Universidad Tecnológica Nacional
Facultad Regional Buenos Aires

PROGRAMA ANALÍTICO DE ASIGNATURA

DEPARTAMENTO: Ingeniería en Sistemas de Información

CARRERA: Ingeniería en Sistemas de Información

NOMBRE DE LA ACTIVIDAD CURRICULAR: Procesamiento lenguaje natural

Año Académico: 2023
Área: Sistemas Inteligentes
Bloque: Electivas
Tipo: Electiva
Modalidad: Cuatrimestral

Cargas horarias totales:

Horas reloj	*Horas cátedra*	*Horas cátedra semanales*
72	96	6

FUNDAMENTACIÓN

El ser humano se comunica y expresa a través del lenguaje. El Procesamiento del Lenguaje Natural se ocupa de la investigación y aplicación de mecanismos computacionales para la interacción entre personas y máquinas. Tiene por objetivo hacer posible la comprensión y el procesamiento de mensajes, tanto en el análisis de textos como en el de voz, ya que nos ofrece una enorme oportunidad de avanzar en la gestión del conocimiento.

El tesoro más valioso de la raza humana es el conocimiento, es decir, la información procesada. Existen en el mundo volúmenes inmensos de información en forma de lenguaje natural: los libros, los periódicos, los informes técnicos, *papers* científicos, entre otros. Sin embargo, el verdadero tesoro implica la habilidad de comprender esa información acumulada con criterios útiles para el razonamiento lógico.

Actualmente, gracias a algoritmos más centrados en capturar generalidades más que particularidades, se ha abierto todo un nuevo campo a la aplicación y desarrollo de estas técnicas en problemáticas diversas.

La asignatura electiva PROCESAMIENTO DEL LENGUAJE NATURAL propone dotar al futuro Ingeniero en Sistemas de Información, analizando los grandes volúmenes de información generados por la humanidad desde una mirada ingenieril, computacional y aplicada para la

resolución de problemas mediante uso de algoritmos, metodologías, métodos y herramientas que son la aplicación de diversos campos de las ciencias.

OBJETIVOS

- Distinguir los conocimientos relacionados a la Sintaxis y Semántica del Lenguaje, Probabilidad, Álgebra Lineal, Algoritmos y Estructuras de Datos y Paradigmas de Programación.
- Utilizar técnicas y herramientas modernas que se utilizan en el desarrollo de sistemas robustos y prácticos para que puedan comunicarse con los usuarios en lenguaje natural.
- Reconocer modelos, teorías, y técnicas de los enfoques computacionales para el tratamiento computacional del lenguaje natural.
- Reconocer procesos, algoritmos y modelos propuestos por la Ingeniería del Lenguaje Natural.
- Describir las características de diferentes modelos del Lenguaje Natural, sus desafíos y limitaciones actuales.
- Utilizar herramientas de desarrollo de software de uso libre y gratuito para el análisis crítico de metodología de investigación aplicada a un problema concreto basado en las competencias del Ingeniero en Sistemas de Información.
- Identificar nociones sobre investigación aplicada a la industria (I+D+i) y su implementación y comunicación efectiva a través de un caso real.
- Utilizar campos avanzados de Procesamiento de Lenguaje Natural, tales como: *Clasificación de artículos, Generación de lenguaje escrito, Análisis de sentimientos de texto y voz, generación de Respuestas a través de Preguntas, Descripción automática de imágenes con lenguaje natural, Resúmenes Automáticos, Bots de Texto, Procesamiento de Lenguaje Natural Robusto*, entre otros.

CONTENIDOS

Contenidos analíticos

Unidad Temática 1: Introducción al Procesamiento del Lenguaje Natural. Estado del Arte y Aplicaciones actuales

Introducción formal a la asignatura. Definiciones. Problemas frecuentes. Diferentes tipos de categorías. Estado del Arte y aplicación del mismo a la industria desde la mirada del Ingeniero en Sistemas de Información. Limitaciones.

Unidad Temática 2: Análisis morfológico y etiquetación morfosintáctica

Definiciones de Lenguaje Natural Español. Etiquetado morfosintáctico. Etiquetado basado en reglas. Etiquetado estocástico. Etiquetado basado en transformación. Etiquetado POST (Part of Speech Tagging).

Unidad Temática 3: Unidades lingüísticas y representaciones vectoriales de palabras

Palabra como unidad básica. Repaso Álgebra Vectorial. Multipalabras o unidades léxicas. Acortamiento. Composición. Clases de palabras. Lematización. One hot encoding. Bag of Words (BoW). CBOW. Skip-gram. Autoencoders. Transformadores. Otros modelos, estado del arte.

Unidad Temática 4: Modelos Secuencia a Secuencia

Seq2seq network. Procesamiento a nivel de caracteres. Mecanismos de Atención. Procesamiento a nivel de palabra. Embeddings. Introducción traducción de máquinas. Métricas para evaluación de modelos de Procesamiento del Lenguaje Natural.

Unidad Temática 5: Clasificación de Textos. Minado de Texto.

Modelos computacionales de clasificación y etiquetado de textos. Revisión de corpus. Práctica y laboratorio intensivo en Python. Preparación de textos para análisis y etiquetado.

Descubrimiento de patrones en texto no estructurado. Análisis automático de texto.

Unidad Temática 6: Modelos de Lenguaje Condicionales

Breve repaso a probabilidad y estadística. Gramática estocástica: N-gramas. Modelo de consulta del lenguaje probabilístico. Limitaciones y ventajas del modelado a través de N-gramas

Unidad Temática 7: Reconocimiento de Entidades Nombradas (NER)

Detección y extracción de Entidades. Expresiones Regulares. Limitaciones. Algoritmos y técnicas para su detección. Desambiguación de Entidades. Métodos de Evaluación: valor-F.

Ejemplos de importancia en la industria.

Unidad Temática 8: Modelado de Preguntas y Respuestas

Definición del concepto de “Recuperación de la Información” (R-I). Método de Filtrado, Encaminado, Routeo. Definición del concepto “Preguntas y Respuestas” (P-R) y “Búsqueda de Respuestas” (B-R). Dominios libres y dominios específicos. Fuentes de información. Tipos de usuarios. Precisión, Completitud y Relevancia de la respuesta. Taxonomía de preguntas. P-R.

Repaso unidades anteriores para implementación de sistemas de P-R.

Unidad Temática 9: Generación automática de resúmenes

Recuperación de la Información (R-I). Minería de textos. Repaso Estadística y Procesamiento Lenguaje Natural. Seguimiento de tópicos. Agrupamiento. Vinculación conceptual. Generación automática de texto de dominio restringido. Competencias internacionales de Generación Automática de Resúmenes. Condiciones y Formas de evaluación: ROUGE, BLUE score, Evaluación Humana.

Unidad Temática 10: Reconocimiento de voz. Texto a voz

Sistemas y algoritmos de código abierto para reconocimiento de voz. Estado del arte y limitaciones del reconocimiento de voz. Palabras aisladas, Palabras conectadas, Palabras continuas. Problema de múltiples oradores en simultáneo.

Procesamiento de Texto, generación de la prosodia, generación de voz sintética. Grafemas. Fonemas. Silabificación. Errores de cuadrados medios y Coeficiente de correlación como métodos de evaluación de algoritmos y modelos.

BIBLIOGRAFÍA OBLIGATORIA

- Allen J. (1987) Natural Language Understanding. Ed- Addison-Wesley.
- Austin J. (1988) How to do things with words. Ed. Oxford University Press.
- Bing Liu (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Ed. Cambridge University Press.
- Candalija Reina, J.A., (1998), "Sobre la cientificidad de la gramática: el uso de corpora informatizados como método de análisis lingüístico". Estudios de Lingüística Cognitiva. Ed. J.L
- Christopher D. Manning y Hinrich Schctze (1999). Foundations of Statistical Natural Language Processing. Ed. The MIT Press.
- Cristian Cardellino, Serena Villata, Laura Alonso Alemany, Elena Cabrio (2015): Information Extraction with Active Learning: A Case Study in Legal Text. Ed. CICLing.
- Dale R., Moisl H., Somers H. , (2000) Handbook of Natural Language Processing. Ed. Dekker.
- Daniel Jurafsky & James H. Martin (2009). Speech and Language Processing, 2nd Edition. Ed. Pearson.
- Fellbaim C. (1998) WordNet: An Electronic Lexical Database. Ed. MIT Press.
- Gazdar G., Mellish C. (1989) Natural Language Processing in Prolog: an introduction to computational linguistics. Ed. Addison-Wesley.
- Grishman, R., (1986), Computational Linguistics: An Introduction. Ed. Cambridge University Press, Cambridge.
- Indurkhya, N. y Damerau, F. J. (2010). Handbook of natural language processing. Ed. Chapman and Hall/CRC.
- Iwanska I., Shapiro Stuart (2000). Natural Language Processing and Knowledge Representation. Ed. MIT Press.
- Manoiloff, L. Carando, M. Defagó, M. Alonso, L. Alemany, Ferrero, C. Cesaretti, D. Ramirez, A. and Seguí, J. (2015). The cognitive processing. Ed. MIT Press.
- Manning C. Schutze H. (1999) Fundations of Statistical Natural Language Processing. Ed. MIT Press.
- Marcu D. The theory and Practice of discourse parsing and summarization. Ed. MIT Press 2000.

- Mariani, Joseph; Francopoulo, Gil; Paroubek, Patrick; Vernier, Frédéric (2019), «The NLP4NLP Corpus (I): 50 Years of Research in Speech and Language Processing», Frontiers in Research.
- Mitkov, R. (ed.) (2003): The Oxford Handbook of Computational Linguistics, Oxford. ED. Oxford University Press.
- Moreno Fernandez, F. (1990), "Lingüística informática e informática lingüística", Ed. Lingüística Española Actual.
- P. Estrella and Nikos Tsourakis (s.f), "Migrating from ISO/IEC 9126 to SQUARE: A case study on the evaluation of medical speech translation systems", book chapter in "Design Development".
- Roche E., Schabes Yves (1997). Finite-state Language Processing. Ed. MIT Press
- Rodríguez, H. (2000) "Técnicas básicas en el tratamiento informático de la lengua". Ed. Quark. Ciencia, Medicina, Comunicación y Cultura.
- Rojo, G. (2005-2006), "Informática y Lingüística: Las lenguas en la sociedad del conocimiento". Ed. Boletín de RedIRIS.
- Smith G. (1991) Computers and Human Language. Ed. Oxford University Press.
- Smith, N. (2011). Linguistic Structure Prediction. Ed. Morgan & Claypool Publishers.
- Tordera, J.C. (2012), El abecé de la Lingüística computacional. Ed.Arco/Libros.
- Torres Moreno, J. (2014). Automatic Text Summarization. Ed. Wiley-ISTE.
- Yoav Goldberg (2017). Neural Network Methods for Natural Language Processing. DXC. Ed. Morgan & Claypool Publishers.
- Zulaica Hernandez, I. (2016), "Lingüística de Corpus", en J. Gutiérrez-Rexach (ed.) Enciclopedia de Lingüística Hispánica, Vol. 1, London: Routledge.

PÁGINAS WEB DE INTERÉS

- Documentación y papers con el estado del arte de la materia a ser entregados por el profesor.
- ARES, F. (2008), El robot enamorado. Una historia de la Inteligencia Artificial, Barcelona: Ariel. [caps. 7 e 11]
- Stanford CoreNLP: https://stanfordnlp.github.io/CoreNLP/
- UDPipe: https://github.com/ufal/udpipe
- Apache OpenNLP: https://opennlp.apache.org/
- Natural Language Toolkit (NLTK): http://www.nltk.org/
- Revista Procesamiento Lenguaje Natural en Español: http://rua.ua.es/dspace/handle/10045/1112
- Jacob Eisenstein. Natural Language Processing
- Yoav Goldberg. A Primer on Neural Network Models for Natural Language Processing

CORRELATIVAS

Para cursar y rendir

- Cursadas:
    - Análisis de Sistemas de Información
    - Sintaxis y Semántica de los Lenguajes
    - Paradigmas de Programación