Tarea 2 de Minería de Textos: *Clustering*

En esta tarea los estudiantes utilizarán algoritmos de clustering con una subcolección de textos de *Newsgroups* y tendrán que realizar la evaluación y análisis de los resultados.

El conjunto de textos con el que se va a trabajar es un subconjunto de la colección 20_newsgroup del CMU Text Learning Group. El subconjunto está compuesto de 7 grupos temáticos, cada uno en una carpeta, que contiene una serie de ficheros con comentarios sobre noticias. Este subconjunto se facilita junto con el enunciado.

Un ejemplo recortado de uno de los ficheros sería el siguiente:

Newsgroups: talk.politics.guns
Path:cantaloupe.srv.cs.cmu.edu!magnesium.club.cc.cmu.edu!news.sei.cmu.edu!cis.ohio-state.edu!magnus.acs.ohio-state.edu!usenet.ins.cwru.edu!howland.reston.ans.net!wupost!uunet!pmafire!mica-nsf.inel.gov!guinness!gozer.idbsu.edu!betz
From: betz@gozer.idbsu.edu (Andrew Betz)
Subject: Re: My Gun is like my....
Message-ID: <1993Apr18.222951.10744@guinness.idbsu.edu>
Sender: usenet@guinness.idbsu.edu (Usenet News mail)
Nntp-Posting-Host: gozer
Organization: SigSauer Fan Club
References: <1993Apr16.194708.13273@vax.oxford.ac.uk>
Date: Sun, 18 Apr 1993 22:29:51 GMT
Lines: 84

In article <1993Apr16.194708.13273@vax.oxford.ac.uk> jaj@vax.oxford.ac.uk writes:
>What all you turkey pro-pistol and automatic weapons fanatics don't seem to
>realize is that the rest of us *laugh* at you. You don't make me angry, you
>just make me chuckle - I remeber being in Bellingham, Washington and seeing a

[Warning: Flammage to follow...]

Ah, that British sense of humor. Probably got a real gut-buster going
when the IRA blew that kid up a couple of weeks ago, huh? Of course,
in Britain, your government has ordered you defenseless, so your way
of coping with violent criminals is to laugh at victims.

>pick-up truck in front of the car that my friend and I were in. It had a bumper
>sticker proclaiming "Gun Control is a firm grip on a .45." Now I'm sure that
>that wanker thought he was pretty cool.

I don't know about a .45. My own preference is for 9mm.

>What he didn't realize was that we took a photo of the back of his truck, and
>showed it to our friends when we got back to Vancouver, Canada (where I'm from
>originally). People were guffawing at the basic stupidity of such a
>sticker, and the even greater stupidity of the person who put it there in the
>first place! :)

Ah, Canada. Where the criminals don't bother with checking to see
if the victims are home. They just break on in. America's a little
different, you see. Criminals worry a bit more about getting shot,
so they more frequently check to see if anyone's home.

>I knew somebody else who went to one of your "Gun-mart" superstore places, just
>so he could experience the sight of people putting guns and ammo into shopping

>carts! I didn't believe it myself until I drove by one in Vegas last year!!!
…

Drew
--
betz@gozer.idbsu.edu
*** brought into your terminal from the free state of idaho ***
*** when you outlaw rights, only outlaws will have rights ***
*** spook fodder: fema, nsa, clinton, gore, insurrection, nsc,
semtex, neptunium, terrorist, cia, mi5, mi6, kgb, deuterium

Como se puede observar, aparece una cabecera que identifica la noticia y la caracteriza, después los comentarios y finalmente la firma. La parte que aparece en rojo en este documento, la cabecera, habría que eliminarla del fichero; en el caso de la firma, al final del ejemplo, se deja a elección del estudiante. Al observar varios ficheros podrá comprobar que la firma puede ser muy simple en unas ocasiones y no tan simple en otras.

La realización de la práctica consiste en lo siguiente:

- Preparación de la colección de textos. La estructura de las carpetas sirve para establecer el *goldstandard*, es decir, la asignación correcta de cada fichero a uno de los 7 posibles grupos para utilizarla como solución de referencia en la evaluación externa. El nombre de la carpeta puede servir como etiqueta de cada grupo. También la etiqueta de grupo aparece en la cabecera, pero esta información debe eliminarse para realizar el clustering. El estudiante deberá establecer una relación entre cada fichero y la etiqueta del grupo al que pertenece de forma que le permita posteriormente realizar una evaluación externa.
- Preprocesamiento. Se deberá eliminar la cabecera de cada fichero, en rojo en el ejemplo. Además, se realizarán la tokenización y lematización/truncado y cualquier otro preprocesamiento que el estudiante estime conveniente.
- Representación. Se utilizarán 2 tipos de funciones de pesado del modelo del espacio vectorial.
- Clustering. Realizar el clustering y hacer el análisis de los resultados del clustering de la citada colección en 7 clústeres con:
    - Un algoritmo de partición, Kmeans, y otro algoritmo a elección del estudiante que corresponda a otro tipo de clustering.
    - Se deberán combinar las 2 funciones de pesado con los 2 algoritmos, siempre que sea posible.
    - Se utilizarán las medidas de evaluación Precisión, Cobertura y Medida-F para comparar la solución obtenida por los algoritmos con respecto a la solución de referencia o *goldstandard.*
    - Se realizará una matriz de confusión de cada ejecución que permita determinar los grupos más difíciles de caracterizar y entre los que hay más confusión en cada caso.
- Preparar una memoria en la que:
    - Se describa y caracterice la colección en cuanto a número de documentos por grupo, número medio de palabras por grupo y su desviación estándar.
    - Se describa cómo se ha formado el *goldstandard* para realizar la evaluación externa.

- Se detalle el preprocesamiento que se ha llevado a cabo.
- Se indique qué tipos de representación se han utilizado.
- Se justifique la elección del segundo algoritmo de clustering.
- Se muestren los resultados parciales de cada algoritmo/representación y una tabla final que incluya todos los resultados. Se deben analizar los resultados en términos de calidad de las medidas de evaluación externas con los 2 algoritmos y las 2 representaciones. Se presentarán las matrices de confusión para determinar los grupos más difíciles de caracterizar y entre los que hay más confusión en cada caso.
- Se presentarán las conclusiones de la tarea realizada.

Parte opcional

Con la misma subcolección que se ha utilizado en la parte obligatoria, se pretende utilizar una técnica para determinar el número óptimo de grupos teniendo solo en cuenta la información proporcionada por la propia colección.

La realización de la parte opcional de la práctica consiste en lo siguiente:

- Utilización del método Elbow -- (https://en.wikipedia.org/wiki/Elbow_method_(clustering)-- para determinar el número óptimo de clusters. Ténganse en cuenta que este valor puede ser diferente del indicado en la solución de referencia. En la web se puede encontrar información de ayuda para su implementación o el uso de una herramienta.
- En caso de que el valor de k obtenido sea diferente de 7, realizar el clustering utilizando dicho valor k.
- La memoria correspondiente a esta parte deberá contener:
  - Una descripción del método Elbow.
  - Indicar la fuente de la que se ha partido para su implementación o la herramienta utilizada.
  - Descripción del algoritmo y en su caso los parámetros utilizados para realizar el clustering.
  - Mostrar el resultado del clustering utilizando las etiquetas de clase de la parte obligatoria como identificadores de los ficheros.