Ciencias de la Computación
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

PROGRAMA DE CURSO
PROCESAMIENTO DE LENGUAJE NATURAL

A. Antecedentes generales del curso:

Departamento	Ciencias de la Computación					
Nombre del curso	Procesamiento de Lenguaje Natural					
Nombre del curso en inglés	*Natural Language Processing*					
Código	CC6205		Créditos	6		
Horas semanales	Docencia	3,0	Auxiliares	0	Trabajo personal	7,0
Carácter del curso	Obligatorio	0		Electivo	**X**	
Requisitos	CC3001 Algoritmos y Estructuras de Datos ó MA3403 Probabilidades y Estadística.					

B. Propósito del curso:

El propósito del curso Procesamiento de Lenguaje Natural es introducir a los estudiantes a la disciplina del procesamiento de lenguaje natural (PLN). El estudiante utilizará este método para resolver tareas (task), con soluciones delimitadas respecto a su pertinencia y tamaño. Esta disciplina estudia el diseño de métodos y algoritmos que reciben como entrada y/o producen como salida datos en forma de lenguaje natural (e.g., texto, voz). El curso se centra en el procesamiento de texto aunque se mencionan aplicaciones en procesamiento de voz.

PLN abarca varias tareas (*tasks*) como por ejemplo: la traducción automática de documentos, el análisis de sentimientos, la detección de entidades y el parsing de árboles sintácticos. Cada tarea (*task*) se resuelve y se evalúa mediante técnicas y métricas que son propias a esta. El grueso de estas técnicas involucra el uso de algoritmos, métodos estadísticos y redes neuronales artificiales.

Durante el desarrollo del curso el estudiante se verá expuesto a las tareas más relevantes en PLN y reconocerá el funcionamiento detrás de las técnicas más efectivas para solucionar estas tareas como la forma de evaluar cuantitativamente la calidad de una solución. Es importante mencionar que PLN está fuertemente relacionado a la lingüística computacional. Por lo tanto, varios métodos de PLN serán motivados introduciendo conceptos lingüísticos.

El contexto del desarrollo de habilidades de aplicación metodológica será a través del planteamiento de tareas en PLN, con soluciones bien delimitadas respecto a su alcance y tamaño.

Desde el punto de vista teórico, el curso busca que los estudiantes sean capaces de leer artículos científicos en inglés con avances recientes en el área. Desde el punto de vista práctico, se apunta a que los estudiantes sean capaces de implementar soluciones a tareas de PLN utilizando la programación.

En resumen, se espera que los estudiantes desarrollen una metodología de trabajo que los lleve resolver problemas en PLN en base al razonamiento algorítmico, lingüístico y estadístico.

Las competencias específicas (CE) y genéricas (CG) a las que tributa el curso son:

CE1: Analizar problemas computacionales, construir modelos, expresándolos en representaciones y lenguajes formales adecuados.

CE4: Extraer información relevante, utilizando el proceso de descubrimiento de conocimiento de datos.

CG1: Comunicación académica y profesional

> Comunicar en español de forma estratégica, clara y eficaz, tanto en modalidad oral como escrita, puntos de vista, propuestas de proyectos y resultados de investigación fundamentados, en situaciones de comunicación compleja, en ambientes sociales, académicos y profesionales.

CG2: Comunicación en inglés

> Leer y escuchar de manera comprensiva en inglés variados tipos de textos e informaciones sobre temas concretos o abstractos, comunicando experiencias y opiniones, adecuándose a diferentes contextos de acuerdo a las características de la audiencia.

CG3: Compromiso ético

> Actuar de manera responsable y honesta, dando cuenta en forma crítica de sus propias acciones y sus consecuencias, en el marco del respeto hacia la dignidad de las personas y el cuidado del medio social, cultural y natural.

CG4: Trabajo en equipo

> Ejecutar con su equipo, de forma estratégica, diversas actividades formativas propuestas, considerando la autogestión de sí mismo y la relación con el otro, asumiendo diversos roles: de líder, colaborador u otros, según requerimientos y objetivos, sin discriminar por género u otra razón.

C. Resultados de aprendizaje:

Competencias Específicas	Resultados de aprendizaje
CE1	RA1: Reconoce, analíticamente, los componentes principales de una tarea (tasks) en PLN, deduciendo sus datos de entrada y salida, con el fin de descomponer, de forma irreductible, dicha tarea.
CE1	RA2: Plantea y explica un modelo de solución, delimitado respecto a su alcance y tamaño, para una tarea (task), mediante la identificación de componentes principales con el fin de procesar la entrada y generar la salida deseada, usando técnicas y algoritmos específicos de PNL.
CE4	RA3: Implementa y ejecuta, a nivel básico, un programa computacional, según la tarea de PLN a resolver y el modelo propuesto, con el fin de obtener una solución ejecutable de dicho modelo.
CE4	RA4: Evalúa la solución implementada, delimitada según alcance y tamaño, usando métricas de evaluación con el fin de validar y/o rectificar el modelo propuesto para la tarea.
Competencias Genéricas	**Resultados de aprendizaje**
CG1, CG4	RA6: Presenta de manera oral y escrita propuestas de solución a problemas en PLN, a fin de explicar, de manera sintética y precisa, las soluciones propuestas y su pertinencia.
CG2	RA7: Lee en inglés, de manera analítica y comprensiva, artículos científicos del estado del arte en PLN, a fin de relacionar dicha información y generar nuevos conocimientos atingentes y aplicables a temas de procesamiento de Lenguaje Natural.
CG3, CG4	RA8: Realiza, con su equipo, las actividades comprometidas, de manera responsable y honesta, en los plazos comprometidos, citando fuentes y referencias de donde se extrae la información, a fin de elaborar propuestas propias sin incurrir en plagio.

D. Unidades temáticas:

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
1	RA1-RA2-RA3-RA4-RA5-RA6-RA7	Fundamentos	4 semanas
Contenidos		**Indicador de logro**	
1. Modelo vectorial de texto. 2. Preprocesamiento: stemming y borrado de stopwords. 3. Naive Bayes para clasificación de documentos. 3.1. Aplicación: Análisis de sentimientos. 4. Modelos lineales para clasificación de documentos. 4.1. N-gramas de palabras y caracteres. 4.2. Funciones de pérdida. 4.3. Entrenamiento basado en gradiente. 5. Redes neuronales 5.1. Grafo Computational. 5.2. Backpropagation. 5.3. Derivación Automática. 6. Vectores de palabra (word embeddings). 6.1. Matrices palabra-contexto. 6.2. Modelo Skip-gram. 6.3. Modelo Continuos Bag of Words. 6.4. Negative Sampling. 6.5. Glove. 6.6. FastText. 7. Modelos de Lenguaje (language models) 7.1. Procesos de Markov. 7.2. Modelo de unigramas, bigramas y trigramas. 7.3. Evaluación de modelos de lenguaje: perplejidad. 7.4. Interpolación de modelos de lenguaje 7.5. Katz Back-Off Models. 7.6. Modelos de lenguaje neuronales.		El estudiante: 1. Reconoce de forma analítica, una tarea (*task*) de clasificación de texto. 2. Extrae características a partir de texto, usando modelos vectoriales. 3. Resuelve tareas (*tasks*) de clasificación de texto usando Naive Bayes, modelos lineales y redes neuronales. 4. Compara diversos modelos, considerando sus ventajas y desventajas para la clasificación de texto. 5. Implementa modelos de clasificación de texto usando *Scikitlearn*, NLTK y Pytorch. 6. Extrae vectores de palabra, usando matrices palabra-contexto y redes neuronales a partir de un corpus de documentos. 7. Implementa y evalúa vectores de palabra, usando la librería Gensim. 8. Explica la tarea (*task*) de modelamiento de lenguaje, considerando su definición formal y sus propiedades. 9. Determina el funcionamiento de modelos de lenguaje basados en n-gramas y redes neuronales. 10. Lee de manera comprensiva diversos textos y artículos en inglés sobre procesamiento de lenguaje natural, determinando sus ideas principales. 11. Cumple obligaciones y acuerdos, respetando los compromisos adquiridos en sus actividades académicas.	

	12. Planifica y presenta sus trabajos, basándose en sus capacidades, sin incurrir en plagio, copia, suplantación de identidad. 13. Respeta las ideas y opiniones de otros para definir acuerdos comunes, compartiendo ideas para dar cumplimiento a la meta. 14. Compone su texto, considerando las diferentes audiencias posibles (especializada, no especializada, pares, académicos, entre otros) para una adecuada recepción del mensaje y cumplimiento del propósito comunicativo.
Bibliografía de la unidad	[1] Capítulo 4. [2] Capítulos 1-11. [3] Capítulos 2 y 4.

Número	RA al que tributa	Nombre de la unidad	Duración en Semanas
2	RA1-RA2-RA3-RA4-RA7	Etiquetado de Secuencias	2 semanas
Contenidos		**Indicador de logro**	
1. Problemas de Etiquetado. 1.1. POS tagging. 1.2. Reconocimiento de entidades nombradas. 2. Cadenas de Markov Ocultas. 2.1 Algoritmo de Viterbi. 3. Etiquetado con modelo lineales. 3.1. Conditional random fields.		El estudiante: 1. Reconoce de forma analítica, una tarea (*task*) de etiquetado de secuencias 2. Identifica aplicaciones del etiquetado de secuencias a problemas de PLN. 3. Explica el funcionamiento de las cadenas de Markov ocultas y de modelos lineales para el problema de etiquetado de secuencias, identificando las ventajas y desventajas de estos métodos. 4. Resuelve problemas de etiquetado de secuencia usando las librerías computacionales NLTK y SpaCy. 5. Cumple obligaciones y acuerdos, respetando los compromisos adquiridos en sus actividades académicas. 6. Planifica y presenta sus trabajos, basándose en sus capacidades, sin incurrir en plagio, copia, suplantación de identidad.	

	7. Plantea a su equipo, de manera clara, precisa y constructiva, su posición acerca de un tema para cumplir de forma efectiva la tarea emprendida. 8. Relaciona, jerarquiza e integra en sus escritos información proveniente de múltiples fuentes.
Bibliografía de la unidad	[1] Capítulo 5 y 6. [3] Capítulo 3.

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
3	RA1-RA2-RA3-RA4-RA5-RA6-RA7	Arquitecturas Especializadas de Redes Neuronales	5 semanas

Contenidos	Indicador de logro
1. Redes Neuronales Convolucionales para la detección de n-gramas. 1.1. Operación convolución. 1.2. Operación de pooling. 2. Redes Neuronales Recurrentes. 2.1. Aplicaciones: clasificación, etiquetado y modelos de lenguaje. 2.2. Redes recurrentes bidireccionales. 2.3. Arquitecturas recurrentes con compuertas: LSTM y GRU. 3. Modelos Secuencia-Secuencia. 3.1. Arquitectura codificador-decodificador. 3.2. Capas de atención. 3.3. Decodificación aproximada usando Beam search. 3.4. Aplicaciones: Generación de resúmenes, traducción automática. 3.5. Transformer. 4. Transferencia de aprendizaje usando modelos de lenguaje neuronales. 4.1. ELMO. 4.2. BERT.	El estudiante: 1. Reconoce tareas (*tasks*) de PLN donde es conveniente usar redes neuronales convolucionales. 2. Explica el funcionamiento de las redes convolucionales para problemas de clasificación de texto. 3. Implementa redes neuronales convolucionales para clasificación de texto, usando Pytorch. 4. Reconoce dominios de aplicación en el PLN, adecuados para el uso de redes neuronales recurrentes. 5. Explica el funcionamiento de las redes neuronales recurrentes. 6. Implementa redes neuronales recurrentes para clasificación de texto y etiquetado de secuencias usando Pytorch. 7. Explica el funcionamiento de modelos secuencia-secuencia usando redes recurrentes y redes de atención para codificar la entrada y decodificar la salida. 8. Reconoce problemas de PLN adecuados para abordar usando arquitecturas secuencia-secuencia. 9. Implementa arquitecturas secuencia-secuencia usando AllenNLP. 10. Explica el concepto de transferencia de aprendizaje y el funcionamiento de arquitecturas modernas en ese ámbito como ELMO y BERT.

	11. Cumple obligaciones y acuerdos, respetando los compromisos adquiridos en sus actividades académicas. 12. Plantea a su equipo, de manera clara, precisa y constructiva, su posición acerca de un tema para cumplir de forma efectiva la tarea emprendida. 13. Planifica y presenta sus trabajos, basándose en sus capacidades, sin incurrir en plagio, copia, suplantación de identidad. 14. Relaciona, jerarquiza e integra en sus escritos información proveniente de múltiples fuentes. 15. Utiliza de manera pertinente los recursos verbales y no verbales para guiar a la audiencia en la interpretación y en los aspectos clave de su propuesta. 16. Plantea soluciones e ideas de forma clara, precisa y coherente, a través de una argumentación oral consistente.
Bibliografía de la unidad	[2] Capítulos 13-17.

Número	RA al que tributa	Nombre de la unidad	Duración en semanas
4	RA1-RA2-RA3-RA4-RA5-RA6-RA7	Parsing y Gramáticas	4 semanas

Contenidos	Indicador de logro
1. Introducción al problema de parsing: extracción de estructura sintáctica. 2. Árboles de Parsing y Treebanks. 3. Gramáticas libre de contexto. 4. Una gramática simplificada del inglés. 5. Gramáticas libre de contexto probabilísticas.	El estudiante: 1. Reconoce problemas de análisis sintáctico en PLN 2. Describe la estructura de un árbol sintáctico lingüístico. 3. Explica una gramática libre de contexto probabilística. 4. Extrae estructuras sintácticas a partir de un TreeBank, usando algoritmos de PLN específicos para la tarea (task).

6. Algoritmo CYK para parsing. 7. Gramáticas libre de contexto lexicalizadas. 8. Parsing de dependencias. 9. Parsing usando redes neuronales.	5. Reconoce las limitaciones de las gramáticas libre de contexto probabilísticas, en la tarea (*task*) de Parsing. 6. Analiza el funcionamiento de gramáticas libre de contexto lexicalizadas. 7. Identifica la tarea (*task*)de parsing de dependencias. 8. Explica el funcionamiento de arquitecturas basadas en redes neuronales para parsing. 9. Implementa modelos de parsing, usando AllenNLP. 10. Cumple obligaciones y acuerdos, respetando los compromisos adquiridos en sus actividades académicas. 11. Plantea a su equipo, de manera clara, precisa y constructiva, su posición acerca de un tema para cumplir de forma efectiva la tarea emprendida. 12. Planifica y presenta sus trabajos, basándose en sus capacidades, sin incurrir en plagio, copia, suplantación de identidad. 13. Utiliza de manera pertinente los recursos verbales y no verbales para guiar a la audiencia en la interpretación y en los aspectos clave de su propuesta. 14. Plantea soluciones e ideas de forma clara, precisa y coherente, a través de una argumentación oral consistente.
Bibliografía de la unidad	[2] Capítulos 12-14. [3] Capítulo 3.

E. Estrategias de enseñanzas:

La metodología de enseñanza y aprendizaje fomenta la participación del estudiante en el aula, las clases son principalmente:

- Clase expositiva, en donde el estudiante identifica los problemas fundamentales en PLN así como modelos y técnicas para abordarlos.
- Resolución de problemas. En cada unidad el estudiante es expuesto a librerías de programación que permiten implementar modelos de solución a problemas en PLN.

A lo anterior se le suman dos tareas individuales que deben ser desarrolladas con el computador y enviadas a través de U-cursos, además de un proyecto grupal donde se desarrollará una solución a un problema de PLN.

F. Estrategias de evaluación:

El curso tiene las siguientes instancias de evaluación:

Práctica:

- 2 tareas individuales de programación:
  - Tarea 1: clasificación de texto.
  - Tarea 2: etiquetado de secuencias.

  En esta actividad se evalúan los resultados de aprendizaje RA3 y RA4.

- 1 proyecto grupal:
  - En el proyecto cada grupo escogerá una tarea (task) en PLN de este sitio: http://nlpprogress.com/, y deberá desarrollarlo en dos hitos. Los problemas deben ser distintos a los estudiados en las tareas.
    - Hito 1: presentar al curso la *task*, las soluciones del estado del arte, las métricas de evaluación y los datasets existentes. Se evalúa el resultado de aprendizaje RA1 y RA6.
    - Hito 2: presentar al curso una implementación computacional propia de una solución al problema. Se evalúan los resultados de aprendizaje RA1, RA2, RA3, RA4, RA5, RA6 y RA7.

Teórica

- 5 ejercicios cortos de 15 minutos de duración que serán tomados al principio de la clase. Se eliminará el ejercicio con peor nota.

La nota final se calcula así: 30% para la nota tareas (nota promedio de las dos tareas), 30% para la nota proyecto (nota promedio del hito 1 y el hito 2) y 40% para la nota de ejercicios (nota promedio de los 4 mejores ejercicios). Para aprobar el curso se debe tener una nota final igual o superior a 4.0.

G. Recursos bibliográficos:

Bibliografía obligatoria:

[1] Dan Jurasfky and James H, Martin. Speech and Language Processing, (2nd Edition). Pearson, 2014.

[2] Joav Goldberg. Neural Network Methods for Natural Language Processing, Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2017.

[3] Christopher Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT press, 1999.

Bibliografía Complementaria:

- Apuntes de clases de Michael Collins: http://www.cs.columbia.edu/~mcollins/
- Borrador de la Tercera Edición de [1]: https://web.stanford.edu/~jurafsky/slp3/
- Notas sobre PLN de Jacob Eisenstein: https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

H. Datos generales sobre elaboración y vigencia del programa de curso:

Vigencia desde:	Primavera 2019
Elaborado por:	Felipe Bravo
Validado por:	Jorge Pérez –Sergio Ochoa
Revisado por:	Área de Gestión Curricular