Ingeniería de Sistemas y Computación
Posgrado: ISIS 4221 Procesamiento de Lenguaje Natural
https://cursos.virtual.uniandes.edu.co/isis4221/
Semestre 2023-20

PROGRAMA DEL CURSO

INFORMACIÓN GENERAL

Instructor	Email	Virtual Office Hours
Rubén Francisco Manrique	rf.manrique@uniandes.edu.co	Appointment via email

TAs	Email
Arturo Hurtado	ja.hurtado905@uniandes.edu.co

PROCESAMIENTO DE LENGUAJE NATURAL

El Procesamiento del Lenguaje Natural es una disciplina de la Inteligencia Artificial que se ocupa de la formulación e investigación de mecanismos computacionales para la comunicación entre personas y máquinas mediante el uso de Lenguajes Naturales. El objetivo principal del curso es desarrollar una comprensión profunda de los algoritmos disponibles para procesar información lingüística y las propiedades computacionales subyacentes de los lenguajes naturales.

Plan de temas

- Introducción
  - Definición de procesamiento de lenguaje natural. el Lenguaje y sus componentes. Aplicaciones de PLN. ¿Por qué el entendimiento del lenguaje es una tarea compleja?
  - Linea de evolución del PLN.
  - Pipeline del procesamiento de lenguaje natural.
  - Operaciones de procesamiento de texto: Tokenización, Normalización, Stemming/Lemmatización, Palabras de Parada. Construcción de vocabularios y conteo de palabras.
- Introducción a recuperación de información textual (IR). Definición de colección, consulta
  - Búsquedas binarias (matrices termino-documento e índice invertido)
  - Búsquedas ranqueadas
  - Representaciones vectoriales de documentos
  - Modelo de bolsa de palabras: poderacipon tf-idf
- Métricas de Evaluación en recuperación de información.
  - Conjunto de recuperación: Relevantes vs Recuperados
  - Precision, y Recall
  - F-score
  - Recuperación ranqueada: P@k, R@K
  - Average Precision
  - MAP: Mean average precision
  - Niveles de relevancia no binarios: DCG, NDCG
- Introducción a modelos de lenguaje probabilísticos

    - Modelos de n-gramas: unigramas, brigramas
    - Perplejidad
    - Problema de generalización (división por cero)
    - Laplace Smoothing
- Clasificación de texto usando aprendizaje de máquina
    - Aplicación de aprendizaje de máquina en información textual
    - Clasificadores generativos vs discriminativos
    - Usando representaciones vectoriales como espacio de características.
    - Creando representaciones reducidas de características basadas en lexicones (diseñadas a mano).
    - Repaso de métricas de evaluación de clasificadores binarios y multinomiales.
- Semántica vectorial (incrustaciones)
    - Recapitulando las representaciones vectoriales vistas: Vectores Dispersos
    - Vectores Densos: Word2Vec (SkipGram, CBOW)
    - Repaso de red neuronal (solo hasta feedforward FNN)
    - Entrenamiento y construcción de vectores densos
    - Concepto de Auto-supervisión y muestreo negativo
    - Propiedades semanticas de las incrustaciones (visualización)
- Modelos de lenguaje neuronales
    - FNN como clasificadores de texto: usando vectores densos como espacio de características.
    - FNN Redes neuronales como modelo de lenguaje.
    - Proceso de entrenamiento de redes neuronales como modelos de lenguaje.
    - Problemas de las redes neuronales convencionales.
- Modelos secuenciales
    - Redes neuronales recurrentes
    - LSTMs y GRUS
    - Modelos Seq2Seq
- Arquitecturas enconder/decoder
    - Tarea de traducción
    - Problemas de los modelos secuenciales
- Mecanismos de atención
    - Introducción a mecanismos de atención
    - Arquitectura transformer
- Incrustaciones contextuales
    - BERT
    - Variaciones BERT relevantes.
- Grandes Modelos de Lenguaje Preentrenados.

Evaluación y aspectos académicos

- Proyecto: 30% (Primera Entrega 10% Octubre 6 - Segunda Entrega 20% Diciembre 4)
- Talleres: 30% (prácticos en clase y casa)
- Examen 1 (28 septiembre): 20%
- Examen 2 (30 noviembre): 20%

Política de aproximación de notas finales

- Para aprobar el curso es indispensable lograr una nota sin aproximar de 3.0 o superior.
- No se hace aproximación de notas finales

Reclamos

- Si se trata de una prueba escrita, el estudiante deberá dirigir el reclamo por escrito, dentro de los ocho (8) días hábiles siguientes al que conoció la calificación en cuestión. El profesor cuenta con diez (10) días hábiles para responderle. Si el estudiante considera que la decisión no corresponde a los criterios de evaluación, podrá solicitar la designación de un segundo calificador ante el Consejo de Facultad, dentro de los ocho (8) días hábiles al conocimiento de la decisión (Art. 62 y 63 del RGEPr).
- En caso de reclamo por una calificación obtenida en una prueba oral, el estudiante podrá exponer la razón de su desacuerdo a los profesores evaluadores en el mismo momento en que tiene conocimiento de la nota. Si el grupo evaluador mantiene la calificación, la realización de un nuevo examen quedará a discreción del Consejo de Facultad al que pertenece la materia, previa solicitud escrita del estudiante (Art. 64 del RGEPr).

Generalidades

- Clases: 3 horas semanales, en dos sesiones de asistencia obligatoria. El estudiante que no asista al menos al 80% de las clases y sesiones de trabajo supervisado no podrá aprobar el curso, de acuerdo con el artículo 42 y 43 del RGRPr.
- Para que una ausencia sea justificada deberá presentarse el soporte valido correspondiente dentro de los (8) días hábiles siguientes, de acuerdo con el artículo 45 del RGRPr.
- Solo serán excusas válidas las estipuladas en el artículo 45 del RGRPr.
- El curso tiene como canales oficiales de comunicación:
    - Correo electrónico Uniandes
    - Bloque neon (https://bloqueneon.uniandes.edu.co)

Bibliografía

- Speech and Language Processing (3rd ed. draft) Dan Jurafsky and James H. Martin download via (https://web.stanford.edu/~jurafsky/slp3/)
- Deep Learning in Natural Language Processing by Li Deng, Yang Liu
- Neural Network Methods in Natural Language Processing. Yoav Goldberg , Graeme Hirst.
- Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. Delip Rao y Brian McMahan.
- Pattern Recognition and Machine Learning (Information Science and Statistics). Christopher M. Bishop.
- Machine Learning: A Probabilistic Perspective. Kevin P. Murphy.

Lecturas Adicionales

- Peter F Brown, et al.: Class-Based n-gram Models of Natural Language, 1992.
- Tomas Mikolov, et al.: Efficient Estimation of Word Representations in Vector Space, 2013.
- Tomas Mikolov, et al.: Distributed Representations of Words and Phrases and their Compositionality, NIPS 2013.
- Quoc V. Le and Tomas Mikolov: Distributed Representations of Sentences and Documents, 2014.
- Jeffrey Pennington, et al.: GloVe: Global Vectors for Word Representation, 2014.
- Ryan Kiros, et al.: Skip-Thought Vectors, 2015.
- Piotr Bojanowski, et al.: Enriching Word Vectors with Subword Information, 2017.
- Thomas Hofmann: Probabilistic Latent Semantic Indexing, SIGIR 1999.

- David Blei, Andrew Y. Ng, and Michael I. Jordan: Latent Dirichlet Allocation, J. Machine Learning Research, 2003.
- Yoon Kim: Convolutional Neural Networks for Sentence Classification, 2014.
- Christopher Olah: Understanding LSTM Networks, 2015.
- Matthew E. Peters, et al.: Deep contextualized word representations, 2018.