Tarea 3 - Minería de Datos y Procesamiento de Lenguaje Natural

Profesor: Alejandro Figueroa
Ayudante: Nicolás Olivares
Fecha de Entrega: Viernes 13 de septiembre de 2013 23:59hrs
Ponderación: 1
Vía de Entrega: Correo electrónico al ayudante (nicolivares@gmail.com)

Enunciado Parte 1

En esta tarea exploraremos el uso del algoritmo de agrupamiento K-means. Tome la representación vectorial que ha utilizado en las tareas anteriores, y entréguesela como entrada a alguna implementación del algoritmo en Java o en algún otro lenguaje. En este paso considere la mejor representación obtenida hasta ahora entre: 1) bag-of-words (BoW); 2) BoW con las palabras lematizadas; o 3) palabras perteneciente a ciertos categorías sintácticas prominentes. Hay varias librerías que implementan este algoritmo, como javaml o WEKA.

http://www.cs.waikato.ac.nz/ml/weka/ http://java-ml.sourceforge.net/

El algoritmo necesita como entrada el número de clases, que en nuestro caso de estudio es k=3 (informacional, navegacional y de recurso). Después de ejecutar K-means, asigne a todos los elementos de cada clúster la etiqueta manual predominante. Y después de eso calcule: Accuracy, Recall, Precisión y F-Score. Además, haga un análisis utilizando diferentes números de iteraciones, y un estudio del error observando la matriz de confusión (vea http://es.wikipedia.org/wiki/Matriz_de_confusi%C3%B3n).

Enunciado Parte 2

En la segunda parte de esta tarea exploraremos el uso del algoritmo de agrupamiento Fuzzy C-means. Tome la misma representación vectorial utilizada en la parte anterior, y entréguesela como entrada a algoritmo Fuzzy C-means. El algoritmo Fuzzy C-Means es como sigue:

1. Inicializar $U=[u_{ij}] = U^{(0)}$
2. Calcular los centros $C^{(k)}=[c_j]$ con $U^{(k)}$
3. Actualizar $U^{(k)}=U^{(k+1)}$
4. Si se cumple el criterio de término, terminar, sino volver a 2.

En este algoritmo U es una matriz de membrecía donde cada celda representa el grado de pertenencia del ejemplo i a la clase j. Fuzzy C-means al contrario de K-means, hace un ranking de las clases para cada ejemplo. La matriz U puede ser inicializada de manera aleatoria, asegurándose que $U_{i1}+U_{i2}+U_{i3}=1$. En este algoritmo k es el número de iteración, y como condición de término utilizar un número de iteraciones fijo K=100, 500, 1000, 5000, 10000. Las ecuaciones de este algoritmo están dadas por:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

Utilice m=2. Para cada uno de los casos calcule: Accuracy, Recall, Precisión y F-Score. Compare con K-Means.