MASTER	Master en Data Science para Finanzas
ASIGNATURA	***Data Science para la Gestión de Información no Estructurada***
Nº de ECTS	**2.1**
Nº de horas docentes	**21**
Nº de horas actividades académicas dirigidas	**21**
Profesor responsable de la asignatura	**Pedro Santos**
Curso académico	**2016 / 2017**
Cuatrimestre	**2º Cuatrimestre**

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

DATA SCIENCE PARA LA GESTIÓN DE INFORMACIÓN NO ESTRUCTURADA

Web mining

Text mining

2.- FORMA DE EVALUACIÓN PREVISTA:

Participación y asistencia	21h
Actividades académicas dirigidas	21h
Prueba objetiva final	2h

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	1.1 Introduction to the course (outline, grading and case studies) 1.2. Introduction to text and web mining for economics and finance	1. Bholat et al. (2015). **Text Mining for Central Banks**. CCBS, Bank of England. 2. Lecture notes (to be provided)
2	2.1 Data sources and corpus pre-processing 2.2 N-gram representations 2.3 Case studies	1. Bholat et al. (2015). **Text Mining for Central Banks**. CCBS, Bank of England. 2. Lecture notes (to be provided)
3	3.1 Introduction to Boolean methods 3.2 Introduction to dictionary methods 3.3 Term weighting and the Zipf's Law 3.4 The vector space model 3.5 Similarity measures	1. Baker, Scott R. and Bloom, Nicholas and Davis, Steven J., **Measuring Economic Policy Uncertainty** (January 1, 2013). Chicago Booth Research Paper No. 13-02. Available at SSRN: http://ssrn.com/abstract=2198490 or http://dx.doi.org/10.2139/ssrn.2198490

		2. Tetlock, Paul C., **Giving Content to Investor Sentiment: The Role of Media in the Stock Market**. Journal of Finance, Forthcoming. Available at SSRN: http://ssrn.com/abstract=685145 or http://dx.doi.org/10.2139/ssrn.685145 3. Loughran, Tim and McDonald, Bill, **When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks** (March 4, 2010). Journal of Finance, Forthcoming. Available at SSRN: http://ssrn.com/abstract=1331573 4. Hansen, Stephen, McMahon, Michael and Prat, Andrea (2014) **Transparency and deliberation within the FOMC: a computational linguistics approach**. CFM discussion paper series, CFM-DP2014-11. Centre For Macroeconomics, London, UK. 5. Hoberg, G and Phillips, G M (2010), '**Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis**', *The Review of Financial Studies*, Vol. 23, No. 10, pages 3773-3811. 6. Lecture notes (to be provided)
4	4.1 Practical session: text mining with Python and R	1. Sample code (to be provided)
5	5.1 Latent Variable Models 5.1.1 Single membership models (K-means algorithm, multinomial misture and the EM algorithm) 5.1.2 Mixed-membership models (latent semantic indexing and singular value decomposition; probabilistic latent semantic indexing)	1. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990), **Indexing by latent semantic analysis**. J. Am. Soc. Inf. Sci., 41: 391–407. 2. Hendry, Scott and Madeley, Alison, **Text Mining and the Information Content of Bank of Canada Communications** (November 1, 2010). Available at SSRN: http://ssrn.com/abstract=1722829 or http://dx.doi.org/10.2139/ssrn.1722829 3. Acosta, J M (2014), '**FOMC Responses to Calls for Transparency: Evidence from the Minutes and Transcripts Using Latent Semantic Analysis**', Honours Thesis, Department of Economics, Stanford University. Available at http://economics.stanford.edu/content/honors-thesis-2014 4. Hofmann, T. (1999), **Probabilistic Latent Semantic** Indexing, SIGIR '99 Proceedings of the 22nd anual International ACM SIGIR conference on Research and development in information retrieval, Pages 50-57.

		5. Lecture notes (to be provided)
6	6.1 Bayesian inference for discrete data 6.2 The Black Swan paradox in text mining 6.3 Dirichlet prior	1. Lecture notes (to be provided)
7	5.1 Graphical models (Bayesian networks)	1. Koller, D., Friedman (2009), N., Probabilistic Graphical Models, MIT Press. 2. Lecture notes (to be provided)
8	8.1 Latent Dirichlet Allocation 8.1.1 Gibbs sampling 8.1.2 Formalizing interpretability 8.1.3 Chain convergence and selection; perplexity	1. Blei, D. M., Ng, A. Y. and Jordan, M. I., Latent Dirichlet Allocation (2003). Journal of Machine Learning Research 3: 993-1022. 2. Chang et al. (2009), **Reading Tea Leaves: How Humans Interpret Topic Models**, NIPS. 3. Lecture notes (to be provided)
9	9.1 Supervised learning 9.1.1 Rocchio classification 9.1.2 K-nearest neighbor classification 9.1.3 Discriminative classification 9.1.4 OLS and overfitting	1. Blei, D.M., McAuliffe, J.D. (2008), **Supervised topic models**, Advances in Neural Information Processing Systems 20 (NIPS 2007). 2. Lecture notes (to be provided)
10	10.1 Supervised learning (continuation) 10.1.5 Ridge regression 10.1.6 Lasso regression 10.1.7 Gamma/Lasso 10.1.8 Naive Bayes Classifier 10.1.9 Supervised LDA	1. Lecture notes (to be provided)
11	11.1 Variational inference 11.2 Variational Bayes and LDA	1. Lecture notes (to be provided)
12	12.1 Practical session: implementation of topic models and supervised learning algorithms.	1. Lecture notes (to be provided)
13/14	13.1 Morning session (1.5h): short Project aimed at parsing textual data from the web, creating a supervised LDA model for classification and measure its performace. 13.2 Afternoon sessions (3h): revision for exam.	1. Lecture notes (to be provided)
15	Exam.	

INFORMACION ADICIONAL	
Bibliografía básica	1. Manning, Raghavan, and Schütze (2009). **An Introduction to Information Retrieval**. Cambridge University Press. 2. Murphy (2012). **Machine Learning: a Probabilistic Perspective**. MIT Press. 3. Lecture notes (to be provided)
Bibliografía Complementaria	See references in Bholat et al. (2015). **Text Mining for Central Banks**. CCBS, Bank of England.
Actividades Complementarias	
Localización del profesor	**pedromiguel.pocas@cunef.edu**