Professorship of Finance and Data Science
Prof. Dr. Alexander Hillert

# "Textual Analysis" – Final Exam

***Exam Date: August 2, 2022.***

## General instructions

- This exam is a 72-hour take home exam. The start date is Tuesday, August 2, 2022, at 12 pm (noon). The solutions must be handed in by Friday, August 5, 2022, at 12 pm (noon). Please, send your solutions by e-mail to hillert@safe-frankfurt.de.
- The exam consists of two parts. While the first part asks you to answer questions on textual analysis methods, the second parts deals with the practical implementation using Python.
  Please note that in this year's exam the second part also includes one question about the interpretation of your results.
- You can achieve a maximum of 34 points. For each task, the maximum number of points is indicated in parentheses.
- At the end of the exam, you find the list of references.
- The answers to the questions of the first part as well as the answers to the two interpretation questions of the second part should be written in a document in pdf format ("Solutions_Part_1.pdf").
- Your solutions to the second part should be .py (Python) files with your programming code. Please submit a separate file for each programming task (e.g., "Solution_Part_2_Problem_1.py", "Solution_Part_2_Problem_2.py", etc.).
- If the result of a programming problem is a csv or a txt file, please also submit these files. The exact structure as well as the names of the csv/txt files are specified in the instructions of each programming problem.
- You are also allowed to use xls or xlsx files instead of csv files.
- Please submit all your solution documents in a combined zip file. Please name the file as "Exam_Student_ID.zip", e.g., "Exam_123456789.zip". If you do not have a student ID, include your first name and last name in the file name (e.g., "Exam_Jane_Doe.zip").
- Your programming code should be as easy to comprehend as possible. So, please add detailed comments to your code to describe what you are doing and how the code is structured.
  Please use "reasonable" variable names (e.g., "text_10K", "text_MDA_section") but not cryptic names (e.g., "var1", "var248_v29").

- You are allowed to use Python packages that we have not used in class. However, you should list these packages at the top of your programming code and explain how to install them.
- Your programming code should run on a different computer after the directory has been adjusted to the respective computer.
- You are allowed to use the programs and/or parts of the programs that we have developed in class.
- You are not allowed to forward this exam to any third party (neither during nor after the exam). Students not participating in this exam are considered to be a third party.
- **By submitting your solutions, you declare that you have written your textual analysis exam by yourself and that you have not copied (parts of) answers from others.**

# Part 1 – Questions on textual analysis research (9 points)

This part consists of a single problem, in which you should review a textual analysis paper for an academic journal.

**Referee task – Mangers' and investors' reaction to unionization activities at publicly listed U.S. companies (9 points)**

Assume that you have been asked by the editor of a top finance journal to review a paper submitted to her journal. The paper analyzes investors' response to unionization activities at publicly listed U.S. companies. The authors of the paper motivate their research question by the increasing number of unionization activities seen at companies like Amazon and Starbucks (see, e.g., the two recent New York Times articles https://www.nytimes.com/2022/04/08/business/economy/starbucks-union-new-york-vote.html and https://www.nytimes.com/2022/04/15/briefing/amazon-union-warehouse.html).

To measure unionization activities, the authors use Form 10-K filings and determine the percentage of sentences that discuss unionization. Moreover, they analyze the tone – measured by the Loughran and McDonald (2011) dictionary of negative words – of unionization sentences to gauge management's views about unionization. To investigate investors' response to unionization activities, the authors regress the abnormal filing return on the percentage of unionization sentences, their tone, and control variables. The abnormal filing return is measured over the 4-day event window from the filing date (day t) to day t+3. I.e.,

the authors use the same dependent variable as in Loughran and McDonald (2011). The set of control variables includes the firms' total assets, total sales, and the stock return over the previous fiscal year. Furthermore, the authors include industry fixed effects and annual time fixed effects (i.e., a dummy variable for every calendar year).

Sample selection and textual analysis approach

The authors use all Form 10-K filings of publicly listed U.S. companies that are available in the EDGAR system of the Securities and Exchange Commission (SEC). Like Loughran and McDonald (2011), they only include firms with non-missing CRSP (stock market database) and Compustat (accounting database) data. The authors' sample period starts in 1994 and ends in 2021. The final sample of the authors comprises 81,678 observations/Form 10-K filings.

Regarding the textual analysis approach the authors write:

*"We identify all filings that have the form type "10-K" in the quarterly SEC overview files. After downloading these files, we perform several editing operations to get from the complete submission files of the Form 10-Ks to the final version of the text that we use to compute the text-based measures. First, we delete all exhibits from the 10-K filings, eliminate all html code, and remove tables. Afterwards we transform the edited text to lower case. All of the word lists used are also transformed to lower case.*

*After these editing operations, we identify sentences that cover the topic of unionization. We start with the most obvious key words: "union," "unions," "unionization," "unionizations," "unionize," "unionizes," "unionized," and "unionizing." We use an inflected word list to avoid any imprecision (e.g., "odd" vs. "odds," see Loughran and McDonald (2011)). Next, we add the word "united" to cover references to the "United Auto Workers," "United Steelworkers," and "United Food and Commercial Workers," which are all among the most important labor unions in North America. Moreover, the word "united" is often part of union names. So, by including it in our list, we make sure that we capture references to unions comprehensively. Besides, we also include the words "worker," "workers," and "force" to capture statements about the "labor force" and the "work force." We do not include the plural ("forces"), as it is rarely used in the context of workers but rather refers to the armed forces, i.e., the U.S. military. Similarly, "employee(s)" is also not frequently used in the union context. So, we neither include these two words in our word list. We count all sentences as union-related if they include at least one of these key words.*

*To identify sentences, we rely on popular software packages which have been shown to yield good results on typical texts like newspaper articles. So, we do not expect any problems with the sentence identification in our Form 10-K filings.*

*In the next step, we compute the negativity of union-related sentences to measure a company's attitude towards the unions and unionization activities. We also compute the negativity in all non-union-related sentences, i.e., in all sentences that do not contain any of the key words. We expect that the negativity in union-related sentences is much higher than the negativity in non-union-related sentences, as the management typically holds quite negative opinions about unions.*

*Next, we test whether investors also react negatively to information about unionization activities at firms. As investors learn the details about these activities from firms' annual reports, for example, from Management's Discussion and Analysis, we analyze their response to the filing of Form 10-K filings. More specifically, we regress the abnormal filing CAR (as in Loughran and McDonald (2011)) on (1) the percentage of union-related sentences, (2) the negativity of union-related sentences, (3) the negativity of non-union-related sentences, (4) the base set of controls (i.e., the firm's total assets, total sales, and the stock returns over the previous fiscal year), and year and industry fixed effects. In a second specification, we additionally control for other text-based measures including the Fog Index (see, e.g., Li (2008)) and the percentage of Loughran and McDonald (2011) uncertainty words. In a third specification, we interact union-related negativity with uncertainty.*

*To measure negativity, we rely on the well-established Loughran and McDonald (2011) dictionary of negative words. We rely on the original version of their word list and do not consider words that have been added or removed since then. We do not include positive words, as they often carry a more ambiguous meaning (see Loughran and McDonald (2016)).”*

The summary statistics of the key text-based measures and the regression results are shown in the two tables below:

Table 1 – Summary statistics

| Variable | mean | std. dev. | p10 | p50 | p90 | N |
|---|---|---|---|---|---|---|
| Percentage of union-related sentences | 5.04% | 4.41% | 1.24% | 3.82% | 7.52% | 81,678 |
| Negativity in union-related sentences | 2.04% | 1.24% | 0.97% | 1.89% | 4.62% | 81,678 |
| Negativity in non-union-related sentences | 1.30% | 0.49% | 0.64% | 1.25% | 2.87% | 81,678 |
| Uncertainty | 0.70% | 0.28% | 0.32% | 0.68% | 1.12% | 81,678 |
| Modal weak words | 0.59% | 0.29% | 0.18% | 0.55% | 0.92% | 81,678 |
| Fog index | 15.63 | 5.34 | 10.41 | 14.89 | 24.35 | 81,678 |

*This table shows the mean, median, standard deviation, 10th percentile, 90th percentile, and the number of observations. The percentage of union-related sentences is the percentage of sentences with at least one union-related key word. The list of key words includes: "union," "unions," "unionization," "unionizations," "unionize," "unionizes," "unionized," "unionizing," "united," "worker," "workers," and "force."*
*Union-related negativity is the percentage of Loughran and McDonald (2011) negative words in sentences with at least one union key word. Non-union-related negativity is the negativity in all remaining sentences of the 10-Ks. Uncertainty (modal weak words) is the percentage of uncertainty (modal weak) words according to the Loughran and McDonald (2011) words. Fog Index is the well-established readability measure and is computed as Fog Index = 0.4 x (words per sentence + percentage of complex words).*

Table 2 – Main results

| Dependent variable | Abnormal return from t to t+3 | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| Negativity in union-related sentences | -0.0514 | -0.0415 | -0.0387 |
| | (-0.81) | (-0.74) | (-0.56) |
| Negativity in non-union-related sentences | -0.2185*** | -0.2101*** | -0.1853** |
| | (-2.70) | (-2.63) | (-2.15) |
| Percentage of union-related sentences | 0.0351 | 0.0319 | 0.0289 |
| | (0.70) | (0.55) | (0.46) |
| Sales ($ million) | 0.0003** | 0.0002** | 0.0003** |
| | (2.08) | (1.99) | (2.01) |
| Total assets ($million) | -0.0007* | -0.0006* | -0.0006* |
| | (-1.70) | (-1.69) | (-1.71) |
| Stock return previous year | 0.0109** | 0.0102** | 0.0103** |
| | (2.30) | (2.10) | (2.12) |
| Fog index | | -0.0024** | -0.0025** |
| | | (-2.04) | (-2.06) |
| Uncertainty | | -0.2176 | -0.1834 |
| | | (-1.40) | (-1.07) |
| Interaction of uncertainty and negativity in union-related sentences | | | -0.0061 |
| | | | (-0.36) |
| Year fixed effects | Yes | Yes | Yes |
| Industry fixed effects | Yes | Yes | Yes |
| $R^2$ | 0.023 | 0.023 | 0.019 |
| N | 76,924 | 76,924 | 76,924 |

*This table shows regressions of abnormal 4-day filing period returns from the filing day t up to day t+3 on tone measures and control variables. The abnormal filing period return is the return of stock i over the four-day period from t (i.e., the filing day) to t+3 minus the market return over the same four-day period (see Loughran and McDonald (2011)). The text-based variables are defined as in Table 1 (see above). Column (1) is the main specification. In column (2), we add uncertainty and readability (measured by the Fog Index) as control variables. Column (3) includes an interaction term between negativity in union-related sentences and uncertainty.*

*Total assets and sales are obtained from Compustat (items "AT" and "SALE"). Stock return is the firm's stock return over the previous fiscal year, i.e., the year that is covered in the Form 10-K filing. The abnormal filing day return is expressed in decimals, i.e., a one percent abnormal return corresponds to a value of 0.01.*
*t-statistics are provided in parentheses and are based on robust standard errors. The regressions include industry and year fixed effects.*

The authors interpret their results as follows:

*"The summary statistics in Table 1 show that the negativity in union-related sentences is indeed much higher than in non-union-related sentences. The difference between union-related negativity (2.04%) and non-union-related negativity (1.30%) is highly statistically significant (t-statistic of 4.23 [not reported in Table 1]). Thus, management views unionization efforts at their firms very critical and seem to be opposed to any unionization efforts.*

*Looking at the percentage of union-related sentences, we find that managers talk a lot about unionization. For the average form 10-K filing 5.04% of sentences deal with unions. The median is a bit lower 3.82% but even that number seems very large especially given our small list of key words.*

*Next, we turn to investors' response to unionization activities. The results in Table 2 show that investors do neither respond to the amount of union-related information in the 10-K nor to whether the information is good or bad. Both, the percentage of union-related sentences and the negativity in union-related sentences are not significantly related to abnormal returns.*

*The negativity in non-union-related sentences, i.e., the negativity in the rest of the document is significantly negative associated with filing CARs. This result is in line with Loughran and McDonald (2011): investors respond more negatively to more qualitative bad news.*

*For our base set of controls, we obtain a positive relation between previous year's stock returns and CARs, i.e., the annual reports of winner stocks are perceived more positively. Sales carries a positive coefficient, while firm size – measured by total assets – shows a negative relation. This may indicate that higher cash flows are good news, while having a larger balance sheet might not.*

*The R-squared (R2) is slightly above 2% which is within the normal range of regressions of individual stock returns on firm characteristics. For comparison, Loughran and McDonald (2011) report an R2 of around 2.5% in their Table IV, which is very similar to our setting.*

*Focusing on column (2), we find a significantly negative relation between readability and investors' response indicating that investors prefer annual reports that are well written and easily understood. This result is in line with the SEC's (1999) "Plain English Initiative," which asks firm to write clear documents.*

*We find no significant relation between the percentage of uncertainty words and filing CARs. While this result might be surprising – Loughran and McDonald (2011) – find a negative relation between uncertainty and filings CARs, it might be explained by the sample period. The period after 2008, which is the end of the Loughran and McDonald (2011) sample, was a period of steady economic growth and little uncertainty except for the recent Covid crisis.*

*Most importantly, adding these additional controls does not change our main result. There is still no significant relation between union-related negativity and filing CARs.*

*In specification (3), we interact union-related negativity with uncertainty to test whether unionization threats matter to investors in times of high uncertainty. However, the interaction term is not statistically significant. Thus, there is no evidence for investors being concerned about unionizations at publicly listed companies. The conclusion of our paper is that managers might be exaggerating the threat of unionization to some extent, while shareholders take a neutral stance on unionization.*

*For completeness, note that our result regarding the relation between uncertainty and filing CARs changes when we use the percentage of modal weak words instead of the uncertainty words. The coefficient becomes significantly negative. (The interaction effect in column (3) is still insignificant.) However, as (1) these two word lists are quite different and (2) the uncertainty list is better able to capture uncertainty in the sense of not being able to exactly quantify things, we focus on the uncertainty word list."*

Please discuss the authors' empirical approach critically and comment on their results and conclusions.

In this problem, you should describe the paper's weaknesses (with respect to both methodology and economic interpretation and with a focus on the text-based variables). Please clearly explain why you believe that a certain task has not been executed accurately and/or needs to be improved. Please also explain why you think that an interpretation of a result is flawed. If you believe that information on the authors' approach is missing describe which information is missing and why it is important to provide this information to the readers.

For each of the paper's weaknesses that is correctly explained, you will get 1.5 points (0.5 points for identifying a problem and 1 additional point for explaining the problem). Consequently, you need to correctly identify and describe at least six weaknesses to get the full number of points.

Note that it is not required to write a formal referee report, i.e., you do neither need a summary of the paper nor do you need to use any polite or formal language. You can simply describe the weaknesses you identify in the paper. If you want, you can provide a recommendation to the editor (you don't have to).

*Hint: it may be helpful to compare the authors' approach and results with papers that we have discussed in our course.*

# Part 2 – Implementing a textual analysis in Python (22 points for programming + 3 points for interpretation)

The second part comprises three programming problems that are based on two samples of earnings conference calls.

The first sample comprises the full transcripts of the earnings calls from Facebook, Bank of America, and Amazon for the period from 2013 to 2017, i.e., for each of the three companies there are 20 transcripts (five years and four quarterly calls per year). The attached csv file "Overview_File_Problem_1.csv" list all 60 transcripts. Column 1 is the call ID (from 1 to 60), column 2 shows the file name of the transcript, column 3 is the companies' ticker symbol, and column 4 displays the company name from the transcript. The transcripts are obtained from Thomson Reuters. In Problem 1, you are asked to preprocess the documents, create an overview file on the call participants, and split the transcripts into presentation, questions, and answers.

Problems 2 and 3 are based on a second set of earnings conference calls, for which I provide you with preprocessed files. For each call, you get a file containing all questions and a file containing all answers. The sample also comprises 60 earnings conference calls but is based on different companies: Coca-Cola, Target, and J.P. Morgan Chase. The preprocessed files assure that you can complete Problems 2 and 3 even if you do not manage to solve Problem 1.

***The problems have not been ordered based on the level of difficulty. It is well possible that you find problem 2 and/or problem 3 easier than problem 1. Often the perceived level of difficulty is***

***subjective. Thus, I recommend that you look at all three problems and then start with the problem that you find most comfortable working on.***

**Programming templates**

For all three problems, you get programming templates. The templates are similar to the ones you know from class, and they may help you to arrive at the solution.

It is not required that you use these templates. It is neither required that you complete all commands in the template. Feel free to approach a problem in the way that you like best. There are often many different correct approaches to solve a programming problem.

## Please complete the following programming problems:

1. **Data preparation – Creating an overview file and identification of call segments (10 points)**

   This problem is split into two subquestions (5 points and 5 points).

   You find the text documents for this problem in the folder "Problem_1_Sample."

   **a) Create an overview file (5 points)**

   Please create a csv file named "Problem_1_Overview_Calls.csv" that provides information on (1) the fiscal quarter and year, (2) the date and time, and (3) the participants of each of the 60 earnings conference calls. One call corresponds to one line in the csv file. The first line should contain the column names.

   More precisely, the csv file should comprise the following information:

   - Column 1: the ID of the call ranging from 1 to 60.
   - Column 2: the filename of the call. For example, "2013-Jan-29-AMZN.OQ-139057386295-Transcript.txt" for the first call.
   - Column 3: the fiscal quarter of the call. For example, "Q4" for the first call (ID=1).
   - Column 4: the fiscal year of the call. For example, "2012" for the first call (ID=1).
   - Column 5: the date of the call in the format YYYYMMDD. For example, "20130129" for the first call (ID=1).
   - Column 6: the time of the call (time + am/pm + time zone). For example, "05:00 PM GMT" for the first call (ID=1).

- Column 7: the total number of non-corporate call participants according to the header of the document. For example, in the first call (ID=1), there are 15 non-corporate participants.
- Column 8, 10, 12, etc.: the name of the first, second, third, etc. corporate participant. For example, for the first call (ID=1), "Sean Boyle" and "Tom Szkutak" should be listed in columns 8 and 10, respectively.
  Note: the highest number of corporate participants in this sample are 4 people.
- Column 9, 11, 13, etc.: the position of the first, second, third, etc. corporate participant. For example, for the first call (ID=1), "VP of IR" and "CFO" should be listed in columns 9 and 11, respectively. A corporate participant can have more than one position (e.g., "President & CEO"). Include all positions in the same cell.

*Hint: Use the attached file "Overview_File_Problem_1.csv" to open the transcripts. You can iterate lines 2 to 61 (line 1 is the column header) and open the transcripts using their file name (column (2)).*

**b) Identification of call segments and preprocessing (5 points)**

In this problem you should identify the three segments of an earnings conference call: (1) managers' presentation, (2) analysts' questions, and (3) managers' answers. The edited text of each segment should be written to an output txt file. The individual files should be named "call_XXX_presentation.txt", "call_XXX_questions.txt", "call_XXX_answers.txt", where XXX is the ID of the call ranging from 1 to 60.

In terms of editing, you should:

1. Remove the general information about the call including the list of call participants at the beginning of the document. I.e., you should start after "Presentation."
2. Delete the "Definitions" and "Disclaimer" at the end of the documents.
3. Exclude all operator statements.
4. Delete technical remarks like, for example, "(inaudible)" and "(technical difficulty)."

For the presentation, keep only the pure text, i.e., remove speaker names and their position. For example, in the first call (ID=1), you should drop "Sean Boyle, Amazon.com, Inc. - VP of IR [2]," "Tom Szkutak, Amazon.Com Inc - CFO [3]," and "Sean Boyle, Amazon.com, Inc. - VP of IR [4]."

For the questions, you should number all questions from 1 to N, where N is the total number of questions. More specifically, each question should start with "Question_XXX:" followed by a line break and the text of the analyst's question. (XXX is the number of the question.) For example, the first two entries in the question file for the first call (ID=1) should read as follows:

- Question_1:
  Tom, it looks to us that you have successfully begun […] fixed cost in the future? Thanks.
- Question_2:
  Just wanted to ask about the shift to third-party and in particular, […] where you made a similar shift? Thanks.

For the management's answers, you should number all answers from 1 to N, where N is the total number of answers. More specifically, each answer should start with "Answer_XXX:" followed by a line break and the text of the manager's answer. (XXX is the number of the answer.) For example, the first two entries in the answer file for the first call (ID=1) should read as follows:

- Answer_1:
  In terms of fulfillment question, you're right […] going forward to do that.
- Answer_2:
  We did see a good expansion, as you've mentioned, […] you're seeing that, but certainly you're seeing it there.

Note that some analyst questions will be answered by more than one corporate participant. In these cases, it is fine to label each manager's statement as a separate answer, i.e., you do not need to combine the different parts of the total answer into a single one.

*Please put all 180 txt files (60 files for the presentation, 60 files for the analysts' questions, and 60 files for management's answers) into a zip file ("Problem_1_Conference_Call_Segments.zip") and attach this zip file to the solutions that you hand in.*
*In addition to the zip file, please submit your program code.*

**In the subsequent two problems (Problem 2 to 3), please use the files from the second sample of 60 conference calls that contain the calls' questions and answers.**

The files are named "XXX_questions.txt" and "XXX_answers.txt," where XXX is the ID of the call ranging from 1 to 60. You find these files in the folder "Problem_2_3_Sample_QandA."

The attached file "Overview_File_Problem_2.csv" provides an overview on the files. Column 1 is the call ID (ranging from 1 to 60), column 2 is the name of the full transcript, column 3 is the name of the question file ("1_questions.txt" to "60_questions.txt"), column 4 is the name of the answer file ("1_answers.txt" to "60_answers.txt"), column 5 shows the company's ticker symbol, and column 6 displays the company name.

The full transcripts are not needed for this problem. They are, however, included for the case that you would like to look something up in the full transcript of the call (folder "Problem_2_3_Sample_full_transcripts").

The sample comprises 20 quarterly earnings conference calls from each Coca-Cola, Target, and J.P. Morgan Chase. The sample period is 2013 to 2017.

The question and the answer files are organized in a similar way as the question and the answer documents that you have created in Problem 1. Question and answers are numbered from 1 to N, where N is the number of questions and answers in the call. If a question has been answered by more than one corporate participant, all parts of the total answer are combined to a single answer in the answer document. This grouping of answers is different to the one in Problem 1 but allows you to directly link questions and answers. You can be sure that answer number X ("Answer_X:") corresponds to question number X ("Question_X:").

All operator statements and the speaker names have been removed. The text said by analysts and corporate participants has <u>not</u> been edited in any way.

## 2. Determine the percentage of Loughran and McDonald (2011) positive words per manager statement (8 points)

This problem comprises two subquestions (4 points + 4 points).

### a) Determine the percentage of Loughran and McDonald (2011) positive words (4 points)

Please determine the absolute number as well as the percentage of positive words according to the Loughran and McDonald (2011) dictionary for each manager answer. You find the managers' answers in the answer files of the 60 calls. When counting positive words, you

should control for negations as in Loughran and McDonald (2011). The Loughran and McDonald (2011) word list is included in the exam materials. The list is a txt file ("LMD_pos_master_dictionary_2018.txt") and contains one word per line.

The solution to this problem should be a csv file with six columns:

- Column 1: the ID of the call ranging from 1 to 60.
- Column 2: the ID of the answer ranging from 1 to N, where N is the number of answers in the call.
- Column 3: the total number of words in the answer.
- Column 4: the number of positive words according to the Loughran and McDonald (2011) dictionary (controlling for negations) in the answer.
- Column 5: the percentage of positive words in the answer, i.e., the number of positive words of the answer divided by the total number of words of the answer.
- Column 6: the text of the answer.

  Note that you need to replace line breaks ("\n") by whitespaces (" ") and semicolons (or whatever column delimiter you are using in the csv file) by a placeholder (e.g., "SEMICOLON"). Otherwise, the text will mess up your csv file.

The csv file should be named "Problem_2a_Percentage_Positive_Words.csv".

The unit of observation in this problem is the individual answer level. If there are, for example, 15 answers in a call, you should determine the positivity for each of the 15 answers and you will get 15 entries (rows) for the csv file.

**Important note on Problems 2a) and 2b)**

The template contains a placeholder for further editing operations to improve the measurement of positive tone.

I recommend that you first ignore this part and look at your output file for Problem 2b) (see below). The result from Problem 2b) will clearly show you what the problem is. With this knowledge, go back to Problem 2a) and work on the additional editing. Finally, return to Problem 2b) with the improved output from Problem 2a) and provide your final answer to Problem 2b)

*In addition to this csv file, please submit your program code.*

*Hints:*

- *In Problem 8 of our course, which is similar to this problem, we used a text corpus that we had carefully edited in advance. So, in this problem, you may want to include some editing operations to make sure to only count actual words.*
- *Be careful to consider both contracted (e.g., "don't") and long (e.g., "do not") forms.*
- *The solution to Problem 8 of our course shows you how to determine the percentage of positive words controlling for negations.*

**b) Discuss the top 10 positive answers (1 point + 3 points)**

*You can earn 1 point for the csv file and 3 points for the interpretation question below.*

Based on the csv files (preliminary and final) from part a) of this problem, create a csv or an Excel file that lists the top 10 most positive answers by management, i.e., the 10 statements with the highest percentage of positive words.

The solution to this problem should be a csv file with the following 7 columns:

- Column 1: the rank of the answer from 1 (most positive) to 10 ($10^{th}$ most positive).
- Column 2: the ID of the call (1 to 60) where the answer is from.
- Column 3: the ID of the answer within the call (1 to N, where N is the number of answers in the call).
- Column 4: the total number of words in the answer.
- Column 5: the number of positive words according to the Loughran and McDonald (2011) dictionary (controlling for negations) in the answer.
- Column 6: the percentage of positive words in the answer, i.e., the number of positive words in the answer divided by the total number of words in the answer.
- Column 7: the text of the answer.

The csv/Excel file should be named "Poblem_2b_Top_10_Positive_Answers.csv/.xlsx".

There is no need to recalculate any information (e.g., number of words, number of positive words). You can rely on the information from the csv file from part a).

It is not required to solve this task in Python. You are allowed to sort the answer on positivity and copy the top 10 manually.

**Interpretation question**: Look at the top ten positive statements from the csv files (preliminary and final). Do managers provide very positive information for the firm in these

statements? In other words, does the high percentage of positive words indicate very good news about the firm? Explain your answer!

In general, do positive dictionary words perform better, the same, or worse than negative dictionary words in capturing sentiment? Why? Explain your answer!

*Please write your answer to this question to the pdf document from Part 1 of this exam.*

**3. Identifying the answers to questions about the market (7 points)**

This problem comprises two subquestions (3 points + 4 points).

**a) Identification of answers to market-related questions (3 points)**

This problem is based on the same set of transcripts as Problem 2. You should identify all questions that are related to (financial) markets. A question is considered to be market related if it contains the word "market" or its plural form "markets." For this problem, it is fine if the question is about a firm's "product market" (or some other "market") and not about the financial market. In other words, we only rely on the two key words "market" and "markets" without any further conditions.

In the first step, you should create a csv file that lists the number and the percentage of market-related questions per earnings conference call.

For this task, we consider the entire statement of an analyst as a single question, i.e., you do not need to count the question marks in the analyst's statement. As soon as the word "market" or "markets" is found in the analyst's statement, the statement counts as a market-related question.

Example: "Question_XXX: *I have one for Mark and one for Sheryl. […] How is Facebook thinking about sharing in the economics of when brands use celebrities or influencers to market their products using their Facebook posts? Is it by bringing more transactions onto the platform, building on the shopping experience on Facebook. How is Facebook going to share in those economics?"* → *this statement would count as one market-related question.*

The solution to this problem should be a csv files with four columns:

- Column 1: the ID of the call ranging from 1 to 60.
- Column 2: the total number of questions in the call.
- Column 3: the number of market-related questions in the call.
- Column 4: the percentage of market-related questions in the call, i.e., the number of market-related questions divided by the total number of questions.

The csv file should be named "Problem_3a_Market-related_Questions.csv"

*Hint: To identify the questions (according to our definition), you can split the text of the question files ("1_questions.txt" to 60_questions.txt") using the regex "Question_[0-9]{1,}:".*

**b) Most frequent trigrams in the answers to market-related questions (4 points)**

In the second step, you should identify managers' answers to market-related questions and then, create a list of trigrams (i.e., 3-word combinations) based on the text of managers' answers. Finally, you should write the 30 most frequent trigrams to a csv file.

Trigrams should not include NLTK stop words and should be formed only within a sentence (i.e., the last word of the first sentence does not form a trigram with the first two words of the second sentence). For example, the two sentences "*And once we* get *that to a* big base, *I* think *there are* going *to be a* lot *of* opportunities *to* build *the* business. *And the* business *will be* proportional *to the* amount *of that* activity *that* people want *to do* organically." (NLTK stop words are *italicized*) result in the trigrams "get big base," "big base think," "base think going," "think going lot," "going lot opportunities," "lot opportunities build," "opportunities build business," (all from the first sentence) "business proportional amount," "proportional amount activity," "amount activity people," "activity people want," and "people want organically."

The solution to this problem should be a csv file with three columns:

- Column 1: the rank ranging from 1 (most frequent trigram) to 30 (30$^{th}$ most frequent trigram).
- Column 2: the 30 most common trigrams in the answers to market-related questions.
- Column 3: the frequency of the trigram, i.e., how often the trigram appears in the answers to market-related questions.

The csv file should be named "Problem_3b_Most_Frequent_Trigrams.csv".

*In addition to the two csv files (one for part a) and one for part b), please submit your program code (either one file with the entire code for parts a) and b) or two separate code files).*

*Hints:*

- *You must work with both the question and the answer file. If question X in a call is a market-related question, you need to obtain answer X to form the trigrams.*
- *The solution to programming Problem 9 and/or NLTK's sentence tokenizer (see our slides) as well as the solutions to Problem 12 might be helpful.*

*Note that due to the small sample, the trigrams may not be very meaningful.*

**References**


- Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics, 45(2-3), 221-247.
- Loughran, T., and B. McDonald, 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance 66, 35–65.
- Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. Journal of Accounting Research, 54(4), 1187-1230.
- Securities and Exchange Commission (1999). A Plain English Handbook. Available at https://www.sec.gov/reportspubs/investor-publications/newsextrahandbookhtm.html