



Universidad Nacional Autónoma de Mexico.

Instituto de Ingeniería.

Grupo de Ingeniería Lingüística.

**Manual de Usuario del Sistema de Categorización
de Definiciones.**

Involucrados:

Ignacio Arroyo Fernández.

Ramón Pantoja Velasco.

México D.F. a 31 de Noviembre de 2015.

Índice.	Pag.
Introducción.....	3
Objetivos.....	3
Instalación del programa.....	3
Requerimientos de los archivos.....	4
Procedimientos del Sistema.....	5

Introducción.

El motivo de este sistema no es que sea utilizado por un usuario de forma separada, va incorporado en un sistema más grande de anotación de definiciones, por lo tanto su uso por separado solo es considerado para pruebas.

El sistema de categorización de definiciones permite al usuario crear una asignación de definiciones obtenidas de Describe con artículos obtenidos de Wikipedia y sus acepciones con el fin de saber a qué acepción del término pertenecen. El programa maneja 6 opciones para relacionarlos y son con TFIDF para unigramas, con TFIDF para bigramas, con TFIDF para trigramas, sin TFIDF para unigramas, sin TFIDF para bigramas y sin TFIDF para trigramas. Al inicio del programa se te mostrarán estas opciones y podrás elegir cualquiera de ellas.

Objetivos.

- Apoyar a los anotadores de definiciones con sugerencias para asignar una definición a cierto grupo.
- Tener una alternativa de Categorización.
- Verificar el correcto trabajo de los anotadores.

Instalación del programa.

Para que puedas utilizar con éxito el sistema de clasificación es necesario que cumplas con los siguientes requisitos:

- Distribución Fedora 22 (Linux).- Se desarrolló completamente en un ambiente Linux, cualquier distribución de Linux es útil solo tendrá pequeñas diferencias en la instalación de las herramientas que han sido utilizadas para crear este sistema. Esto se decidió ya que es más sencillo instalar las herramientas y bibliotecas de programación requeridas en un sistema Linux.
- Python 2.7 instalado de base en Fedora y en la mayoría de distribuciones de kernel Linux, se puede verificar esto con el comando "\$ python -V" desde línea de comandos de Linux y en caso de no estar instalado se instala (en Fedora 22) con el comando "sudo apt-get install python2.7".
- Se utilizaron las bibliotecas de scikit learn el cual se encuentra un manual de instalación en la dirección: <http://scikit-learn.org/stable/install.html> y en el caso de fedora solo fue necesario ejecutar el comando "sudo dnf install python-scikit-learn". Esto nos sirve para poder utilizar las clases TfidfVectorizer para análisis con TFIDF y para usar la clase CountVectorizer y analizar sin TFIDF.
- Se utilizaron así mismo las bibliotecas de scipy con el instructivo que define la página oficial de scipy, la cual es: <http://www.scipy.org/install.html>, y es con el comando "sudo yum install numpy scipy python-matplotlib ipython python-pandas sympy python-nose". Esto sirve para utilizar la función de similitud coseno.

Manual de Usuario del Sistema Categorizador de Definiciones Apoyo a Sistema Anotador de Definiciones

- Archivos de entrada que son archivo .csv de Definiciones, archivo .csv de definiciones de Wikipedia y archivo .csv de Aceptaciones de Wikipedia.

Una vez cumplidos estos requisitos solo basta con tener el código fuente en Python Categorizador.py y ejecutarlo.

Requerimientos de los archivos.

Los archivos deben de cumplir ciertas reglas para que pueda ser manejado adecuadamente por el programa, aquí se lista lo que cada archivo de entrada debe contener.

Archivo de Definiciones:

Debe ser un archivo en formato .csv de dos campos, el primero no contiene nada y lleva una etiqueta 'nada' y el segundo campo lleva las definiciones y una etiqueta llamada 'definicion'. Estas etiquetas van hasta arriba dentro del documento separados por comas y sin espacios. Los valores de definiciones van entre comillas y después de una coma, pondré un ejemplo gráfico para que sea más claro. Como se puede observar en la ilustración 1.

Este archivo es donde se reciben los archivos que se quieren tratar, por ello la importancia del mismo, es nuestra "Materia prima" por llamarle de algún modo.

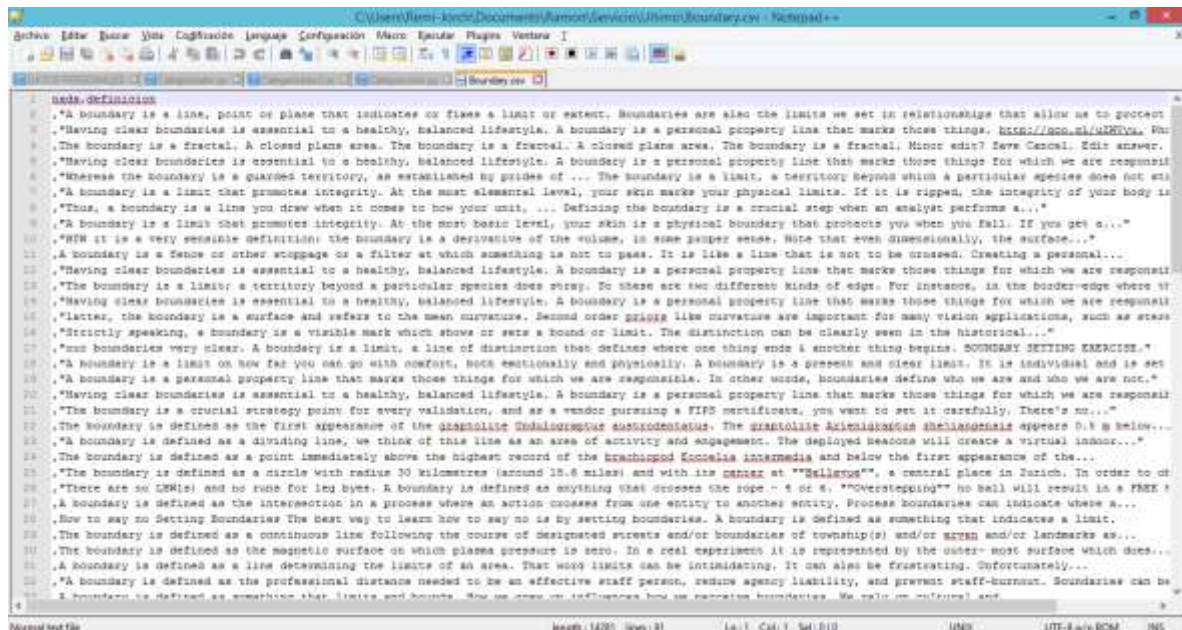


Ilustración 1 Ejemplo de archivo de Definiciones

Archivo de Artículos de Wikipedia.

Este archivo debe ser un .csv de dos campos, el primero contiene el ID de la acepción a la que pertenece esa definición y lleva una etiqueta 'id' y el segundo campo lleva los artículos y una etiqueta llamada 'definicion'. Estas etiquetas van hasta arriba dentro del documento separados por comas y sin espacios. Los valores de Id van uno por línea, seguidos de una coma ',' los valores de definiciones van entre comillas y después de la coma. Observar

Manual de Usuario del Sistema Categorizador de Definiciones Apoyo a Sistema Anotador de Definiciones

Ilustración 2. Es importante que las definiciones tengan en promedio 40 palabras ya que sino se vuelve más difícil calcular la similitud coseno, pondré un ejemplo gráfico para que sea más claro.

Este archivo tiene como finalidad obtener y organizar los artículos obtenidos de Wikipedia relacionándolos con su acepción para poder asignar una definición del archivo de definiciones con un artículo de Wikipedia.

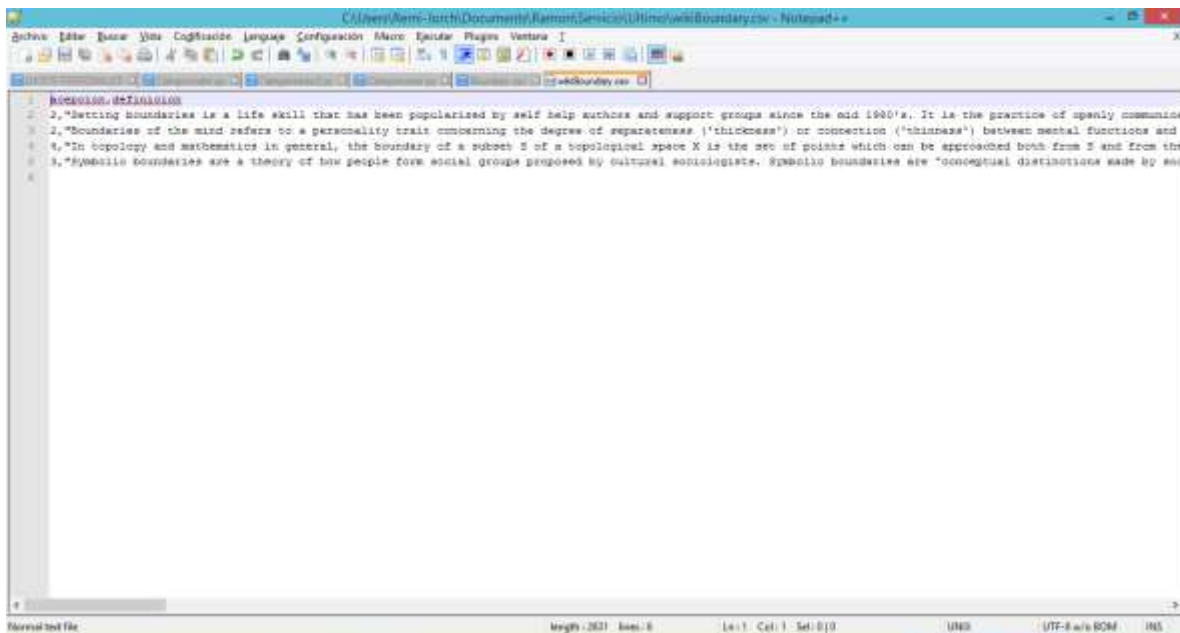


Ilustración 2 Ejemplo de Archivo de Artículos de Wikipedia.

Archivo de Acepciones.

Este archivo deberá ser un .csv donde se almacenará el nombre de la acepción y la definición, esto debe de ser relacionado con las definiciones que se obtuvieron de Wikipedia y las definiciones a procesar, esto significa que deben ser referentes al mismo término. El .csv consiste de dos campos, el primero consiste en el ID de la acepción en la cual recomiendo que sea secuencial, en el segundo se encuentra el nombre de la acepción entre comillas. Estas etiquetas van hasta arriba dentro del documento separados por comas y sin espacios. Los valores de Id van uno por línea, seguidos de una coma ',' los valores de nombres de acepciones van entre comillas y después de la coma. Como se puede ver en la ilustración 3.

Este documento tiene la finalidad de organizar las acepciones y asociarlas con los artículos de Wikipedia para finalmente poder tratarlos y asignar definiciones en ellos.

Manual de Usuario del Sistema Categorizador de Definiciones Apoyo a Sistema Anotador de Definiciones

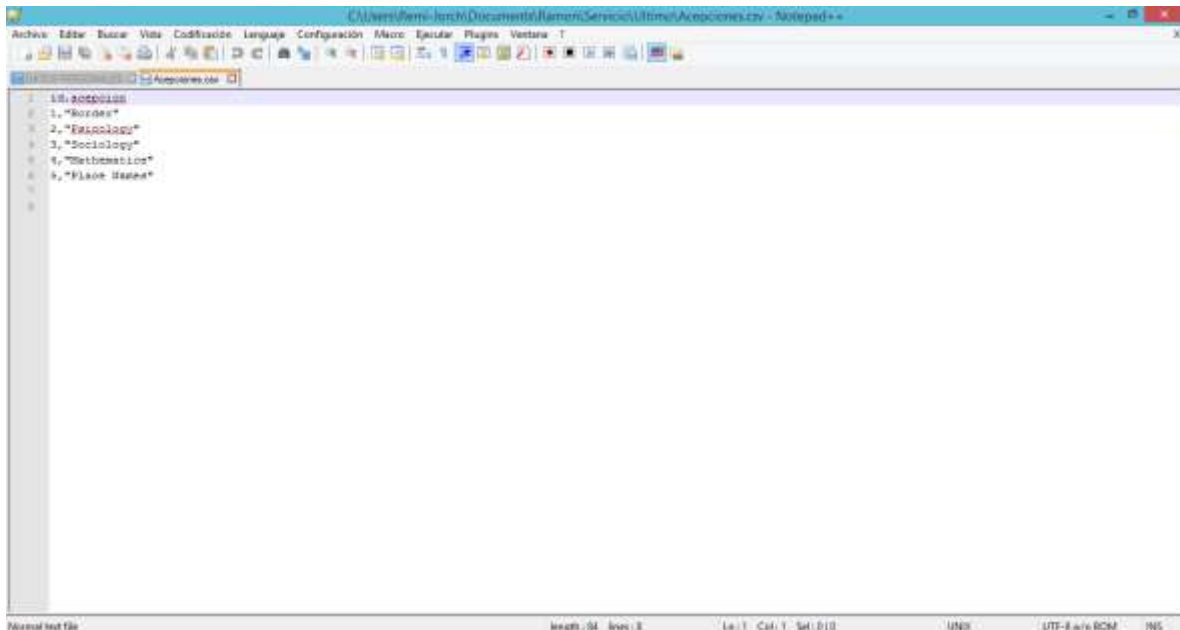


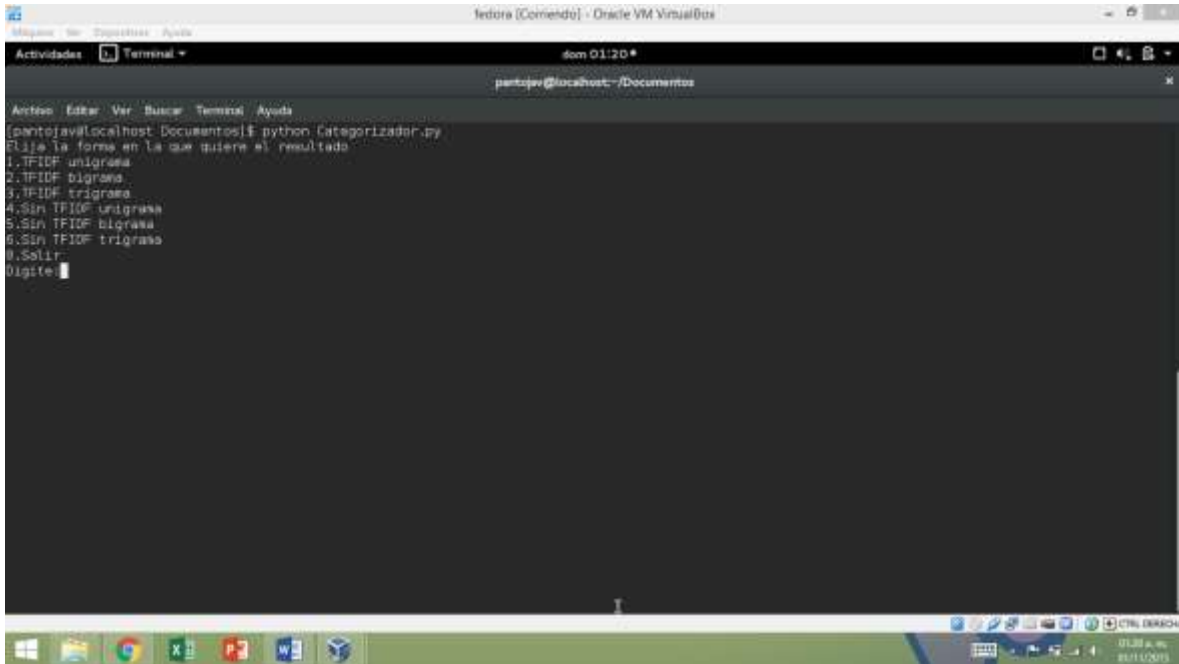
Ilustración 3 Ejemplo de archivo de acepciones.

Archivo de Salida.

El archivo de Salida está compuesto por un Diccionario (en programación) que sirve para ordenar las asignaciones de definición con acepción, este archivo lo puedes leer de la siguiente manera:

Primero se indica que es un archivo de acepciones: "{ass:}", después viene la primer acepción que se lee junto con lo anterior (en este caso usaremos border): "{ass:{Border:}}", posteriormente viene el artículo de Wikipedia que pertenece a la acepción, de esta manera: "{ass:{Border:{art1: setting boundaries is a life skill that has been popularized by self help authors and support groups since the mid 1980's. it is the practice of openly communicating and asserting personal values as way to preserve and protect against having them compromised or violated.[1] the term 'boundary' is a metaphor – with in-bounds meaning acceptable and out-of-bounds meaning unacceptable.[1] without values and boundaries our identities become diffused and often controlled by the definitions offered by others.[2] the concept of boundaries has been widely adopted by the counseling profession.[3]}}", Después del artículo vienen por fin las definiciones pertenecientes a dicho artículo, identificadas por un ID de definición y ordenadas en parejas de datos donde el primer elemento es el valor de la similitud y el segundo la definición, de tal forma que: Def_N:(<valorDeSimilitud>, "Definicion"), por ejemplo: {ass:{Border:{Border:{art1:setting boundaries is a life skill that has been popularized by self help authors and support groups since the mid 1980's. it is the practice of openly communicating and asserting personal values as way to preserve and protect against having them compromised or violated.[1] the term 'boundary' is a metaphor – with in-bounds meaning acceptable and out-of-bounds meaning unacceptable.[1] without values and boundaries our identities become diffused and often controlled by the definitions offered by others.[2] the concept of boundaries has been widely adopted by the counseling profession.[3]},defs1:{{0.296362767222,how to say no setting boundaries the best way to learn how to say no is by setting boundaries. a boundary is defined as something that indicates

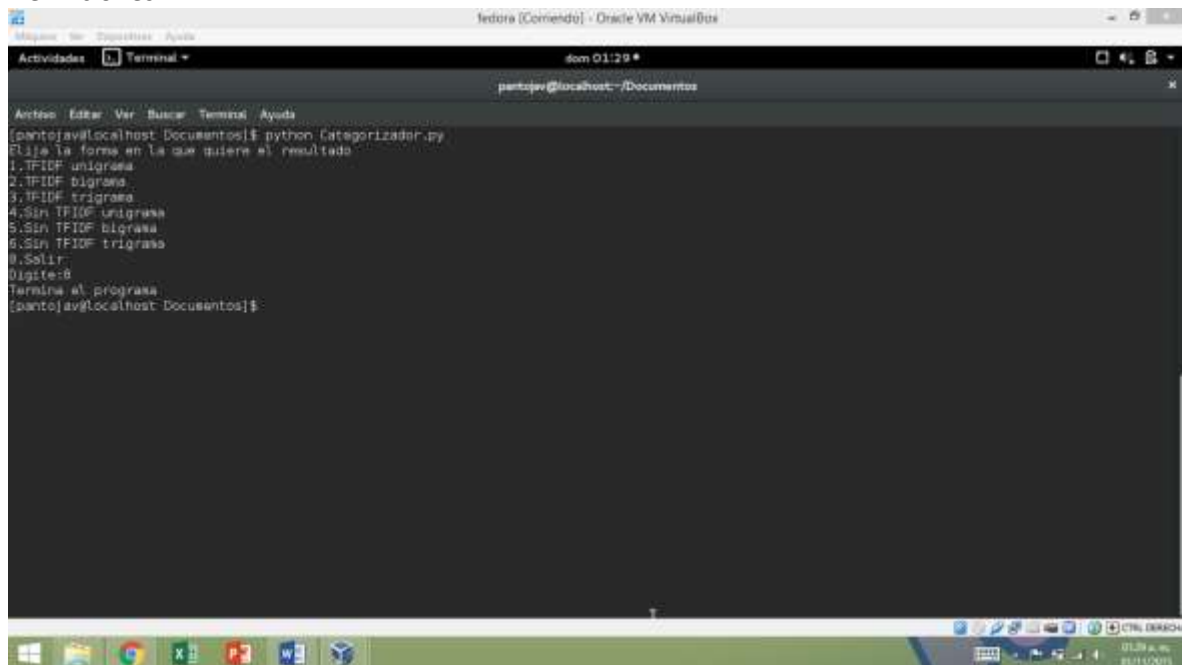
Manual de Usuario del Sistema Categorizador de Definiciones Apoyo a Sistema Anotador de Definiciones



```
Actividades Terminal *  
dom 01:20 *  
pantojav@localhost: ~/Documentos  
Antes Editar Ver Buscar Terminal Ayuda  
pantojav@localhost: Documentos$ python Categorizador.py  
Elija la forma en la que quiere el resultado  
1. TFIDF unigramas  
2. TFIDF bigramas  
3. TFIDF trigramas  
4. Sin TFIDF unigramas  
5. Sin TFIDF bigramas  
6. Sin TFIDF trigramas  
0. Salir  
Digite 2
```

- 2.- Al seleccionar la opción 2 se generará el archivo de salida con TFIDF para bigramas.
- 3.- Al seleccionar la opción 3 se generará el archivo de salida con TFIDF para trigramas.
- 4.- Al seleccionar la opción 4 se generará el archivo de salida con TFIDF para unigramas.
- 5.- Al seleccionar la opción 5 se generará el archivo de salida con TFIDF para bigramas.
- 6.- Al seleccionar la opción 6 se generará el archivo de salida con TFIDF para trigramas.
- 7.- Al seleccionar la opción 0 se terminará el programa y te regresará al Shell de la terminal de comandos \$.

Manual de Usuario del Sistema Categorizador de Definiciones Apoyo a Sistema Anotador de Definiciones

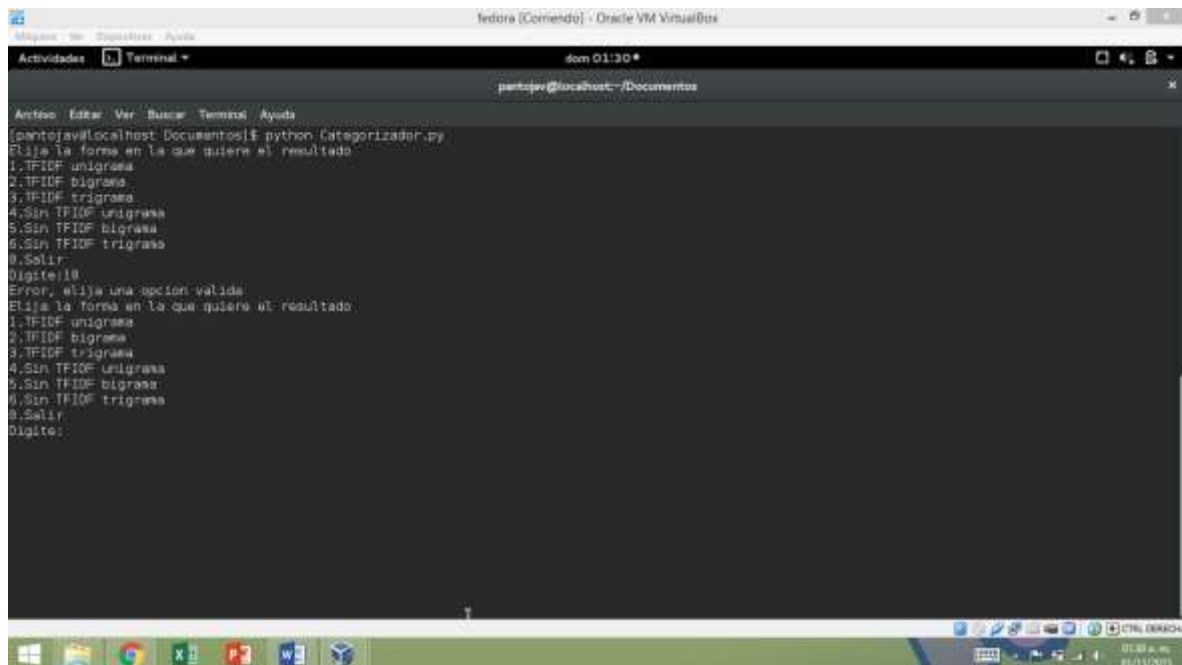


```

fedora (Comiendo) - Oracle VM VirtualBox
Actividades Terminal *
dom 01:29 *
pantojav@localhost: ~/Documentos

Antes Editar Ver Buscar Terminal Ayuda
(pantojav@localhost: Documentos) $ python Categorizador.py
Elija la forma en la que quiere el resultado
1. TFIDF unigramas
2. TFIDF bigramas
3. TFIDF trigramas
4. Sin TFIDF unigramas
5. Sin TFIDF bigramas
6. Sin TFIDF trigramas
0. Salir
Digite: 1
Termina el programa
(pantojav@localhost: Documentos) $
  
```

8.- Si se selecciona alguna opción que no está dentro de las especificadas se mandará un mensaje al usuario de que se ingrese una opción válida y se pedirá de nuevo que elija una opción.



```

fedora (Comiendo) - Oracle VM VirtualBox
Actividades Terminal *
dom 01:30 *
pantojav@localhost: ~/Documentos

Antes Editar Ver Buscar Terminal Ayuda
(pantojav@localhost: Documentos) $ python Categorizador.py
Elija la forma en la que quiere el resultado
1. TFIDF unigramas
2. TFIDF bigramas
3. TFIDF trigramas
4. Sin TFIDF unigramas
5. Sin TFIDF bigramas
6. Sin TFIDF trigramas
0. Salir
Digite: 10
Error, elija una opción válida
Elija la forma en la que quiere el resultado
1. TFIDF unigramas
2. TFIDF bigramas
3. TFIDF trigramas
4. Sin TFIDF unigramas
5. Sin TFIDF bigramas
6. Sin TFIDF trigramas
0. Salir
Digite:
  
```

La opción que elija el usuario debe ser obligatoriamente un dígito.

Los archivos generados se podrán observar en la misma carpeta donde están guardados el programa y los archivos que se utilizaron para crearlo si es que no se ha modificado el código fuente.