udp FACULTAD DE INGENIERÍA

Tarea 6: Minería de Datos y Procesamiento de Lenguaje Natural

Universidad Diego Portales

Profesor: Alejandro Figueroa

Ayudante: Nicolás Olivares

Fecha de Entrega: Viernes 25 de octubre de 2013 23:59hrs

Ponderación: 1

Vía de Entrega: Correo electrónico al ayudante (nicolivares@gmail.com)

Enunciado

El objetivo de esta tarea es comparar dos clasificadores multi-clases: SVM y MaxEnt. La idea es realizar lo mismo de la tarea dos, pero esta vez utilizando modelos de máxima entropía. Muestre una tabla comparativa.

Nótese que debe mantener la consistencia de los folders de datos, es decir, utilizar los mismos folders de entrenamiento y predicción para ambos clasificadores.

Tarea Recuperativa

Explicit Semantic Analysis (ESA) es un método para representar un documento en un espacio vectorial donde las componentes son páginas (conceptos) de Wikipedia. Para un texto (consulta), ESA determina un conjunto de k conceptos relacionados mediante una función de rankeo. Para determinar los k conceptos relacionados con una consulta se puede utilizar la implementación en:

http://ticcky.github.io/esalib/

Por ejemplo, para la consulta "pancreas cancer" se obtiene los siguientes top k=9 conceptos de Wikipedia:

0. Pancreatic cancer, docID=363559, ranking=0.19825363
1. Pancreas, docID=38300, ranking=0.106568076
2. Zollinger-Ellison syndrome, docID=357810, ranking=0.10141763
3. Multiple endocrine neoplasia type 1, docID=2574340, ranking=0.098277286
4. Kidney transplantation, docID=1584036, ranking=0.09377637
5. Blood sugar, docID=289406, ranking=0.09067796
6. Frederick Banting, docID=75076, ranking=0.08955204
7. The Awful Truth, docID=884265, ranking=0.08439648
8. Regnier de Graaf, docID=1446750, ranking=0.081688836
9. Insulinoma, docID=785061, ranking=0.07852207

El objetivo de la tarea consiste en cambiar la representación vectorial provista por la bolsa de palabras (BoW) por el espacio semántico dado por ESA. Para ésto, la idea es usar una representación con los k=0...30 primeros documentos que entrega ESA.

Para k=2, tendriamos por ejemplo (asumiendo que "pancreas cancer" es una consulta informacional, y para este tipo se usa "2") en formato SVM_LIGHT:

2 363559:1 38300:1357810:1 #pancreas cancer

Nótese que se recomienda utilizar features binarios. Elaborar un informe que contenga como mínimo:

1) Una explicación detallada de lo que hace ESA. Compare con la vista dada por BoW. Como referencia use: http://www.cs.technion.ac.il/~gabr/papers/ijcai-2007-sim.pdf

2) Hacer 10 cross-validation para MaxEnt y SVM, manteniendo la consistencia de los folders (véase la tarea 5). Muestre los resultados en términos de accuracy y mean reciprocal rank. Discuta y ejemplifique sus resultados.

Nótese que para la pregunta 1) no se está pidiendo una traducción del artículo recomendado. Tampoco se puede utilizar las figuras ahí provistas.