

IN5526-1 Web Intelligence. Primavera 2016

Profesores: Juan Velásquez, Gaspar Pizarro. **Auxiliar:** Tomás Valdivia.

Auxiliar 2

P1. El Departamento de Educación Continua (DEC) de la Universidad de Chile, preocupados por brindar un mejor servicio a sus potenciales clientes, ha preparado una encuesta a los egresados de todas las carreras de las universidades Chilenas con el fin de saber de primera mano qué competencias o nuevos conocimientos se requieren para enfrentar un mercado laboral cada vez más competitivo y cambiante.

Para lograr lo anterior, se preparó una encuesta que puede ser completada en línea a partir de un sitio web que se crea para tales

efectos. La encuesta posee preguntas de selección múltiple, más espacios para que los encuestados puedan escribir sus opiniones en texto libre. Como forma de incentivar el llenado de la encuesta, se sortearán 1.000 cupos en cursos de especialización.

La campaña para llenar la encuesta fue totalmente exitosa, recibándose más de 400 mil respuestas durante el período que duró la consulta. El problema ahora es cómo sacar información y conocimiento valioso a partir de los datos, que vayan más allá que la estadística.

Consultando algunos expertos de Data Science, se llegó a la conclusión de que se debían aplicar algoritmos de clustering para agrupar encuestas con características comunes y a partir de los centroides identificados, extraer patrones que permitieran analizar qué quieren los egresados.

Al respecto:

- a) Explique el proceso KDD ejemplificando cómo serían las etapas en este contexto.
- b) Establezca un vector de características para el análisis de clustering explicando claramente los contenidos de sus componentes.
- c) Establezca una distancia o medida de similitud para comparar vectores, justificando claramente su formulación.
- d) Explique cómo usaría los algoritmos k-Means y SOM para la extracción de los patrones desde las encuestas.

P2. Para los siguientes documentos:

- Leí un libro de la primera guerra mundial.
- En la primera guerra mundial se quemaron muchos libros.
- El libro "La guerra de dos mundos" fue el primero que leí.

y usando n -gramas para $n \in \{1, 2\}$:

- a) Genere una tokenización en base a unigramas y bigramas.
- b) Utilice lematización como parte del preprocesamiento. ¿Qué diferencia importante habría ocurrido al utilizar stemming? ¿Cuál es el objetivo de aplicar estas técnicas? De un ejemplo.
- c) ¿Comparando los n -grams lematizados, cómo cree usted que afectará el largo de los n -gramas del criterio al cálculo de similitud entre los tres textos? Explique.
- d) Genere los vectores Tf-Idf en ambos casos.
- e) Utilice la medida del coseno para determinar cuáles textos son más parecidos entre ellos.
- f) Discuta el resultado.