Tarea 2: Minería de Datos y Procesamiento de Lenguaje Natural

Universidad Diego Portales

Profesor: Alejandro Figueroa
Ayudante: Nicolás Olivares

Fecha de Entrega: Viernes 30 de agosto de 2013 23:59hrs
Ponderación: 1
Vía de Entrega: Correo electrónico al ayudante (nicolivares@gmail.com)

Enunciado

En la tarea anterior, el estudiante asocio 3,000 consultas web con un elemento dentro de un conjunto de tres etiquetas. Esta tarea trabaja sobre ese conjunto de datos etiquetados. Nótese que cada alumno debe trabajar exclusivamente sobre los datos que el mismo etiqueto.

El objetivo principal de esta tarea es familiarizar al estudiante con el concepto de espacio vectorial. Este concepto es un pilar en prácticamente todas las técnicas de aprendizaje automático, ya que todo problema debe pasarse de cierta manera a un espacio vectorial para poder trabajar con los diferentes algoritmos de aprendizaje.

Para realizar esta tarea, es necesario definir lo que se llama el modelo de espacio vectorial (MEV o VSM en su sigla en inglés), que también los investigadores llaman bag-of-words (BoW o bolsa de palabras en español). Este modelo mapea una secuencia de palabras (e.g., una consulta web) a un vector de frecuencia de palabras. Por ejemplo, la consulta "*acrobat 6.0 plug in*" tiene cuatro palabras, por ende su representación en términos de frecuencias es {<acrobat,1>; <6.0,1>; <plug,1>; <in,1>}.

Es una práctica común, utilizar los espacios en blanco para tokenizar, es decir encontrar los límites de las palabras. En esta tarea basta con usar esa regla simple para construir la representación vectorial de cada una de las 3,000 consultas. Sin embargo, se le sugiere al estudiante utilizar MontyLingua para separar cada consulta en los respectivos tokens.

El siguiente paso, consiste en unir los vectores de cada clase para generar un vector global para la respectiva clase. En otras palabras, el alumno va a tener tres vectores, uno para las consultas de navegación, otro para las de información y otro para las de recurso. Por ejemplo, las dos consultas de recursos: "free music lyrics search" y "lyrics to luther vandross" generarían un vector global para la clase recurso de {<free,1>; <music,1>; <lyrics,2>; <search,1>; <to,1>; <luther,1>; <vandross,1>}.

Después, a cada vector global, y también a cada uno de los vectores de consultas, se le deben remover la puntuación y las stop-words. Las stop-words son definidas como palabras altamente frecuentes que no tienen un significado en ellas mismas. El estudiante puede obtener una lista para el idioma inglés en:

http://norm.al/2009/04/14/list-of-english-stop-words/

Teniendo los tres vectores globales y los 3,000 vectores asociados a cada consulta depurados, se pide lo siguiente:

1. Calcule el vector centroide para una de las tres clases. El vector centroide es el vector global dividido por el número de ejemplos que tiene cada clase. ¿Qué observa en las palabras más frecuentes en cada una de las clases?
2. Calcula la distancia de cada punto con respecto a cada uno de los centroides, y asígnele la etiqueta del centroide más cercano. Es decir, para la consulta X, calcule la distancia entre su respectivo vector y cada uno de los tres vectores centroides. Asígnele la etiqueta correspondiente al centroide que obtenga la menor distancia. Utilice la distancia de Manhattan (h=1):

$$dist(x_i, x_j) = \sqrt[h]{\left|x_{i1} - x_{j1}\right|^h + \ldots + \left|x_{ir} - x_{jr}\right|^h}$$

3. Compare las etiquetas automáticas con las etiquetas que asigno manualmente en la clase anterior, y calcule:
    a. Accuracy: el número de etiquetas que coinciden divido el número de ejemplos (3,000).
    b. El Recall de cada clase, es decir, del total de etiqueta de una clase cuantas fueron identificadas automáticamente.
    c. La Precisión de cada clase, es decir, de las que fueron asignadas a una clase, cuantas realmente son de esa clase.
    d. El F-Score, dado el Recall y la Precisión obtenidas anteriormente, calcule el F-score de cada una de las tres clases. Para esto utilice β=1. http://en.wikipedia.org/wiki/F1_score
4. Realice el mismo procedimiento con las siguientes métricas de distancia:
    a. Canberra

$$dist(x_i, x_j) = \frac{|x_{i1} - x_{j1}|}{x_{i1} + x_{j1}} + \ldots + \frac{|x_{ir} - x_{jr}|}{x_{ir} + x_{jr}}$$

    b. Squared Cord

$$dist(x_i, x_j) = (\sqrt{x_{i1}} - \sqrt{x_{j1}})^2 + \ldots + (\sqrt{x_{ir}} - \sqrt{x_{jr}})^2$$

    c. Squared Chi-squared

$$dist(x_i, x_j) = \frac{(x_{i1} - x_{j1})^2}{x_{i1} + x_{j1}} + \ldots + \frac{(x_{ir} - x_{jr})^2}{x_{ir} + x_{jr}}$$

5. Comente qué métrica da mejores resultados ¿Por qué? ¿En qué aspectos?
6. Considere la mejor métrica de distancia obtenida, y estudie el impacto de la lematización provista por Montylingua. ¿Qué observa en los tres vectores globales / centroides?
    a. Tamaños.
    b. ¿Qué sucede con el Recall, Precisión, Accuracy y F-Score? ¿Mejoran o empeoran los resultados?
7. Compare la distribución de las categorías sintácticas de cada clase. ¿Qué pasa si repite el procedimiento sólo con las más frecuentes en términos globales (3,000 consultas)?