CC66Q: Diplomado Inteligencia Artificial
NLP - Evaluación 1

Profs. Mauricio Cerda, Fabian Villena

22 noviembre 2020

En Chile, cada Servicio de Salud genera una lista de espera, en base a las derivaciones de los médicos de los servicios primarios de salud (CESFAM). Al momento de derivar, el médico general indica una sospecha diagnóstica y sugiera una derivación a especialista. Ud. dispone de una base de datos de lista de espera entre 2012 y 2017 para la región de aysen (datos obtenidos vía ley de transparencia). A continuación se le presenta una serie de tareas, con el fín de visualizar y predecir las derivaciones en base a la sospecha diagnóstica.

La siguiente actividad se puede realizar de manera individual o grupal (grupos del trabajo del diploma). Se debe entregar un documento (pdf, jupiter notebook, o ambos) seún la fecha indicada en u-cursos, con sus resultados. Los criterios de evaluación son 20 % legibilidad de su entrega (texto, código, o ambos), 20 % resolver el problema solicitado, 60 % utilizar técnicas estudiadas en el curso.

1. Limpieza de la base de datos (2 pts)

Utilizando las columnas "SOSPECHA_DIAG" (texto libre) y "PRESTA_EST" (categoría):

- Cuente las derivaciones a cada especialidad, haga un ranking e indique las que tienen mas 10,000 filas asociadas. A esta selección le llamaremos las D10K.
- Para las D10K, construya una o varias expresiones regulares que eliminen signos de puntuación, cambien todo el texto a minúsculas, y cualquier símbolo no-ASCII.

- Utilice alguna lista de stopwords de su preferencia y elimine dichas palabras del texto.
- Realice lematización o stemming del texto con la libreria de su preferencia.

2. Análisis no-supervisado y visualización (2 pts)

Utilice en esta sección la salida de la sección de limpieza, o el texto crudo.

- Visualice utilizando wordcloud las palabras más comunes para cada categoría D10K.
- Calcule las matrices término-documento, y TF-IDF para D10K.
- Realice SVD sobre la matriz TF-IDF, e indique los términos más comunes en las primeras 8 agrupaciones detectadas. Aproximadamente ¿coinciden las categorías de las derivaciones? Justifique su respuesta.
- Proponga otra visualización alternativa.

3. Análisis supervisado (2 pts)

Utilice en esta sección la salida de la sección de limpieza, o el texto crudo.

- Calcule la probabilidad de cada categoría en D10K.
- Construya un diccionario, y para cada palabra la probabilidad $P(\text{palabra}|\text{categoria})$. No olvide entrenar con una parte de la base de datos para evitar el sobreajuste.
- Construya un clasificador Bayesiano Naive. Reporte el resultado de 3 métricas de clasificación ¿Detecto alguna categoría de mayor dificultad? ¿puede sugerir alguna interpretación a su resultado?
- ¿Cómo varía su resultado al utilizar bigramas?