

# DIRECCIÓN DE DESARROLLO DIGITAL

**Unidad de Científicos de Datos**



**DNP** Departamento  
Nacional  
de Planeación



**GOBIERNO DE COLOMBIA**

## Clasificación de proyectos de inversión para la estimación del presupuesto estatal en función del cumplimiento de los Objetivos de Desarrollo Sostenible.

### Resumen

La relevancia de los Objetivos de Desarrollo Sostenible se da por la urgencia de responder a los problemas primordiales de las naciones y mejorar la calidad de vida de sus habitantes. De lo anterior se desprende la importancia de realizar el seguimiento al cumplimiento de los 17 Objetivos, en el caso de Colombia, se hará a través del análisis de 46776 documentos correspondientes a proyectos de inversión referentes al Sistema General de Regalías (SGR), Presupuesto General de la Nación (PGN) y del Sistema Unificado de Inversiones y Finanzas Públicas (SUIFP) territorial. Se busca clasificarlos en 16 grupos de ODS diferentes, representados por cada uno de los Objetivos de Desarrollo Sostenible (ODS). Usando un modelo Tf-Idf para vectorizar documentos y haciendo una reducción de dimensionalidad con el método SVD truncado para buscar similitudes entre vectores con la distancia coseno, se establece la pertenencia de un documento a uno o varios de los 16 grupos establecidos por 16 documentos específicos.

*Palabras clave:* ODS, expresiones regulares, vectorizar, SVD, divergencia Kullback-Liebr.

### Abstract

The relevance of the Sustainable Development Goals is given by the urgency of responding to the primordial problems of nations and thus improving the quality of life for all. Due to their importance, it's clear the need to trace the accomplishment of all 17 Goals, which in case of Colombia, will be done through the analysis of 24379 documents with information about investment schemes from three different sources: *Sistema General de Regalías (SGR)*, *Presupuesto General de la Nación (PGN)* and territorial *Sistema Unificado de Inversiones y Finanzas Públicas (SUIFP)*. The purpose of this document is to classify them in 16 different groups, each one of them defined by each Sustainable Development Goals (SDG). Through the Tf-Idf model, the investment documents are vectorized, and by applying the truncated SVD method to reduce the vectorial space where said vectors are, the cosine measurement is used to find similarities between them so a belonging relationship of a document to one or more SDG can be set.

*Key words:* SDG, regular expressions, vectorize, SVD, Kullback-Liebr divergence.

## Tabla de contenido

<b>1. Introducción</b> .....	<b>3</b>
<b>2. Objetivos</b> .....	<b>4</b>
2.1. Objetivos generales .....	4
2.2. Objetivos específicos .....	4
<b>3. Estado del arte</b> .....	<b>4</b>
<b>4. Marco Teórico</b> .....	<b>5</b>
4.1. Expresiones regulares.....	5
4.2. Procesamiento del lenguaje natural ( <i>Text mining</i> o NLP).....	5
4.3. Métodos de extracción de información.....	6
4.3.1. Vectorización Tf-Idf .....	6
4.3.2. Latent Semantic Analysis (LSA) o Análisis Factorial de dos modos .....	7
4.4. t-Distributed Stochastic Neighbor Embedding (t-SNE) .....	8
4.5. Similitud coseno .....	9
<b>5. Descripción de los datos</b> .....	<b>9</b>
<b>6. Metodología</b> .....	<b>10</b>
6.1. Procesamiento de texto.....	10
6.2. Vectorización Tf-Idf.....	11
6.3. Reducción de dimensionalidad .....	11
6.4. Bondad de Ajuste	
6.4.1. Error de entrenamiento .....	16
6.4.2. Frecuencias de palabras .....	16
6.5. Clasificación .....	16
<b>7. Resultados</b> .....	<b>17</b>
7.1. Resultados bajo el modelo Tf-Idf.....	17
7.2. Resultados bajo la metodología LSA .....	20
7.3. Clasificación multiclase .....	22
<b>8. Conclusiones</b> .....	<b>23</b>
<b>9. Referencias</b> .....	<b>24</b>

## 1. Introducción

“El 25 de septiembre de 2015, los líderes mundiales adoptaron un conjunto de objetivos globales para erradicar la pobreza, proteger el planeta y asegurar la prosperidad para todos como parte de una nueva agenda de desarrollo sostenible. Cada objetivo tiene metas específicas que deben alcanzarse en los próximos 15 años. Para alcanzar estas metas, todo el mundo tiene que hacer su parte: los gobiernos, el sector privado y la sociedad civil.” (UN 2018)

El Departamento Nacional de Planeación como “entidad eminentemente técnica que impulsa la implantación de una visión estratégica del país en los campos social, económico y ambiental, a través del diseño, la orientación y evaluación de las políticas públicas colombianas, el manejo y asignación de la inversión pública y la concreción de las mismas en planes, programas y proyectos del Gobierno” (DNP 2018) debe tener presente que la creación de políticas públicas deben tener un enfoque hacia el cumplimiento de los ODS. El compromiso anterior implica que, algún porcentaje del presupuesto nacional debe ser destinado a proyectos que además de suplir necesidades, también respondan a las metas que deben alcanzarse para los diferentes objetivos, dado que estos objetivos hacen parte de algún sector de la economía que se ve directamente relacionado con las metas.

Como primer acercamiento, el Departamento Nacional de Planeación quiere estimar la cantidad de dinero invertida, a través de diferentes proyectos de inversión, en los temas referentes a los ODS y aunque para algunos sectores como Comercio, industria y turismo o Minas y energía, no puede asegurarse que los proyectos sigan los lineamientos de sostenibilidad que buscan cumplir los ODS, sí puede tenerse una estimación de cuánto dinero está siendo invertido, según las necesidades de la nación, en los temas referentes a los ODS.

Colombia como parte de las Naciones Unidas, tiene como fecha límite, igual que los demás países, el año 2030 para verificar el cumplimiento de los ODS, y la Dirección de Seguimiento y Evaluación de Políticas Públicas del DNP como brazo técnico del gobierno hará una revisión de la inversión del presupuesto para tener para el control de la asignación del mismo para el cumplimiento de los ODS.

Los 17 ODS agrupados según el documento expedido por el Programa de las Naciones Unidas para el Desarrollo (PNUD) para Colombia son:




















Grupo 1	Grupo 2	Grupo 3	Grupo 4
	 		   
Grupo 5	Grupo 6	Grupo 7	Grupo 8
 	 	   	  

Tabla 1. ODS agrupados

Una característica de los ODS es que no son grupos disyuntos, en los grupos 5 y 6 se incluye el objetivo 12 de producción y consumo responsable, también se observa un traslape del objetivo 2 de hambre cero en los grupos 2 y 5. Además, se tiene que en sí mismos los objetivos tampoco lo son, pues los grupos están formados de tal manera que cada uno tiene una temática que se ve reflejada en uno o más objetivos.

## 2. Objetivos

### 2.1. Objetivos generales

- Realizar un análisis presupuestal del gasto del Gobierno Colombiano en los ODS.

### 2.2. Objetivos específicos

- Clasificar 46776 documentos de inversión en los 16 grupos planteados de acuerdo con los ODS.
- Hacer una clasificación multiclase para considerar el grado de similitud de un proyecto de inversión con más de un objetivo de desarrollo sostenible.
- Determinar medidas de ajuste de clasificación

## 3. Estado del arte

Debido a los avances de la tecnología y en especial en términos de capacidad computacional que permite el manejo de grandes bases de datos, en el análisis textual se cuenta con herramientas que van más allá de los conteos de palabras y frecuencias de las mismas.

Para el análisis estadístico de textos por lo general se plantean descripciones o inferencias. En el caso de la estadística descriptiva, se tienen métodos multivariados para tablas de contingencia de términos, como el análisis de correspondencias múltiples; éste fue el caso de *El análisis estadístico de datos textuales. La lectura según los escolares de enseñanza primaria* publicado por Mónica Bécue, Ludovic Lebart y Núria Rajadell en 1992 donde se hizo una encuesta con dos preguntas abiertas a una muestra de 895 alumnos, un análisis de

correspondencias múltiples en seis grupos diferentes de estudiantes para encontrar diferencias o similitudes en las respuestas a las preguntas ya mencionadas.

En el caso del análisis de inferencia para datos textuales, por lo general se busca la clasificación de documentos, lo cual supone diferencias, aunque no muy grandes, en la metodología dependiendo de si se trata de un análisis supervisado o no supervisado. Los métodos comúnmente usados para clasificación supervisada, son Naive Bayes, Rocchino, Vecino más próximo y Knn, mientras que para la clasificación no supervisada, se tiene k-means, clustering en general y redes neuronales. (Ray 2018)

Actualmente en el análisis textual se tienen algoritmos de machine learning que además de utilizar métodos de conteos y frecuencias, hacen uso de redes neuronales para analizar los pesos en términos o frases representativas. Un ejemplo de esto es la teoría desarrollada por Quoc Le y Tomas Mikolov para el aprendizaje no supervisado de representaciones continuas para grandes bloques de texto, el objetivo es que palabras con significado similar estén cerca en un espacio vectorial, que llevado a representaciones por documento, logran que documentos con contenido similar tengan una distancia mínima entre sí, comparado con documentos con contenido diferente, esto ocurre porque los documentos similares están rodeados por el mismo contexto.

Más allá de las diferencias que hay entre los análisis supervisados y no supervisados, existe un factor común a estos que es la representación numérica o vectorial de los datos textuales y aunque en algunos casos, ésta es tan sencilla como conteos de palabras, en otros casos, para ir más allá de la neta presencia de éstas en los documentos, se tienen pesos de grupos de términos. Una vez se tienen éstos vectores como representación de documentos, pueden usarse para su clasificación medidas probabilísticas como es el caso del método *Naive Bayes* en el que en una de sus versiones, usa el método de máxima verosimilitud para determinar si un documento pertenece o no a una clase determinada, o medidas de distancia que en su principio de mínima o máxima determina la cercanía o similitud de dichos vectores.

Anna Huang (2008) presentó una comparación de algunas medidas de similitud para clustering de documentos textuales, tomó en consideración las métricas: distancia euclidiana, similitud coseno, coeficiente de Jaccard, coeficiente de correlación de Pearson y divergencia de Kullback-Leibler. La comparación se hizo para siete conjuntos de datos, cada uno con sus documentos previamente clasificados y posteriormente vectorizados con el modelo Tf-Idf. La autora concluye que la distancia euclidiana es la que peor desempeño tiene, mientras que el coeficiente de Pearson y la divergencia de Kullback-Leibler tienen resultados cercanos a la clasificación original de los documentos con grupos balanceados, pero en general para todas las medidas, excepto la euclidiana, se obtiene una efectividad similar aceptable. (Huang 2008)

## 4. Marco Teórico

### 4.1. Expresiones regulares

Se define expresión regular como una expresión que describe un conjunto de cadenas de caracteres (strings) o a un conjunto de parejas ordenadas de *strings*. Los operadores más comunes son, concatenación, unión, intersección, complemento, iteración y composición. (Mitkov 2003)

Computacionalmente “una expresión regular es un tipo específico de patrón de texto que puede ser usado en diferentes aplicaciones modernas y lenguajes de programación. Puede usarse para verificar si una entrada corresponde a un patrón de texto, o para encontrar texto que coincida con un patrón dentro de un cuerpo de texto grande, o para reemplazar texto que coincide con el patrón de otro texto, o para dividir un texto en subtítulos.” (Goyvaerts & Levithan 2009)

### 4.2. Procesamiento del lenguaje natural (*Text mining* o NLP)

Se le llama lenguaje natural al lenguaje que es usado cotidianamente por los seres humanos para comunicarse; lenguajes como inglés, hindú o portugués. Se denomina Procesamiento del Lenguaje Natural al cubrimiento de cualquier tipo de manipulación computacional del lenguaje natural. (Steven Bird 2009)

Un modelo comúnmente usado en métodos de clasificación de texto es el Bag of Words (BoW). Como parte de este modelo, una pieza de texto (frase o documento) es representado como una bolsa o conjunto múltiple de palabras, sin tener en cuenta su gramática e incluso el orden de las palabras, la frecuencia de cada palabra es usada como una característica para entrenar al clasificador. (Waldron 2018)

El modelo Tf-Idf sigue el mismo principio de BoW, la diferencia está en el peso que les da a las frecuencias de las palabras del corpus, pues estas son multiplicadas por un factor de importancia que está en función de la ocurrencia de la palabra y el número de documentos en los que aparece.

En el análisis de texto se hace importante no sólo la identificación de términos frecuentes, por el hecho de caracterizar un documento, sino también en algún grado la semántica del texto. Es por esto que se define "colocación como una secuencia de palabras que se encuentran juntas frecuentemente de manera inusual. Así, se tiene que *red wine* es una colocación, mientras que *wine* no lo es" (Steven Bird 2009).

De lo anterior se tiene entonces de manera relevante, el cálculo de bigramas. Estos son listas de palabras consecutivas por pares, que para el caso de la frase: "Análisis textual en Python", quedan determinados por los siguientes pares de palabras: (Análisis, textual), (textual, en), (en, Python). Por consiguiente, si se le hace un análisis de frecuencias por bigramas a un texto, se tendría que para pares de palabras como "New York" o "institución educativa" que comúnmente se encuentran juntas, la frecuencia sería alta y por lo tanto podría considerarse como un sólo término. De manera análoga, se puede estar interesado en trigramas o n-gramas.

### 4.3. Métodos de extracción de información

#### 4.3.1. Vectorización Tf-Idf

Para el estudio de minería de datos textuales se cuenta con la medida TF.IDF para determinar la importancia de palabras en varios documentos.

La diferencia entre palabras poco comunes que nos dicen algo sobre los documentos y aquellas que no lo hacen tiene que ver con concentración de las palabras útiles en pocos documentos.

La medida formal de qué tan concentradas, en relativamente pocos documentos, son las ocurrencias de una palabra dada, es llamada TF.IDF (*Term Frequency times Inverse Document Frequency*). Normalmente, es computada como sigue. Supóngase que se tiene una colección de  $N$  documentos. Defínase  $f_{ij}$  como la frecuencia de la palabra  $i$  en el documento  $j$ . Entonces se define la *frecuencia de término*  $TF_{ij}$  como:

$$TF_{ij} = \text{"Peso de la palabra } i \text{ en el texto } j\text{"} = \frac{f_{ij}}{\max_k \{f_{kj}\}}$$

El IDF para un término se define como sigue. Suponga que el término  $i$  aparece  $n_i$  de los  $N$  documentos de la colección. Entonces :

$$IDF_i = \text{"Importancia del término } i \text{ en la colección } N \text{ de documentos"} = \log_2 \left( \frac{N}{n_i} \right)$$

Los términos con los puntajes más altos de Tf-Idf (que está configurado por el producto de los dos términos anteriormente mencionados) por lo general, son los términos que mejor describen contenido del documento. (Jure Leskovec 2011)

La medida anterior se usa para vectorizar los documentos, pues debido a que el objetivo de este trabajo es la clasificación de los proyectos de inversión en 16 diferentes objetivos, se tiene como criterio de agrupación la distancia entre vectores, que resultan ser los 46776 proyectos y 16 ODS. Esta vectorización se hace a través de la medida Tf-Idf que se explica a continuación.

La dimensión de cada vector es el número de palabras únicas del texto del total de documentos, sin embargo, debido a que la aplicación de ésta teoría se hace en el lenguaje de programación Python, se tienen parámetros que pueden variar e incluso la medida Tf.Idf calculada por las librerías usadas, tiene algunas variaciones. Bajo la librería de python, sklearn versión 0.19.1, se presenta el siguiente cambio en el cálculo de la medida Tf.Idf (Pedregosa et al. 2011):

$$\begin{aligned} \text{tf-idf}(t,d) &= \text{tf}(t,d) \times \text{idf}(t) \\ &= \text{tf}(t,d) \times \log\left(\frac{1+n_d}{1+\text{df}(d,t)}\right) + 1, \end{aligned}$$

En donde  $\text{tf}(t,d)$  es la frecuencia absoluta del término  $t$  en el documento  $d$ ,  $n_d$  el número total de documentos y  $\text{df}(d,t)$  el número de documentos que contienen el término  $t$ .

Adicional a lo anterior, para la vectorización de los documentos se tiene bajo la función “*tfidfVectorizer*” el siguiente conjunto de parámetros:

1. *analyzer*: si la caracterización debe ser hecha por palabras o por caracteres.
2. *ngram\_range*: los límites inferior y superior del rango de los  $n$  valores para diferentes  $n$ -gramas a ser extraídos.
3. *max\_df*: Al construir el vocabulario ignorar los términos que tenga una frecuencia de documentos estrictamente mayor al límite.
4. *min\_df*: Al construir el vocabulario ignorar los términos que tenga una frecuencia de documentos estrictamente menor al límite.

Aunque la medida per se, hace la discriminación de palabras frecuentes en todos los documentos y les da un peso correspondiente a una baja importancia, los parámetros de la función permite ignorar en el corpus términos que sean altamente frecuentes en todos los documentos, lo que para el objetivo de clasificación resulta útil, pues si un término es frecuente en la mayoría de documentos, éste no hace un aporte a la caracterización y diferenciación de los proyectos.

El parámetro *ngram\_range* permite especificar el número de palabras consecutivas a analizar a la vez.

Es necesario notar que la vectorización Tf-Idf se hace bajo los 16 ODS, por lo que los vectores tienen dimensión del número único de palabras disponible en esos 16 textos, pues cuando se procede a vectorizar los documentos correspondientes a los proyectos de inversión, se usa la propiedad de transformación de la función anteriormente mencionada.

En cuanto se tiene un modelo Tf.Idf entrenado bajo un determinado número de textos, pueden inferirse vectores para la representación de nuevos documentos, esto quiere decir que se usa la matriz de términos de los documentos con los que se entrenó el modelo. Lo anterior implica que, si una palabra de un documento que se va a vectorizar bajo el modelo no se encuentra en la matriz de términos, esta no se va a ver reflejada en el vector, pues sólo se tienen en cuenta los términos de los documentos de entrenamiento. Para los documentos nuevos que se vectoricen bajo el modelo ya entrenado, se van a tener vectores representativos de la misma dimensión de los vectores resultantes para los documentos de entrenamiento.

#### 4.3.2. Latent Semantic Analysis (LSA) o Análisis Factorial de dos modos

Ésta técnica fue diseñada para hacer una reducción de dimensionalidad en las técnicas de extracción de información, pues en tablas léxicas se tienen grandes cantidades de información en las que puede haber un alto porcentaje que es irrelevante, sobre todo por la naturaleza del lenguaje en el que distintas palabras tienen significados similares y por lo tanto representan un mismo concepto. “LSA simultáneamente modela la relación entre documentos basado en las palabras que lo constituyen y las relaciones entre palabras basado en sus ocurrencias en documentos. Al usar menos dimensiones para la representación que el número de palabras



únicas que hay, LSA encuentra similitudes entre términos que son útiles para resolver el problema de extracción de información.” (Dumais 2008)

De acuerdo con Dumais, la reducción de dimensionalidad es resultante de la reducción por rango de la descomposición en valores singulares hecha a la matriz de términos, en la que se retienen los  $k$  valores propios más altos, por lo que la matriz resultante de rango  $k$  resulta ser la mejor aproximación a la matriz original en el sentido de mínimos cuadrados. A continuación, se muestra el primer proceso de descomposición de la matriz de términos en valores singulares (SVD):

La descomposición en valores singulares de  $A$ , denotada por  $SVD(A)$ , es definida como:

$$A = U\Sigma V^T$$

$$U^T U = V^T V = I_n, \quad \Sigma = \text{diag}\{\sigma_1, \dots, \sigma_n\}, \quad \begin{cases} \sigma_i > 0, & 1 \leq i \leq r \\ \sigma_j = 0, & j \geq r + 1 \end{cases}$$

En donde las primeras  $r$  columnas de las matrices  $U$  y  $V$  definen los vectores propios ortonormales asociados a los  $r$  valores propios diferentes de cero de  $AA^T$  y  $A^T A$ , respectivamente. Las columnas de  $U$  y  $V$  son los vectores singulares izquierdos y derechos, respectivamente, y los valores singulares de  $A$  son definidos como los elementos de la diagonal de  $\Sigma$  que son las raíces no negativas de los  $n$  valores propios de  $AA^T$ .

El método truncado SVD captura la estructura subyacente más importante en la asociación de términos a documentos y al mismo tiempo remueve el ruido o variabilidad en el uso de palabras que limitan los métodos basados en palabras de extracción de información (M.W.Berry et al. 1995).

Finalmente, en el espacio reducido se cuenta con la representación de los documentos y términos, por lo que resulta sencillo hacer un análisis de comparación para cualquier combinación de estos.

#### 4.4. t-Distributed Stochastic Neighbor Embedding (t-SNE)

La metodología t-SNE hace una reducción de dimensionalidad de grandes vectores de información a dos o tres dimensiones con fines de visualización. Este método “es capaz de capturar la estructura local de datos en altas dimensiones, mientras también revela la estructura global como clusters en varias escalas” (van der Maaten & Hinton 2008).

Para el método SNE, se tiene que la similitud de un punto  $x_j$  a un punto  $x_i$  es la probabilidad condicional  $p_{ji}$  de que  $x_i$  escoja a  $x_j$  como su vecino, si los vecinos fueran escogidos en proporción a una función de probabilidad normal, centrada en  $x_i$ . La probabilidad  $p_{ij}$  está dada por:

$$p_{i|j} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-\|x_i - x_k\|^2}{2\sigma_i^2}\right)}$$

en donde  $\sigma_i$  es la varianza. Para sus pares  $y_i, y_j$  se computa una probabilidad similar, notada por  $q_{ji}$ .

Para ambas se establece una varianza para la normal de  $\frac{1}{\sqrt{2}}$ . La similitud de un punto  $y_i$  con  $y_j$  está dada por

$$q_{j|i} = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}$$

Como el interés es modelar las similitudes por pares, se establecen los valores de  $p_{ji}$  y  $q_{ji}$  en 0.

Para medir la fidelidad del modelo para el que  $q_{ji}$  modela  $p_{ji}$ , se usa la divergencia de Kullback-Liebr, medida que es minimizada por el método SNE por cada punto usando el método de descenso de gradiente.

Para regiones densas, un valor más pequeño de  $\sigma_i$  por lo general es más apropiado que en regiones dispersas, no obstante, es claro que, dependiendo de la escogencia de este valor, la distribución de  $P_i$  cambia, por lo que se hace una búsqueda binaria con una perplejidad dada por el usuario. La perplejidad se define como

$$\text{Perp}(P_i) = 2H(P_i)$$

en donde  $H(P_i)$  es la entropía de Shannon medida en bits (debido al uso de  $\log_2$ ).

La diferencia del método SNE tradicional simétrico con el t-SNE viene dada por el uso de una distribución t-student en lugar de una distribución normal, pues en un espacio de bajas dimensiones, se puede usar una distribución que tenga colas mucho más pesadas que una normal para convertir distancias en probabilidades. Lo anterior permite que una distancia moderada en el espacio de alta dimensionalidad sea fielmente modelada por una distancia mucho más grande en el espacio. Al hacer el cambio a un t-student de un grado de libertad, las ecuaciones quedan definidas de la siguiente forma

$$q_{j|i} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_i - y_l\|^2)^{-1}}$$

Se usa esta distribución por la propiedad de que  $(1 + \|y_i - y_j\|^2)^{-1}$  se aproxima a la ley del cuadrado inverso para grandes distancias para pares  $\|y_i - y_j\|$  en el espacio de baja dimensionalidad, lo que hace que en ese nuevo espacio, la representación de las probabilidades conjuntas sean casi invariantes a los cambios de escala del espacio para puntos alejados. (van der Maaten & Hinton 2008)

#### 4.5. Similitud coseno

Se define al ángulo entre dos vectores diferentes de cero como la medida de similitud entre dos vectores. "El ángulo  $\phi$  entre  $u$  y  $v$  está definido como el ángulo no negativo más pequeño. Si  $\phi$  es el ángulo entre ellos, entonces" (Grossman 2007)

$$\cos(\phi) = \frac{u \cdot v}{\|u\|_2 \|v\|_2}$$

Para documentos con la misma temática, su representación vectorial debe ser similar, por lo que la medida coseno entre éstos debe ser cercana a 1 o -1 y para documentos con temáticas diferentes, se espera una medida de coseno cercana a 0, sin embargo, bajo la librería *Scipy* de python, en su versión 1.1.0, la distancia coseno es definida como

$$\cos(u, v) = 1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2}$$

en donde  $\|\cdot\|_2$  es la norma  $l_2$ , por lo que para los resultados, se dice que dos documentos son similares, si el coseno es cercano 0 o 2, se habla de temáticas contrarias si el coseno es cercano a 1.

#### 5. Descripción de los datos

Se cuenta con una base de datos de 46776 textos: 24379 documentos de proyectos del Sistema General de Regalías (SGR), Presupuesto General de la Nación (PGN) y SUIFP territorial; 14731 documentos de funcionamiento de hacienda y cooperación internacional; 7580 documentos de proyectos privados de Prosperidad Social (Mapa Social) y 86 Documentos de SGP y Entidad Territorial Descentralizada. Cada documento cuenta con un código único, nombre del proyecto, descripción, problema central, causas, efectos, objetivos, productos y alternativas. En el caso específico de 10530 documentos, se tiene además la información del sector y subsector al que pertenecen.

Otra base de datos contiene un documento por cada uno de los 17 ODS, con el nombre, objetivos, metas y los indicadores con los que van a medir el impacto de cada uno, siendo cada uno de los campos de tipo textual.

## 6. Metodología

### 6.1. Procesamiento de texto

*Nota: En el caso de los ODS, se tuvieron en cuenta sólo los 16 primeros objetivos, pues el 17 no tiene indicadores de medición y además hace alusión a los 16 objetivos anteriores.*

De acuerdo con la información disponible, se formaron dos bases de datos, la primera contiene la información de los ODS y la segunda la información de los documentos a clasificar.

Ambas bases de datos quedaron reducidas a un solo campo de información en el que se une todo el texto disponible por "individuo", sin embargo, en algunos casos, los textos habían sido extraídos de páginas web, por lo que la información de interés estaba en sintaxis html y por lo tanto se tenía el texto a analizar, pero con ruido. Una vez se tienen las dos tablas consolidadas, se procede a una primera depuración del texto.

Debido a la forma en la que se recibió la información de los proyectos fue necesario el uso de expresiones regulares, y para los casos en que la información estaba en sintaxis html se hizo una depuración adicional. Para el caso de texto con estructura html debe eliminarse todo aquello que fuera una etiqueta, una declaración de fuente y declaración de parámetros, es decir, todo lo que se encontrara dentro de los símbolos `<...>`, `@...>`, `{...}`. Además, se tenía que afuera de los símbolos mencionados, donde se encuentra la mayor parte de código html, debían eliminarse también nombres de entidades y clases. Las primeras siempre inician con el carácter "&" y las segundas tienen en medio de caracteres alfanuméricos un punto.

Una vez se tenían establecidos los patrones del código html que contenían información irrelevante para el análisis, se usó la librería de python "re" (versión 2.2.1), pues "esta librería provee operaciones de coincidencia de expresiones regulares" (Python 2018), luego se reemplazan los patrones encontrados y los espacios de 2 o más campos por un sólo espacio en blanco con el fin de que al tokenizar (separar por términos) los textos se hagan de la manera correcta. Para hacer la depuración semántica del texto, en el que se remueven términos irrelevantes para la clasificación de cada proyecto, se embebe el texto disponible por palabra en una lista.

Por ejemplo, para la frase "Python, como lenguaje de programación, es el mejor para el análisis de texto", los tokens correspondientes serían los siguientes: ["Python", ",", "como", "lenguaje", "de", "programación", ",", "es", "el", "mejor", "para", "el", "análisis", "de", "texto"], así se tiene separado por palabras el texto de los proyectos. Como puede verse en el ejemplo, para caracteres que no son palabras se hace distinción, así, en el resultado de la depuración se garantiza la conservación de las palabras sin caracteres adicionales que puedan afectar la comparación entre términos iguales.

Una vez se tiene una lista de tokens por proyecto, se procede a eliminar toda la información irrelevante de los textos.

Debido a su naturaleza, como proyectos de inversión, se tiene información que contiene números, ya sea para hacer referencia a cantidades monetarias, o mediciones técnicas. Los números per se no son relevantes en la clasificación del texto, pues para frases como "empedrado con mortero 1:4" o "refuerzo  $F_y = 4200\text{Kg/cm}^2/8 - 1/2$ ", los números pierden importancia al no ser comparables con otros textos.

Computacionalmente se crea una función que hace un barrido a través de todas las 46776 listas de tokens, entrando en cada una de éstas y revisando token por token la presencia de números. Finalmente, desecha todo token que contenga únicamente números, quedándose con todos los demás.

Después de tener el texto sin caracteres numéricos, se eliminan las palabras irrelevantes del texto, llamadas "stopwords". Estas en español hacen referencia a los artículos, preposiciones y similares. Algunas de éstas son: "porque", "entre", "muy", "desde", "fueron", etc. Adicionalmente se hace un filtro de palabras que son comunes para todos los ODS, pues naturalmente, si una palabra está presente en los 16 ODS, no existe razón para discriminar algún proyecto por la misma. Algunas de las palabras filtradas son: "zona", "población", "sistema", "comunidad", "proyecto", entre otras.

De lo anterior se espera que palabras como “la”, “de”, “a” entre otras, ya no hagan parte del corpus, sin embargo, después de hacer el procesamiento de texto anterior, y teniendo en cuenta el formato en el que estaban los textos, sigue habiendo casos de palabras como “ef”, “cj”, “vs” e incluso siglas como “am” y “fm” y abreviaciones como “cm” e “in”. Por lo anterior entonces se hace necesario el filtro de palabras con longitud menor o igual a 2.

Una nota final en éste proceso de depuración es que usualmente en el análisis de texto es común el uso de un “lematizador”, que busca llevar todas las palabras a su raíz, es decir, en el caso de los verbos, llevarlos todos al tiempo infinitivo, en el caso de los plurales a singulares y para sinónimos, usar una sola palabra que los represente. Esto es una práctica que simplifica la clasificación al tener el texto “estandarizado”.

En este caso no se usó lematizador puesto que había errores significativos que comprometían la precisión del clasificador, pues se encontró que para el término “menores” que en el texto hacía referencia a niños menores, pasado por lematizador era reemplazado por la palabra “pequeño”, y en el caso del documento del ODS 4 que se trata de educación de calidad, el término es frecuente y al cambiarlo afecta el peso asignado por el método Tf-Idf.

Aunque se decidió no usar un lematizador, sí se creó una función que trabaja como uno, pues como se verá más adelante, para un análisis de frecuencias era necesario que, sobre todo, para palabras en plural que significaban lo mismo, se viera su relevancia, luego bajo un diccionario creado a partir de resultados encontrados en una fase de prueba previa, la función hacía el reemplazo en términos similares. Algunos ejemplos de las palabras incluidas en éste diccionario, son: menores: niño, aguas: agua, redes: red.

## 6.2. Vectorización Tf-Idf

Una vez el texto de los ODS y los proyectos es depurado, se usó la metodología Tf-Idf para vectorizar cada documento. Para ello se utiliza la librería “*sklearn*” (versión 0.19.1) de python.

Como el objetivo es clasificar 46776 proyectos en 16 diferentes objetivos de desarrollo, se usa el modelo Tf-Idf para crear 16 vectores correspondientes a cada ODS, por lo que se tiene una matriz de 16 filas con 9277 columnas en la que cada columna hace referencia a una palabra o bigrama único de todo el corpus de los 16 ODS, matriz que se usa como referencia para luego crear los vectores representativos para los 24379 proyectos de inversión.

De lo anterior se tiene que cada documento de los ODS y de los proyectos queda representado por un vector en el espacio de dimensión 9278.

## 6.3. Reducción de dimensionalidad

Usando la función “*TruncatedSVD*” de la librería *sklearn* en su versión 0.19.1 de python, con un número de componentes a conservar igual a 2000, usando el método aleatorizado para solución del SVD, que hace referencia a un algoritmo probabilístico desarrollado por Halko et al. (2011) con un número de 5 iteraciones se hace la reducción de dimensionalidad de los vectores de la matriz de términos obtenida por el modelo Tf-Idf, de 9277 columnas a 2000.

Una vez realizado el proceso anterior, se obtiene una matriz de tamaño 24393×2000, en la que se tienen los vectores de los 16 objetivos y 24377 correspondientes a los proyectos de inversión. Dos de los documentos a clasificar no son usados en la reducción de dimensionalidad por ser vectores nulos. Estos dos vectores serán usados en una segunda fase de este proyecto para determinar una nueva categoría que represente a los documentos que no pertenecen a ninguna de las 16 categorías.

## 6.4. Bondad de ajuste

### 6.4.1. Error de entrenamiento

Para este criterio de ajuste, se tienen 20 sectores de la economía. Estos son: Agricultura, Agua potable y saneamiento básico, Ambiente y desarrollo sostenible, Ciencia y tecnología, Comercio, industria y turismo, Comunicaciones, Cultura, deporte y recreación, Defensa, Educación, Estadística, Inclusión social y reconciliación, Interior, Justicia y del derecho, Minas y energía, Planeación, Relaciones exteriores, Salud y protección social, Trabajo, Transporte y Vivienda.

Puede hacerse una relación de las temáticas de los 16 Objetivos, con los 20 sectores, pues se esperaría que para proyectos de inversión del sector de educación, el objetivo al que pertenezcan sea el 3 (educación de calidad), no obstante, hay objetivos y proyectos de inversión con temáticas relacionadas, por lo que un proyecto de inversión del sector de ciencia y tecnología podría ser asignado a casi cualquier objetivo debido a que la clasificación podría depender del impacto que se genere en cierto ámbito de la sociedad y no del hecho de este se haga a través de la ciencia y la tecnología.

Tras considerar este tipo de intersecciones entre sectores y Objetivos, se crean grupos de sectores que sean disyuntos entre sí, y que dentro de cada uno mantengan una similitud, por ejemplo, se agrupan los sectores Justicia y del derecho y Defensa en uno solo, pues ambos ocupan temáticas relacionadas con la justicia y paz, que se relaciona directamente con el Objetivo 16.

Para el caso de los ODS se hace una agrupación que cumpla con heterogeneidad entre grupos y homogeneidad dentro de ellos. Esto únicamente con el propósito de determinar el error de entrenamiento del modelo.

Los grupos de ODS quedan determinados de la siguiente manera:

<b>Grupo 1</b>  <b>1 FIN DE LA POBREZA</b>  <b>9 INDUSTRIA, INNOVACIÓN E INFRAESTRUCTURA</b>  <b>11 CIUDADES Y COMUNIDADES SOSTENIBLES</b>		<b>Grupo 2</b>  <b>2 HAMBRE CERO</b>  <b>12 PRODUCCIÓN Y CONSUMO RESPONSABLES</b>  <b>13 ACCIÓN POR EL CLIMA</b>  <b>14 VIDA SUBMARINA</b>  <b>15 VIDA DE ECOSISTEMAS TERRESTRES</b>		<b>Grupo 3</b>  <b>8 TRABAJO DECENTE Y CRECIMIENTO ECONÓMICO</b>  <b>10 REDUCCIÓN DE LAS DESIGUALDADES</b>	
<b>Grupo 4</b>  <b>3 SALUD Y BIENESTAR</b>	<b>Grupo 5</b>  <b>4 EDUCACIÓN DE CALIDAD</b>	<b>Grupo 6</b>  <b>5 IGUALDAD DE GÉNERO</b>	<b>Grupo 7</b>  <b>6 AGUA LIMPIA Y SANEAMIENTO</b>	<b>Grupo 8</b>  <b>7 ENERGÍA ASEQUIBLE Y NO CONTAMINANTE</b>	<b>Grupo 9</b>  <b>16 PAZ, JUSTICIA E INSTITUCIONES SÓLIDAS</b>

Tabla 1: Grupos de Objetivos de Desarrollo Sostenible

Para la agrupación de sectores por el criterio t-SNE se tomaron en cuenta sólo los que tuvieran una cantidad de documentos superior al 1% del total

Sector	% Doc.	Sector	% Doc.
Transporte	32.51	Ciencia y tecnología	2.90
Cultura, deporte y recreación	12.45	Planeación	1.50
Educación	10.98	Comercio, industria y turismo	0.75
Vivienda	10.18	Interior	0.46
Agua potable y saneamiento básico	8.72	Defensa	0.30
Agricultura	4.31	Justicia y del derecho	0.20
Inclusión social y reconciliación	4.28	Trabajo	0.18
Ambiente y desarrollo sostenible	3.53	Comunicaciones	0.15
Salud y protección social	3.32	Estadística	0.05
Minas y energía	3.14	Relaciones exteriores	0.009

Tabla 2: Documentos por sector

es decir que no se tienen en cuenta *comercio, interior, defensa, justicia, trabajo, comunicaciones, estadística y relaciones exteriores*, mientras que para los restantes 11 sectores, se buscan las superposiciones de puntos en el gráfico de t-SNE mostrado a continuación:

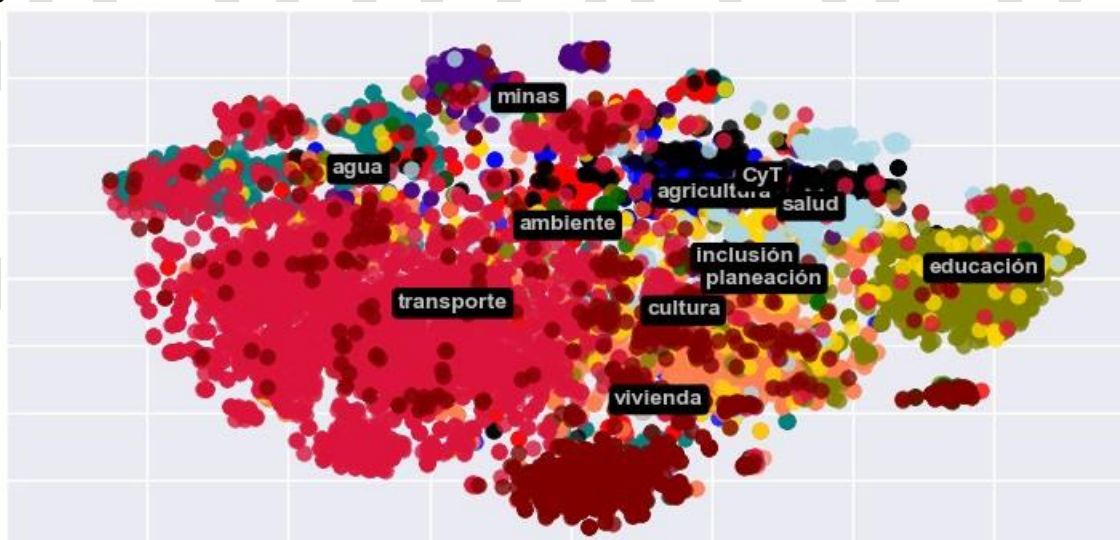


Figura 1: t-SNE

Se muestra en el gráfico a los documentos que tienen sector, en donde cada color representa uno diferente. Se ve en el gráfico, como algunos sectores, a pesar de que tienen puntos dispersos, también tienen una nube concentrada de puntos, además esas nubes concentradas se encuentran alejadas unas de otras. Usando la librería *“matplotlib”* (versión 2.1.2) de python, se escogen por separado los sectores para identificar la superposición de unos con otros, verificando así, que para los 5 sectores predominantes que se muestran con una grande cantidad de puntos concentrados (*Transporte, Vivienda, Educación, Agua y Minas*) se tiene lo que se llamará *“independencia”*. En la Imagen 1, se muestran estos sectores.

Se toman esos 5 sectores como 5 grupos independientes de sectores. Por otro lado, como se ve en la Figura 3, hay documentos de sectores que están dispersos a través de la nube completa de puntos; una vez más,

identificándolas con *matplotlib*, se tienen 3 sectores que podrían estar relacionados con cualquier otro. Se muestra en la Imagen 2. la distribución de los sectores *Inclusión*, *Ambiente* y *Planeación*.

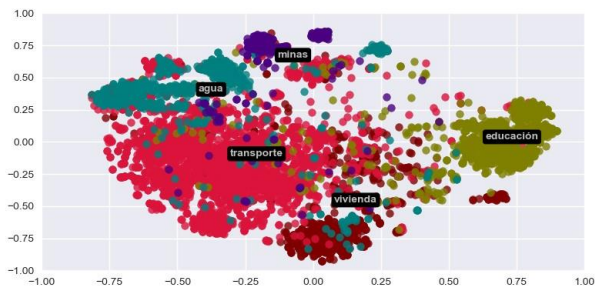


Imagen 1. t-SNE: Nubes concentradas

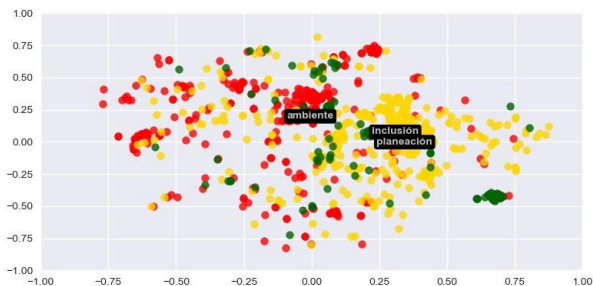


Imagen 2. t-SNE: Nube dispersa

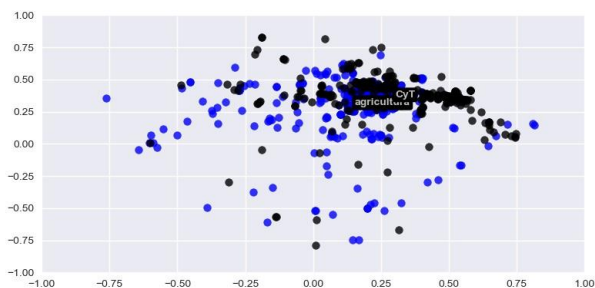


Imagen 3. t-SNE: Agrupación

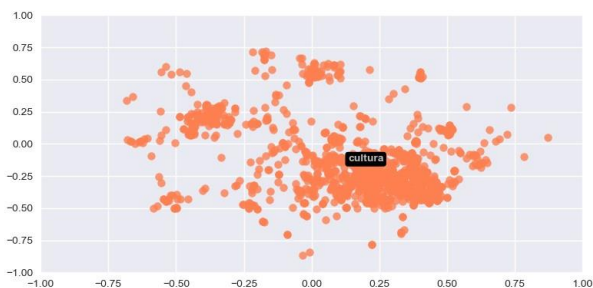


Imagen 4. t-SNE: Cultura

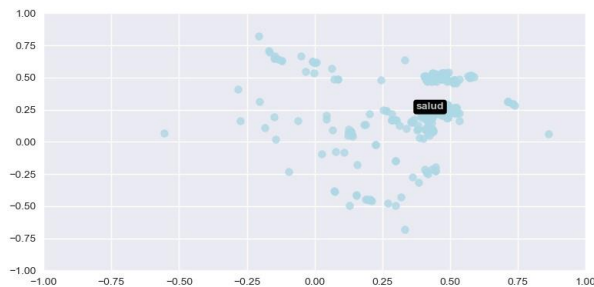


Imagen 5. Salud

Se identificaron dos sectores que se encuentran en el mismo lugar en el plano y que comparten una superposición alta, ver Imagen 3. Aunque para la nube de puntos de los sectores *Cultura* y *Salud* se tiene una alta dispersión, también se tiene que la nube más concentrada de los mismos se ubica en un lugar en el plano en la que éstas tienen prelación, por lo que se forman tres grupos de sectores, uno con *Agricultura* y *Ciencia y tecnología*, otro con *Salud* y el tercero con *Cultura*.

Se establecieron los grupos de sectores como sigue:

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Transporte	Cultura	Educación	Vivienda	Agua
Grupo 6	Grupo 7	Grupo 8	Grupo 9	Grupo 10

Minas	Salud	Agricultura Ciencia y tecnología Trabajo	Defensa Justicia Interior	Inclusión Ambiente Planeación Comercio
-------	-------	--	---------------------------------	---

Tabla 3: Caption

Para los documentos de los sectores *Comunicaciones, Estadística y Relaciones exteriores*, y para el Grupo 10 de sectores, no se va a hacer ningún supuesto, pues para los primeros se cuenta con muy pocos proyectos que no tienen relación con las temáticas de los Objetivos, y para los segundos la dispersión es alta y se espera que queden clasificados en los ODS de la misma manera.

Una vez hechos los grupos para ODS y para sectores, se hace una asignación de correspondencia entre ellos de acuerdo con sus temáticas, por lo que se espera una clasificación de la siguiente manera:

Grupos de sectores	Grupos de ODS	Grupos de sectores	Grupos de ODS
1	→ 1	6	→ 8
2	→ 6	7	→ 4
3	→ 5	8	→ 1,2
4	→ 3,1	9	→ 9
5	→ 7		

Tabla 4: Correspondencia de grupos de sectores a grupos de ODS

Al analizar la dispersión de los grupos de sectores, se optó por no tener en cuenta los documentos que estuviesen a una distancia menor a 0.05 de otros documentos que pertenecieran a grupos de sectores diferentes. En la siguiente Figura, se muestra para el Grupo 1 de sectores, que corresponde a *Transporte*, a los proyectos que no se tomaron en cuenta, delineados de verde.



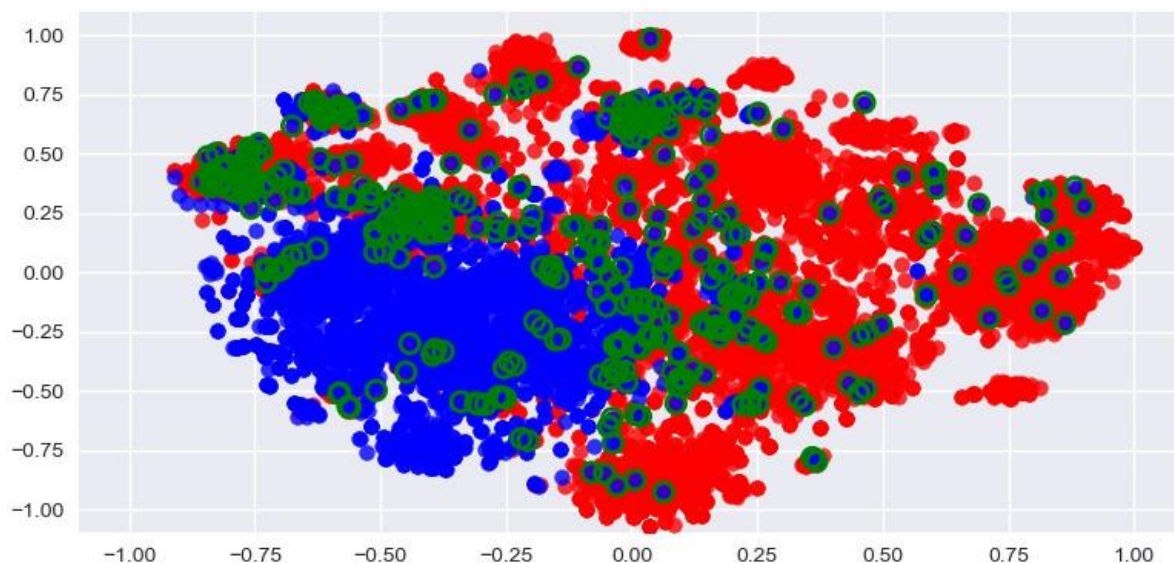


Figura 2: Documentos del sector *Transporte* excluidos del error de entrenamiento

El número de documentos que se conserva por grupo de sector se muestra en la siguiente Tabla:

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
2857	681	973	653	719
Grupo 6	Grupo 7	Grupo 8	Grupo 9	
268	210	376	3	

Tabla 5: Número de documentos para cálculo de error

Se tiene un total de 7136 documentos para calcular un error de entrenamiento. Este error está dado por

$$Err(Sec) = 1 - \left( \frac{\# \text{ de documentos clasificados en el ODS correspondiente a Sec}}{\# \text{ de documentos del Sec}} \right)$$

#### 6.4.2. Frecuencias de palabras

Al asignar cada proyecto de inversión a un ODS, se tomaron los grupos de documentos pertenecientes a cada objetivo y se hizo un conteo de palabras para el conjunto total. Se tomó como buen ajuste de clasificación por ODS, si a las 20 palabras más frecuentes del grupo de documentos correspondiente, tuviera relación conceptual con la temática de ODS.

#### 6.5. Clasificación

Una vez se tienen los documentos en un espacio vectorial en el que son comparables, se usa el ángulo entre vectores como regla de clasificación que, en primera medida, le asigna a un documento el ODS con quien tenga

la menor medida de coseno. En segunda instancia se determina un umbral para la asignación de un documento a uno o más objetivos según la distancia entre la mínima y la inmediatamente siguiente.

## 7. Resultados

### 7.1. Resultados bajo el modelo Tf-Idf

Habiendo procesado los datos textuales con la metodología de expresiones regulares, los textos son modelados con Tf-Idf, con lo que se obtuvo una matriz de 9006 términos y bigramas únicos en las columnas y 46792 filas, en las que las primeras 16 son representaciones vectoriales de los 16 ODS, y las restantes 46776 de los proyectos de inversión.

Tras vectorizar todos los proyectos, se hace la clasificación de cada documento a un solo objetivo, tomando como criterio la similitud coseno. La clasificación es calculada para vectores estrictamente positivos, pues los elementos de la representación vectorial de documentos siempre tienen números mayores a cero, al tratarse de conteos de palabras, ésta medida está comprendida en el intervalo  $[0,1]$ , sin embargo, en un espacio de tan alta dimensión, como en el que se encuentran los vectores del modelo Tf-Idf, está presente el efecto Hughes en el que se tienen distancias concentradas. En la Figura 3. se muestra el coseno mínimo para cada documento:

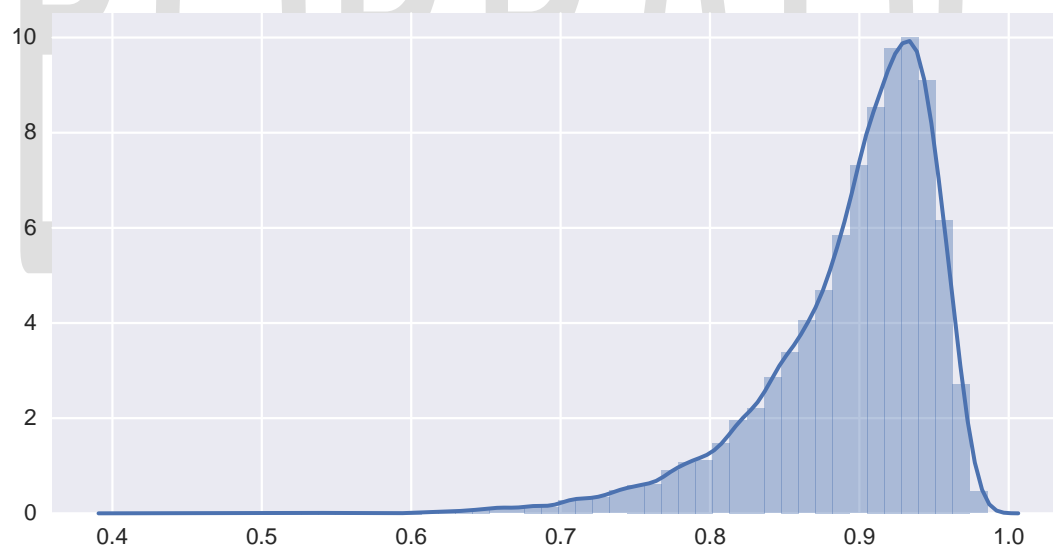


Figura 3: Histograma coseno mínimo por documento

Como puede verse, hay una alta concentración del valor mínimo de coseno en el intervalo  $(0.88,0.95)$  y son muy pocas distancias menores a 0.5, aun así, se hace la clasificación de los documentos. Clasificación que se presenta en la Figura 4. Se observa que el 25% del total de los documentos quedaron clasificados en el Objetivo 11: Ciudades y comunidades sostenibles, seguido por el 16,4% en el Objetivo 4: Educación de calidad y el 10,8% en el Objetivo 6: Agua limpia y saneamiento.

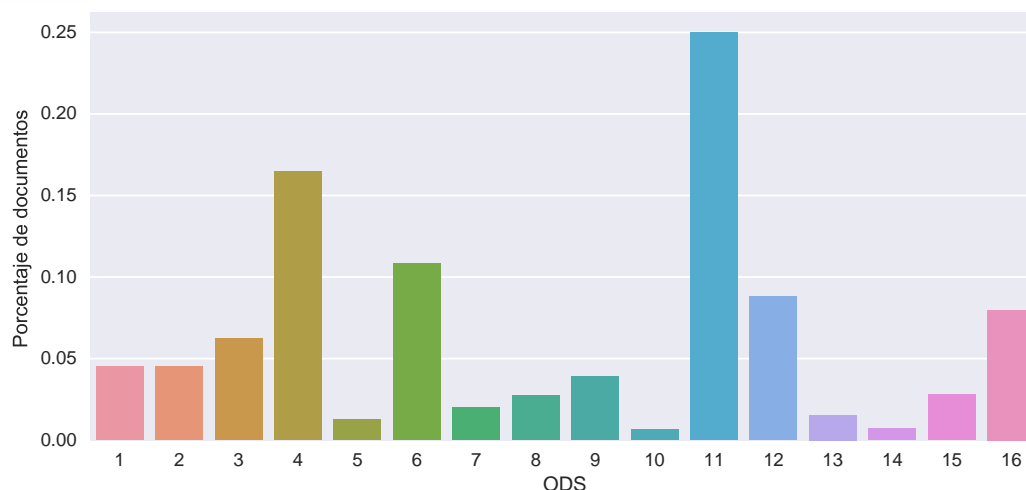


Figura 4: Clasificación por ODS

Teniendo en cuenta que se trata de un modelo no supervisado, se cuenta únicamente con dos criterios para evaluar de error del ajuste del modelo. El primero consiste en verificar la correspondencia conceptual de las frecuencias de palabras más altas de todos los documentos clasificados en cada categoría y el segundo en obtener un error de entrenamiento a través del subconjunto de datos con el que se tiene la información del sector de la economía al que pertenece el proyecto de inversión.

ODS	Palabra	Frec.	ODS	Palabra	Frec.	ODS	Palabra	Frec.	ODS	Palabra	Frec.
1	vivienda	2403	2	productores	2428	3	salud	10457	4	educativa	18388
	familias	1877		producción	1635		servicios	2693		escolar	6143
	rural	1858		asistencia	1435		atención	2142		educación	5851
	social	1536		niño	1324		social	1785		estudiantes	5415
	condiciones	1072		técnica	1286		entemedade	1423		instituciones	4999
5	mujer	1344	6	agua	11328	7	red	1800	8	placa	1245
	derechos	374		alcantarillado	4485		energía	1533		huella	1096
	género	337		red	3724		gas	992		actividades	686
	actividades	322		suministro	3386		eléctrica	958		persona	603
	social	266		instalación	2908		tensión	521		discapacidad	602
9	red	2644	10	ingresos	186	11	vivienda	7480	12	material	5229
	innovación	1469		social	182		actividades	5597		actividades	2428
	investigación	1329		actividades	119		vial	4445		pavimento	2293
	transporte	1223		generación	116		condiciones	4435		suministro	1936
	nacional	784		nacional	112		municipal	4357		ambiental	1814
13	cambio	874	14	marina	361	15	conservación	1620	16	información	3677
	gestión	774		costera	352		áreas	1507		nacional	3130
	climático	619		recursos	324		ambiental	1275		gestión	2946
	nacional	486		pesca	264		manejo	1144		servicios	2255
	ambiental	460		nacional	261		recursos	1110		derechos	2243

Tabla 7: Frecuencias más altas de palabras por ODS

Puede verse en la Tabla 7 cómo los temas correspondientes a cada ODS tienen relación con los conceptos que describen las palabras con más frecuencia, incluso para objetivos con pocos documentos como es el caso del Objetivo 5 de igualdad de género, en donde las palabras más frecuentes son “*mujer*”, “*derecho*”, “*género*”, “*actividades*” y “*social*”. Este resultado da indicios de que la clasificación se está haciendo de manera adecuada.

Bajo el segundo criterio, se obtuvieron los siguientes errores de entrenamiento:

<b>Grupo 1</b>	<b>Grupo 2</b>	<b>Grupo 3</b>
0.6212	0.9926	0.8293
<b>Grupo 4</b>	<b>Grupo 5</b>	<b>Grupo 6</b>
0.5849	0.8748	0.9776
<b>Grupo 7</b>	<b>Grupo 8</b>	<b>Grupo 9</b>
0.9761	0.4281	1

Tabla 8: Errores de entrenamiento

Según estos resultados, las suposiciones que se hicieron de correspondencia de los grupos de los sectores a los grupos de los Objetivos son erradas o que la clasificación del modelo no funciona. Sin embargo, tras un análisis de las clasificaciones por grupos de sectores, se encontró la sobreestimación del modelo con respecto al Objetivo 11 de Ciudades y comunidades sostenibles. Se muestran a continuación las clasificaciones por grupos de sectores.

ODS	G1	G2	G3	G4	G5	G6	G7	G8	G9
1	124	44	48	32	42	15	9	18	0
2	126	29	43	26	35	12	5	16	0
3	101	15	28	19	16	9	5	14	1
4	463	117	168	109	120	46	38	66	0
5	29	5	7	6	7	3	4	5	0
6	414	89	133	81	85	38	27	52	2
7	64	30	31	28	19	6	5	7	0
8	94	18	27	18	18	6	5	10	0
9	128	34	39	29	28	6	5	12	0
10	19	6	6	5	4	2	2	2	0
11	839	190	308	182	208	83	72	111	0
12	344	71	104	75	80	25	22	48	0
13	29	9	13	9	11	9	5	11	0
14	14	5	5	4	3	1	1	1	0
15	36	9	11	6	8	4	4	4	0
16	47	12	20	11	13	4	3	2	0

Tabla 9: Clasificación por grupos de sectores

Aun cuando los errores de entrenamiento son altos, y dan a entender que existe un mal ajuste, de acuerdo con las frecuencias de palabras, las clasificaciones en los Objetivos (exceptuando el 11) son correctas, aunque sean pocos los documentos clasificados en estos. Por esta razón se contempla un análisis de clasificación basado en las distancias entre vectores, pero teniendo en cuenta las dos menores distancias de un documento a los Objetivos

## 7.2. Resultados bajo la metodología LSA

Una vez más se hace uso de la matriz de términos Tf-Idf para vectorizar los documentos, con la que se tiene la matriz anterior de 46792 filas con 9006 columnas, que a través de la metodología LSA, será reducida sustancialmente.

Tomando las primeras 300 componentes que recogen el 70.32% de la varianza de los datos, se reduce la matriz a 24393 filas y 300 columnas, de esta manera se espera reducir el efecto de Hughes y como consecuencia se puede dar que las distancias sean más determinantes para clasificar los documentos.

A continuación, se muestra la comparación de los histogramas del coseno mínimo para todos los documentos con y sin reducción de dimensionalidad

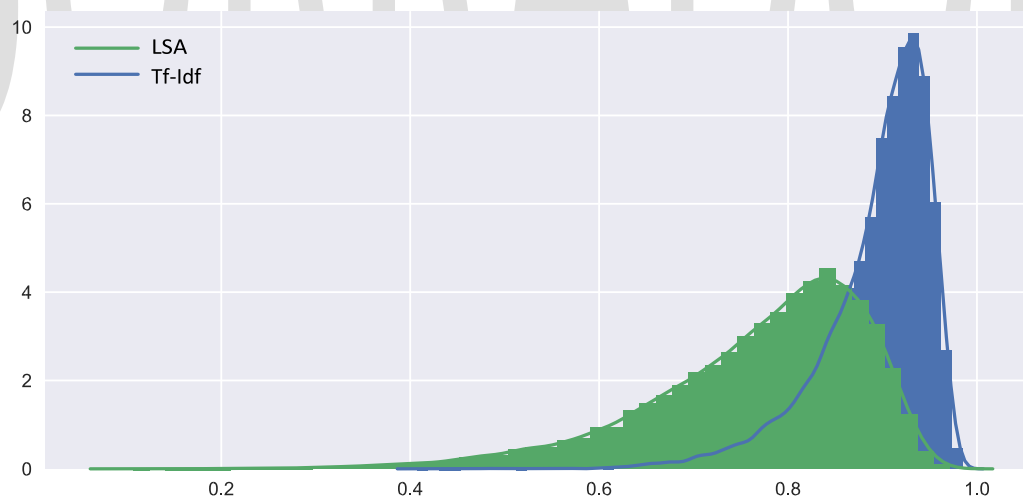


Figura 5: Comparación de histogramas de coseno mínimo con LSA y con Tf-Idf

Puede verse que existe una diferencia en las distancias generadas bajo los vectores de dimensión completa y los reducidos, se tiene una mayor dispersión de las distancias y por lo tanto una mejor clasificación. La clasificación de los 24377 proyectos de inversión con los vectores de dimensión 300, se presenta en la Figura 6.

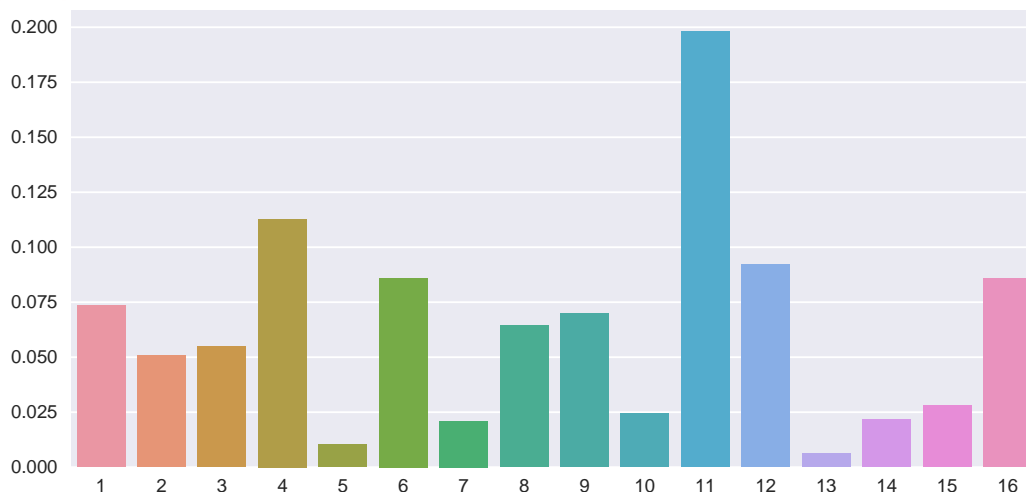


Figura 6: Clasificación con vectores reducidos

La diferencia entre las clasificaciones con los vectores originales y los reducidos, es evidente, pues ahora, el 19.79% quedaron clasificados en el Objetivo 11, el 11% en el Objetivo 4 y el 9.21% en el 12, por lo que los porcentajes se redujeron y el tercer objetivo con más documentos clasificados cambió de ser el 6 al 12.

Se muestran a continuación los dos criterios de clasificación:

ODS	Palabra	Frec.	ODS	Palabra	Frec.	ODS	Palabra	Frec.	ODS	Palabra	Frec.
1	vivienda	5650	2	productores	2465	3	salud	9922	4	educativa	16528
	rural	2876		niño	2283		servicios	2482		educación	5253
	familias	2527		producción	1566		atención	1856		escolar	4922
	social	2100		asistencia	1488		medicinas	1488		estudiantes	4868
	condiciones	1771		medicinas	1427		medicinas	1427		instituciones	4129
5	mujer	1240	6	agua	10461	7	red	1899	8	placa	2023
	género	346		alcantarillado	3944		energía	1639		actividades	1839
	violencia	335		red	1121		energía	1121		discapacidad	1821
	actividades	275		gas	1089		energía	1089		material	1770
	equidad	246		acueducto	2485		tensión	560		persona	1699
9	red	4399	10	actividades	524	11	actividades	4612	12	material	4729
	transporte	3369		educativa	438		vivienda	4362		ambiental	2793
	investigación	1892		nacional	432		municipal	4019		actividades	2591
	innovación	1880		obras	428		espacios	3975		suministro	2369
	vial	1513		gestión	1063		vial	3617		pavimento	2039
13	cambio	674	14	recursos	914	15	conservación	1749	16	información	3939
	climático	546		nacional	864		áreas	1654		nacional	3282
	ambiental	256		información	634		ambiental	1279		gestión	2993
	mitigación	229		actividades	451		manejo	1203		derechos	2434
	gestión	226					especies	1130		atención	2414

Tabla 11: Frecuencias más altas de palabras por ODS

De acuerdo con este criterio, la clasificación se está haciendo de manera adecuada, pues las palabras más frecuentes tienen mayor relación con las temáticas de cada ODS.

El criterio de error de entrenamiento arroja lo siguiente:

Tabla 12: Errores de entrenamiento

Aun cuando los errores de entrenamiento siguen siendo altos, por la misma razón de la clasificación con los vectores completos, algunos de ellos bajaron y otros subieron, pero en general, se mantuvieron con respecto a los errores de entrenamiento bajo el análisis con vectores completos.

### 7.3. Clasificación multiclase

Considerando el hecho de que un proyecto de inversión esté relacionado con más de una temática de los ODS, no se tiene en cuenta sólo la menor distancia de un documento a un ODS, sino las menores, dependiendo de un umbral establecido y de qué tan cerca éstas estén entre sí.

Para este análisis se usó la matriz reducida LSA.

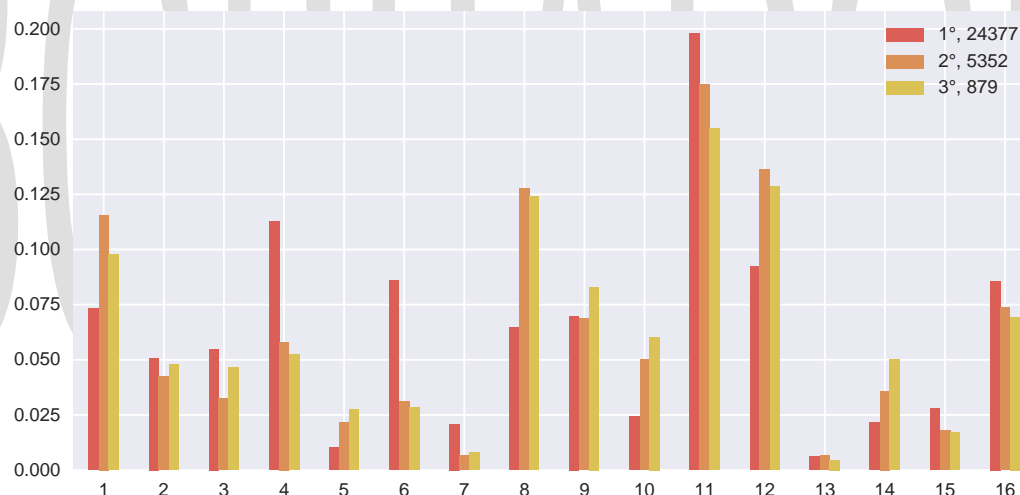


Figura 7: Orden de pertenencia de los documentos a cada ODS (Gamma=0.1)

Bajo este análisis, para 5352 documentos se tiene una segunda clasificación, y para 879, una tercera. Todo esto bajo un límite de distancia de  $1 - \text{Gamma}$ , es decir de 0.9.

Lo anterior quiere decir que existen 879 documentos que fueron clasificados en 3 ODS diferentes de acuerdo con tres distancias que estaban dentro del intervalo  $(0.9, 1)$  (recordando una vez más que se está trabajando con la similitud coseno en donde 1 es textos iguales y 0 textos contrarios). Así pues, se tiene en consideración la relación que puede tener un proyecto con más de una temática de los ODS, de acuerdo con las agrupaciones de ODS que se hizo para la bondad de ajuste, se puede esperar que un documento tenga multiclase correspondiente a los ODS de un grupo ya conformado.

## 8. Conclusiones

- La sobreestimación del objetivo 11, se debe a la información disponible que hay de éste, pues en su descripción están palabras que se esperan sean exclusivas para otros, como: *cultura*, *ciencia*, *infraestructura* y *pobreza*, entre otras.
- Los proyectos de inversión quedan clasificados en el objetivo 11, por la naturaleza de las descripciones en las que se menciona a la comunidad y a la creación de infraestructura, por lo que aun cuando un proyecto de inversión se trate de la construcción de un colegio, que se esperaría fuera clasificado en el ODS 4 de educación, por la descripción técnica de éste, queda clasificado en el 11.
- Los errores tan altos de clasificación son una consecuencia de la poca información que se tiene de los Objetivos de Desarrollo Sostenible, claramente los proyectos de inversión los superan en cantidad de palabras únicas por tratarse de proyectos.
- La reducción de dimensionalidad de los vectores reflejó la disminución del efecto Hughes en las distancias.
- Se propone para líneas de investigación a futuro, hacer un filtro previo de palabras en el que se consideren los casos como el expuesto en la segunda conclusión.

BORRADOR



## 9. Referencias

- DNP (2018), 'Acerca de la entidad'. \*<https://www.dnp.gov.co/DNP/Paginas/acerca-de-la-entidad.aspx>
- Dumais, S. T. (2008), 'Latent semantic analysis', *Annual Review of Information Science and Technology* 38, 188–230.
- Goyvaerts, J. & Levithan, S. (2009), *Regular expressions cookbook*.
- Grossman, S. (2007), *Álgebra Lineal*.
- Halko, N., Martinsson, P. G. & Tropp, J. A. (2011), 'Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions', *Society for Industrial and Applied Mathematics Review* 53, 217–288. \*<http://users.cms.caltech.edu/~jtropp/papers/HMT11-Finding-Structure-SIREV.pdf>
- Huang, A. (2008), 'Similarity measures for text document clustering', *New Zealand Computer Science Research Student Conference*.  
\*<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.332.4480&rep=rep1&type=pdf>
- Jure Leskovec, Anand Rajaraman, J. D. U. (2011), *Mining of Massive Datasets*.  
\*<http://infolab.stanford.edu/~ullman/mmds/book.pdf>
- Mitkov, R. (2003), *The Oxford Handbook of Computational Linguistics*.
- M.W.Berry, Dumais, S. & O'Brien, G. (1995), 'Using linear algebra for intelligent information retrieval', *Society for Industrial and Applied Mathematics Review* 37, 573–595.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* 12, 2825–2830.
- Python (2018), 're — regular expression operations'. \*<https://docs.python.org/3.0/library/re.html>
- Ray, S. (2018), 'Essentials of machine learning algorithms (with python and r codes)'. \*<https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>
- Steven Bird, Ewan Klein, E. L. (2009), *Natural Language Processing with Python*.
- UN (2018), 'Objetivos de desarrollo sostenible. 17 objetivos para transformar nuestro mundo'. \*<https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/>
- van der Maaten, L. & Hinton, G. (2008), 'Visualizing data using t-sne', *Journal of Machine Learning Research* 9, 2579–2605. \*<http://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>
- Waldron, M. (2018), '10 common nlp terms explained for the text analysis novice'. \*<http://blog.aylien.com/10-common-nlp-terms-explained-for-the-text/>



BORRADOR