

Silabo de Asignatura - Semana Internacional

Procesamiento de Lenguaje Natural / Analisis Textual

Universidad del Pacifico - Facultad de Economia y Finanzas

1. Informacion General

Nombre de la asignatura:

Procesamiento de Lenguaje Natural / Analisis Textual

Nombre del docente:

Fernando Diaz H., PhD. - Universidad Tecnica Federico Santa Maria (UTFSM)

2. Introduccion

El curso de Procesamiento de Lenguaje Natural (PLN) / Analisis Textual ofrece una exploracion rigurosa y practica de las principales tecnicas de analisis cuantitativo de texto, con aplicaciones en ciencias sociales, economia y finanzas. Los estudiantes seran introducidos a dos paradigmas complementarios: el enfoque clasico basado en R/Quanteda, ideal para tareas donde la transparencia estadistica y la escalabilidad son prioritarias, y el enfoque basado en modelos de lenguaje de gran escala (LLM) en Python, que supera al enfoque tradicional en tareas que requieren comprension contextual profunda, como analisis de sentimiento financiero o similitud semantica.

Un objetivo central del curso es desarrollar en el estudiante la capacidad critica para elegir la herramienta adecuada segun el corpus, la tarea y el contexto de investigacion, comprendiendo que no existe un enfoque universalmente superior. El curso es eminentemente practico: cada unidad incluye codigo ejecutable en Jupyter Notebook o RMarkdown, con ejercicios aplicados a corpus reales del ambito economico y financiero.

3. Logro de Aprendizaje Final

Al termino del curso, el estudiante **evaluara y aplicara tecnicas de procesamiento de lenguaje natural** -tanto clasicas (R/Quanteda) como basadas en modelos transformer (Python/HuggingFace)- sobre corpus textuales reales, **justificando metodologicamente la eleccion de cada enfoque** segun las caracteristicas de la tarea y los datos.

4. Unidades de Aprendizaje

Unidad de aprendizaje 1: Limpieza y Transformacion de Textos

Logro de aprendizaje: Al finalizar la unidad, el estudiante construira un corpus estructurado y aplicara operaciones de normalizacion textual usando R/Quanteda, comprendiendo el impacto de cada decision de preprocesamiento sobre los analisis posteriores.

Contenidos:

- Corpus, tokens y metadatos en Quanteda
- Eliminacion de puntuacion, numeros y stopwords
- Stemming, lematizacion y normalizacion
- Paquetes clave: quanteda, stringr

Unidad de aprendizaje 2: Document Feature Matrix (DFM)

Logro de aprendizaje: Al finalizar la unidad, el estudiante construira y manipulara una DFM con ponderacion TF-IDF, comprendiendo su rol como representacion matricial base del PLN clasico y prerequisite conceptual para los enfoques basados en embeddings.

Contenidos:

- Definicion y construccion de la DFM
- Frecuencias, TF-IDF, trimming y agrupacion por docvars
- Exportacion de la DFM para uso en modelos externos

Unidad de aprendizaje 3: Estadística Descriptiva de Corpus

Logro de aprendizaje: Al finalizar la unidad, el estudiante calculara e interpretara metricas de diversidad lexica y producira visualizaciones descriptivas que caractericen el corpus de analisis.

Contenidos:

- Resúmenes de corpus y DFM: tokens, types, TTR
- Ley de Zipf y distribuciones de frecuencia
- Visualizaciones: wordclouds y graficos de frecuencia comparativos

Unidad de aprendizaje 4: Keyness y Dispersion Textual

Logro de aprendizaje: Al finalizar la unidad, el estudiante identificara terminos estadisticamente distintivos entre grupos de textos e interpretara su dispersion lexica, incorporando un LLM para contextualizar los hallazgos.

Contenidos:

- Test de chi-cuadrado y log-likelihood para keyness
- Visualizacion con `textplot_keyness()` y `textplot_xray()`
- Uso de LLM para narrar e interpretar los terminos clave detectados

Unidad de aprendizaje 5: Modelamiento de Topicos

Logro de aprendizaje: Al finalizar la unidad, el estudiante estimara modelos de topicos con LDA y BERTopic sobre el mismo corpus, comparando su coherencia tematica y usando un LLM para etiquetar los topicos automaticamente.

Contenidos:

- Supuestos del modelo LDA: mezcla de topicos y generacion de palabras
- Algoritmo LDA: Gibbs vs VEM; seleccion de k optimo
- Introduccion a BERTopic (Python): embeddings para agrupamiento de documentos
- LLM para asignacion de etiquetas interpretables a los topicos

Unidad de aprendizaje 6: Analisis de Sentimiento Financiero

Logro de aprendizaje: Al finalizar la unidad, el estudiante cargara, aplicara y evaluara modelos transformer especializados para analisis de sentimiento financiero, comparando su desempeno con un baseline lexico clasico.

Contenidos:

- Limitaciones de los diccionarios financieros (Loughran-McDonald)
- Modelos HuggingFace: FinBERT, RoBERTa-Financial, DistilRoBERTa, DeBERTa
- Evaluacion con F1 macro, precision, recall y matrices de confusion
- Dataset: Financial PhraseBank (Sentences_AllAgree)

Unidad de aprendizaje 7: Similitud y Distancias Semanticas

Logro de aprendizaje: Al finalizar la unidad, el estudiante calculara similitud entre documentos usando representaciones TF-IDF y embeddings densos, articulando la diferencia conceptual entre similitud lexica y semantica.

Contenidos:

- Calculo de similitud cosine, Jaccard y euclidiana sobre DFM (R)
- Embeddings con sentence-transformers: similitud semantica profunda (Python)
- Contraste TF-IDF vs embeddings sobre el mismo corpus
- Visualizaciones: clustering jerarquico y MDS

5. Estrategias Didacticas

Estrategia didactica	Seleccion
Exposicion dialogante: Explicacion y demostracion de contenidos con intervencion de los estudiantes mediante preguntas.	x
Resolucion de ejercicios y problemas: Los estudiantes resuelven ejercicios aplicando algoritmos sobre corpus reales en R y Python.	x
Aprendizaje activo: Participacion continua mediante ejecucion de codigo en Jupyter/Colab y discusion de resultados.	x
Aprendizaje basado en problemas: Cada unidad plantea un problema textual real que los estudiantes resuelven con las herramientas del curso.	x
Aprendizaje basado en proyectos: Proyecto integrador final donde el estudiante aplica multiples tecnicas a un corpus de su eleccion.	x
Debates y discusiones estructuradas: Discusion sobre cuando usar NLP clasico vs. LLM, con argumentacion metodologica.	x

6. Sistema de Evaluacion

Consideraciones: La asistencia es indispensable para la calificacion de las actividades en clase. El proyecto integrador puede desarrollarse sobre cualquier corpus textual de interes del estudiante.

Nombre evaluacion	%	Criterios	Comentarios
Ejercicios en clase	40	Aplicacion practica de herramientas (R y Python), participacion activa y resolucion guiada de problemas textuales.	Evaluacion continua durante el curso.
Proyecto integrador final	60	Aplicacion de al menos 3 tecnicas del curso a un corpus propio. Comparacion entre enfoque clasico y LLM con justificacion metodologica.	Presentacion individual o en pareja. Entregable: notebook documentado.

7. Referencias

Obligatorias

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Muller, S. & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3(30), 774. <https://doi.org/10.21105/joss.00774>

Grimmer, J., Roberts, M. E. & Stewart, B. M. (2022). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press.

Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. arXiv:1908.10063. <https://arxiv.org/abs/1908.10063>

Malo, P., Sinha, A., Korhonen, P., Wallenius, J. & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782-796.

Recomendadas

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP 2019. <https://www.sbert.net>

Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.

Manning, C. D. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Silge, J. & Robinson, D. (2017). Text Mining with R. O'Reilly. <https://www.tidytextmining.com>

8. Cronograma Referencial

Unidades de aprendizaje	Contenidos, actividades y recursos	Evaluaciones	Referencias
Día 1 - Fundamentos: Corpus y Representación Matricial			
U1: Limpieza y Transformación (R/Quanteda)	Contenidos: - Corpus, tokens y metadatos - Stopwords, stemming, lematización Actividades: - Exposición dialogante - Ejercicios en Quanteda con corpus real Recursos: - Script R / RMarkdown - Google Colab (R kernel)	Ejercicio en clase	Benoit et al. (2018). quanteda. JOSS.

U2: Document Feature Matrix (R)	Contenidos: - Construcción de la DFM - TF-IDF, trimming, docvars Actividades: - Construcción guiada de DFM - Comparación de parametrizaciones Recursos: - Script R - Presentación	Ejercicio en clase	Grimmer et al. (2022). Text as Data. Cap. 4-5.
Día 2 - Estadística Textual: Descripción y Comparación			
U3: Estadística Descriptiva (R)	Contenidos: - Tokens, types, TTR - Ley de Zipf, wordclouds Actividades: - Análisis descriptivo de corpus real - Discusión de métricas Recursos: - Script R - ggplot2	Ejercicio en clase	Manning & Schütze (1999). Foundations of Statistical NLP. Cap. 1-2.
U4: Keyness y Dispersion (Híbrido)	Contenidos: - Chi-cuadrado, log-likelihood - textplot_keyness(), textplot_xray() - LLM para narrar hallazgos Actividades: - Comparación de grupos de textos - Debate: ¿qué agrega el LLM al análisis estadístico? Recursos: - Script R - API LLM (opcional)	Ejercicio en clase	Gabrielatos (2018). Keyness analysis. Routledge.
Día 3 - Temáticas: del Modelo Generativo a los Embeddings			
U5: Topic Modeling (Híbrido)	Contenidos: - LDA: Gibbs/VEM, selección de k - BERTopic (Python) - LLM para etiquetar tópicos Actividades: - Estimación de LDA sobre corpus real - Comparación LDA vs BERTopic sobre mismo corpus - Etiquetado automático de tópicos con LLM Recursos: - R: topicmodels - Python: BERTopic - Jupyter Notebook	Ejercicio en clase	Blei et al. (2003). LDA. JMLR. / Grootendorst (2022). BERTopic. arXiv.
Día 4 - Sentimiento Financiero: LLM en Acción			
U6: Análisis de Sentimiento Financiero (Python/LLM)	Contenidos: - Limitaciones de diccionarios financieros (Loughran-McDonald) - Modelos HuggingFace: FinBERT, RoBERTa, DeBERTa - Evaluación con F1 macro y matrices de confusión Actividades:	Ejercicio en clase	Araci (2019). FinBERT. arXiv. / Malo et al. (2014). JASIST.

	- Carga y aplicacion de modelos HuggingFace - Evaluacion comparativa de 5 modelos - Discusion: cuando supera el LLM al enfoque lexico? Recursos: - Python - HuggingFace Transformers - Jupyter Notebook - Financial PhraseBank		
Dia 5 - Similitud Semantica y Cierre del Curso			
U7: Similitud Semantica (Hibrido)	Contenidos: - Cosine/Jaccard sobre DFM (R) - Embeddings con sentence-transformers (Python) - Clustering jerarquico y MDS Actividades: - Contraste TF-IDF vs embeddings sobre mismo corpus - Discusion: similitud lexica vs semantica - Presentacion de proyectos integradores Recursos: - R + Python - sentence-transformers - Jupyter Notebook	Proyecto integrador final	Reimers & Gurevych (2019). Sentence-BERT. EMNLP.