Enunciado del Proyecto de fin de Curso

INF656, Minería Web
2019-II

Análisis de documentos textuales

La Minería de la Web es una disciplina que permite generar información acerca del comportamiento de los usuarios en un Sitio Web, mediante el análisis de los datos que ellos mismos van dejando a medida que visitan los Sitios Web. Antes de analizar este comportamiento, primeramente será necesario construir un sitio Web que permita el acceso de los internautas. Este trabajo del fin de curso esta dividido en dos partes:

1. La construcción de una página web. En esta página web los alumnos mostrarán la información obtenida (resultados) del uso de cualquier técnica de *text mining* sobre cualquier "corpus", en cualquier lengua.

2. El análisis de comportamiento de los visitantes de la página web construida anteriormente.

Con respecto al primer punto, el grupo puede poner en práctica cualquier método aprendido en el curso o cualquier otro método o métodos que se encuentren en la literatura (classification, clustering, event summarization, topic extraction, topic models, visualization, spatial entities identification, sentiment analysis, information retrieval, search engines, etc.). Para este trabajo, los alumnos deben utilizar Python + Flask

La cantidad de documentos textuales (el tamaño del corpus) puede variar de 1 a muchos. Esta cantidad está en relación al tipo de documento a analizar (un tweet o una novela). Por ejemplo. si analizamos una novela, podemos crear documentos a partir los capítulos, cosa que no podemos hacer con tweets. La selección del método dependerá del corpus a analizar y de lo que se desea tener como resultado.

Con respecto al segundo punto, los detalles le será entregado en las siguiente semanas.

Los grupos: los trabajos serán presentados en grupos de 3 a 5 personas. Obviamente, la complejidad de los trabajos presentados por 5 personas deberá ser mayor que la complejidad de aquellos entregados en grupos de 3.

Acerca del entregable: la forma de evaluación consistirá en una defensa de la proposición y un reporte de tres hojas como máximo. Para la primera parte, los grupos contarán con 15 minutos, en los cuales, deberán exponer ante sus compañeros la metodología utilizada y hacer una demostración de su proposición. El grupo podrá utilizar el material que crea conveniente (slides, videos, animaciones, hologramas, etc.). Con respeto al reporte, éste deberá mostrar una breve descripción del corpus, la metodología utilizada y los resultados obtenidos del análisis de comportamiento de los visitantes al sitio Web (esto se verá durante el resto del curso).

Sobre la calificación: la nota del trabajo práctico se calculará bajo los siguientes criterios: 1) la selección del corpus, la forma cómo se representaron los documentos, los métodos utilizados, la novedad de la propuesta y los resultados obtenidos conformarán el 70 % del total de la nota, y; 2) el otro 30 % de la nota está asociada a la defensa de la proposición.

Fechas importantes: la fecha de entrega y exposición de los proyectos será el último día de clases (5 de diciembre del 2019).

Finalmente, el alumno está invitado a consultar con el docente del curso sobre su proyecto, a fin de poder determinar la complejidad de la propuesta que se desea desarrollar y para aclara cualquier duda.

Buena suerte...

–
Hugo Alatrista Salas
PUCP