MASTER	Master en Data Science para Finanzas
ASIGNATURA	***Data Science para la Gestión de Información No Estructurada***
Nº de ECTS	**2**
Nº de horas docentes	**Parte II: 15h (2 ECTS, 10 sesiones)**
Nº de horas actividades académicas dirigidas	
Profesor responsable de la asignatura	**Francisco José Izquierdo Catalán**
Cuatrimestre	**2º Cuatrimestre**

1.- DESCRIPCIÓN GENERAL DE LA ASIGNATURA Y OBJETIVOS DE DOCENCIA:

El objetivo de esta asignatura es introducir a los alumnos en las necesidades que aparecen en el entorno financiero con la gestión de información en lenguaje natural, así como la introducción a las diferentes herramientas y técnicas de minería de textos. Se cubrirán de forma teórica y práctica aspectos desde la captura y procesamiento de diferentes fuentes, la clasificación de la información, extracción de entidades y finalmente la visualización y explotación de la misma. Para cada uno de los diferentes puntos se verá la forma de abordarlos mediante programación. Así mismo se realizará una instalación de una herramienta de Minería de Textos de Código Abierto.

Durante las horas docentes se cubrirán los siguientes puntos:

- Introducción a la Gestión de Información No Estructurada y Minería de Textos
- Pre-procesamiento de la información y extracción de entidades
- Modelos de almacenamiento y representación de textos
- Clasificación
- Clusterización y modelado de tópicos
- Análisis semántico. Análisis de sentimientos
- UIMA y herramientas de Minería de Textos de Código Abierto.
- Chatbots y asistentes virtuales
- Deep Learning en NLP

2.- FORMA DE EVALUACIÓN PREVISTA:

La evaluación de la asignatura se compondrá en un 40% de la entrega y evaluación de las actividades académicas dirigidas, y en un 60% del resultado del examen final de la asignatura. Para aprobar la asignatura **será imprescindible obtener al menos un 5 en el examen final**; las actividades académicas dirigidas no serán reevaluables. La asistencia a clase es obligatoria, admitiéndose hasta un 20% de ausencias sin justificación; será criterio del profesor admitir o no la justificación; una asistencia menor del 80% supondrá la pérdida del derecho a examen en convocatoria ordinaria. Durante el curso se propondrán 4 actividades académicas dirigidas. Estas actividades se propondrán al final de la clase semanal, y se deberán entregar al final de la segunda semana posterior, es decir si la actividad se propone un miércoles, el alumno contará con los 4 días de esa semana y dos semanas adicionales. **El plazo de entrega no es prorrogable**.

Actividades académicas dirigidas	40%
Prueba objetiva final	60%

PROGRAMA DETALLADO		
Nº de sesión	**Detalle del contenido docente: temas, casos prácticos, actividades académicas dirigidas que se verán en dicha sesión,…**	**Lecturas recomendadas o referencias bibliográficas relativas a los conceptos-temas desarrollados en la sesión**
1	Sesión 1: Introducción a la Minería de Textos • Historia • Objetivos • Campos de aplicación de la Minería de Textos • Aplicación de Minería de Textos en Economía y Finanzas	*1.* Ronen Feldman y James Sanger (2007). *The Text Mining Handbook, Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press *2.* Presentación y notas de clase
2	Sesión 2: Pre-procesamiento y extracción de entidades • Pre-procesamiento y normalización • Detección de idioma • Tokenización • Lematización • Análisis gramatical • Extracción de Entidades	1. Sarkar D. (2019). Text Analytics with Python. Apress, Berkeley, CA 2. Presentación y notas de clase
3	Sesión 3: Modelos de almacenamiento y representación de textos • Bag of Words • Bag of N-grams • El modelo TF-IDF • Word2Vec • GloVe • FastText	1. Sarkar D. (2019). Text Analytics with Python. Apress, Berkeley, CA 2. Presentación y notas de clase
4	Sesión 4: Clasificación • Medidas de frecuencias • Ley de Zipf • Concepto de Clasificación • Clasificación Bayesiana • Regresión Logística • Support Vector Machine • Random Forest	1. Gareth James, Daniela Witten, Trevor Hastie y Robert Tibshirani (2013). *An Introduction to Statistical Learning with applications in R.* Springer 2. Ted Kwartler (2017). Text Mining in Practice with R. Wiley 3. Ashish Kumar y Avinash Paul (2016). *Mastering Text Mining with R.* Packt Publishing Ltd

		4. Sarkar D. (2019). Text Analytics with Python. Apress, Berkeley, CA 5. Presentación y notas de clase
5	Sesión 5: Clustering y modelado de tópicos • Latent Dirichlet Allocation • Latent Semantic Analysis • Medidas de Similitud • K-means	*1.* Ronen Feldman y James Sanger (2007). *The Text Mining Handbook, Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press *2.* Ashish Kumar y Avinash Paul (2016). *Mastering Text Mining with R.* Packt Publishing Ltd *3.* Sarkar D. (2019). Text Analytics with Python. Apress, Berkeley, CA *4.* Presentación y notas de clase
6	Sesión 6: Análisis semántico y análisis de Sentimientos • Conceptos • Relaciones semánticas • Desambiguación • Representación del análisis semántico • Modelos de análisis de sentimiento no supervisados • Modelos supervisados	1. Ashish Kumar y Avinash Paul (2016). *Mastering Text Mining with R.* Packt Publishing Ltd. 2016 2. Sarkar D. (2019). Text Analytics with Python. Apress, Berkeley, CA *3.* Presentación y notas de clase
7	Sesión 7: UIMA y herramientas de código abierto • Arquitectura UIMA • Soluciones de código abierto • Instalación • Carga de documentación • Búsqueda • Otras funcionalidades	1. Página de documentación del Proyecto UIMA: https://uima.apache.org/documentation.html 2. Apache SolR Tutorial: http://lucene.apache.org/solr/guide/7_4/solr-tutorial.html 3. Rafał Kuć (2013). Apache SolR 4 Cookbook. Packt Publishing Ltd 4. Presentación y notas de clase
8	Sesión 8: Chatbots y asistentes virtuales • Concepto de asistente virtual • Intenciones, entidades y diálogos • Diálogos complejos	1. Chiara Martino (2019). Conversation Design Workflow: How to design your chatbot in 10(basic) steps: https://medium.com/@chiara.martino/conversation-design-workflow-how-to-design-your-chatbot-in-10-basic-steps-721652b056d

	• Integraciones	2. Presentación y notas de clase
9	Sesión 9: Deep learning en NLP • Word-embedding • Text Classification en DL	1. Sarkar D. (2019). Text Analytics with Python. Apress, Berkeley, CA 2. Presentación y notas de clase
10	Sesión 10: Revisión general anterior al examen. Resolución de dudas	

INFORMACION ADICIONAL	
Bibliografía básica	1. Bholat et al (2015). *Text Mining for Central Banks*. CCBS, Bank of England 2. Sarkar D. (2019). Text Analytics with Python. Apress, Berkeley, CA 3. Manning, Raghavan, and Schütze (2009). *An Introduction to Information Retrieval.* Cambridge University Press *4.* Murphy (2012). *Machine Learning: a Probabilistic Perspective.* MIT Press 5. Hércules Antonio do Prado y Edilson Ferneda (2008). *Emerging Technologies of Text Mining: Techniques and Applications.* Information Science Reference 6. Héctor Cuesta (2013). *Practical Data Analysis.* Packt Publishing Ltd 7. Michael W. Berry y Malu Castellanos (2008). *Survey of Text Mining: Clustering, Classification and Retrieval.* Springer
Bibliografía Complementaria	1. Scott R. Baker, Nicholas Bloom y Steven J. Davis. *Measuring Economic Policy Uncertainty* (Enero de 2013). Chicago Booth Research Paper No. 13-02. Disponible en SSRN: http://ssrn.com/abstract=2198490 or http://dx.doi.org/10.2139/ssrn.2198490 2. Paul C. Tetlock. *Giving Content to Investor Sentiment: The Role of Media in the Stock Market*. Journal of Finance, Forthcoming. Disponible en SSRN: http://ssrn.com/abstract=685145 o http://dx.doi.org/10.2139/ssrn.685145

	3. Tim Loughran y Bill McDonald. *When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks* (Marzo de 2010). Journal of Finance, Forthcoming. Disponible en SSRN: http://ssrn.com/abstract=1331573 4. Stephen Hansen,, Michael McMahon y Andrea Prat. (2014) *Transparency and deliberation within the FOMC: a computational linguistics approach.* CFM discussion paper series, CFM-DP2014-11. Centre For Macroeconomics, London, UK. 5. G. Hoberg, y G. M. Phillips. (), '*Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis*', 2010. *The Review of Financial Studies*, Vol. 23, No. 10, pages 3773-3811. 6. Scott Hendry y Alison Madeley. *Text Mining and the Information Content of Bank of CanadaCommunications* (Noviembre, 2010). Disponible en SSRN: http://ssrn.com/abstract=1722829 o http://dx.doi.org/10.2139/ssrn.1722829
Actividades Complementarias	
Localización del profesor	francisco.izquierdo@cunef.edu