BIG DATA SOURCES – WEB, SOCIAL MEDIA AND TEXT ANALYTICS

COURSE LEADER	**Piet Daas** (CBS)
TARGET GROUP	Official statisticians who already have knowledge about big data and its tools and who will start to work in practice on the use of web, social media and other natural language content for the production of statistics
ENTRY QUALIFICATIONS	▪ Sound command of English. Participants should be able to make short interventions and to actively participate in discussions ▪ Preferentially the participants should have followed the ESTP course "Hands-on immersion on big data tools" ▪ The participants should be computer literate and able to programme in R and/or Python
OBJECTIVE(S)	Main objectives of the course: ▪ Learn how to apply web scraping and other techniques to collect texts from the web; ▪ Learn how to analyse and mine texts in order to determine their content and sentiment; ▪ Learn how to deal with privacy and personal data
CONTENTS	▪ Text from the web and social media messages as a potentially rich big data source; ▪ Web scraping and other techniques to collect texts from the web; ▪ Text mining techniques applied to the content of web pages and social media; ▪ Sentiment and other emotion determination in texts; ▪ Extract and profile units to assess selectivity; ▪ Examples of the use of information derived from texts relevant for official statistics; ▪ Exercises and demonstrations.
EXPECTED OUTCOME	At the end of the course, participants will be able to: ▪ Apply web scraping techniques to extract texts from web pages and use API's to collect social media data. ▪ Mine texts to determine their content and sentiment. ▪ Study and profile units to assess selectivity.

	▪ Initiate big data case studies.
TRAINING METHODS	*Example (please insert what applies to your course):* ▪ Presentations and lectures ▪ Exchange of views and experiences on national practices ▪ Exercises and demonstrations
REQUIRED READING	▪ *I.H Witten (2005) Text mining. Link: http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf*
SUGGESTED READING	▪ *Ten Bosch, O. Windmeijer, D. (2014) On the use of internet robots for official statistics, Unece MSIS conference, Dublin. http://urlz.fr/5JH9* ▪ *Griffioen, R. de Haan, J., Willeborg, L. (2014) Collecting clothing data from the web. Paper for the Group of Experts on Consumer Price Indices meeting, Unece, Geneva. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/UNECE-ILO_2014_Griffioen_deHaan_Willenborg.pdf* ▪ *Russel, M.A. (2015) Mining the Social Web, 2nd edition. O'Reilly, Sebastopol, USA. In particular Chapter 1.* ▪ Abbott, D. (2013) Introduction to Text Mining. Presentation at the Virtual Data Intensive Summer School, July 10, 2013. http://www.vscse.org/summerschool/2013/Abbott.pdf ▪ *Daas, P.J.H., Puts, M.J.H. (2014) Social Media Sentiment and Consumer Confidence. European Central Bank Statistics Paper Series No. 5, Frankfurt, Germany. https://www.ecb.europa.eu/pub/pdf/scpsps/ecbsp5.en.pdf* ▪ *Shah, D.V., Capella, J.N., Neuman, W.R. (2015) Toward Computational Social Science: Big Data in Digital environments. Special issue of the Annals of the American Academy of Political and Social Science, vol. 659, May.*
REQUIRED PREPARATION	▪ Create a Twitter account (if you not already have one), see: https://twitter.com/signup, and take a mobile phone with you to the course. Both are needed for some of the exercises in the course. ▪ Search the web for a list of 'stop words' specific for your language. The following page provides a good start: https://en.wikipedia.org/wiki/Stop_words. You will need the list for some of the exercises in the course.
TRAINER(S)/ LECTURER(S)	Piet Daas (CBS-NL), Olav ten Bosch (CBS-NL), Marco Puts (CBS-NL), Antonino Virgillito (ISTAT-IT).

PRACTICAL INFORMATION				
WHEN	**DURATION**	**WHERE**	**ORGANISER**	**APPLICATION VIA NATIONAL CONTACT POINT**
1 – 4 October 2018	4 days	The Hague, Netherlands	Expertise France	Deadline**:** **6 August 2018**