.165
4	2.67923	0.596 %	0.13
5	2.66807	0.177 %	0.55
6	2.66117	-0.083 %	0.832
7	2.63444	-1.086 %	0.114
8	2.62171	-1.564 %	0.119
9	2.70215	1.456 %	0.157
10	2.64001	-0.877 %	0.066
11	2.68919	0.97 %	0.156
12	2.69685	1.257 %	0.03
13	2.67402	0.4 %	0.606
14	2.69691	1.26 %	0.089
15	2.64636	-0.639 %	0.746

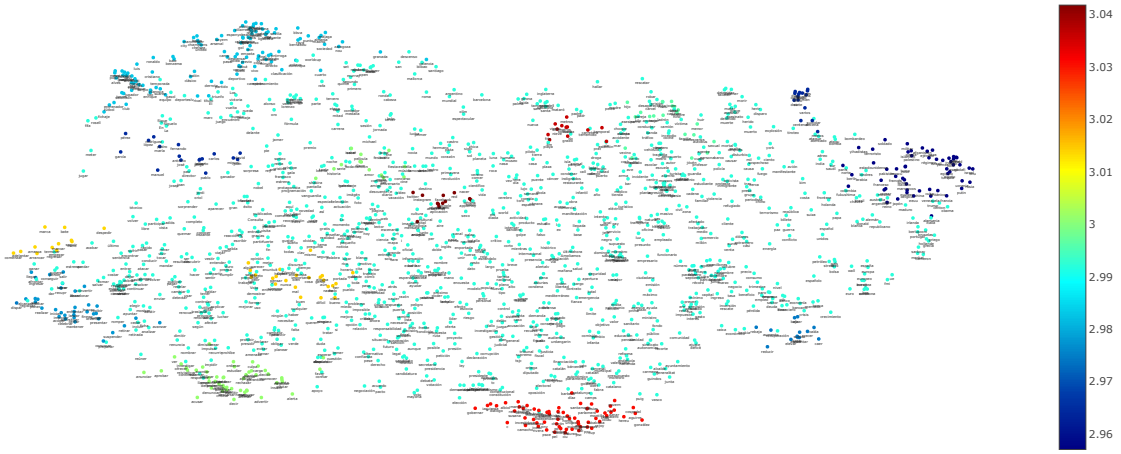
Cluster2AMI table. Mean number of AMI per day of population between 40 and 59 years old.



Cluster2AMI scatter plot. Weighted mean number of AMI per day of population between 40 and 59 years old.

Cluster	Mean	Var. rate	P-Value
0	2.66554	-0.082 %	0.043
1	2.74574	2.925 %	0.0
2	2.64255	-0.943 %	0.278
3	2.68964	0.822 %	0.098
4	2.66707	-0.024 %	0.202
5	2.65621	-0.432 %	0.175
6	2.65784	-0.37 %	0.939
7	2.60492	-2.354 %	0.042
8	2.60713	-2.271 %	0.042
9	2.72458	2.131 %	0.097
10	2.62808	-1.486 %	0.018
11	2.6975	1.116 %	0.166
12	2.702	1.285 %	0.03
13	2.64958	-0.68 %	0.863
14	2.69842	1.151 %	0.482
15	2.65072	-0.637 %	0.457

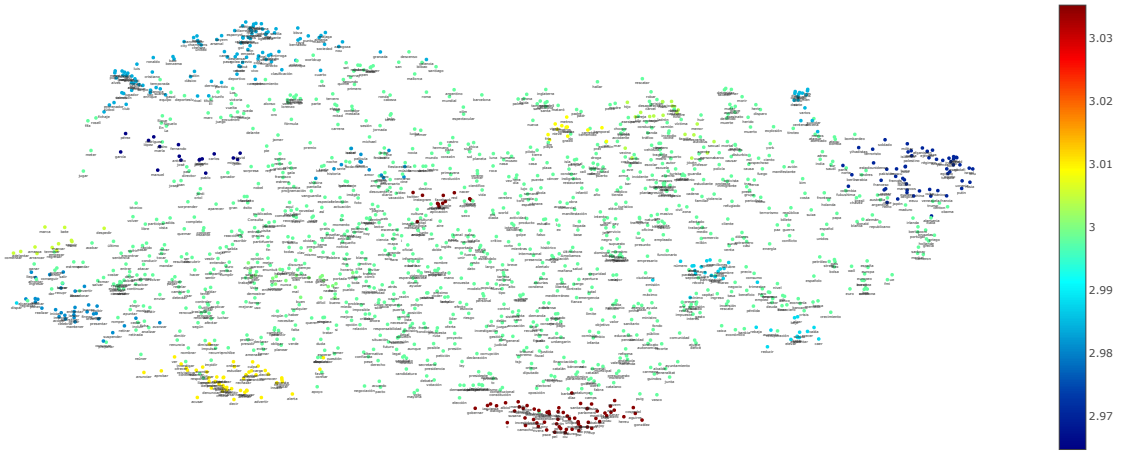
Cluster2AMI table. Weighted mean number of AMI per day of population between 40 and 59 years old.



Cluster2AMI scatter plot. Mean number of AMI per day of population between 60 and 79 years old.

Cluster	Mean	Var. rate	P-Value
0	2.99245	-0.107 %	0.237
1	2.98131	-0.479 %	0.221
2	2.99207	-0.12 %	0.647
3	3.02444	0.961 %	0.034
4	2.97384	-0.728 %	0.176
5	2.97378	-0.73 %	0.018
6	2.98994	-0.191 %	0.979
7	2.994	-0.055 %	0.474
8	2.98334	-0.411 %	0.894
9	2.96619	-0.984 %	0.222
10	2.98451	-0.372 %	0.08
11	3.01131	0.522 %	0.231
12	3.00183	0.206 %	0.471
13	3.00747	0.394 %	0.412
14	3.0117	0.535 %	0.82
15	3.0239	0.943 %	0.167

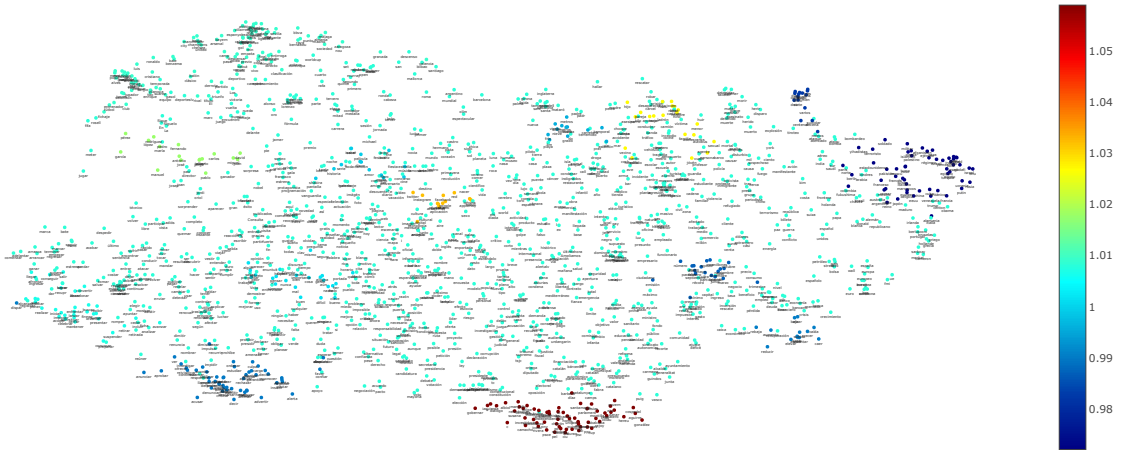
Cluster2AMI table. Mean number of AMI per day of population between 60 and 79 years old.



Cluster2AMI scatter plot. Weighted mean number of AMI per day of population between 60 and 79 years old.

Cluster	Mean	Var. rate	P-Value
0	3.00035	0.006 %	0.525
1	2.98125	-0.63 %	0.239
2	2.98758	-0.419 %	0.488
3	3.0427	1.418 %	0.042
4	2.98732	-0.428 %	0.578
5	2.99232	-0.261 %	0.121
6	2.9976	-0.085 %	0.828
7	3.00415	0.133 %	0.66
8	2.97965	-0.684 %	0.787
9	2.96547	-1.156 %	0.228
10	2.98161	-0.618 %	0.107
11	2.98708	-0.436 %	0.991
12	2.99723	-0.098 %	0.855
13	2.98143	-0.624 %	0.885
14	2.98771	-0.415 %	0.328
15	3.01613	0.532 %	0.292

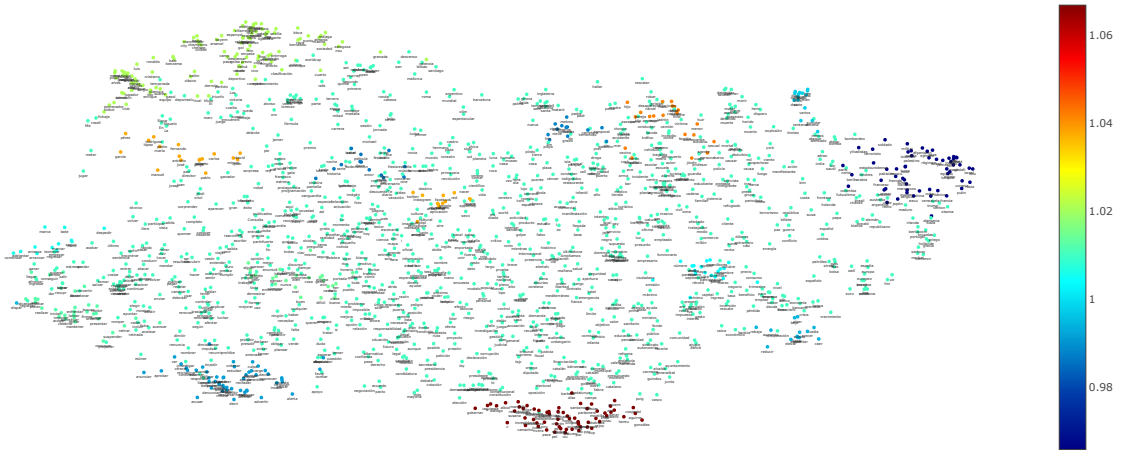
Cluster2AMI table. Weighted mean number of AMI per day of population between 60 and 79 years old.



Cluster2AMI scatter plot. Mean number of AMI per day of population between 80 and 99 years old.

Cluster	Mean	Var. rate	P-Value
0	1.00715	-0.328 %	0.254
1	1.00606	-0.436 %	0.826
2	0.99524	-1.507 %	0.0
3	1.04457	3.375 %	0.002
4	0.99016	-2.01 %	0.008
5	0.97936	-3.078 %	0.0
6	1.02656	1.593 %	0.229
7	0.99018	-2.007 %	0.152
8	0.98965	-2.06 %	0.096
9	1.01712	0.658 %	0.551
10	1.00234	-0.804 %	0.704
11	1.00278	-0.761 %	0.459
12	0.99954	-1.081 %	0.852
13	0.98913	-2.111 %	0.461
14	0.99647	-1.385 %	0.361
15	1.02077	1.02 %	0.357

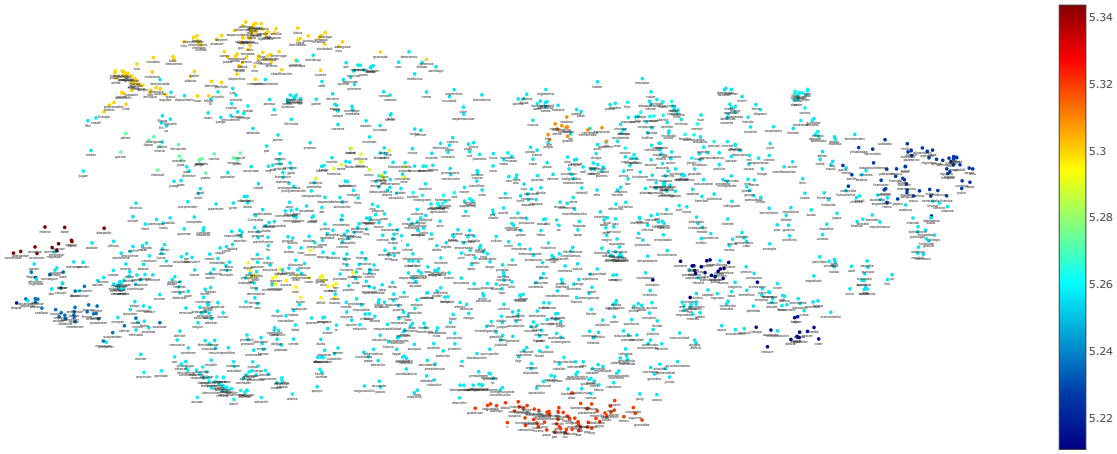
Cluster2AMI table. Mean number of AMI per day of population between 80 and 99 years old.



Cluster2AMI scatter plot. Weighted mean number of AMI per day of population between 80 and 99 years old.

Cluster	Mean	Var. rate	P-Value
0	1.01228	-0.249 %	0.055
1	1.0183	0.343 %	0.491
2	0.9974	-1.716 %	0.001
3	1.05941	4.395 %	0.002
4	1.00695	-0.775 %	0.188
5	0.97445	-3.978 %	0.0
6	1.03725	2.211 %	0.109
7	0.98835	-2.608 %	0.149
8	1.00583	-0.885 %	0.605
9	1.03705	2.191 %	0.165
10	1.00971	-0.503 %	0.822
11	1.01149	-0.327 %	0.953
12	0.99242	-2.207 %	0.58
13	0.99108	-2.338 %	0.268
14	0.98819	-2.623 %	0.033
15	1.02192	0.701 %	0.389

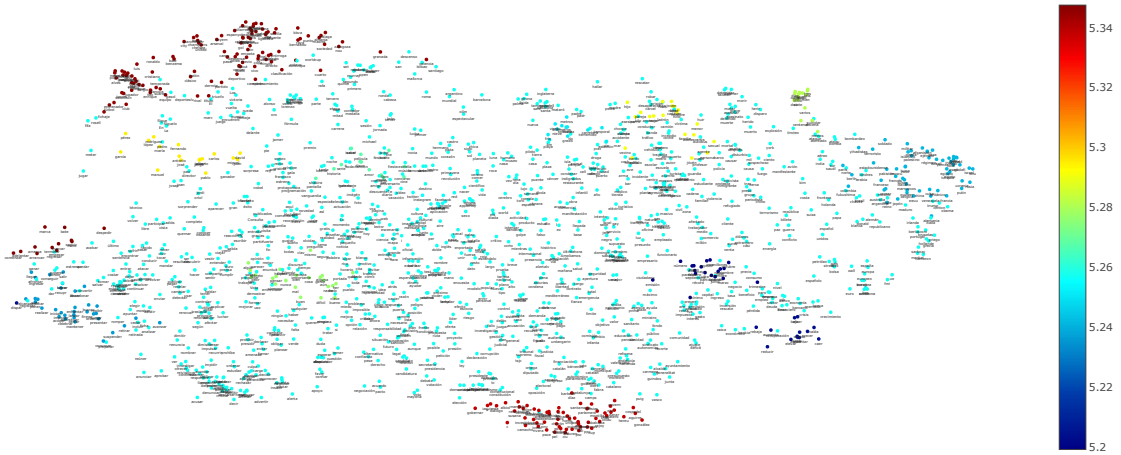
Cluster2AMI table. Weighted mean number of AMI per day of population between 80 and 99 years old.



Cluster2AMI scatter plot. Mean number of AMI per day of men population.

Cluster	Mean	Var. rate	P-Value
0	5.25951	-0.033 %	0.296
1	5.28983	0.544 %	0.027
2	5.24971	-0.219 %	0.862
3	5.30603	0.852 %	0.054
4	5.24308	-0.345 %	0.919
5	5.25111	-0.192 %	0.202
6	5.26054	-0.013 %	0.966
7	5.22648	-0.66 %	0.24
8	5.20993	-0.975 %	0.251
9	5.27301	0.224 %	0.756
10	5.2387	-0.428 %	0.192
11	5.29353	0.614 %	0.102
12	5.30437	0.82 %	0.076
13	5.28523	0.456 %	0.465
14	5.30773	0.884 %	0.269
15	5.24275	-0.351 %	0.988

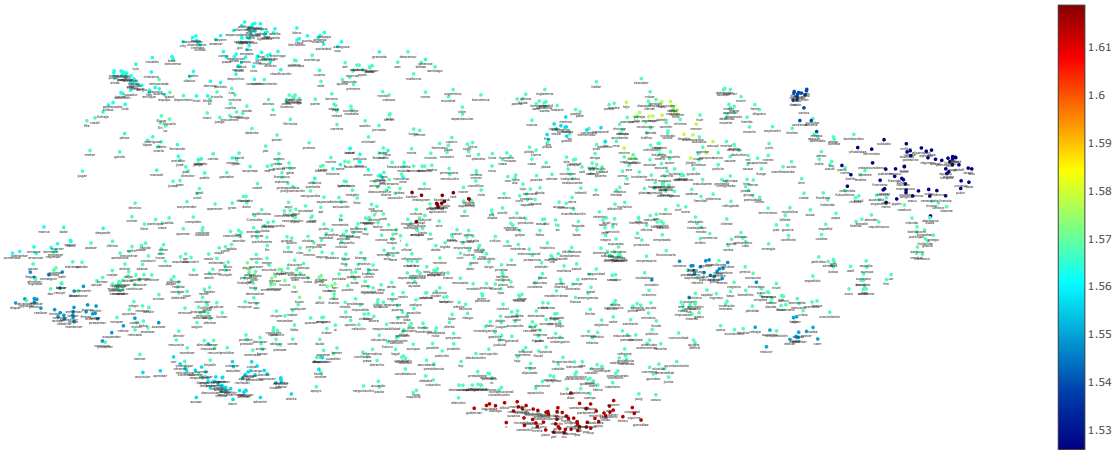
Cluster2AMI table. Mean number of AMI per day of men population.



Cluster2AMI scatter plot. Weighted mean number of AMI per day of men population.

Cluster	Mean	Var. rate	P-Value
0	5.26389	-0.079 %	0.036
1	5.32977	1.172 %	0.002
2	5.22671	-0.785 %	0.539
3	5.33327	1.238 %	0.033
4	5.25484	-0.251 %	0.658
5	5.26997	0.036 %	0.38
6	5.27481	0.128 %	0.401
7	5.1966	-1.356 %	0.131
8	5.19935	-1.304 %	0.223
9	5.3038	0.679 %	0.562
10	5.22498	-0.818 %	0.12
11	5.27173	0.07 %	0.712
12	5.28866	0.391 %	0.105
13	5.23551	-0.618 %	0.808
14	5.28786	0.376 %	0.971
15	5.24848	-0.372 %	0.765

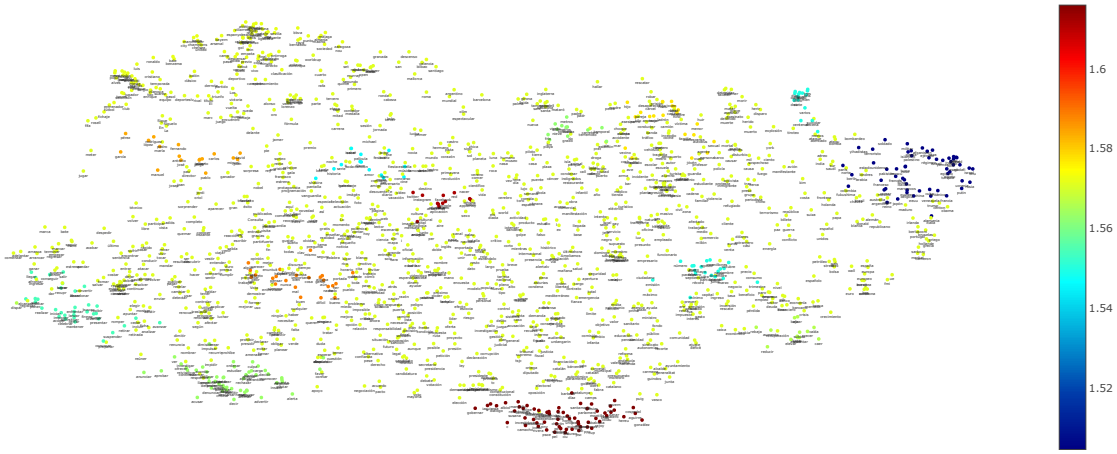
Cluster2AMI table. Weighted mean number of AMI per day of men population.



Cluster2AMI scatter plot. Mean number of AMI per day of women population.

Cluster	Mean	Var. rate	P-Value
0	1.5667	-0.121 %	0.663
1	1.56761	-0.063 %	0.473
2	1.56354	-0.322 %	0.156
3	1.60193	2.125 %	0.001
4	1.55649	-0.772 %	0.121
5	1.5377	-1.97 %	0.001
6	1.57258	0.254 %	0.372
7	1.55863	-0.635 %	0.234
8	1.54349	-1.6 %	0.353
9	1.56982	0.078 %	0.822
10	1.55142	-1.095 %	0.039
11	1.56762	-0.062 %	0.855
12	1.55895	-0.615 %	0.737
13	1.54024	-1.808 %	0.706
14	1.55806	-0.672 %	0.708
15	1.60325	2.209 %	0.1

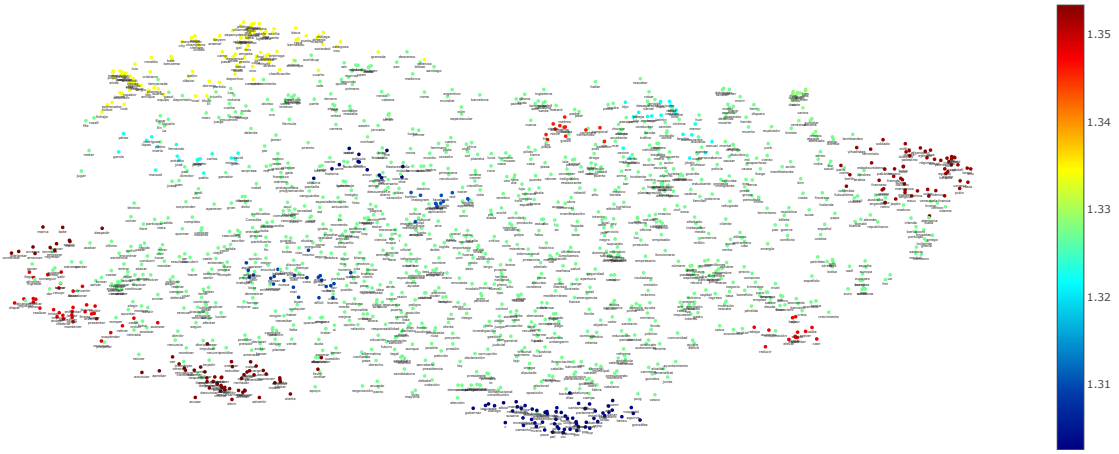
Cluster2AMI table. Mean number of AMI per day of women population.



Cluster2AMI scatter plot. Weighted mean number of AMI per day of women population.

Cluster	Mean	Var. rate	P-Value
0	1.57287	0.006 %	0.452
1	1.57267	-0.006 %	0.905
2	1.56333	-0.6 %	0.303
3	1.61115	2.441 %	0.008
4	1.56277	-0.636 %	0.359
5	1.51635	-3.587 %	0.0
6	1.56929	-0.221 %	0.709
7	1.56642	-0.403 %	0.66
8	1.54569	-1.722 %	0.411
9	1.58663	0.881 %	0.532
10	1.55687	-1.011 %	0.073
11	1.57737	0.293 %	0.36
12	1.56585	-0.44 %	0.929
13	1.54458	-1.792 %	0.732
14	1.54683	-1.649 %	0.212
15	1.59377	1.335 %	0.215

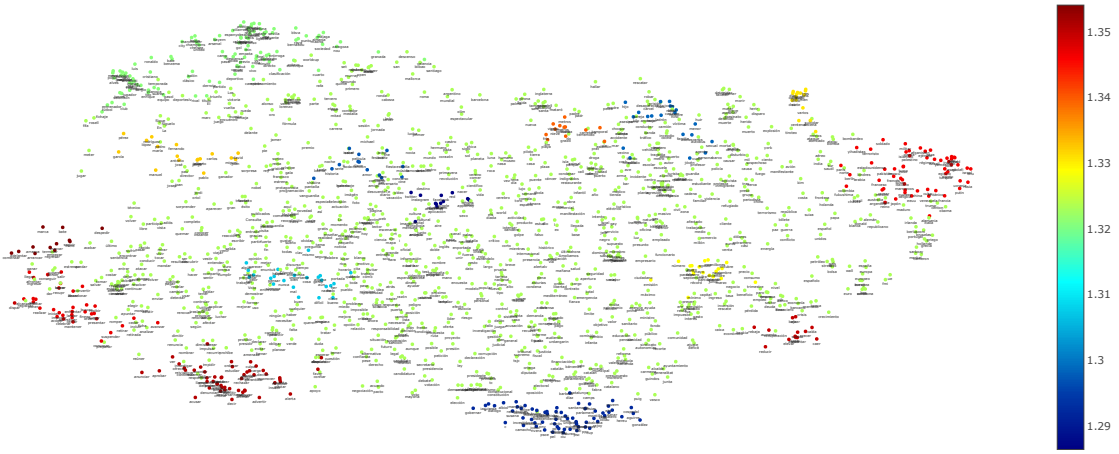
Cluster2AMI table. Weighted mean number of AMI per day of women population.



Cluster2AMI scatter plot. Mean number of mortal AMI per day.

Cluster	Mean	Var. rate	P-Value
0	1.33241	0.641 %	0.154
1	1.32972	0.438 %	0.225
2	1.35388	2.264 %	0.0
3	1.32069	-0.243 %	0.183
4	1.32707	0.238 %	0.707
5	1.34636	1.695 %	0.008
6	1.32448	0.043 %	0.823
7	1.34565	1.642 %	0.189
8	1.32059	-0.251 %	0.864
9	1.32707	0.239 %	0.93
10	1.34517	1.605 %	0.019
11	1.30649	-1.316 %	0.287
12	1.34484	1.58 %	0.272
13	1.33548	0.874 %	0.531
14	1.30949	-1.089 %	0.323
15	1.327	0.233 %	0.561

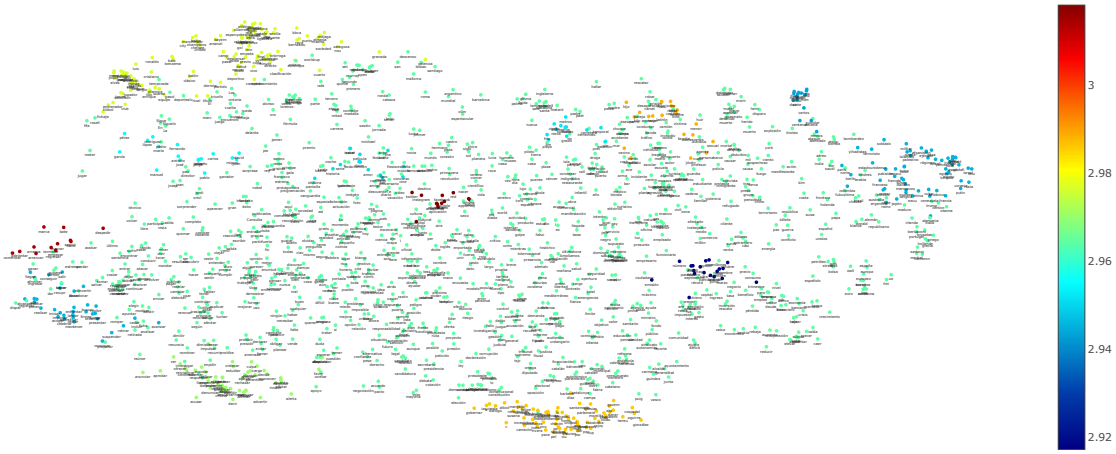
Cluster2AMI table. Mean number of mortal AMI per day.



Cluster2AMI scatter plot. Weighted mean number of mortal AMI per day.

Cluster	Mean	Var. rate	P-Value
0	1.32513	0.497 %	0.114
1	1.31289	-0.431 %	0.831
2	1.34661	2.127 %	0.0
3	1.3023	-1.234 %	0.153
4	1.32446	0.447 %	0.421
5	1.34145	1.735 %	0.035
6	1.29889	-1.493 %	0.171
7	1.34179	1.761 %	0.123
8	1.31938	0.061 %	0.693
9	1.33643	1.355 %	0.434
10	1.34039	1.655 %	0.016
11	1.29957	-1.441 %	0.446
12	1.33509	1.253 %	0.239
13	1.33061	0.913 %	0.607
14	1.29599	-1.712 %	0.28
15	1.30488	-1.038 %	0.356

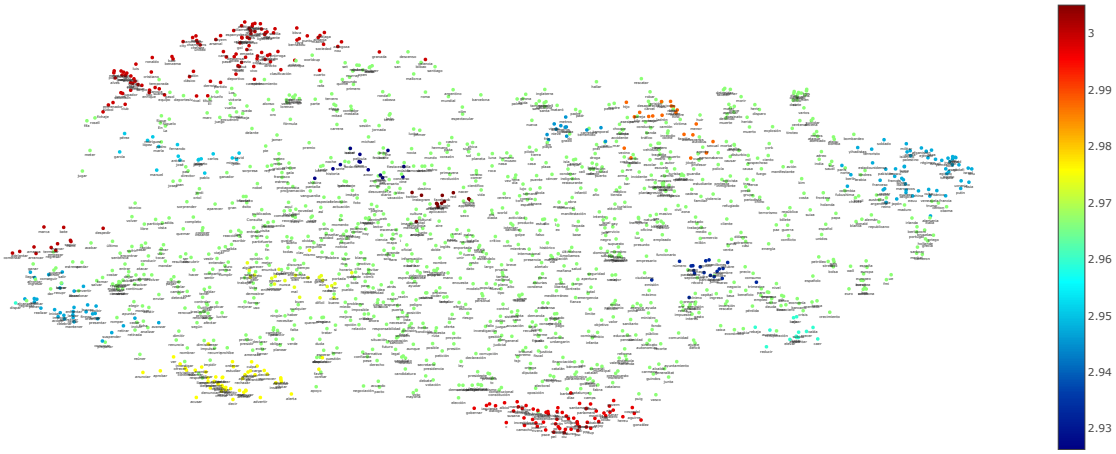
Cluster2AMI table. Weighted mean number of mortal AMI per day.



Cluster2AMI scatter plot. Mean number of AMI, diagnosed as anterior Q-wave, per day.

Cluster	Mean	Var. rate	P-Value
0	2.96503	-0.051 %	0.444
1	2.97176	0.176 %	0.415
2	2.96659	0.002 %	0.747
3	2.98098	0.487 %	0.322
4	2.9467	-0.668 %	0.133
5	2.95293	-0.459 %	0.235
6	2.97237	0.197 %	0.177
7	2.95869	-0.264 %	0.938
8	2.91564	-1.715 %	0.144
9	2.95095	-0.525 %	0.694
10	2.94458	-0.74 %	0.224
11	2.95495	-0.39 %	0.87
12	2.98244	0.536 %	0.124
13	2.94363	-0.772 %	0.682
14	2.96667	0.005 %	0.406
15	3.00451	1.28 %	0.028

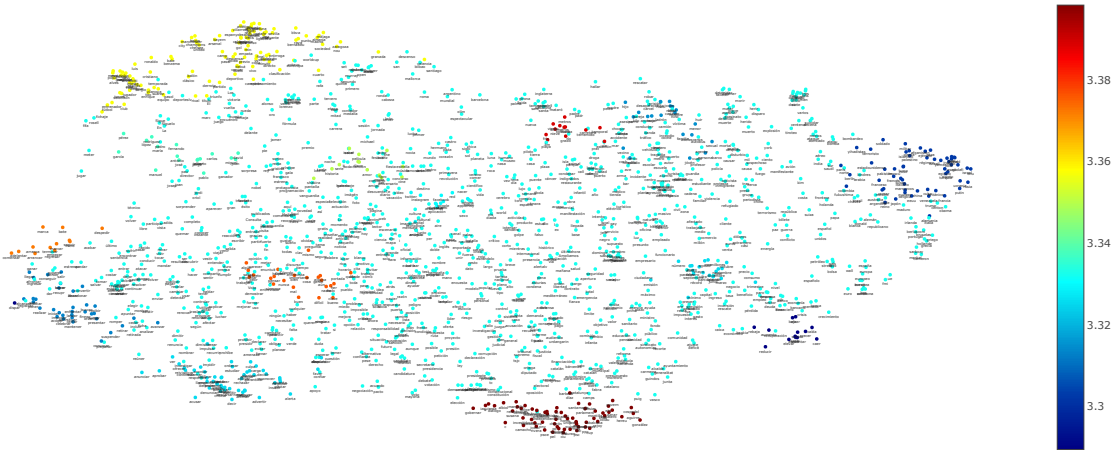
Cluster2AMI table. Mean number of AMI, diagnosed as anterior Q-wave, per day.



Cluster2AMI scatter plot. Weighted mean number of AMI, diagnosed as anterior Q-wave, per day.

Cluster	Mean	Var. rate	P-Value
0	2.97023	-0.029 %	0.222
1	2.99034	0.648 %	0.07
2	2.95679	-0.481 %	0.73
3	3.00004	0.975 %	0.232
4	2.95715	-0.469 %	0.76
5	2.95778	-0.448 %	0.247
6	2.96892	-0.073 %	0.329
7	2.9467	-0.821 %	0.634
8	2.93001	-1.383 %	0.335
9	2.94592	-0.847 %	0.521
10	2.94804	-0.776 %	0.234
11	2.96874	-0.079 %	0.906
12	2.96754	-0.119 %	0.344
13	2.93487	-1.219 %	0.52
14	2.93873	-1.089 %	0.035
15	3.00055	0.992 %	0.133

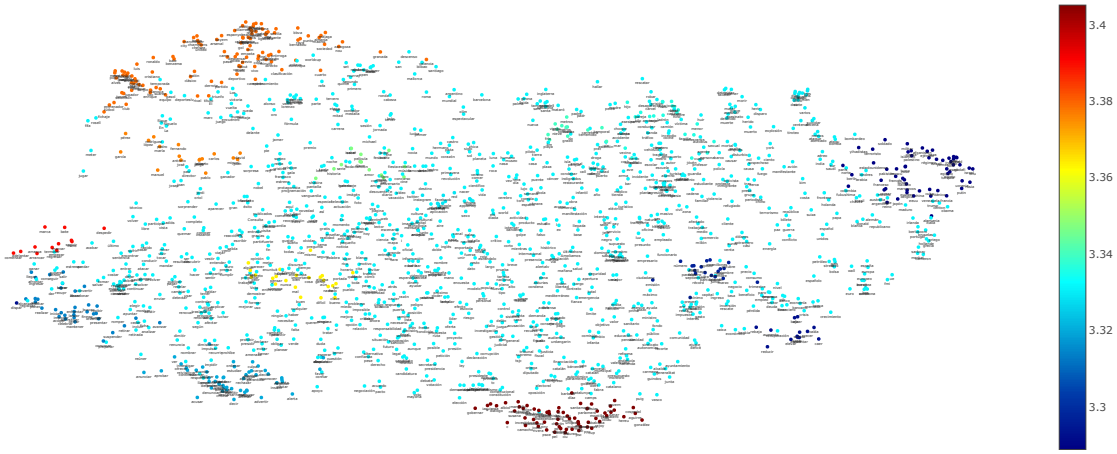
Cluster2AMI table. Weighted mean number of AMI, diagnosed as anterior Q-wave, per day.



Cluster2AMI scatter plot. Mean number of AMI, diagnosed as inferior Q-wave, per day.

Cluster	Mean	Var. rate	P-Value
0	3.33545	0.024 %	0.444
1	3.35647	0.654 %	0.05
2	3.32568	-0.269 %	0.383
3	3.38375	1.472 %	0.002
4	3.33381	-0.025 %	0.918
5	3.32451	-0.304 %	0.059
6	3.32801	-0.199 %	0.295
7	3.3132	-0.643 %	0.067
8	3.32252	-0.364 %	0.824
9	3.34419	0.286 %	0.933
10	3.3198	-0.445 %	0.039
11	3.37458	1.197 %	0.003
12	3.36042	0.773 %	0.089
13	3.36095	0.789 %	0.269
14	3.34636	0.351 %	0.396
15	3.3127	-0.658 %	0.908

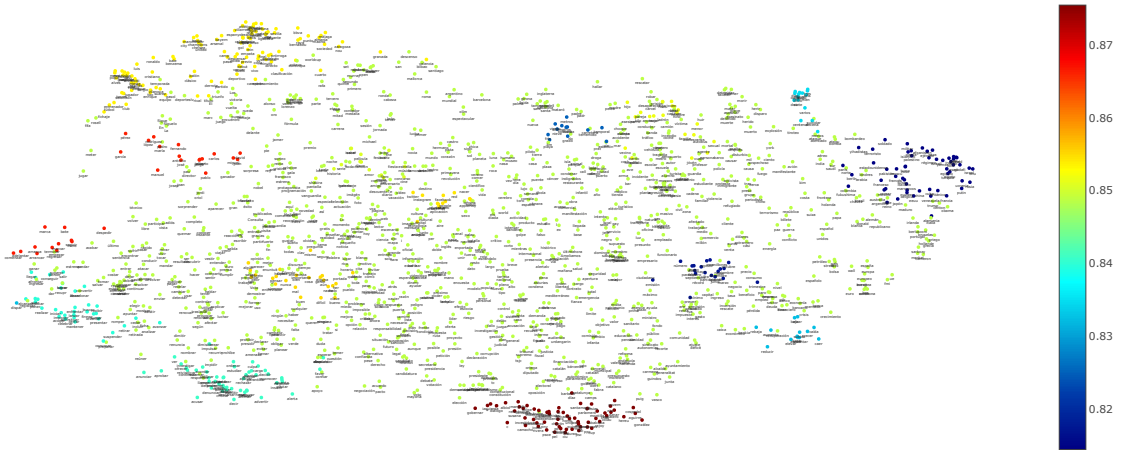
Cluster2AMI table. Mean number of AMI, diagnosed as inferior Q-wave, per day.



Cluster2AMI scatter plot. Weighted mean number of AMI, diagnosed as inferior Q-wave, per day.

Cluster	Mean	Var. rate	P-Value
0	3.33537	-0.032 %	0.163
1	3.36643	0.899 %	0.017
2	3.30788	-0.856 %	0.2
3	3.39527	1.763 %	0.004
4	3.32973	-0.201 %	0.825
5	3.31184	-0.737 %	0.015
6	3.33795	0.045 %	0.962
7	3.29915	-1.118 %	0.064
8	3.29747	-1.168 %	0.412
9	3.39272	1.687 %	0.216
10	3.30793	-0.854 %	0.066
11	3.34946	0.39 %	0.031
12	3.35812	0.65 %	0.12
13	3.32095	-0.464 %	0.953
14	3.34706	0.318 %	0.742
15	3.31853	-0.537 %	0.854

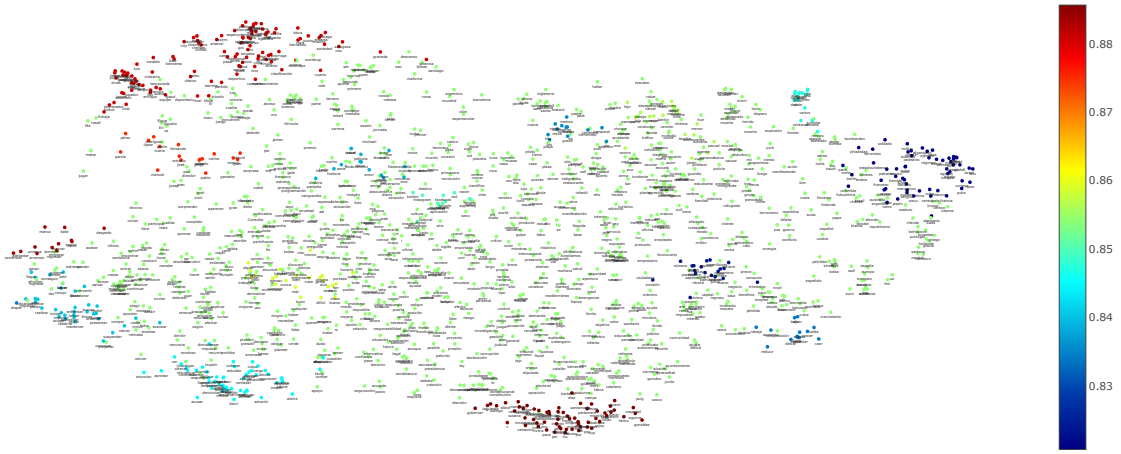
Cluster2AMI table. Weighted mean number of AMI, diagnosed as inferior Q-wave, per day.



Cluster2AMI scatter plot. Mean number of AMI, diagnosed as lateral Q-wave, per day.

Cluster	Mean	Var. rate	P-Value
0	0.84592	-0.332 %	0.878
1	0.85174	0.354 %	0.522
2	0.84274	-0.707 %	0.252
3	0.86793	2.262 %	0.014
4	0.83053	-2.146 %	0.148
5	0.82577	-2.706 %	0.001
6	0.84716	-0.185 %	0.797
7	0.82997	-2.211 %	0.203
8	0.82006	-3.378 %	0.005
9	0.8685	2.329 %	0.139
10	0.83944	-1.096 %	0.19
11	0.85395	0.615 %	0.473
12	0.85904	1.214 %	0.173
13	0.82377	-2.942 %	0.15
14	0.85607	0.864 %	0.915
15	0.84404	-0.553 %	0.81

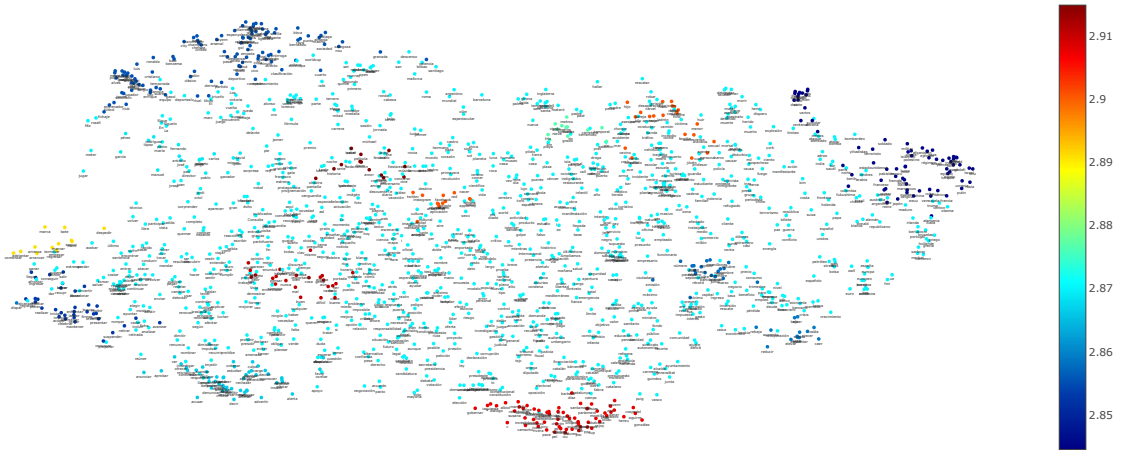
Cluster2AMI table. Mean number of AMI, diagnosed as lateral Q-wave, per day.



Cluster2AMI scatter plot. Weighted mean number of AMI, diagnosed as lateral Q-wave, per day.

Cluster	Mean	Var. rate	P-Value
0	0.85491	-0.181 %	0.287
1	0.88126	2.896 %	0.007
2	0.84638	-1.176 %	0.119
3	0.88158	2.933 %	0.015
4	0.83856	-2.09 %	0.502
5	0.83537	-2.462 %	0.002
6	0.846	-1.22 %	0.98
7	0.82846	-3.268 %	0.111
8	0.82787	-3.338 %	0.034
9	0.87514	2.182 %	0.192
10	0.83898	-2.04 %	0.049
11	0.8571	0.076 %	0.864
12	0.87505	2.171 %	0.056
13	0.83284	-2.757 %	0.088
14	0.85033	-0.715 %	0.425
15	0.83673	-2.302 %	0.783

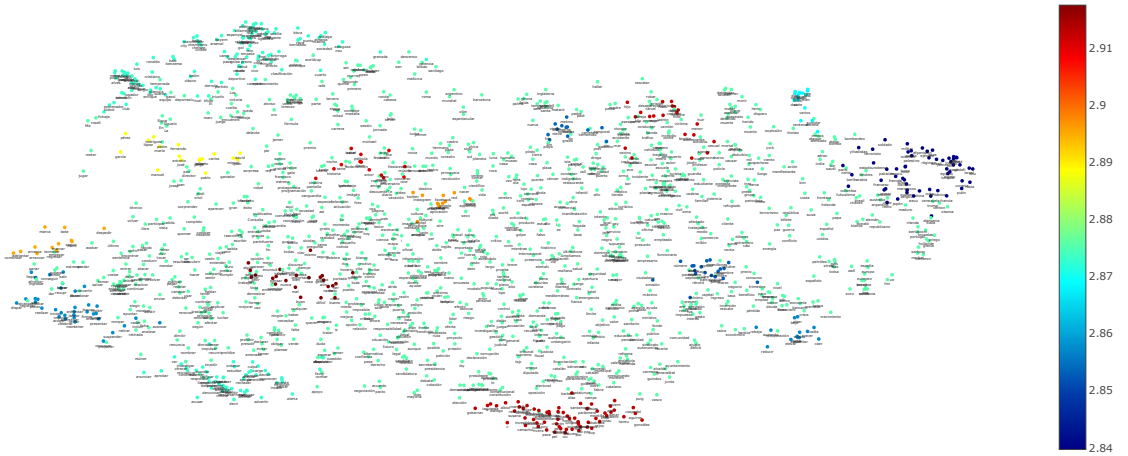
Cluster2AMI table. Weighted mean number of AMI, diagnosed as lateral Q-wave, per day.



Cluster2AMI scatter plot. Mean number of AMI per day, with initial pain between 6 a.m. and 2 p.m.

Cluster	Mean	Var. rate	P-Value
0	2.86872	-0.139 %	0.392
1	2.85799	-0.513 %	0.124
2	2.86596	-0.235 %	0.491
3	2.89688	0.841 %	0.058
4	2.84933	-0.815 %	0.113
5	2.85736	-0.535 %	0.104
6	2.88355	0.377 %	0.118
7	2.86351	-0.321 %	0.514
8	2.85338	-0.673 %	0.591
9	2.87477	0.071 %	0.863
10	2.8485	-0.843 %	0.128
11	2.90456	1.108 %	0.013
12	2.87309	0.013 %	0.483
13	2.86927	-0.12 %	0.905
14	2.90596	1.157 %	0.126
15	2.87872	0.209 %	0.33

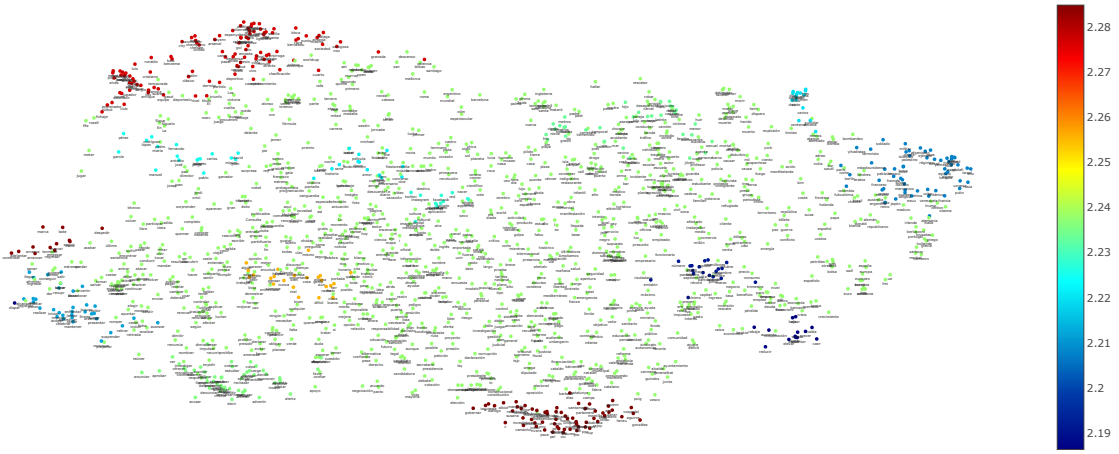
Cluster2AMI table. Mean number of AMI per day, with initial pain between 6 a.m. and 2 p.m.



Cluster2AMI scatter plot. Weighted mean number of AMI per day, with initial pain between 6 a.m. and 2 p.m.

Cluster	Mean	Var. rate	P-Value
0	2.87898	0.002 %	0.296
1	2.86273	-0.563 %	0.669
2	2.86768	-0.391 %	0.55
3	2.91176	1.14 %	0.093
4	2.87015	-0.305 %	0.625
5	2.85954	-0.674 %	0.064
6	2.88883	0.344 %	0.066
7	2.85492	-0.834 %	0.383
8	2.84913	-1.035 %	0.398
9	2.89894	0.695 %	0.707
10	2.84844	-1.059 %	0.134
11	2.90043	0.747 %	0.039
12	2.87662	-0.08 %	0.572
13	2.85017	-0.999 %	0.376
14	2.90171	0.791 %	0.229
15	2.89189	0.45 %	0.537

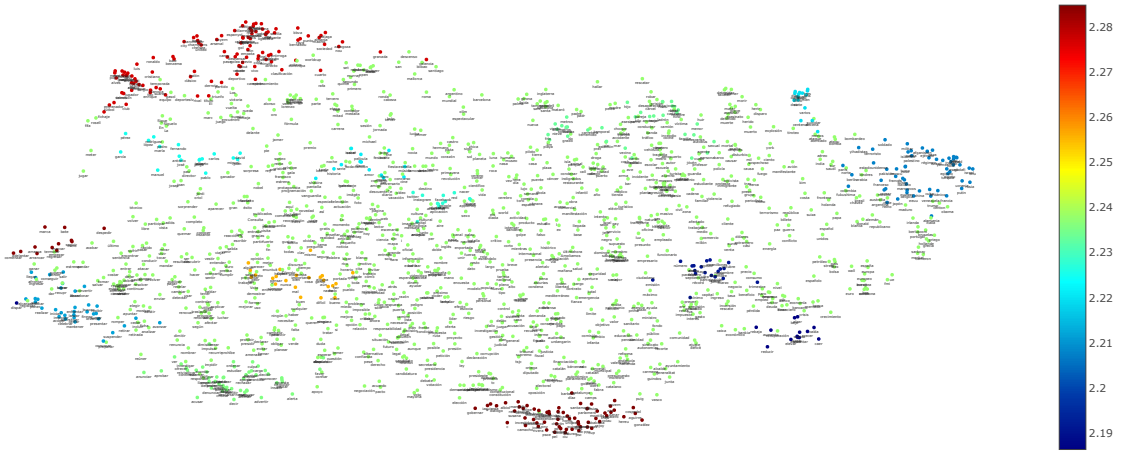
Cluster2AMI table. Weighted mean number of AMI per day, with initial pain between 6 a.m. and 2 p.m.



Cluster2AMI scatter plot. Mean number of AMI per day, with initial pain between 2 p.m. and 10 a.m.

Cluster	Mean	Var. rate	P-Value
0	2.23916	0.004 %	0.573
1	2.26295	1.067 %	0.001
2	2.23139	-0.343 %	0.759
3	2.26795	1.29 %	0.026
4	2.2249	-0.632 %	0.355
5	2.22573	-0.596 %	0.062
6	2.23766	-0.063 %	0.708
7	2.20589	-1.482 %	0.054
8	2.19239	-2.085 %	0.028
9	2.22354	-0.693 %	0.587
10	2.21945	-0.876 %	0.019
11	2.25147	0.554 %	0.223
12	2.25764	0.829 %	0.1
13	2.21154	-1.229 %	0.895
14	2.23951	0.02 %	0.403
15	2.2308	-0.369 %	0.705

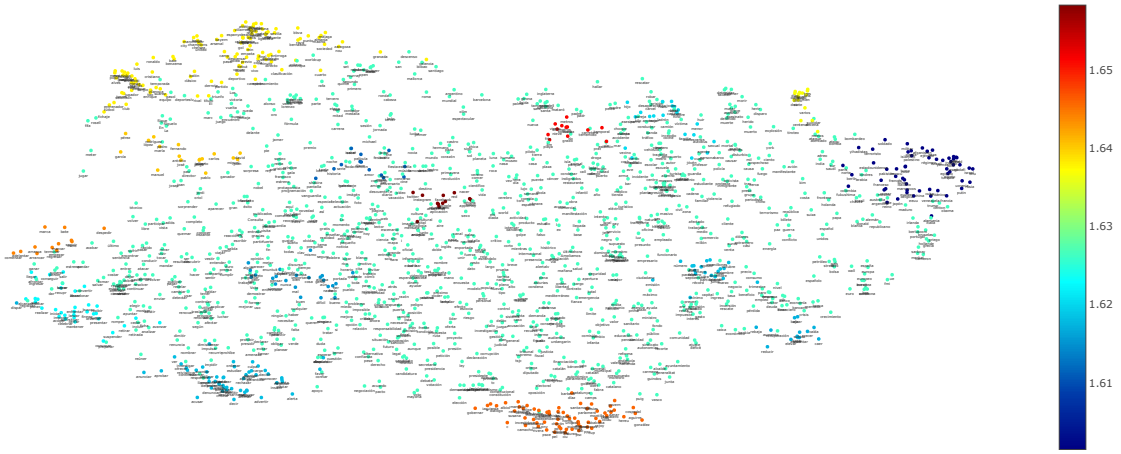
Cluster2AMI table. Mean number of AMI per day, with initial pain between 2 p.m. and 10 a.m.



Cluster2AMI scatter plot. Weighted mean number of AMI per day, with initial pain between 2 p.m. and 10 a.m.

Cluster	Mean	Var. rate	P-Value
0	2.24375	-0.136 %	0.033
1	2.2943	2.114 %	0.0
2	2.21669	-1.34 %	0.452
3	2.29275	2.045 %	0.013
4	2.22726	-0.87 %	0.327
5	2.23408	-0.567 %	0.127
6	2.24764	0.037 %	0.953
7	2.2042	-1.896 %	0.033
8	2.1952	-2.297 %	0.043
9	2.26171	0.663 %	0.653
10	2.21687	-1.333 %	0.003
11	2.24939	0.115 %	0.464
12	2.25805	0.5 %	0.226
13	2.20679	-1.781 %	0.671
14	2.24956	0.122 %	0.749
15	2.23469	-0.539 %	0.722

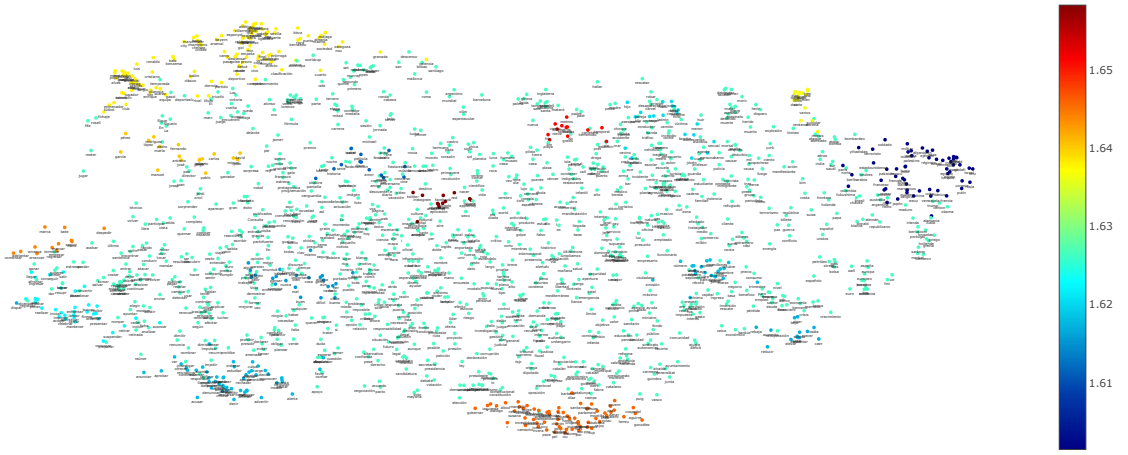
Cluster2AMI table. Weighted mean number of AMI per day, with initial pain between 2 p.m. and 10 a.m.



Cluster2AMI scatter plot. Mean number of AMI per day, with initial pain between 10 p.m. and 6 a.m.

Cluster	Mean	Var. rate	P-Value
0	1.62902	0.176 %	0.945
1	1.64422	1.111 %	0.146
2	1.62261	-0.218 %	0.265
3	1.64906	1.409 %	0.124
4	1.63178	0.346 %	0.235
5	1.61073	-0.949 %	0.059
6	1.6257	-0.028 %	0.636
7	1.61576	-0.639 %	0.533
8	1.6161	-0.618 %	0.716
9	1.63953	0.823 %	0.456
10	1.62625	0.006 %	0.629
11	1.61771	-0.519 %	0.447
12	1.64524	1.173 %	0.27
13	1.64109	0.918 %	0.47
14	1.62031	-0.36 %	0.419
15	1.63911	0.796 %	0.286

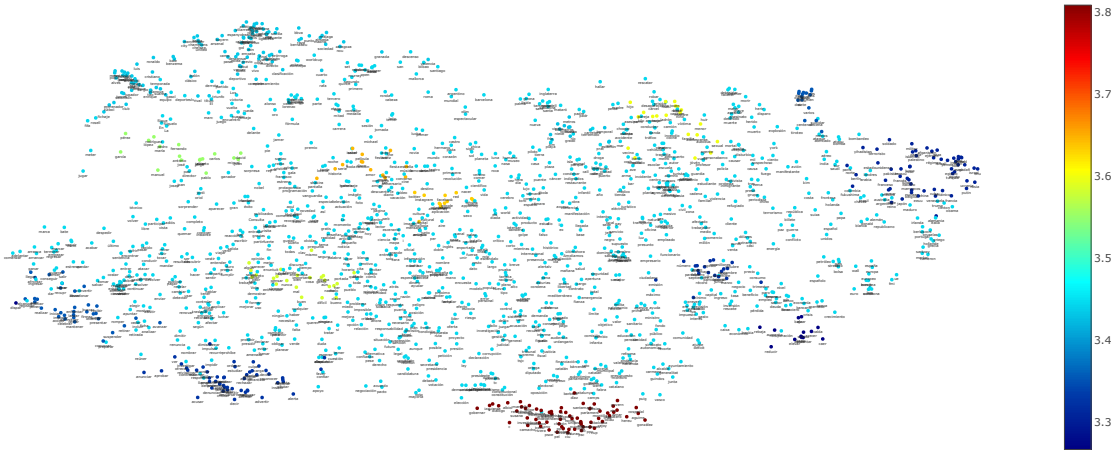
Cluster2AMI table. Mean number of AMI per day, with initial pain between 10 p.m. and 6 a.m.



Cluster2AMI scatter plot. Weighted mean number of AMI per day, with initial pain between 10 p.m. and 6 a.m.

Cluster	Mean	Var. rate	P-Value
0	1.62441	0.052 %	0.471
1	1.65608	2.002 %	0.009
2	1.6113	-0.756 %	0.229
3	1.64794	1.501 %	0.175
4	1.62453	0.059 %	0.476
5	1.59308	-1.878 %	0.005
6	1.6199	-0.226 %	0.942
7	1.59929	-1.496 %	0.413
8	1.60661	-1.045 %	0.631
9	1.63014	0.405 %	0.691
10	1.61622	-0.453 %	0.692
11	1.61106	-0.771 %	0.214
12	1.63504	0.706 %	0.188
13	1.61738	-0.381 %	0.908
14	1.573	-3.115 %	0.034
15	1.6168	-0.418 %	0.526

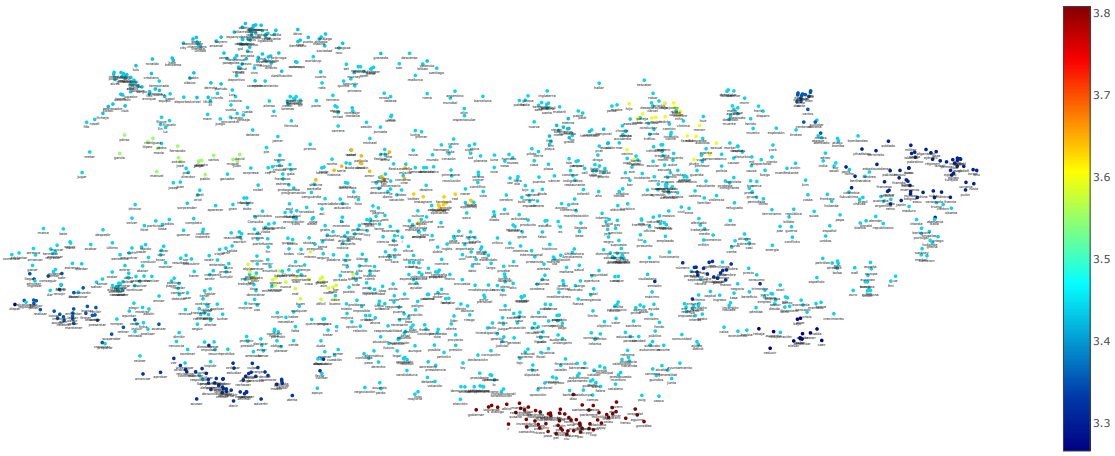
Cluster2AMI table. Weighted mean number of AMI per day, with initial pain between 10 p.m. and 6 a.m.



Cluster2AMI scatter plot. Mean number of AMI per day, from people with medical history.

Cluster	Mean	Var. rate	P-Value
0	3.41226	-2.409 %	0.0
1	3.45423	-1.209 %	0.231
2	3.29192	-5.851 %	0.0
3	3.62786	3.757 %	0.028
4	3.36387	-3.793 %	0.002
5	3.35978	-3.91 %	0.004
6	3.5517	1.579 %	0.093
7	3.25689	-6.853 %	0.007
8	3.32182	-4.996 %	0.013
9	3.54999	1.53 %	0.551
10	3.36045	-3.891 %	0.0
11	3.55528	1.681 %	0.31
12	3.37828	-3.381 %	0.685
13	3.4017	-2.711 %	0.36
14	3.62671	3.724 %	0.141
15	3.48824	-0.236 %	0.469

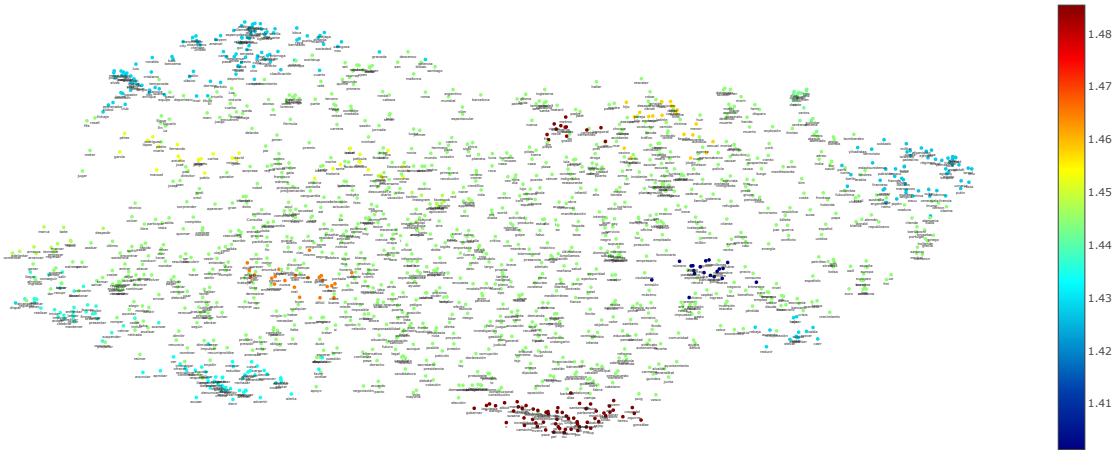
Cluster2AMI table. Mean number of AMI per day, from people with medical history.



Cluster2AMI scatter plot. Weighted mean number of AMI per day, from people with medical history.

Cluster	Mean	Var. rate	P-Value
0	3.51285	-1.629 %	0.0
1	3.61496	1.23 %	0.685
2	3.32565	-6.872 %	0.0
3	3.85395	7.923 %	0.017
4	3.51394	-1.599 %	0.107
5	3.44487	-3.533 %	0.015
6	3.72578	4.333 %	0.009
7	3.25365	-8.888 %	0.004
8	3.39261	-4.996 %	0.035
9	3.60082	0.834 %	0.85
10	3.45002	-3.389 %	0.001
11	3.65746	2.42 %	0.189
12	3.47743	-2.621 %	0.681
13	3.39673	-4.881 %	0.048
14	3.71805	4.117 %	0.186
15	3.5513	-0.553 %	0.492

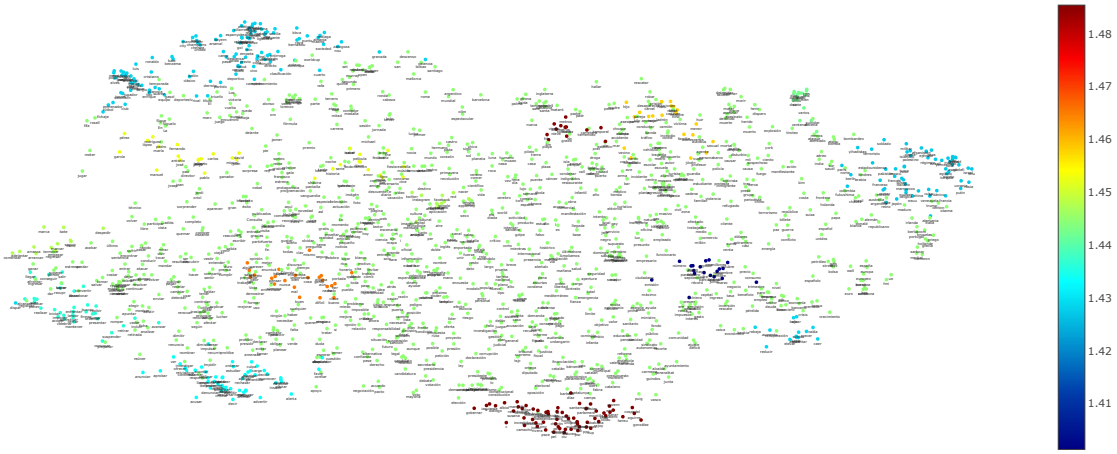
Cluster2AMI table. Weighted mean number of AMI per day, from people with medical history.



Cluster2AMI scatter plot. Mean number of AMI per day, from people with diabetes.

Cluster	Mean	Var. rate	P-Value
0	1.4441	-0.188 %	0.31
1	1.43432	-0.864 %	0.017
2	1.4311	-1.086 %	0.092
3	1.46654	1.363 %	0.012
4	1.4444	-0.168 %	0.807
5	1.4345	-0.852 %	0.091
6	1.4579	0.765 %	0.379
7	1.42719	-1.357 %	0.265
8	1.40231	-3.076 %	0.034
9	1.45732	0.725 %	0.768
10	1.43559	-0.776 %	0.313
11	1.46405	1.191 %	0.135
12	1.43662	-0.705 %	0.978
13	1.47199	1.739 %	0.099
14	1.46291	1.112 %	0.8
15	1.45984	0.9 %	0.988

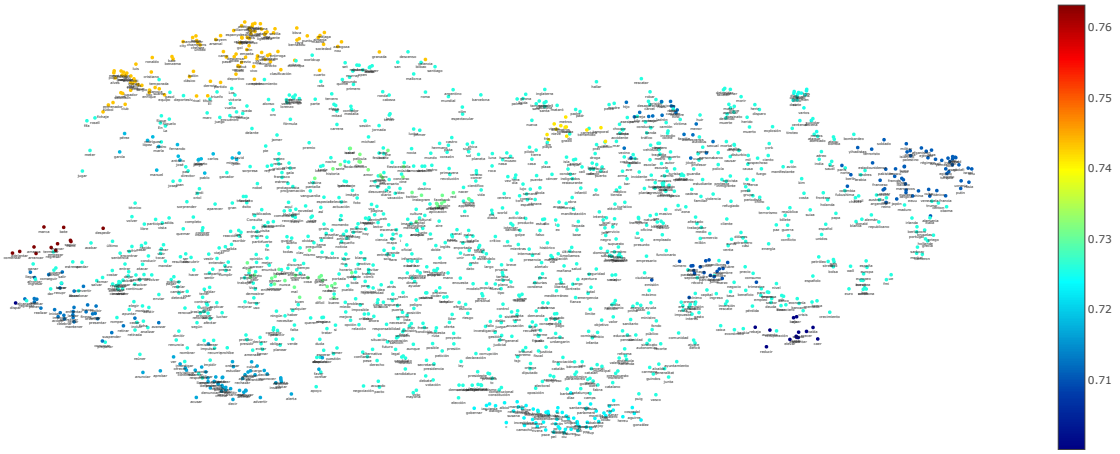
Cluster2AMI table. Mean number of AMI per day, from people with diabetes.



Cluster2AMI scatter plot. Weighted mean number of AMI per day, from people with diabetes.

Cluster	Mean	Var. rate	P-Value
0	1.45016	-0.137 %	0.125
1	1.44319	-0.618 %	0.037
2	1.43403	-1.249 %	0.389
3	1.49279	2.798 %	0.014
4	1.46745	1.053 %	0.557
5	1.43599	-1.113 %	0.101
6	1.45396	0.124 %	0.971
7	1.42884	-1.606 %	0.439
8	1.40078	-3.538 %	0.028
9	1.46934	1.183 %	0.693
10	1.43785	-0.986 %	0.125
11	1.48108	1.992 %	0.106
12	1.43257	-1.349 %	0.928
13	1.45567	0.242 %	0.439
14	1.46625	0.971 %	0.845
15	1.44926	-0.2 %	0.274

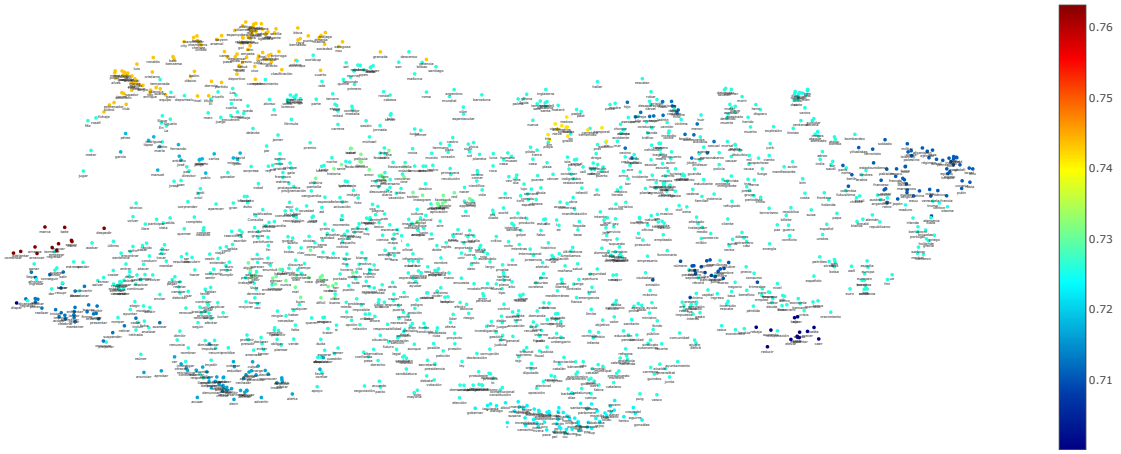
Cluster2AMI table. Weighted mean number of AMI per day, from people with diabetes.



Cluster2AMI scatter plot. Mean number of AMI per day, from people with a previous AMI.

Cluster	Mean	Var. rate	P-Value
0	0.72446	-0.287 %	0.572
1	0.74074	1.952 %	0.001
2	0.71983	-0.925 %	0.095
3	0.71979	-0.931 %	0.597
4	0.7218	-0.654 %	0.837
5	0.71435	-1.679 %	0.048
6	0.71797	-1.181 %	0.217
7	0.70412	-3.088 %	0.05
8	0.70905	-2.409 %	0.116
9	0.71989	-0.917 %	0.498
10	0.71769	-1.22 %	0.103
11	0.73249	0.817 %	0.613
12	0.75336	3.69 %	0.01
13	0.73684	1.416 %	0.413
14	0.74503	2.544 %	0.683
15	0.7306	0.558 %	0.793

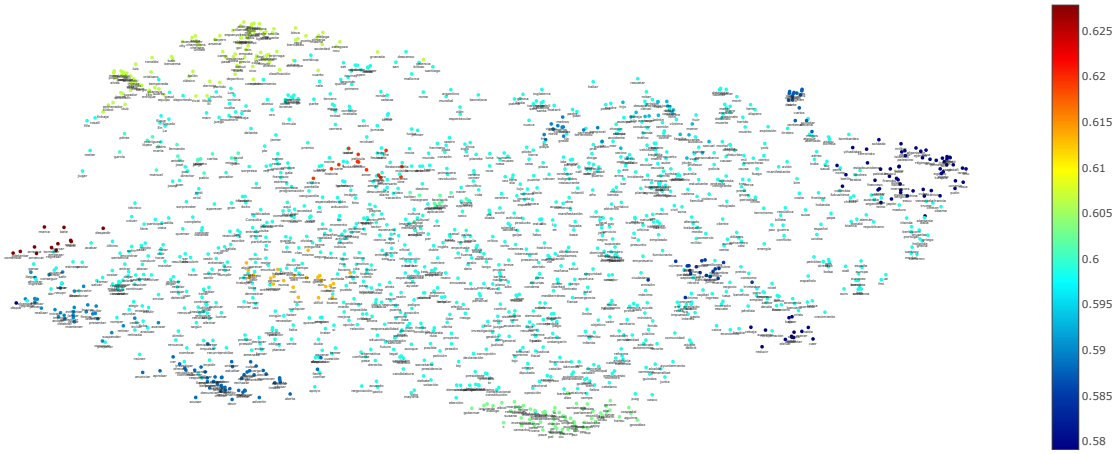
Cluster2AMI table. Mean number of AMI per day, from people with a previous AMI.



Cluster2AMI scatter plot. Weighted mean number of AMI per day, from people with a previous AMI.

Cluster	Mean	Var. rate	P-Value
0	0.72719	-0.387 %	0.571
1	0.75777	3.802 %	0.0
2	0.71809	-1.634 %	0.059
3	0.73006	0.006 %	0.625
4	0.72553	-0.614 %	0.881
5	0.70387	-3.583 %	0.032
6	0.72113	-1.218 %	0.164
7	0.70381	-3.59 %	0.069
8	0.71329	-2.292 %	0.16
9	0.72663	-0.465 %	0.771
10	0.72528	-0.649 %	0.374
11	0.73804	1.099 %	0.958
12	0.76323	4.549 %	0.005
13	0.73725	0.99 %	0.628
14	0.76719	5.092 %	0.274
15	0.7267	-0.455 %	0.856

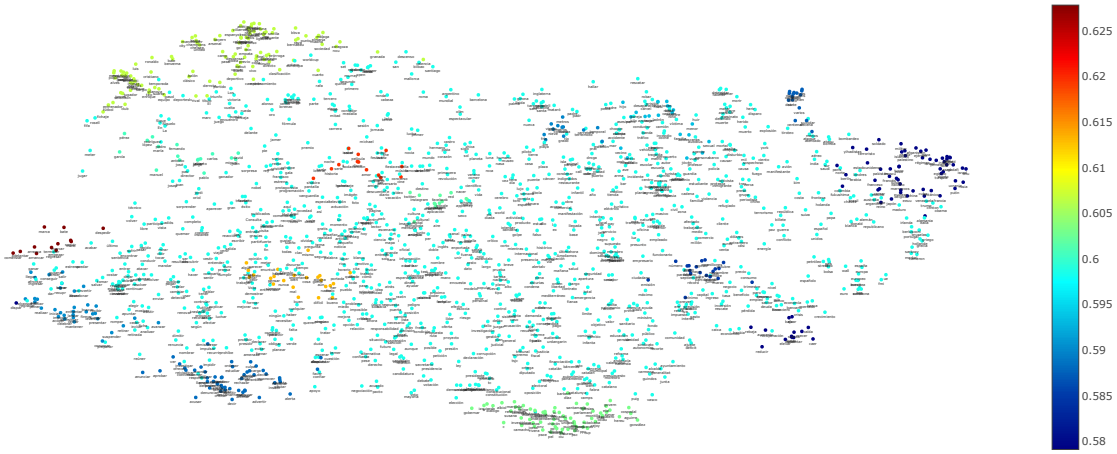
Cluster2AMI table. Weighted mean number of AMI per day, from people with a previous AMI.



Cluster2AMI scatter plot. Mean number of AMI per day, from people with a previous angioplasty.

Cluster	Mean	Var. rate	P-Value
0	0.59656	-0.776 %	0.069
1	0.60503	0.633 %	0.33
2	0.58837	-2.139 %	0.023
3	0.5993	-0.321 %	0.843
4	0.58877	-2.071 %	0.018
5	0.58498	-2.702 %	0.011
6	0.59377	-1.241 %	0.406
7	0.5814	-3.297 %	0.004
8	0.58593	-2.543 %	0.24
9	0.59987	-0.225 %	0.957
10	0.59126	-1.658 %	0.075
11	0.60897	1.288 %	0.109
12	0.615	2.29 %	0.143
13	0.58879	-2.068 %	0.424
14	0.62583	4.092 %	0.097
15	0.59581	-0.901 %	0.979

Cluster2AMI table. Mean number of AMI per day, from people with a previous angioplasty.



Cluster2AMI scatter plot. Weighted mean number of AMI per day, from people with a previous angioplasty.

Cluster	Mean	Var. rate	P-Value
0	0.60262	-0.61 %	0.132
1	0.6273	3.461 %	0.005
2	0.58303	-3.84 %	0.014
3	0.61751	1.846 %	0.461
4	0.5973	-1.487 %	0.531
5	0.5804	-4.276 %	0.007
6	0.6035	-0.464 %	0.591
7	0.58754	-3.096 %	0.012
8	0.59002	-2.688 %	0.28
9	0.60571	-0.101 %	0.971
10	0.60229	-0.664 %	0.608
11	0.6142	1.301 %	0.165
12	0.61709	1.777 %	0.08
13	0.57058	-5.894 %	0.035
14	0.63603	4.9 %	0.05
15	0.59735	-1.479 %	0.91

Cluster2AMI table. Weighted mean number of AMI per day, from people with a previous angioplasty.

Universitat Politècnica de Catalunya

Barcelona, 2019