Individual Assignment: Text Analytics

Dr Nikolaos Korfiatis
(n.korfiatis@warwick.ac.uk)

Associate Professor of Business Analytics
Warwick Business School, MSc in Business Analytics

Instructions

Please read the instructions carefully and provide your solutions at the instructed period.

1. Overview and Pedagogical Goal

The goal of this assignment is to familiarize you with the complete process of *extracting*, *refining* and *delivering* insights of particular financial value that are extracted from unstructured data of non-conventional size from company reports. This is an individual assignment where you are supposed to work alone in order to extract insights from company financial statements filed at the Electronic Data Gathering, Analysis, and Retrieval system used at the U.S. Securities and Exchange Commission (SEC). The assignment maps to level 7 qualification level and aims to establish your ability to handle *the development of in-depth and original solutions to a domain specific problem of a high business value.*

The tasks is structured in three (3) parts. The first part (Part A) covers your ability to construct and demonstrate the handling of text data. It aims to familiarize you with the principles of text mining, the bag-of-words model and the development of metrics that can be used to analyse structural elements of text, such as normalizing and cleaning textual corpora. The core of this assignment involves the translation of these insights to actionable features that can be used to predict an outcome variable of financial interest: the stock price value. Therefore the second and third parts (Part B and Part C) are concerned with the identification of features and in particular (a) polarity – whether the text under consideration is positive or negative, (b) sentiment – the extraction of affective states from the text and (c) the evaluation and extraction of important topics that are covered and elaborated in the quarterly and annual financial reports (10-Q, 10-K) and the predictability of these insights on a company, sector and market level (Part C).

The report should be written from the perspective of an analyst involving text mining methods in constructing a well written piece of work. This should be both academic as well as practical and consider possible application scenarios where text mining can be used (e.g., Risk analysis etc).

2. Marking Criteria and Weights

The marking criteria for all parts of the assignment are as follows:

- Part A: 30% - Completeness of the solution, efficiency of the code, interpretation of the results.
- Part B: 25% - Completeness of the solution, efficiency of the code, interpretation of the results.
- Part C: 25% - Completeness of the solution, efficiency of the code, interpretation of the results.

20% is reserved for the whole academic content in the report distributed equally among the motivation for the selection of the companies, the interpretation of the outcomes of this analysis and the convincing line of argument in providing the results.

3. Submission Instructions

The assignment solutions should be submitted as one-file **pdf document** containing both the narrative for each part as well as the code in the form of a compiled R markdown notebook in html. The student shall combine all files in a zip file with the following naming format:

student_number.zip

No other files are going to be considered. It is your responsibility to comply with the requirements of the submission, otherwise this will have repercussions for marking.

Part A: Construction of Corpus – Fetching 10-Q and 10-K forms from EDGAR

The S&P 500 stock market index, maintained by S&P Dow Jones Indices, comprises common stocks issued by companies of large capitalization and traded on the New York Stock Exchange (including the 30 companies that compose the Dow Jones Industrial Average). The index covers about 80 percent of the American equity market by capitalization. All companies are required by law to file quarterly and annual reports through the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) used at the U.S. Securities and Exchange Commission (SEC).

The publicly available page of EDGAR is provided at:

`https://www.sec.gov/edgar.shtml`

Each company files a report using an XML archetypal format, commonly known as Standard Generalized Markup Language (SGML) using a specialized interface which indexes them by filing period (quarter/year) and a unique identifier known as Central Index Key (CIK). The later is by definition of 10 digits in length with preceding zeros when needed. The list of current companies, their CIK codes and Stock Ticker Symbol can be obtained from the accompanying table in the appendix.

You are required to build a portfolio selection of minimum 30 companies. For companies in your portfolio (Selected only from the companies listed in the appendix) you need to download the 10-K forms for the period between 2010 and 2020 (where available).

You need to develop a normalization strategy that will remove the standard parts of the 10-K and 10-Q form and enable the text for further analysis such as stop-word removal. You need to provide outlines of TF-IDF weights for important keywords across the industry level using the MSCI Global Industry Classification Standard (GICS) as a control.

Part B: Sentiment association with Financial Indicators

Using polarity and sentiment you are asked to demonstrate how text derived features connect with the stock price. You can use different aspects of sentiment such as affection categorization as well as the use of context specific keywords. For achieving that you can use several different dictionaries such as the Loughran-McDonald, AFIN, NRC, Wordnet Affect etc. For all cases, a regression model should be used to evaluate the predictability of the stock price at the time of the report against the variables obtained from the sentiment analysis part using the related material discussed in the labs. Your goal is to see how the financial results of the previous year reflect the textual argumentation in the **management's reflection part of the 10-K (and not the other content of the 10-K text)**

Part C: Topic Modelling and Latent Dirichlet allocation

Using a topic model across the MSCI levels you are requested to provide an analysis of the topics that become dominant using both the unsupervised and supervised approach. Your analysis should focus on finding which topics become important across time as well as how the characteristics of particular industries are more likely to influence a dominance of different themes in the 10-K form (**management's reflection part of the 10-K and not the other content of the 10-K text)**.

Your topic solution should evaluate among others:

a. The number of topics (Kappa) that need to be created for the particular corpus. You can opt to use the coherence criterion for the Kappa selection
b. The semantic coherence of the word-topic associations that are going to be created.
c. The additive predictability that some topics add on estimating the stock price.

All solutions need to be properly described and articulated by providing the relevant fragments of code. Your report should reflect the effort and the steps you took into your analysis and the potential insights that can be generated by these textual features.

Appendix: List of Selected Companies from SP500

Symbol	Security	GICS Sector	GICS Sub Industry	CIK
ACN	Accenture plc	Information Technology	IT Consulting & Other Services	1467373
ADBE	Adobe Systems Inc	Information Technology	Application Software	796343
AMD	Advanced Micro Devices Inc	Information Technology	Semiconductors	2488
AKAM	Akamai Technologies Inc	Information Technology	Internet Services & Infrastructure	1086222
ADS	Alliance Data Systems	Information Technology	Data Processing & Outsourced Services	1101215
APH	Amphenol Corp	Information Technology	Electronic Components	820313
ADI	Analog Devices, Inc.	Information Technology	Semiconductors	6281
ANSS	ANSYS	Information Technology	Application Software	1013462
AAPL	Apple Inc.	Information Technology	Technology Hardware, Storage & Peripherals	320193
AMAT	Applied Materials Inc.	Information Technology	Semiconductor Equipment	6951
ANET	Arista Networks	Information Technology	Communications Equipment	1596532
ADSK	Autodesk Inc.	Information Technology	Application Software	769397
ADP	Automatic Data Processing	Information Technology	Internet Services & Infrastructure	8670
AVGO	Broadcom Inc.	Information Technology	Semiconductors	1730168
BR	Broadridge Financial Solutions	Information Technology	Data Processing & Outsourced Services	1383312
CDNS	Cadence Design Systems	Information Technology	Application Software	813672
CDW	CDW	Information Technology	Technology Distributors	1402057
CSCO	Cisco Systems	Information Technology	Communications Equipment	858877
CTXS	Citrix Systems	Information Technology	Application Software	877890
CTSH	Cognizant Technology Solutions	Information Technology	IT Consulting & Other Services	1058290
GLW	Corning Inc.	Information Technology	Electronic Components	24741
DXC	DXC Technology	Information Technology	IT Consulting & Other Services	1688568
FFIV	F5 Networks	Information Technology	Communications Equipment	1048695
FIS	Fidelity National Information Services	Information Technology	Data Processing & Outsourced Services	1136893
FISV	Fiserv Inc	Information Technology	Data Processing & Outsourced Services	798354
FLT	FleetCor Technologies Inc	Information Technology	Data Processing & Outsourced Services	1175454

FLIR	FLIR Systems	Information Technology	Electronic Equipment & Instruments	354908
FTNT	Fortinet	Information Technology	Systems Software	1262039
IT	Gartner Inc	Information Technology	IT Consulting & Other Services	749251
GPN	Global Payments Inc.	Information Technology	Data Processing & Outsourced Services	1123360
HPE	Hewlett Packard Enterprise	Information Technology	Technology Hardware, Storage & Peripherals	1645590
HPQ	HP Inc.	Information Technology	Technology Hardware, Storage & Peripherals	47217
INTC	Intel Corp.	Information Technology	Semiconductors	50863
IBM	International Business Machines	Information Technology	IT Consulting & Other Services	51143
INTU	Intuit Inc.	Information Technology	Application Software	896878
IPGP	IPG Photonics Corp.	Information Technology	Electronic Manufacturing Services	1111928
JKHY	Jack Henry & Associates	Information Technology	Data Processing & Outsourced Services	779152
JNPR	Juniper Networks	Information Technology	Communications Equipment	1043604
KEYS	Keysight Technologies	Information Technology	Electronic Equipment & Instruments	1601046
KLAC	KLA Corporation	Information Technology	Semiconductor Equipment	319201
LRCX	Lam Research	Information Technology	Semiconductor Equipment	707549
LDOS	Leidos Holdings	Information Technology	IT Consulting & Other Services	1336920
MA	Mastercard Inc.	Information Technology	Data Processing & Outsourced Services	1141391
MXIM	Maxim Integrated Products Inc	Information Technology	Semiconductors	743316
MCHP	Microchip Technology	Information Technology	Semiconductors	827054
MU	Micron Technology	Information Technology	Semiconductors	723125
MSFT	Microsoft Corp.	Information Technology	Systems Software	789019
MSI	Motorola Solutions Inc.	Information Technology	Communications Equipment	68505
NTAP	NetApp	Information Technology	Technology Hardware, Storage & Peripherals	1002047
NLOK	NortonLifeLock	Information Technology	Application Software	849399
NVDA	Nvidia Corporation	Information Technology	Semiconductors	1045810
ORCL	Oracle Corp.	Information Technology	Application Software	1341439
PAYX	Paychex Inc.	Information Technology	Data Processing & Outsourced Services	723531
PAYC	Paycom	Information Technology	Application Software	1590955
PYPL	PayPal	Information Technology	Data Processing & Outsourced Services	1633917

QRVO	Qorvo	Information Technology	Semiconductors	1604778
QCOM	QUALCOMM Inc.	Information Technology	Semiconductors	804328
CRM	Salesforce.com	Information Technology	Application Software	1108524
STX	Seagate Technology	Information Technology	Technology Hardware, Storage & Peripherals	1137789
NOW	ServiceNow	Information Technology	Systems Software	1373715
SWKS	Skyworks Solutions	Information Technology	Semiconductors	4127
SNPS	Synopsys Inc.	Information Technology	Application Software	883241
TEL	TE Connectivity Ltd.	Information Technology	Electronic Manufacturing Services	1385157
TXN	Texas Instruments	Information Technology	Semiconductors	97476
VRSN	Verisign Inc.	Information Technology	Internet Services & Infrastructure	1014473
V	Visa Inc.	Information Technology	Data Processing & Outsourced Services	1403161
WDC	Western Digital	Information Technology	Technology Hardware, Storage & Peripherals	106040
WU	Western Union Co	Information Technology	Data Processing & Outsourced Services	1365135
XRX	Xerox	Information Technology	Technology Hardware, Storage & Peripherals	108772
XLNX	Xilinx	Information Technology	Semiconductors	743988
ZBRA	Zebra Technologies	Information Technology	Electronic Equipment & Instruments	877212