NLP in Financial Economics

Fall 2018

Doctoral School in Economics and Finance (DSEF)
Université du Luxembourg

1 General information

By keeping this handout you can remember all this important information:

Class meetings TBD

Instructor: Diego García.

E-mail: `diego.garcia@colorado.edu` (or `financieru@gmail.com`).

Phone number: 1-919-810-0396.

Google folder: TBD.

Office hours: TBD.

2 About the course

This course will discuss some basic ideas around natural language processing (NLP) in research on financial markets. The course is meant for PhD students as an introductory course on textual analysis with particular emphasis on methods and applications in Finance and Accounting.The course will be using Perl, R, and Python

3 Course Material

Grading: homeworks/assignments 30%, final exam 50%, class participation 20%. We shall announce the due date for the homeworks/assignments as well as the date for the final exam as the course progresses. I will occassionally use the homeworks as a way to start a lecture. Note how class participation is an important part of the course grade.

There is no required textbook for this course. Fortunately, there are many online resources for textual analysis. I will try to point you to them as we move through the course. Classes will be a mix of cases, paper discussions, as well as student presentations. The course is meant to be as "hands-on" as possible, in that we will try to solve a few assignments that are related to influential papers. My goal is to at least get you familiar with techniques to deal with "unstructured data," as many empirical projects can be started these days with a little creativity and some sharp code.

4 Topics and outline

The following is a tentative outline of the course. I have tried to list a few papers per class meetings to give a sense of what we will do: there are high chances I will move things around over the semester.

I. Introduction

Outline of the course, pertinent introductions. Quick discussion regarding computer architectures, data requirements, expectations regarding computer hardware, etc.

Readings:

- NLTK chapter one.
- Gentzkow, Shapiro and Taddy, "Text as Data," Journal of Economic Perspectives, forthcoming.

II. Crawling data (I)

Introduction to tools for data analysis: crawling data, automated reading. One of the emphasis in the course will be learning how to get data off the Internet ("crawling"), and this first class we will spend some time discussing open-source databases (in particular EDGAR and newspapers).

Readings:

- Loughran and McDonald, 2016, Textual Analysis in Accounting and Finance: A Survey, Journal of Accounting Research, 54(4), 1187–1230.
- García and Norli, 2012, "Crawling EDGAR," Spanish Review of Financial Economics, 10(1), 1-10.
- Li "Textual Analysis of Corporate Disclosures: A Survey of the Literature." Journal of Accounting Literature 29 (2010a): 143–65.

III. Media and sentiment (I)

Sentiment analysis of newspaper articles.

Readings:

- Tetlock, "Given content to investor sentiment," 2007, Journal of Finance 62, 1139-1168.
- Tetlock, Saar-Tsechansky, and Macskassy, 2008, "More Than Words: Quantifying Language to Measure Firms Fundamentals," 2008, Journal of Finance 63, 1437-1467.
- García, 2013, "Sentiment during recessions," Journal of Finance, 68(3), 12671300.
- Dougal, Engelberg, García and Parsons, 2012, "Journalists and the stock market," Review of Financial Studies, 2012, 25(4), 639679.

IV. Media and sentiment (II)

Sentiment analysis of newspaper articles, tf-idf and other nuances.

Readings:

- García, "The kinks of financial journalism," 2017, working paper.
- Jegadeesh and Wu, 2014, "Word Power: A New Approach for Content Analysis," Journal of Financial Economics.

V. Crawling data (II)

More nuanced crawling exercises.

Readings:

- Linked-in paper (TBD).
- Crawling CVs exercise.

VI. EDGAR documents (I)

Working with EDGAR documents: sectioning, topic classification, scraping different cross-sectional data.

Key papers:

- Loughran and McDonald, 2011, "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks," Journal of Finance, vol. 66, February 2011, 35-65.
- García and Norli, 2012, "Geographic dispersion and stock returns," Journal of Financial Economics, 106(3), 547565.
- Hoberg and Phillips, "Text-Based Network Industries and Endogenous Product Differentiation." Journal of Political Economy (2015): forthcoming.

VII. EDGAR documents (II)

Topics: IPO studies.

Key papers:

- Hanley and Hoberg, 2010, "The Information Content of IPO Prospectuses." Review of Financial Studies 23 (2010): 2821–64.
- Loughran and McDonald "IPO First-Day Returns, Offer Price Revisions, Volatility, and Form S-1 Language." Journal of Financial Economics 109 (2013): 307–26.

VIII. Topic models (LDA, STM)

Topics: LDA models, STM implementation.

Key papers:

- Hansen, McMahon and Prat, "Transparency and deliberation within the FOMC: a computational linguistics approach," working paper, Columbia University.

- Bellstam, Bhagat and Cookson, “Innovation in Mature Firms: A Text-Based Analysis,” working paper, University of Colorado.

IX. Earning calls

Topics: textual analysis of earning calls and earnings press releases.

Key papers:

- Avis, Ge, Matsumoto and Zhang, “The Effect of Manager-Specific Optimism on the Tone of Earnings Conference Calls.” Review of Accounting Studies 20 (2015): 639–73.
- Avis, Piger and Sedor, “Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language.” Contemporary Accounting Research 29 (2012): 845–68.
- Lang and Stice-Lawrence “Textual Analysis and International Financial Reporting: Large Sample Evidence.” Journal of Accounting and Economics 60 (2015): 110–35.

X. Working with unstructured data

Topics: scraping data from OCR output.

Key papers: TBD.

X. Other topics (I)

Topics: small talk, Internet message boards, voice versus text, media during the crisis,.

Key papers:

- Mayew and Venkatachalam “The Power of Voice: Managerial Affective States and Future Firm Performance.” Journal of Finance 67 (2012): 1–43.
- Das and Chen “Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web.” Management Science 53 (2007): 1375–88.
- Antweiler and Frank “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.” Journal of Finance 59 (2004): 1259–94.
- Mamaysky and Glasserman, 2017, “Does Unusual News Forecast Market Stress?” working paper, Columbia University.

XII. Other topics (II)

Topics: measuring readability, strategy corporate press releases, media slant, newspapers and politics.

Key papers:

- Loughran and McDonald, 2014, “Measuring Readability in Financial Disclosures,” Journal of Finance, vol. 69, 2014, 1643-1671.
- Lehavy, Li and Merkley, “The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts.” The Accounting Review 86 (2011): 1087–115.

- Ahern and Sosyura "Who Writes the News? Corporate Press Releases During Merger Negotiations." Journal of Finance 69 (2014): 24191.
- Kim, Verdi and Yost, 2017 "The Feedback Effect of Disclosure Externalities," working paper, MIT.
- Gentzkow and Shapiro, 2010, "What Drives Media Slant? Evidence from U.S. Daily Newspapers," Econometrica, 78 (1), January 2010.
- Gentzko, Shapiro and Sinkinson, "The Effect of Newspaper Entry and Exit on Electoral Politics," American Economic Review. 101 (7). December 2011.
- Cohen et al. "Lazy prices."

Note: the main source of material for this course is the class notes that I provide. Reading through the original research papers is also a must.