

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

MASTER	Master en Data Science para Finanzas
ASIGNATURA	<i>Data Science para el análisis de mercado 3</i>
Nº de ECTS	2.1
Nº de horas docentes	21
Nº de horas actividades académicas dirigidas	21
Profesor responsable de la asignatura	Fracisco J. Izquierdo Catalán
Curso académico	2017 / 2018
Cuatrimestre	2º Cuatrimestre

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

### DATA SCIENCE PARA LA GESTIÓN DE INFORMACIÓN NO ESTRUCTURADA

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

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

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.

## 2.- FORMA DE EVALUACIÓN PREVISTA:

La evaluación de la asignatura se compondrá en un 40% de la entrega y evaluación de las prácticas de clase, y en un 60% del resultado del examen final de la asignatura. La asistencia y participación durante el curso se valorará positivamente en la evaluación de las prácticas de clase.

Participación y asistencia	10%
Actividades académicas dirigidas	40%
Prueba objetiva final	50%

## PROGRAMA DETALLADO

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

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: <ul style="list-style-type: none"><li>• Historia</li><li>• Objetivos</li><li>• Campos de aplicación de la Minería de Textos</li><li>• Aplicación de Minería de Textos en Economía y Finanzas</li></ul>	<ol style="list-style-type: none"><li>1. Ronen Feldman y James Sanger. <i>The Text Mining Handbook, Advanced Approaches in Analyzing Unstructured Data</i>. Cambridge University Press. 2007</li><li>2. Presentación y notas de clase</li></ol>
2	Sesión 2: Ingesta de datos y pre-procesamiento <ul style="list-style-type: none"><li>• Data sources y construcción del corpus</li></ul>	<ol style="list-style-type: none"><li>1. Ashish Kumar y Avinash Paul. <i>Mastering Text Mining with R</i>. Packt Publishing Ltd. 2016</li><li>2. Presentación y notas de clase</li></ol>

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

	<ul style="list-style-type: none"><li>• Pre-procesamiento y normalización</li><li>• Modelo espacio-vector</li><li>• Representación N-gram</li></ul>	
3	Sesión 3: Categorización y Tagging <ul style="list-style-type: none"><li>• Tokenización, POS tagging, análisis sintáctico</li><li>• Modelos de Markov</li></ul>	<ol style="list-style-type: none"><li>1. Ronen Feldman y James Sanger. <i>The Text Mining Handbook, Advanced Approaches in Analyzing Unstructured Data</i>. Cambridge University Press. 2007</li><li>2. Presentación y notas de clase</li></ol>
4	Sesión 4: Clasificación (I) <ul style="list-style-type: none"><li>• Concepto de Clasificación</li><li>• Aprendizaje Inductivo</li><li>• Árboles de Decisión</li></ul>	<ol style="list-style-type: none"><li>1. Gareth James, Daniela Witten, Trevor Hastie y Robert Tibshirani. <i>An Introduction to Statistical Learning with applications in R</i>. Springer, 2013</li><li>2. Ashish Kumar y Avinash Paul. <i>Mastering Text Mining with R</i>. Packt Publishing Ltd. 2016</li><li>3. Presentación y notas de clase</li></ol>

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

5	Sesión 5: Clasificación (II) <ul style="list-style-type: none"><li>• Clasificación Bayesiana</li><li>• K-nearest</li><li>• Vector Machine</li></ul>	<ol style="list-style-type: none"><li>4. Gareth James, Daniela Witten, Trevor Hastie y Robert Tibshirani. <i>An Introduction to Statistical Learning with applications in R</i>. Springer, 2013</li><li>5. Ashish Kumar y Avinash Paul. <i>Mastering Text Mining with R</i>. Packt Publishing Ltd. 2016</li><li>6. Presentación y notas de clase</li></ol>
6	Sesión 6: Sesión práctica: Clasificación de Textos en R	<ol style="list-style-type: none"><li>1. Notas de clase y código de ejemplo</li></ol>
7	Sesión 7: Clustering <ul style="list-style-type: none"><li>• Latent Dirichlet Allocation</li><li>• Latent Semantic Analysis</li><li>• Medidas de Similitud</li><li>• K-means</li></ul>	<ol style="list-style-type: none"><li>1. Ronen Feldman y James Sanger. <i>The Text Mining Handbook, Advanced Approaches in Analyzing Unstructured Data</i>. Cambridge University Press. 2007</li><li>2. Presentación y notas de clase</li></ol>
8	Sesión 8: Sesión práctica: Clustering	<ol style="list-style-type: none"><li>1. Notas de clase y código de ejemplo</li></ol>

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

9	<p>Sesión 9: Extracción de Entidades</p> <ul style="list-style-type: none"><li>• Arquitectura UIMA</li><li>• Sinónimos</li><li>• Antónimos</li><li>• Negación</li><li>• Conceptos Similares</li></ul>	<ol style="list-style-type: none"><li>1. Ashish Kumar y Avinash Paul. <i>Mastering Text Mining with R</i>. Packt Publishing Ltd. 2016</li><li>2. Página de documentación del Proyecto UIMA: <a href="https://uima.apache.org/documentation.html">https://uima.apache.org/documentation.html</a></li><li>3. Presentación y notas de clase</li></ol>
10	<p>Sesión 10: Sesión práctica: Extracción de Entidades</p>	<ol style="list-style-type: none"><li>1. Notas de clase y código de ejemplo</li></ol>
11	<p>Sesión 11: Modelos Gráficos: Redes Bayesianas</p>	<ol style="list-style-type: none"><li>1. Kevin Murhy. <i>A Brief Introduction to Graphical Models and Bayesian Networks</i>. 2001. Disponible en: <a href="https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html">https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html</a></li><li>2. Presentación y notas de clase</li></ol>

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

12	<p>Sesión 12: Técnicas de Visualización</p> <ul style="list-style-type: none"><li>• Facetas</li><li>• Series Temporales</li><li>• Tendencias y Desviaciones</li></ul>	<ol style="list-style-type: none"><li>1. Ronen Feldman y James Sanger. <i>The Text Mining Handbook, Advanced Approaches in Analyzing Unstructured Data</i>. Cambridge University Press. 2007</li><li>2. Presentación y notas de clase</li></ol>
13	<p>Sesión 13: Sesión Práctica: Proyecto con extracción de datos de la web, clasificación supervisada y presentación de resultados</p>	<ol style="list-style-type: none"><li>1. Planteamiento de proyecto</li></ol>
14	<p>Sesión 14: Revisión general anterior al examen. Resolución de dudas</p>	

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

## INFORMACION ADICIONAL

### Bibliografía básica

1. Bholat et al. *Text Mining for Central Banks*. CCBS, Bank of England. 2015
2. Manning, Raghavan, and Schütze. *An Introduction to Information Retrieval*. Cambridge University Press. 2009
3. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press. 2012
4. Hércules Antonio do Prado y Edilson Ferneda. *Emerging Technologies of Text Mining: Techniques and Applications*. Information Science Reference. 2008
5. Héctor Cuesta. *Practical Data Analysis*. Packt Publishing Ltd. 2013
6. Michael W. Berry y Malu Castellanos. *Survey of Text Mining: Clustering,*



# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

	<i>Classification and Retrieval</i> . Springer. 2008
<b>Bibliografía Complementaria</b>	<ol style="list-style-type: none"><li>1. Scott R. Baker, Nicholas Bloom y Steven J. Davis. <i>Measuring Economic Policy Uncertainty</i> (Enero de 2013). Chicago Booth Research Paper No. 13-02. Disponible en SSRN: <a href="http://ssrn.com/abstract=2198490">http://ssrn.com/abstract=2198490</a> or <a href="http://dx.doi.org/10.2139/ssrn.2198490">http://dx.doi.org/10.2139/ssrn.2198490</a></li><li>2. Paul C. Tetlock. <i>Giving Content to Investor Sentiment: The Role of Media in the Stock Market</i>. Journal of Finance, Forthcoming. Disponible en SSRN: <a href="http://ssrn.com/abstract=685145">http://ssrn.com/abstract=685145</a> o <a href="http://dx.doi.org/10.2139/ssrn.685145">http://dx.doi.org/10.2139/ssrn.685145</a></li><li>3. Tim Loughran y Bill McDonald. <i>When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks</i> (Marzo de 2010). Journal of Finance, Forthcoming. Disponible en SSRN: <a href="http://ssrn.com/abstract=1331573">http://ssrn.com/abstract=1331573</a></li><li>4. Stephen Hansen,, Michael McMahon y Andrea Prat. (2014) <i>Transparency and deliberation within the FOMC: a computational linguistics</i></li></ol>

# CUNEF

COLEGIO UNIVERSITARIO DE  
ESTUDIOS FINANCIEROS

	<p><i>approach</i>. CFM discussion paper series, CFM-DP2014-11. Centre For Macroeconomics, London, UK.</p> <p>5. G. Hoberg, y G. M. Phillips. (), '<i>Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis</i>', 2010. <i>The Review of Financial Studies</i>, Vol. 23, No. 10, pages 3773-3811.</p> <p>6. Scott Hendry y Alison Madeley. <i>Text Mining and the Information Content of Bank of Canada Communications</i> (Noviembre, 2010). Disponible en SSRN: <a href="http://ssrn.com/abstract=1722829">http://ssrn.com/abstract=1722829</a> o <a href="http://dx.doi.org/10.2139/ssrn.1722829">http://dx.doi.org/10.2139/ssrn.1722829</a></p>
Actividades Complementarias	
Localización del profesor	fizquierdo@es.ibm.com