



Procesamiento de Lenguaje Natural para CCSS y Gestión Pública

Cantidades enormes de información relevante para las ciencias sociales son producidas diariamente, en reportes de organizaciones, tweets, periódicos, redes sociales, entre otros. Extraer información útil de estas grandes cantidades de texto es una tarea que requiere conocer los principios de procesamiento de texto como análisis de datos y de inteligencia artificial. En este taller se proporcionarán estos fundamentos con un enfoque aplicado a las ciencias sociales. En particular, se aplicará lo aprendido a la extracción de información desde cientos de reportes de la contraloría y de sentencias judiciales para producir información sobre corrupción en el Perú.

Docentes	Ph.D. Cristian Muñoz Villalobos (cmunozv@uni.pe)
Horario	Sábados 2:00 PM - 5:00 PM

Contenido

# Semana	Titulo	Descripción	Fecha	Tipo	Tema
	Introducción al Procesamiento de Lenguaje Natural y Extracción de Información	Fundamentos. Pipeline do proceso de extracción de información y tareas de Procesamiento de Lenguaje Natural.	@July 2, 2022 2:00 PM-5:00 PM	Clase	Introducción
	Herramientas para extracción de información	Formatos de datos no estructurados. Procesamiento de documentos word, pdf, excel. Web Scrapping. Limpieza de datos, Expresiones Regulares.	@July 9, 2022 2:00 PM-5:00 PM	Clase	Extracción de Información
	Análisis de Formatos de Documentos	Reconocimiento Óptico de Caracteres, Detección de Objetos para análisis de documentos. API Comerciales. Exploración de datos estructurados.	@July 16, 2022 2:00 PM-5:00 PM	Clase	Extracción de Información

# Semana	Titulo	Descripción	Fecha	Tipo	Tema
	Tarea 1			Tarea	
	Representation Learning (Parte 1)	Representación binaria y distribuida de palabras. Tipos de representaciones. Aprendizaje de Representaciones. Representación contextual y no contextual.	@July 23, 2022 2:00 PM-5:00 PM	 Clase	Procesamiento de Lenguaje N
	Representation Learning (Parte 2)	Similaridad de documentos.: TF-IDF, LDA, Doc2Vec LDA. T-SNE, UMAP/Tensorboard	@July 30, 2022 2:00 PM-5:00 PM	 Clase	Procesamiento de Lenguaje N
	Modelos Recurrentes y Transformers	Modelamiento de Lenguaje. Modelos de Atención, Estructura de Modelo Transformer. BERT, GPT-3, T5.	@August 6, 2022 2:00 PM-5:00 PM	 Clase	Procesamiento de Lenguaje N
	Análisis de Sentimiento e Intensiones	Fundamentos. Modelamiento de un sistema de análisis de sentimientos. Detección de Intenciones.	@August 13, 2022 2:00 PM-5:00 PM	 Clase	Procesamiento de Lenguaje N
	Tarea 2			Tarea	
	Extracción de Entidades, Relaciones Nombradas y Grafos de Conocimiento y nuevas direcciones PLN	Fundamentos. Modelos pre-entrenados, entrenamiento de modelos, visualización y almacenamiento de conocimiento. Asistentes Virtuales, Sistemas Pregunta-Respuesta, Resúmenes de Texto, etc.	@August 20, 2022 2:00 PM-5:00 PM	 Clase	Procesamiento de Lenguaje N
	Tarea 3			Tarea	
	Proyecto	A elegir	@August 27, 2022 2:00 PM-5:00 PM	 Presentación	

Pre-requisitos

Es recomendable tener los siguientes conocimientos básicos para un correcto entendimiento del curso:

- Redes Neuronales Artificiales
- Conceptos relacionados a Aprendizaje Automático (Machine Learning)
- Programación en Python



Bibliografía

- Karthiek Reddy Bokka, Shubhangi Hora, Tanuj Jain, Monicah Wambugu, Deep Learning for Natural Language Processing (2019)

- V Kishore Ayyadevara, Yeshwanth Reddy, "Modern Computer Vision with PyTorch" (2020)
- Masato Hagiwara, "Real-World Natural Language Processing" (2021)
- Denis Rothman. "Transformers for Natural Language Processing", (2021)
- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, "Introduction to Information Retrieval" (2008)
- Masato Hagiwara , "Real-World Natural Language Processing" (2021)
- Ian Goodfellow and Yoshua Bengio and Aaron Courville, "Deep Learning" (2015)
- Wolf, Thomas, et al. "Transformers: State-of-the-art natural language processing." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020.

Calificación

Participación (P): 20%

Tareas (T): 30% ($T = \frac{T1+T2+T3}{3}$)

Proyecto Final (F): 50%

$$\text{Nota Final} = \frac{20P+30T+50F}{100}$$

Proyecto Final

Para el proyecto final, se deberá preparar una presentación de aproximadamente 15 minutos con el siguiente contenido:
Introducción, Conjunto de datos, Metodología, Resultados y un repositorio con el código fuente (GitHub, GitLab, entre otros).