# **DATATON**

## **INFORME TEÓRICO**

## **Creación de categorías y clasificación**

Aguilar Restrepo William,

Bermúdez Gonzales Heber,

Caicedo Chamorro Edwin Alexander

*Fecha de entrega: 29 – 10 – 2018*

**INTRODUCCIÓN**

El objetivo principal es desarrollar una metodología que permita la creación de diferentes categorías para las cuales se clasifican las transacciones realizadas por los usuarios de PSE.

En primera instancia se utilizó procesamiento de lenguaje natural, este campo combina las tecnologías de la ciencia computacional (como la inteligencia artificial, el aprendizaje automático o la inferencia estadística) con la lingüística aplicada, con el objetivo de hacer posible la comprensión y el procesamiento asistidos por ordenador de información expresada en lenguaje humano para determinadas tareas, como los sistemas de diálogos interactivos, el análisis de opiniones, etc; con la técnica mencionada anteriormente se realiza un análisis de las variables referencia 1, referencia 2, subsector y descripción para establecer las posibles categorías en las cuales los usuarios de PSE realizan sus movimientos financieros.

Una vez se establecieron las categorías (mascotas, comida, industria, pagos recurrentes, comercio, actividades corporativas, educación, servicio, tecnología y comunicación, pago de deudas) las cuales contienen palabras claves, se realizó la clasificación por medio de máquinas de soporte vectorial, con las palabras claves de cada categoría se calcula la frecuencia de ocurrencia del término en la colección de palabras (tf-idf) que es una medida numérica que expresa cuán relevante es una palabra para un documento o en nuestro caso las categorías establecidas.

**MÉTODO DE TRABAJO CON TEXT MINING**

**FORMATO TEXT MINING Y UDPIPE**

El uso de los principios de datos ordenados es una forma poderosa de hacer que el manejo de datos sea más fácil y efectivo (cada variable es una columna, cada observación es una fila y cada tipo de unidad observacional es una tabla), entonces el formato de texto ordenado actúa como una tabla con un token (unidad significativa del texto) por fila; se encuentran varias formas de almacenamiento para enfoques con minería de texto:

· Cadena: Vector de caracteres.

· Cuerpo: Contienen cadenas sin procesar anotada con metadatos.

· Matriz de términos del documento: Una colección de documentos con una fila para cada documento y una columna para cada documento.

**PROCESO DE FILTRADO**

El procesamiento de filtrado de texto consistió en: Separar cada palabra si se encuentra en un conjunto de texto, luego se filtra eliminando aquellas palabras que no representan relevancia (conectores, artículos, entre otros), o conocidos en el mundo de text mining como “Stopwords”. Una vez filtrado el algoritmo analiza su lexema (es el elemento que contiene la significancia de la palabra, tradicionalmente se le llama raíz). En este caso se maneja la librería “tm” para realizar el filtrado en la cadena de caracteres conformada por las variables referencia 1, referencia 2, subsector y descripción de la base de datos (trxpse\_personas).

**CREACION DE CATEGORIAS**

Después de realizar el filtrado del texto se procede a utilizar la librería “udpipe” para extracción de palabras clave las cuales permiten **encontrar las frases para la creación de cada categoría en un texto plano**.

Las técnicas que se utilizan para la extracción de palabras claves son.

1. Encuentre palabras clave basadas en **Colocaciones y Co-ocurrencias**
2. Encuentre palabras clave basadas en el **algoritmo Textrank**
3. Encuentre palabras clave basadas en **RAKE (extracción automática rápida de palabras clave)**
4. Encuentre palabras clave buscando **frases** (frases nominales / frases verbales)
5. Encuentre palabras clave basadas en los **resultados del análisis de dependencia** (obteniendo el tema del texto)

De las cuales la metodología utilizada para este problema es la técnica de “**Colocaciones y Co-ocurrencias”** la cual consiste en obtener expresiones de múltiples palabras observando las colocaciones (palabras seguidas), las coincidencias de palabras dentro de cada oración o las coincidencias de palabras que están cerca una de la otra. La visualización de estas co-ocurrencias se puede hacer usando una gráfica de red de la siguiente manera para los 30 nombres y adjetivos más frecuentes más concurrentes.

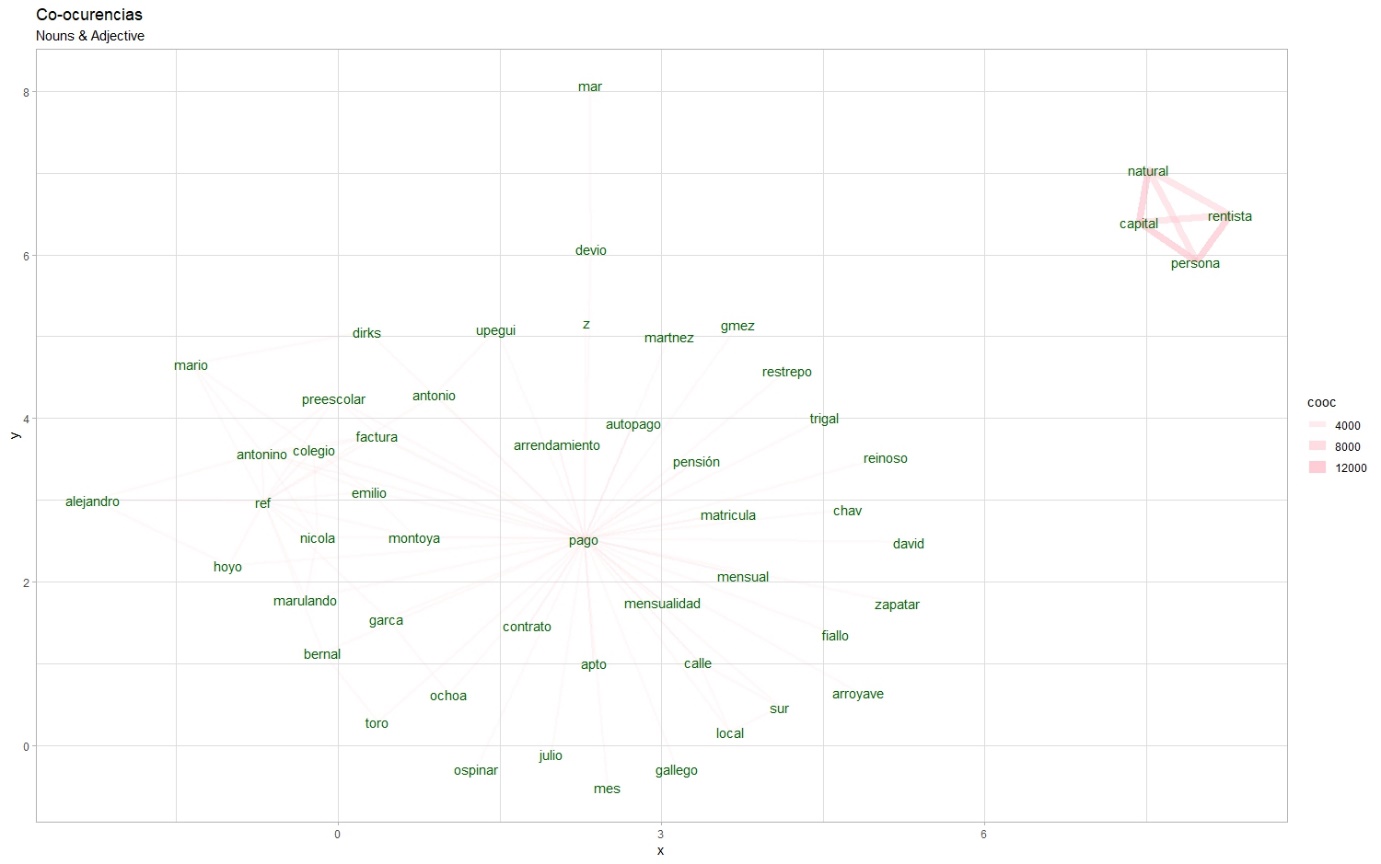


Figura 1: Gráfico de co-ocurrencias

Para la realización de cada co-ocurrencia se hace a partir de cada sector, es decir, para encontrar una mayor cantidad de palabras familiarizadas hacia una categoría se filtra teniendo en cuenta el sector al que se dirige la transacción por PSE.

De cada co-ocurrencia presentada en las gráficas del procesamiento de lenguajes se toman las palabras que se encuentran asociadas al grupo que se encuentre más cercano, y dichas palabras son asignadas a la categoría que se asemeje a dicho conjunto de palabras.

Al momento de dar uso de las categorías presentadas por Bancolombia se observa que algunas no coinciden con los movimientos financieros de PSE, por tanto se prefiere realizar nuevas categorías acordes a lo obtenido en el modelo de procesamiento de lenguaje natural (mascotas, comida, industria, pagos recurrentes, comercio, actividades corporativas, educación, servicio, tecnología y comunicación, pago de deudas).

**CLASIFICACIÓN DE TRANSACCIONES**

Se usó un método de clasificación supervisada como lo es las máquinas de soporte vectorial (SVM) para asignar una categoría a una nueva transacción (PSE). Para esto se define la variable respuesta que corresponde a cada categoría, y se mide el tf-idf que es una medida numérica que expresa cuán relevante es una palabra para una categorías establecida, por ejemplo la categoría mascotas aparecen las palabras claves "mineral", "producción", "agua", "alcohólico", "bebido", "concentrado", "conserva", "alimento", "bebida", "elaboración", "animal", "alimenticio", "aliño" y "producto", la nueva observación tendrá un vector asociado con algunas palabra que coincida con el vector mascota o con cualquier otro posible vector de palabras. En este caso la máquina de soporte vectorial calcula en qué categoría es más probable que se clasifique esta observación.

Este procedimiento es debido a que se saben las categorías y las palabras claves que la definen la cuales ya han sido asignadas a su correspondiente categoría.

En primera medida para poder categorizar una nueva transacción se usa el indicador tf-idf que nos permite encontrar las palabras importantes para el contenido de cada categoría disminuyendo el peso de las palabras de uso común y aumentando el peso de las palabras que no se usan mucho en una las categorías, luego de tener la matriz con los indicadores por cada categoría se procede a entrenar el modelo de máquinas de soporte vectorial tomando como respuesta la categoría y como regresores las palabras clave cuyas observaciones es el indicador tf-idf.