



Control 1

El siguiente control está pensado para una duración de 2:30 horas. Ponga de inmediato su nombre en las hojas de respuesta, si lo desea puede utilizar también las hojas de enunciado para responder.

Pregunta 1

Responda las siguientes preguntas de manera clara y precisa.

- (a) Explicar el proceso KDD y su importancia en el ámbito de Web Intelligence. (0,5 pto).
- (b) Explicar conceptos de Machine Learning, aprendizaje supervisado, no supervisado y por refuerzo. (0,5 pto).
- (c) Dada la siguiente matrix de confusión, calcular: accuracy, precision, recall. (1 pto).

	Actual: Yes	Actual: No
Predicted: Yes	320	105
Predicted: No	110	295

- (d) Explicar cómo funciona el algoritmo del gradiente descendente para encontrar los coeficientes de una regresión lineal. (1 pto).
- (e) Explicar cómo funciona el método de clusterización K-means, y qué consideraciones deben tenerse en cuenta para elegir el valor de K. Además, mencionar cómo se puede resolver el problema de la obtención de diversos óptimos locales. (1 pto).
- (f) Para ingresar a la Escuela de Anacoretas, es necesario rendir dos pruebas, Verbal y Matemáticas. Suponga que cuenta con los resultados de una generación de cierto colegio para ambas pruebas y si cada uno de esos estudiantes fue aceptado en la Escuela o no. Indique cómo haría usted para, mediante una regresión logística, predecir si un estudiante, dados sus resultados en las pruebas, puede ingresar a la Escuela. (1 pto).
- (g) Para la generación del vector de características en un problema de Sentiment Analysis, indique al menos 5 tipos de propiedades que utilizaría para la generación del mismo. (1 pto).

Hoja Respuesta Pregunta 1

Nombre:

Hoja Respuesta Pregunta 1

Nombre:



Pregunta 2

En la clase de WCM se abordó la metodología del *bag of words* donde cada documento se separa en tokens por cada palabra (en el concepto de los n-grams se conoce como unigram). En el ejemplo del blog subido a U-Cursos, pudieron ver que no necesariamente tiene que ser así. Se pueden generar también bi-grams, tri-grams (con dos y tres palabras consecutivas como una sola "token"). Éstos son casos especiales de los n-grams.

1. Con respecto a lo anterior,

- Indique posibles ventajas de utilizar n-grams de mayor longitud y de un ejemplo. (0,5 pto).
- Indique posibles desventajas de utilizar n-grams de mayor longitud y de un ejemplo (0,5 pto).

2. Para los siguientes documentos:

- Leí un libro de la primera guerra mundial el mes pasado.
 - En la primera guerra mundial se quemaron muchos libros.
 - El libro "La guerra de dos mundos" fue el primero que leí.
- Genere una tokenización en base a unigrams y bigrams. (0,2 pto).
 - 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. (1 pto).
 - ¿Comparando los n-grams lematizados, cómo cree usted que afectará el largo de los n-grams del criterio al cálculo de similitud entre los 3 textos? Explique (1 pto).
 - Genere los vectores TF-IDF utilizando los unigrams y el resultado de la lematización. (1 pto).
 - Utilice la medida del coseno para determinar cuáles textos son más parecidos entre ellos. (1 pto).
 - Discuta el resultado. (0,8 pto).

Hint: Para un documento $d = t_1 t_2 t_3 t_4 t_5$, el resultado utilizando unigrams sería t_1, t_2, t_3, t_4, t_5 , mientras que utilizando bigrams sería $t_1 t_2, t_2 t_3, t_3 t_4, t_4 t_5$.

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{|d_1| \times |d_2|} = \frac{\sum_{i=1}^{|V|} (w(t_i, d_1) \times w(t_i, d_2))}{\sqrt{\sum_{i=1}^{|V|} w(t_i, d_1)^2} \times \sqrt{\sum_{i=1}^{|V|} w(t_i, d_2)^2}}$$



INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 - Web Intelligence

Profesores: Juan D. Velásquez
Pablo Tapia
Auxiliar: Gino Slanzi

Hoja Respuesta Pregunta 2

Nombre:



INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 - Web Intelligence

Profesores: Juan D. Velásquez
Pablo Tapia
Auxiliar: Gino Slanzi

Hoja Respuesta Pregunta 2

Nombre:



Pregunta 3

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 fines. 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 400.000 mil 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.

Consultados 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 Establezca un vector de características para el análisis de clustering (feature vector), explicando claramente los contenidos de los contenidos de sus componentes. (1 punto)
- b Establezca una distancia o medida de similitud para comparar vectores, justificando claramente su formulación. (2 puntos)
- c Explique cómo usaría un algoritmo KNN y un SOFM para la extracción de los patrones desde las encuestas. (3 puntos)



INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 - Web Intelligence

Profesores: Juan D. Velásquez
Pablo Tapia
Auxiliar: Gino Slanzi

Hoja Respuesta Pregunta 3

Nombre:



INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5526 - Web Intelligence

Profesores: Juan D. Velásquez
Pablo Tapia
Auxiliar: Gino Slanzi

Hoja Respuesta Pregunta 3

Nombre: