

**Tarea no. 1**

**Curso:** *Analítica Textual*

**Magister en Data Sciences**

**Prof. John Atkinson**

**Universidad Adolfo Ibañez**

**Fecha Publicación: Sábado 12 de Marzo de 2022**

**Fecha de Entrega: Sábado 26 de Marzo de 2022**

**Objetivo**:

Entender y aplicar modelos y técnicas de análisis textual para la resolución de un problema real que involucren datos de naturaleza textual.

**Problema**:

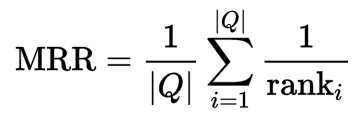
Actualmente, existen varios portales web donde profesionales pueden recurrir a buscar trabajo (ej. *Linkedin, Glassdoor, JD*). Usualmente estos sitios están organizados de tal forma, que un profesional que busca trabajo, puede filtrar opciones de búsqueda según su perfil y/o expectativas, ya sea por salario, funciones, tipo de trabajo, conocimiento requerido, años de experiencia, etc.

Sin embargo, el match entre quien busca trabajo y las descripciones de los trabajos, usualmente debe ser realizado manualmente por el mismo candidato, leyendo cada descripción de cargo buscando las cercanías con su propio perfil. Este proceso es claramente muy ineficiente y no exento de problemas causados por la ambigüedad de las descripciones, términos poco claros, etc.

Con el fin de abordar este problema, Ud. deberá construir un modelo en base a las técnicas vistas en clases que permita realizar un match automático entre un perfil dado y las descripciones de trabajo que se encuentran en una base de datos textual dada por el profesor. El modelo debe permitir que un usuario (i.e., profesional que busca trabajo), especifique algunas características de su perfil, y luego de analizar los textos de descripción se entregue los cargos que mejor calzan ordenados de mayor a menor relevancia.

Para facilitar el trabajo y evitar la recolección de ofertas de trabajo, en webcursos ya se encuentra disponible un pequeño dataset de ofertas (en Inglés) con un poco más de 4400 registros, que se puede utilizar como corpus de entrada.

Para lograr el objetivo anterior, es **MUY IMPORTANTE** que Ud. revise los programas de ejemplos prácticos dados por el profesor en clases pues allí se encuentran muchos módulos que Ud. puede adaptar para elaborar esta tarea y así evitar construirlos desde cero. Luego, se sugiere seguir los siguientes pasos **TENTATIVOS** (Ud. puede adaptar/agregar para mejorar la calidad de sus resultados):

1. Bajar el conjunto de descripciones de trabajo disponibles en WebCursos.
2. Preprocesar los textos y convertirlos en un corpus accesible desde un programa en Python (i.e., eliminar stopwords, lematizar, reconocer entidades, etc).
3. Generar una representación vectorial (i.e., *TF x IDF*) de los textos para poder realizar las actividades posteriores.
4. Solicitar al usuario una cierta descripción que especifique el tipo de trabajo o perfil que está buscando (ej. “data scientist”).
5. Utilizando la representación vectorial de (4) compararla con cada uno de los documentos representados en el paso (3) via alguna medida de similitud (ej. *coseno*).
6. Ordenar el resultado de mayor a menos similitud y mostrar las descripciones, o sea, se generará un ranking.
7. Evaluar **MANUALMENTE** el rendimiento de los resultados generados. Para evaluar la calidad de los resultados que corresponden a una lista ordenados (ranking), Ud. deberá utilizar una métrica conocida como el **MRR** (*Mean Reciprocal Rank*), que mide el promedio de los “rangos” recíprocos (i.e., inversos) de los resultados de una muestra de perfiles **Q**:
8. 

Donde **ranki** se refiere a la posición en el ranking de salida del primer texto de trabajo relevante para la **i**-ésimo perfil de entrada, y **|Q|** es el número de perfiles consultados. Ud. se dará cuenta de que esto “premiará” (i.e., entregará valores altos) a aquellos perfiles de entrada cuya mejor descripción de cargo se encuentra en las posiciones más altas del ranking.

¿Por qué es necesario todo esto?. Podrían darse las siguientes situaciones dado un perfil de entrada **q:**

1. La descripción de cargo más relevante a **q** aparece en la primera posición del ranking. Este es el caso ideal por lo que el MRR será muy alto.
2. La descripción de cargo más relevante a **q** aparece en una posición entre medio del ranking. El MRR es más bajo pero sigue siendo bueno.
3. La descripción de cargo más relevante a **q** aparece al final del ranking. El MRR entregará valores muy bajos indicando que la calidad de los resultados no es muy buena (la respuesta correcta no aparece en las primeras posiciones).

Para medir el MRR, Ud. deberá tomar **MANUALMENTE** una pequeña muestra de perfiles (ej. 10) y según el ranking de salida para cada uno calcular el MRR total de su modelo.

En el caso de que su sistema tenga un MRR no muy alto, deberá volver atrás y realizar ajustes a su modelo para mejorar el rendimiento (ej. preprocesar otros elementos, variar el tipo de modelo vectorial, utilizar otro tipo de limpieza, etc).

Las tareas finales se evaluarán según el MRR obtenido con la muestra analizada manualmente.

***Aspectos Administrativos***:

El trabajo debe ser realizado en **grupos de 2-3 personas** y se debe entregar:

1. El/Los programas fuentes que resuelven el problema
2. Un breve informe (máximo 2 páginas) que muestre y discuta la evaluación obtenida según MRR, sugerencias de mejora, problemas encontrados, etc.
3. Una breve descripción del rol o tareas que realizó cada uno de los miembros del grupo (no más de un párrafo por cada integrante).

La no entrega de una tarea en la fecha oficial, implicará un descuento de ***0.5 ptos por cada día de atraso en la nota final***.