Lancaster University Management School

**Department of Accounting and Finance Lancaster University**

# AcF 351b Career Skills in Accounting and Finance

## Python for Data Analysis

## Stream Assignment

## 2019/20

---

### 1. Overview

Python for Data Analysis stream is designed to provide introductory programming knowledge to students who have no or little prior programming experience. Throughout the five sessions, this module has covered Python language basics, scientific computing and web scraping packages and introductory textual analysis. The assignment therefore is intended to give students the opportunity to practice on what have learned from the course and encourage students to do independent work towards writing a complete script to download public data and conducting basic textual analysis in the fields of accounting and finance.

In this coursework, students are expected to investigate potential **consequences of negative/constraining tone reporting**, i.e., do investors respond strongly to such soft information hidden in companies' financial reporting, if so, in which way and to what extend? Students are also encouraged to be creative and explore further on research questions that are meaningful to the above topic.

This document sets out the details of the stream coursework requirements along with some instructions/tips and core reading list.

### 2. Coursework Submission

- Submission Deadline：**Wednesday 15th January 2020 12:00noon**
- Submission Location: **Moodle**
- Submission Documents：
    - Part 1: A **Jupyter Notebook** file for your Python script. You are required to provide codes to demonstrate the workflow to obtain, process and analyze your data. This part contributes **50%** of your coursework assessment. In general, this coding part should consist of the followings (for detailed instructions see Section 3 Part 1):
        - Sub-section to obtain scrape 10-K data from SEC Edgar Database;

- Sub-section to access financial data from WRDS;
- Sub-section to merge, clean, process and analyze on the above downloaded data;

**NOTE**:

- Make sure that your codes can be run in other's environment and your results can be reproduced;
- Make sure you also put short comments/markdown text to highlight what you are doing for those codes.

- Part 2: A **Microsoft Word** document for your report. You are required to write a report to interpret the results from your analysis. This part contributes **50%** of your coursework assessment. You should include the following:
  - A cover sheet on the first page of your report including the details of your full name and student ID;
  - A short introduction to the topic;
  - Relevant literature review (You are expected to read more than the selected references listed in the end of this documents);
  - Analysis on the data and main findings (for detailed instructions see the Section 3 Part 2);
  - A short conclusion;
  - References (Review of Financial Studies or Journal of Finance Style).

  **NOTE**:

  - The coursework assignment should be kept as short and concise as possible. The overall length of the coursework **MUST NOT EXCEED** 2,000 words (Please note that the word limit *excludes* tables, the list of references, and appendices containing illustrative and supporting material, but *includes* footnotes.)
  - Completed reports that exceed the maximum word limit will be **SEVERELY PENALIZED**.
  - Make sure to use **12 point Times New Roman** font with generous margins and **1.5 line spacing** consistently.
  - For more detailed guidance on writing your report, please refer to the general outline of the AcF351b module.

## 3. Instructions

*Part 1: Gathering and preparing data*

In this part, the assessment will be carried out primarily based on your Python codes in Jupyter Notebook documents. This part has two main tasks and it counts for 50% towards the final mark of this coursework.

- TASK 1 (25 marks)

  Download and scrape 10-K filings for all listed US companies during the period from 2000 to 2018 from SEC Edgar Database and conduct preliminary analysis on the textual data. Some tips (as discussed in details in the fifth session) are:

  - Start with the SEC Edgar's Archives for directory listing of full-index and understand how SEC Edgar database organizes the U.S. companies' filings;
  - Download the crawler.idx file and parse the useful information on the crawler files (for each QTR each Year);
  - Access to the link for 10-K filing summary page and then the actual 10-K filing link in htm format (for each filing);
  - Harvest all the text data in10-K filing report (for each filing).
  - Store all relevant data locally ready for further processing
  - Clean the downloaded textual data (Text Pre-processing, refer to Bodnaruk, Loughran and McDonald, 2015 for detailed process)

- TASK 2 (25 marks)

  Access and download CRSP stock data (and Compustat data if you think it can enhance your analysis) from WRDS and merge with the processed text data above. Some tips (as discussed in details in the third session) are:

  - Go through Fama-French sample codes from WRDS can be enormously helpful to understand CRSP and Compustat data. However you do not need to re-write the codes, just take whatever you think is sensible for our purpose;
  - Merge databases. Pay extra attention to company IDs since CRSP, Compustat and SEC Edgar databases use different identification codes (for instance CRSP has crsp_permno, Compustat uses GVKEY, and SEC Edgar uses CIK). You should have access to a linking table/dataset in WRDS to map these different IDs. There are also plenty resources online demonstrate how to link IDs and you should also be able to find instructions from Bodnaruk, Loughran and McDonald (2015);
  - Equally, you need to figure out how to merge data on the dimension of time. 10-K filings have two sets of timestamps: filing date and report date. Filing date is when an individual 10-K report be sent to SEC and you might want to treat this date as the date when the report is public available. Report data however is meant to indicate the covering period of that report. For instance, a reporting date of 30$^{th}$ September 2018 for a 10-K file covers all the financials of company since the last 10-K file in 2017;
  - Merge multiple datasets into a panel data table ready for further processing and analyzing.

*Part 2: Exploring and analyzing data*

In this part, the assessment will be carried out primarily based on your academic report in Microsoft Word documents. This part has two main tasks and it counts for 50% towards the final mark of this coursework.

- TASK 3 (20 marks)

  Conduct exploratory data analysis (EDA). You need to provide detailed summary statistics on your data (including textual data from SEC Edgar and stock data from WRDS) and explore relevant questions (but certainly not restricted to) and interpret:

  1. How many negative and/or constraining words in the 10-K files? Are there two tones highly correlated?
  2. Are there cross-sectional differences? Who are those companies reporting in the most negative (and/or constraining) tone?
  3. Are companies reporting in a consistent tone over time or change dramatically?

  Note: You can search and follow others' textual analysis workflow if you think those can provide insights to understand your data and of course it is always a solid idea to replicate this part from published journal articles such as Loughran and McDonald (2010) and Bodnaruk, Loughran and McDonald (2015) and others.

- TASK 4 (30 marks)

  Follow Bodnaruk, Loughran and McDonald (2015) to conduct textual analysi on companies' 10-K text and investigate potential consequences of financial reporting with negative/constraining tone. You need to conduct an event-study with 10-K filing disclosure as the time window and examine whether and how the market reacts to such soft information. Answer the following questions and interpret your findings:

  1. How did most (least) constrained companies perform around the disclosure, such as one day/week prior the 10-K filing and one day/week/month/year afterwards? You might want to group stocks into quantiles according to the tone in their reports. Please refer to market event study published in journal articles including (but not limited to) Bernard and Thomas (1989) who documented the post-earnings announcement drift.

     Note: After your attempt to answer the above question, you are encouraged to explore other research questions relate to this topic and interpret your findings. One example would be whether companies' operation performance actually deteriorate (improve) after reporting in a negative/constraining (positive) tone? Be creative in an academic way of course.

## 4. Resources

*Selected Academic References:*


- Bernard, V. L., & Thomas, J. K. (1989). Post-earnings-announcement drift: delayed price response or risk premium?. Journal of Accounting research, 27, 1-36.
- Loughran, Tim and Bill McDonald. (2011). "When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks", Journal of Finance, 66: 67-97.
- Bodnaruk, Andriy, Tim Loughran, and Bill McDonald. (2015) "Using 10-k text to gauge financial constraints." Journal of Financial and Quantitative Analysis 50: 623-646.
- Loughran, Tim and Bill McDonald. (2016). "Textual Analysis in Accounting and Finance: A Survery", Journal of Accounting Research, 54: 1187-1230.


*Selected Online Resources:*

- Software Repository for Accounting and Finance (SRAF): https://sraf.nd.edu/
- Professor Bill McDonald's website: https://www3.nd.edu/~mcdonald/ and his various recent published papers utilising textual analysis