**Código de la Asignatura: IO153**

| | |
|---|---|
| **Nombre del programa académico** | Maestría en Investigación Operativa y Estadística |
| **Nombre completo de la asignatura** | Técnicas de Procesamiento de Lenguaje Natural |
| **Área académica o categoría** | Estadística Enfocada a la Ciencia de Datos |
| **Semestre y año de actualización** | Segundo semestre de 2023 |
| **Tipo de asignatura** | [ ] Obligatoria [ x ] Electiva |
| **Número de créditos ECTS** | 5.22 ECTS |
| **Intensidad horaria** | 10 Horas Semanales – 40 Mensuales |
| **Director o contacto del programa** | Eliana Mirledy Toro Ocampo |
| **Coordinador o contacto de la asignatura** | Julián David Echeverry Correa - jde@utp.edu.co-Google Scholar Julian David Echeverry |

**Descripción y Contenidos**

**1. Breve descripción**

La asignatura de Técnicas de Procesamiento de Lenguaje Natural es de naturaleza teórica-práctica, el propósito que tiene es el estudio de las técnicas probabilísticas, estadísticas y de aprendizaje de máquina relacionadas con el desarrollo de sistemas de procesamiento de lenguaje natural. Se abordan los siguientes temas: procesamiento básico de texto, modelos de lenguaje probabilísticos, modelos de recuperación de información, sistemas de clasificación de textos, sistemas de diálogo (e.g. question-answering, chatbots), sistemas de sentiment analysis y sistemas automáticos de reconocimiento de habla y de traducción.

**2. Objetivos del programa académico**

OP2. Identificar herramientas y enfoques efectivos para el uso de los recursos de las organizaciones con el fin de optimizar sus procesos, combinando las prácticas de investigación de operaciones y estadística, que permitan mejorar la eficiencia, optimizar la toma de decisiones, prever tendencias y mantener su competitividad en un mercado cada vez más complejo y basado en datos. OP3. Analizar y extraer conocimientos significativos a partir de enfoques estadísticos multivariados, tanto cualitativos como cuantitativos, para examinar simultáneamente múltiples variables o factores en un conjunto de datos, con el objetivo de obtener información más completa y profunda sobre las relaciones y patrones entre ellas, aplicados en diferentes campos, como la investigación científica, la ingeniería, la economía, la medicina, entre otras.

**Objetivos de la asignatura**

OA1. Introducir conceptos fundamentales de PLN desde la perspectiva de las ciencias computacionales, empleando el pensamiento crítico para estructurar soluciones que transformen el lenguaje humano en datos estructurados. OA2. Desarrollar habilidades avanzadas de análisis de textos y minería de opinión bajo los principios de la formación humana, garantizando que el tratamiento de la información respete la privacidad y la veracidad de los datos.

OA3. Diseñar y desplegar modelos de aprendizaje de máquina (Machine Learning) y sistemas de recuperación de información orientados a la sostenibilidad, fortaleciendo la ciudadanía mediante el acceso a información transparente y democrática.

**3. Resultados de aprendizaje de la asignatura**
RAA1. Define conceptos de análisis de textos y diseña sistemas de diálogo complejos (Chatbots, agentes virtuales) empleando el pensamiento crítico, asegurando que la arquitectura del sistema sea coherente con la sostenibilidad de los recursos tecnológicos. RAA2. Distingue y aplica modelos de lenguaje avanzados para resolver problemas semánticos y de ambigüedad, validando los resultados bajo el rigor técnico y la ética de la formación humana, evitando sesgos algorítmicos en la interpretación de los textos. RAA3. Identifica y ejecuta algoritmos de clasificación de textos y técnicas de traducción automática, optimizando los modelos para garantizar una operación sostenible y eficiente en contextos de alta demanda de datos. RAA4. Aplica modelos de recuperación de información y análisis de grandes volúmenes de texto para fortalecer la transparencia informativa, comunicando hallazgos que promuevan la ciudadanía y el juicio profesional responsable.

**4. Contenido**● T1. Introducción a las técnicas de procesamiento de lenguaje natural (9 h: 3 h con docente y 6 h independientes). ● T2. Procesamiento de texto (15 h: 4 h con docente y 11 h independientes). ● T3. Modelos de lenguaje probabilísticos (18 h: 5 h con docente y 13h independientes). ● T4. Clasificación de textos (18 h: 5 h con docente y 13h independientes). ● T5. Modelos de recuperación de información (18 h: 5 h con docente y 13h independientes). ● T6: Análisis semántico (12 h: 3 h con docente y 9 h independientes). ● T7: Sistemas de diálogo (12 h: 3 h con docente y 9 h independientes). ● T8: Reconocimiento automático de habla (12 h: 3 h con docente y 9 h independientes). ● T9: Reconocimiento automático de locutor (9 h: 3 h con docente y 6 h independientes). ● T10: Alineamiento estadístico y traducción automática (9 h: 3 h con docente y 6 h independientes). ● T11: Aplicaciones de los sistemas de procesamiento de lenguaje (12 h: 3 h con docente y 9 h independientes)

**5. Requisitos:** ● Teoría de probabilidad ●Inferencia Estadística

**6. Recursos**
**Libros de texto:** Baeza-Yates, R., y Ribeiro-Neto, B. (2011). Modern information retrieval: The concepts and technology behind search (2ª ed.). Addison-Wesley. Bird, S., Klein, E., y Loper, E. (2019). Natural language processing with Python: Analyzing text with the Natural Language Toolkit. O'Reilly Media. Jurafsky, D., y Martin, J. H. (2023). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3ª ed. revisada). Manning, C. D., Raghavan, P., y Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press. Manning, C. D., y Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press. Witten, I. H., Frank, E., Hall, M. A., y Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques (4ª ed.). Morgan Kaufmann. Artículos académicos y lecturas complementarias proporcionadas durante el curso. **Herramientas:** NLTK, Google Cloud Natural Language, Stanford CoreNLP, SpaCy, Gensim, Mallet, HTK y Moses. **Recurso web:** página de la asignatura en Google Classroom.

**7. Herramientas técnicas de soporte para la enseñanza:** Desarrollo de algoritmos y sistemas automáticos para el procesamiento de lenguaje, mediante ejercicios individuales y grupales, apoyados en las herramientas técnicas mencionadas en los recursos.

**8. Trabajos en laboratorio y proyectos:** Se proponen, en el desarrollo del curso, proyectos de laboratorio y trabajos que relacionen las temáticas tratadas en él: taller de procesamiento de texto y modelos de lenguaje (T2 y T3: 12 horas), taller de clasificación, recuperación de información y análisis semántico (T4 a T6: 12 horas), anteproyecto (6 horas) y proyecto final de interacción persona-máquina (T7 a T11: 30 horas).

**9. Métodos de aprendizaje:** Clases magistrales apoyadas en TIC, lecturas especializadas como trabajo autónomo y talleres grupales relacionados con los contenidos del curso.

**10. Métodos de evaluación:** ●Entrega 1 (20%): Taller grupal sobre procesamiento básico de texto y modelos de lenguaje probabilístico . ●Entrega 2 (20%): Taller grupal sobre clasificación de textos, recuperación de información y análisis semántico . ●Entrega 3 (10%): Anteproyecto individual con problema, hipótesis y propuesta metodológica . ●Entrega 4 (50%): Proyecto individual para desarrollar un sistema de interacción persona-máquina en PLN, incluyendo diálogo, reconocimiento de habla y traducción automática