Introduction to Textual Analysis and Natural Language Processing for Financial Market Applications

Workshop Outline
(Version: 30 January 2020)

1. Instructors, venue and timing

The workshop will be delivered by academics from the departments of finance and computing at Lancaster University. The programme convener is Professor Steven Young, who is based at Lancaster University Management School and whose research employs natural language processing tools to study the properties and consequences of textual information in financial markets.

Contact details for Professor Young are as follows:

E: s.young@Lancaster.ac.uk
T: ++44 01524 594242
W: https://www.lancaster.ac.uk/lums/people/steven-young

Support will also be provided by colleagues from the School of Computing and Communications at Lancaster University, including Dr. Paul Rayson who specializes in developing and applying computational linguistics methods (https://www.lancaster.ac.uk/scc/about-us/people/paul-rayson)

The programme will be delivered at The Work Foundation's premises in central London:

The Work Foundation
21 Palmer Street
London
SW1H 0AD

See here for map

Sessions will be delivered one evening per week over a five-week period. Each session will run for three hours and involve a mix of mini contextualizing lectures and practical applications (typically in R and python). Participants will be required to bring their laptop to each session. Pre-reading, datasets and sample code will be disseminated to participants at least one week prior to the start of the workshop. Additional resources will be available at http://ucrel.lancs.ac.uk/cfie/.

2. Workshop objectives and target audience

This workshop provides introductory and intermediate guidance on popular natural language processing (NLP) methods and their application to financial market data. The programme comprises five three-hour sessions designed to introduce participants to the key steps of textual analysis, from extraction and preprocessing through to common machine learning methods and their applications.

Content is targeted at participants interested in learning more about textual analysis method and the opportunities they afford, but who are yet to apply NLP techniques on any routine basis. Programming experience in python or R will be an advantage but is not a requirement. (Code and instructions for all example applications will be provided.) The primary objectives of the workshop are to:

- Provide an introduction to the fundaments of textual analysis and natural language processing as applied to financials market data;
- Distinguish between bag-of-words approaches to analysing text and semantic approaches that better reflect meaning and context;
- Provide an introduction to word sense disambiguation methods to support semantic analysis;
- Introduce supervised machine learning approaches for text classification;
- Review topic modelling methods and associated procedures supporting topic selection.

4. Registration and Attendance Fees

- Registration is FREE for nominated employees of Inquire sponsor firms.
- In the event that any places remain after Inquire sponsors have registered, non-sponsor participants may apply to register.
- The fee for non-sponsors* is £2000 + VAT as applicable, payable in advance.
- Registrations due by 20 March 2020
- Accepted registrations will be confirmed no later than 25 March 2020.
- Registrations may be transferred to a colleague in your organisation. Late cancellations (after 1 April 2020) or no-shows will be charged at £500 + VAT as applicable.

To ensure a positive learning experience, the overall number of registrations for the Workshop will be limited. Workshop places will be allocated based on the number of Inquire sponsor seats held. Standard (i.e., 2 seat) and Premium (i.e., 3 seat) members will receive registration priority until 28 February. Basic (i.e., 1-seat) members wishing to upgrade to Standard or Premium membership in order to secure workshop places should email the Company Secretary, Lesley Flowers (lesley.flowers@inquire.org.uk) before 28 February 2020, stating that they would like to upgrade their INQUIRE membership to Standard membership at a cost *of £1,800 + VAT (as applicable) or Premium membership at a cost of £3,600 + VAT (as applicable).*

Sponsors may nominate more participants Inquire seats held, but if they do so they should indicate their priority ranking of participants in the online registration form. Rankings will be used if it becomes necessary to ration the number of registrations. When registering please provide the contact details of the sponsor manager responsible for nominating participants.

*Firms wishing to join Inquire as sponsors should contact lesley.flowers@inquire.org.uk or complete the following form located: http://inquire.org.uk/membership/how-do-i-join

To register your interest for this training: Please click here

5. Learning outcomes

On completion of the workshop, participants will understand:

- How to construct a term document matrix and apply preprocessing methods aimed at reducing redundant dimensionality;
- Commonly applied bag-of-words proxies for constructs such as reading ease, sentiment, text similarity and thematic content, and the strengths and weaknesses associated with each proxy;

- The distinction between bag-of-words approaches to analysing text and semantic approaches that better reflect meaning and context;
- How and when to apply word sense disambiguation methods to support semantic analysis;
- Topic modelling using latent Dirichlet allocation (LDA) and alternative procedures for incorporating domain-specific expertise;
- Approaches to text classification using a suite of popular supervised machine learning classifiers and associated implementation choices.

6. Session outline

Session 1: Introduction to methods and corpus construction (Wednesday, 8 April 17.30-20.30)
The session will introduce the basic steps and approaches in textual analysis, review the process of text extraction, and discuss approaches to cleaning raw text for further analysis. Topics covered will include:

- Organising framework: Overview of approaches and methods
- Text processing pipeline
- Extraction: SEC Edgar filings, PDF annual reports, HTML earnings announcements, conference calls, social media
- Term-document matrix and text preprocessing: stopword removal, infrequent words and dispersion, sentence splitting, stemming and lemmatization

Session 2: Bag-of-words analyses (Wednesday, 15 April 17.30-20.30)
This session will provide an introduction to popular summary measures of text properties constructed using automated content analysis and simple NLP methods. Topics covered in the session will include:

- Dictionaries and lexicons
- Popular constructs and applications: Sentiment, uncertainty, readability, text similarity
- Resources and pitfalls
- Simple refinements: Term weighting, domain-specific lexicons, aspect-level conditioning

Session 3: Word sense disambiguation (Wednesday, 22 April 17.30-20.30)
This session will explore automated and manual methods for labelling and interpreting text designed to help improve word sense disambiguation and move beyond a bag-of-words approach. Topics covered in the session will include:

- Automatic labelling: Part of speech (POS) tagging, named entity recognition (NER), semantic tagging
- Resources: Stanford NLP; NLTK; SpaCy; CLAWS; USAS
- Manual labelling: Best practice
- Corpus methods: Concordance, collocation, keyness
- Corpus resources: AntConc; #LancsBox; Sketch Engine

Session 4: Topic modelling (Wednesday, 29 April 17.30-20.30)
This session will introduce participants to topic modelling methods for identifying key themes in financial disclosures. The session will address the following issues relating to topic identification and refinement:

- Unsupervised models: Overview, Latent Drichlet Allocation (LDA), correlated topic modelling, structural topic modelling, example applications
- Topic selection: Conceptual validity and common problems
- Semi-supervised models: Correlation explanation (COrEx), supervised LDA, labelled LDA, semantic tagging

Session 5: Machine learning classifiers (Wednesday, 6 May 17.30-20.30)
This session reviews common machine learning algorithms for classifying financial market outcomes such as M&A targets, fraudulent reporting, etc. Topics covered in the session will include:

- Unsupervised versus supervised approaches
- Common supervised classifiers: Naïve Bayes, random forest, support vector machines, neural networks and example applications
- Sample balancing: under-sampling and over-sampling
- Classification performance: accuracy, error rate, recall, precision, F1
- Training samples versus out-of-sample prediction
- Pitfalls and tripwires