Prof. Dr. Alexander Hillert

Introduction to Textual Analysis using Python

VHB Preconference Workshop

March 17, 2020
2020 Annual Meeting of the VHB

Agenda for our Workshop

- Background on Textual Analysis in Accounting/Finance/Economics. How to measure tone?
  - Tetlock (2007)
  - Loughran and McDonald (2011)
- Implementing your first Textual Analysis in Python
  - Installing and starting Python
  - Transcript of earnings announcement call as example
  - Programming first steps in Python
  - Helpful software

If you have questions please use the chat in the conference app.

Tetlock (2007) – Motivation (1)

Abreast of the Market column in the *Wall Street Journal*; January 7, 2004

Title: Sun Microsystems, Brocade Rise; Gateway Loses Large-Cap Status

By Karen Talley, Dow Jones Newswires

- NEW YORK -- Sun Microsystems and Brocade Communications Systems helped the Nasdaq Composite Index hit a two-year high, while the Dow Jones Industrial Average pulled back a bit.
- The Nasdaq gained 10.01 points, or 0.49%, to 2057.37, its highest level in 24 months. The Dow Jones Industrial Average fell 5.41 points, or 0.05%, to 10538.66 after a 134-point rise on Monday, and the S&P 500 index rose 1.45 points, or 0.13%, to 1123.67, a new 20-month high.
- The generally upbeat movement came despite some downbeat economic news. But investors are looking farther out "and buying on what they believe will be an improving economic picture," said Mark Donahoe, managing director, institutional sales trading, at Piper Jaffray. "We're starting to see much more institutional involvement."
- Sun Microsystems gained 33 cents, or 7%, to $5.03 after Merrill Lynch raised its sales and earnings estimates, saying checks, though not complete, suggest the maker of large computer systems experienced a strong close to the latest quarter.

Motivation of Tetlock (2007)

- ‘Abreast of the Market’ column in the WSJ

> One of the more fascinating sections of the *WSJ* is on the inside of the back page under the standing headline “Abreast of the Market.” There you can read each day what the market did yesterday, whether it went up, down or sideways as measured by indexes like the Dow Jones Industrial Average . . . . In that column, you can also read selected post-mortems from brokerage houses, stock analysts and other professional track watchers explaining why the market yesterday did whatever it did, sometimes with predictive nuggets about what it will do today or tomorrow. This is where the fascination lies. For no matter what the market did—up, down or sideways—somebody will have a ready explanation.
>
> Vermont Royster (*Wall Street Journal*, “Thinking Things Over Abaft of the Market,” January 15, 1986)

- What is the relation between the content of the ‘Abreast of the Market’ column and daily stock market activity?

Tone measurement

‘Bag of the word approach’ / dictionary approach:

- Count the number of words of a specific category/list (e.g., negative, positive).
- Calculate the fraction of these words by dividing the category word count by the total number of words.

Which word lists?

→ General Inquirer Harvard IV-4 psychosocial dictionary

- 77 dictionaries, e.g.
  - Negative: 2,291 words
  - Positive: 1,902 words
  - Passive: 911 words
  - Pleasure: 168 words
- The dictionaries are available at: http://www.wjh.harvard.edu/~inquirer/homecat.htm

Tone measurement

How to aggregate the 77 dimensions into a single factor?

→ Principal component analysis (PCA).

- Linear combination of the General Inquirer categories.
- Choose the factor with the largest variance.

- Results of the PCA:
  - Positive weights: negative, weak, fail, and fall categories.
  - Negative weight: positive category.

  → first factor is a pessimism factor.

Main result – Sentiment and market returns

Time-series regressions of returns on sentiment; Tetlock (2007) - Table 2

$$Dow_t = \alpha_1 + \beta_1 \cdot L5(Dow_t) + \gamma_1 \cdot L5(BdNws_t) + \delta_1 \cdot L5(Vlm_t) + \lambda_1 \cdot Exog_{t-1} + \varepsilon_{1t}$$

- Exog.: January dummy, day-of-the-week dummies, October 19, 1987 dummy.
- Coefficients measure the effect of a one std. dev. increase in negative investor sentiment on returns (in bp).

	Regressand: Dow Jones Returns		
News Measure	Pessimism	Negative	Weak
$BdNws_{t-1}$	***−8.1***	***−4.4***	***−6.0***
$BdNws_{t-2}$	0.4	3.6	2.0
$BdNws_{t-3}$	0.5	−2.4	−1.2
$BdNws_{t-4}$	**4.7**	***4.4***	***6.3***
$BdNws_{t-5}$	1.2	2.9	**3.6**
$\chi^2(5)$ [*Joint*]	***20.0***	***20.8***	***26.5***
p-value	0.001	0.001	0.000
Sum of 2 to 5	**6.8**	***9.5***	***10.7***
$\chi^2(1)$ [*Reversal*]	**4.05**	***8.35***	***10.1***
p-value	0.044	0.004	0.002

- Low sentiment predicts low market returns the next day.
- Return reversal on the subsequent four days is about the same magnitude as initial reaction. → media tone predicts sentiment.

Agenda for our Workshop

- Background on Textual Analysis in Accounting/Finance/Economics. How to measure tone?
  - Tetlock (2007)
  - Loughran and McDonald (2011)
- Implementing your first Textual Analysis in Python
  - Installing and starting Python
  - Transcript of earnings announcement call as example
  - Programming first steps in Python
  - Helpful software

Is the Harvard dictionary suitable for a business context?

Analyzing the words in the dictionary shows

- Neutral meaning
  - Examples: tax, costs, expense, liabilities.
  - → tone measurement is noisy.
- Systematic bias
  - Capital → banking and insurance
  - Crude → oil industry
  - Mine → precious metals and coal
  - Illustration of the magnitude of the problem: in the 1999 10-K of Coeur d'Alene Mines Corporation, the word 'mine' accounts for 25% of all negative words.

Main result of the study: almost 75% of the words in the Harvard IV psychosocial dictionary are misclassified in business contexts.

Loughran and McDonald's word lists

1. Negative: 2,337 words
   - 1,121 overlap with Harvard negative
   - Restated, litigation, termination, unpaid, investigation, serious, deterioration, etc.
2. Positive: 353 words
   - Achieve, efficient, improve, profitable, etc.
3. Uncertainty: 285 words
   - General notion on imprecision, not only risk
   - Approximate, depend, fluctuate, indefinite, uncertain, etc.
4. Litigious: 731 words
   - Claimant, deposition, testimony, etc.
5. Modal strong: 19 words
   - Always, highest, must, etc.
6. Modal weak: 27 words
   - Could, depending, might, etc.

Details on the construction of the dictionaries

- How are these lists created?
  1. Take the list of all words contained in the 10-Ks.
  2. Manually classify all words that occur in at least 5% of the filings.

- Word lists can be downloaded from Bill McDonald's webpage.
  https://sraf.nd.edu/textual-analysis/resources/#LM%20Sentiment%20Word%20Lists

- List include inflected versions of the word lists.
  - Accident, accidental, accidentally, and accidents
  - The expand the original Harvard negative list from 2,005 (word stem) to 4,187 words (incl. inflections)
  - Problem with stemming: odd vs. odds, good vs. goods (costs of goods sold).

Loughran and McDonald (2011) – Comparison of word lists (1)

Most frequent words from the Harvard negative dictionary

Full 10-K Document				MD&A Subsection			
Word in Fin-Neg	Word	% of Total Fin-Neg Word Count	Cumulative %	Word in Fin-Neg	Word	% of Total Fin-Neg Word Count	Cumulative %
	TAX	4.83%	4.83%		COSTS	6.45%	6.45%
	COSTS	4.61%	9.44%		EXPENSES	5.51%	11.96%
√	LOSS	3.77%	13.21%		EXPENSE	4.70%	16.66%
	CAPITAL	3.62%	16.83%		TAX	4.68%	21.34%
	COST	3.51%	20.34%		CAPITAL	4.24%	25.58%
	EXPENSE	3.12%	23.46%		COST	3.70%	29.28%
	EXPENSES	2.92%	26.38%	√	LOSS	3.29%	32.57%
	LIABILITIES	2.66%	29.04%		DECREASE	3.06%	35.63%
	SERVICE	2.57%	31.61%		RISK	2.97%	38.60%
	RISK	2.34%	33.95%	√	LOSSES	2.62%	41.22%
	TAXES	2.23%	36.18%		*DECREASED*	2.21%	43.44%
√	LOSSES	2.20%	38.38%		LIABILITIES	2.15%	45.58%
	BOARD	2.13%	40.51%		LOWER	2.10%	47.69%
	FOREIGN	1.68%	42.20%		TAXES	1.95%	49.63%
	VICE	1.52%	43.71%		SERVICE	1.91%	51.55%
	LIABILITY	1.41%	45.12%		FOREIGN	1.87%	53.42%
	DECREASE	1.29%	46.41%	√	IMPAIRMENT	1.63%	55.05%
√	IMPAIRMENT	1.18%	47.59%		CHARGES	1.40%	56.44%
	LIMITED	1.10%	48.69%		LIABILITY	1.16%	57.60%
	LOWER	1.01%	49.70%		CHARGE	1.16%	58.76%
√	AGAINST	1.00%	50.70%		RISKS	1.05%	59.80%
	MATTERS	0.99%	51.69%	√	*DECLINE*	1.00%	60.80%
√	ADVERSE	0.94%	52.63%		DEPRECIATION	0.92%	61.72%
	CHARGES	0.94%	53.57%		MAKE	0.86%	62.58%
	MAKE	0.89%	54.46%	√	ADVERSE	0.84%	63.42%
	ORDER	0.88%	55.33%		BOARD	0.79%	64.21%
	RISKS	0.85%	56.19%		LIMITED	0.78%	64.99%
	DEPRECIATION	0.85%	57.04%		EXCESS	0.71%	65.70%
	CHARGE	0.83%	57.87%		ORDER	0.70%	66.40%
	EXCESS	0.82%	58.69%	√	AGAINST	0.70%	67.10%

Results

- List is dominated by HVD neg. words that are not meaningful in a business context.
- Only 5 (6) of the 30 most frequent HVD neg. words in the overall text (in the MD&A) are included in LMD neg.

Loughran and McDonald (2011) –Table 3, part 1

Most frequent words from the Loughran and McDonald negative dictionary

Full 10-K Document				MD&A Subsection			
Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %	Word in H4N-Inf	Word	% of Total Fin-Neg Word Count	Cumulative %
√	LOSS	9.73%	9.73%	√	LOSS	9.51%	9.51%
√	LOSSES	5.67%	15.40%	√	LOSSES	7.58%	17.10%
	CLAIMS	3.15%	18.55%	√	IMPAIRMENT	4.71%	21.81%
√	IMPAIRMENT	3.04%	21.59%		RESTRUCTURING	2.93%	24.74%
√	AGAINST	2.58%	24.17%	√	DECLINE	2.89%	27.62%
√	ADVERSE	2.44%	26.61%		CLAIMS	2.71%	30.33%
	RESTATED	2.09%	28.70%	√	ADVERSE	2.44%	32.77%
√	ADVERSELY	1.75%	30.45%	√	AGAINST	2.01%	34.78%
	RESTRUCTURING	1.72%	32.17%	√	ADVERSELY	1.94%	36.72%
	LITIGATION	1.67%	33.83%		LITIGATION	1.67%	38.40%
	DISCONTINUED	1.57%	35.40%		CRITICAL	1.63%	40.03%
	TERMINATION	1.35%	36.75%		DISCONTINUED	1.62%	41.64%
√	DECLINE	1.19%	37.93%	√	*DECLINED*	1.30%	42.94%
√	CLOSING	1.08%	39.01%		TERMINATION	1.06%	44.00%
√	FAILURE	0.97%	39.98%	√	NEGATIVE	0.96%	44.96%
	UNABLE	0.84%	40.82%	√	FAILURE	0.93%	45.89%
√	*DAMAGES*	0.82%	41.64%		UNABLE	0.91%	46.80%
√	DOUBTFUL	0.77%	42.41%	√	CLOSING	0.86%	47.65%
√	LIMITATIONS	0.75%	43.17%		*NONPERFORMING*	0.81%	48.47%
√	FORCE	0.74%	43.91%	√	IMPAIRED	0.81%	49.28%
√	VOLATILITY	0.73%	44.64%	√	VOLATILITY	0.79%	50.07%
	CRITICAL	0.73%	45.37%	√	FORCE	0.75%	50.82%
√	IMPAIRED	0.70%	46.07%	√	*NEGATIVELY*	0.73%	51.56%
	TERMINATED	0.70%	46.77%	√	DOUBTFUL	0.72%	52.27%
√	*COMPLAINT*	0.63%	47.39%	√	*CLOSED*	0.70%	52.97%
√	DEFAULT	0.57%	47.96%	√	DIFFICULT	0.69%	53.66%
√	NEGATIVE	0.51%	48.47%	√	*DECLINES*	0.63%	54.29%
√	*DEFENDANTS*	0.51%	48.99%	√	*EXPOSED*	0.60%	54.89%
√	*PLAINTIFFS*	0.51%	49.49%	√	DEFAULT	0.59%	55.48%
√	DIFFICULT	0.50%	50.00%	√	*DELAYS*	0.56%	56.04%

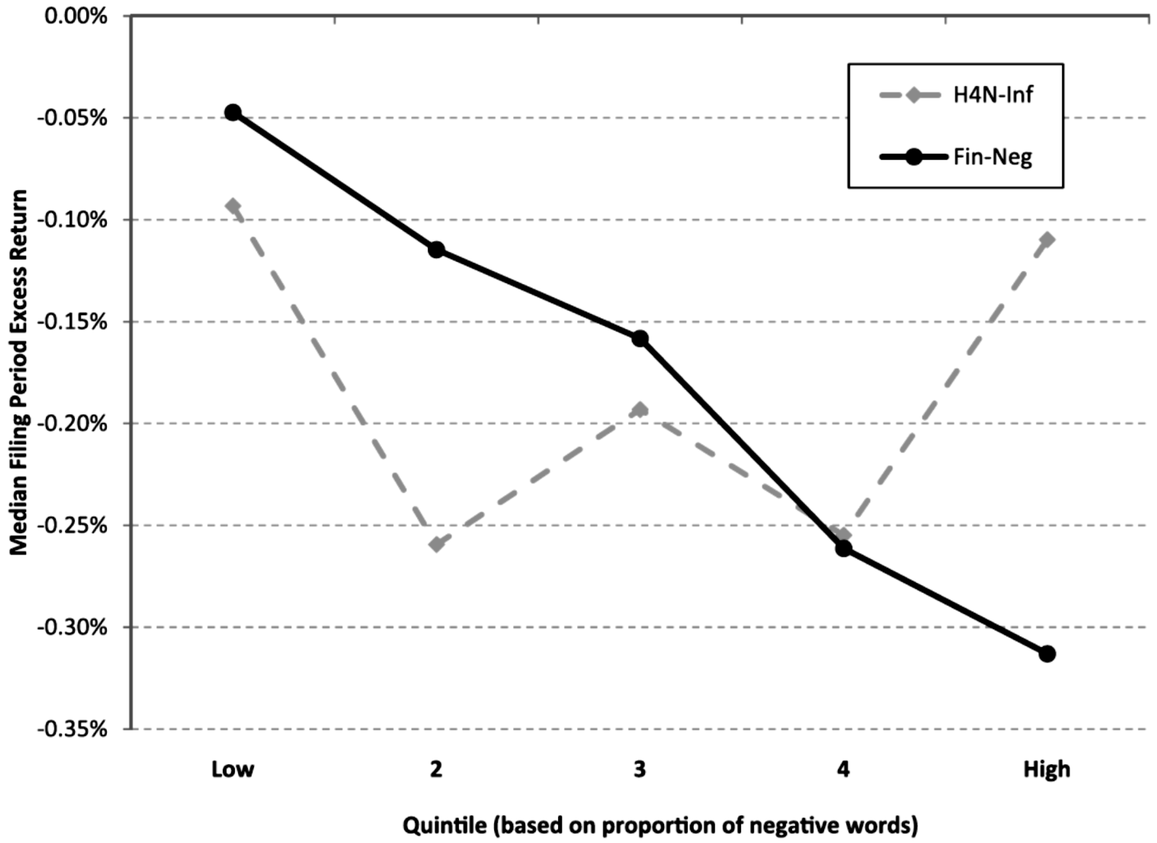
Results

- Words make intuitively sense.
- Large overlap with HVD: only 9 (8) of the 30 most frequent LMD neg. words (in the MD&A) are “new”.
  → LMD neg. is mainly constructed by dropping inappropriate HVD neg. words.

Loughran and McDonald (2011) –Table 3, part 2

Loughran and McDonald (2011) – Comparison of word lists (3)

Relation between HVD/LMD and stock returns



Discussion

- The figure shows the median 3-day market-excess return around the filing date of tone quintiles.
- As 10-Ks are informative, negativity should be negatively related to returns.
- Result
  While HVD neg. does not show a link to returns, LMD neg. is negatively related to returns.

Loughran and McDonald (2011) – Figure 1

Should you use positive words, negative words or net tone?

- Positive words often carry an ambiguous meaning.
- Real-word example: GM's 2007 annual report
  - Available at: https://www.sec.gov/Archives/edgar/data/40730/000095012408000921/k23797e10vk.htm
  - "In 2007, the global automotive industry continued to show strong sales and revenue growth." (p. 48).
  - 2007's net loss(!): $38,732 million (p. 46).

- Negative words are rarely used in an ambiguous way.

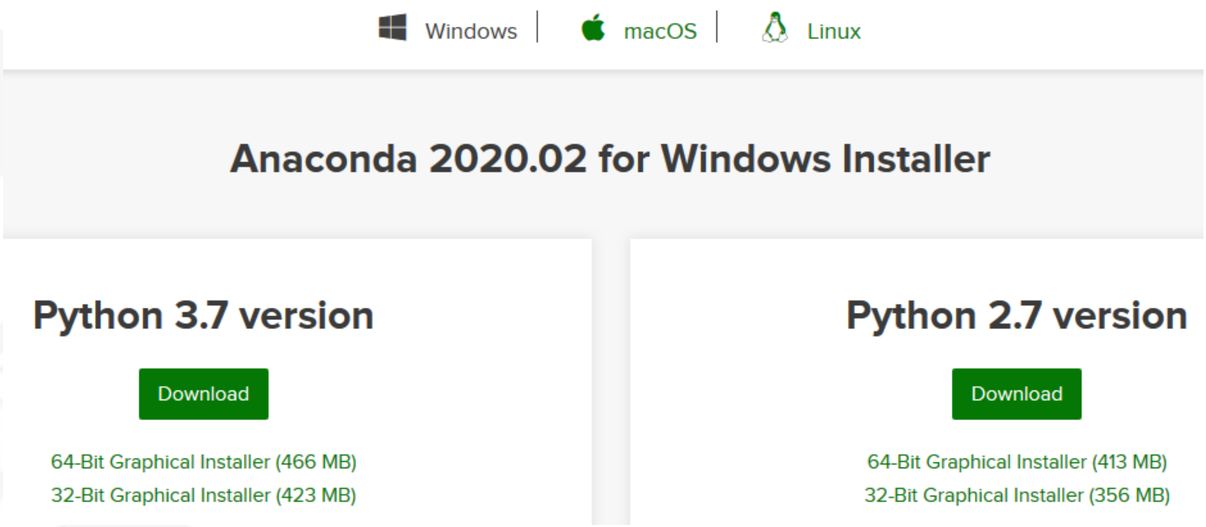
- My and Loughran and McDonald's recommendation: focus on negative words.

Agenda for our Workshop

- Background on Textual Analysis in Accounting/Finance/Economics. How to measure tone?
  - Tetlock (2007)
  - Loughran and McDonald (2011)
- Implementing your first Textual Analysis in Python
  - Installing and starting Python
  - Transcript of earnings announcement call as example
  - Programming first steps in Python
  - Helpful software

Recommendation for Python environment

- Anaconda is a popular and very convenient Python environment.
- Available: https://www.anaconda.com/distribution/

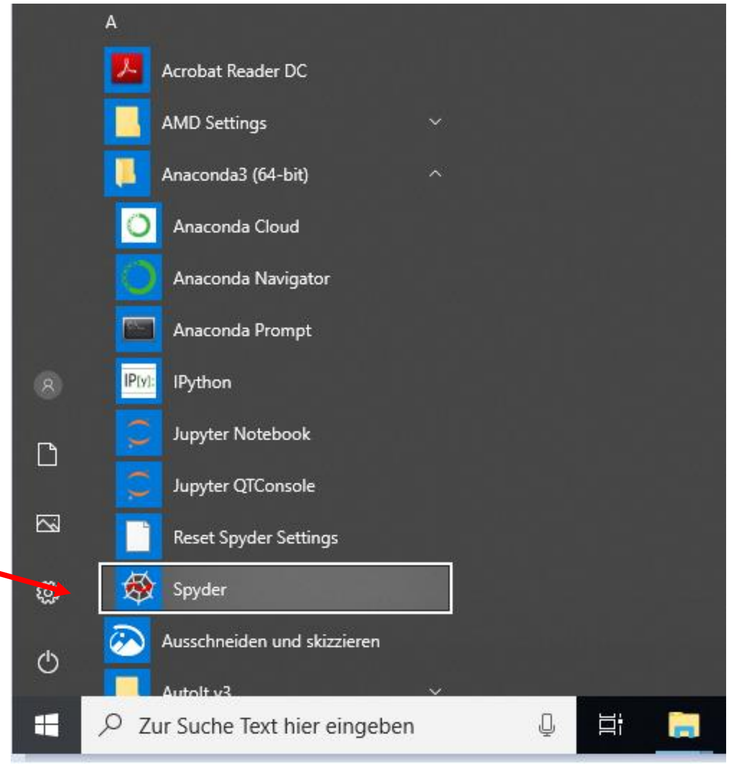


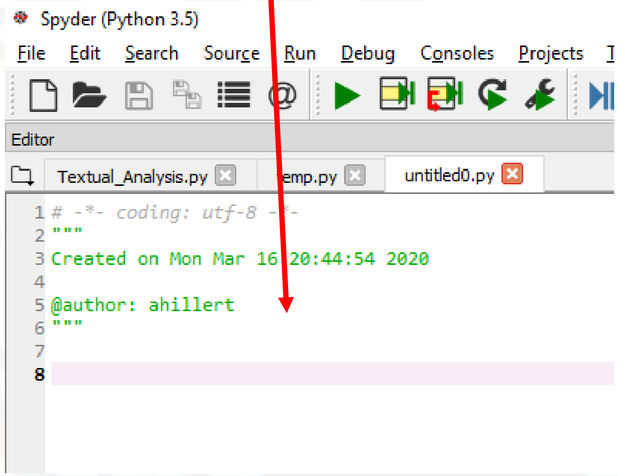
- Use Python 3.7.
- Python 2.7 no longer supported and updated.

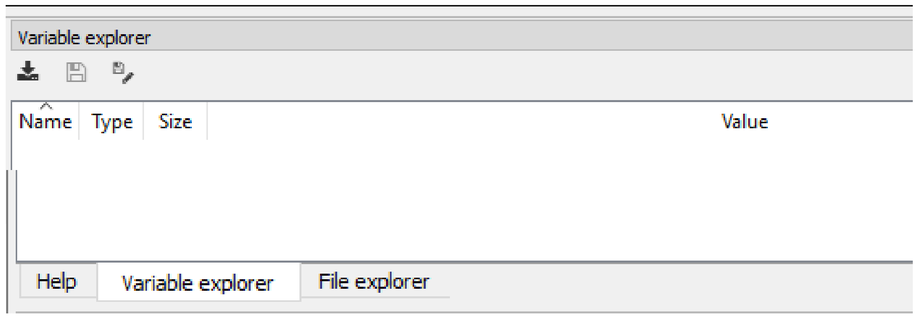
Starting Anaconda/Python

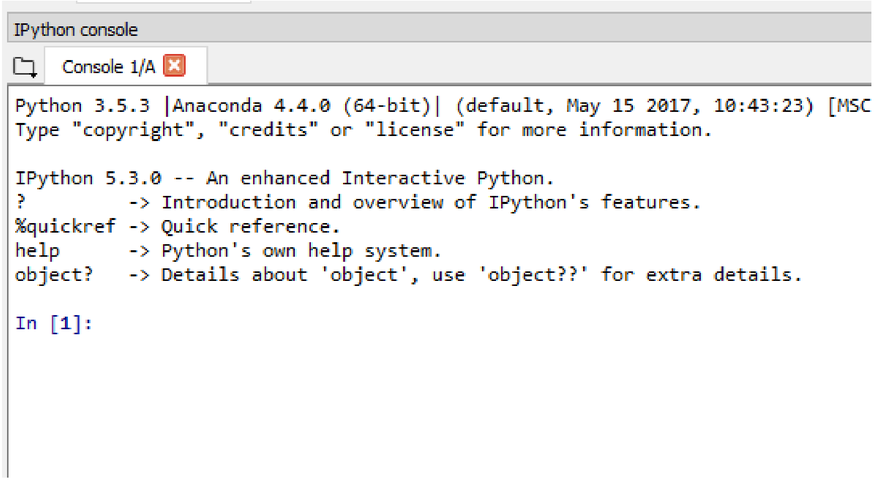
- The program is called “Spyder”.

- Three main parts
1. IPython console
2. Variable explorer
3. Programming editor









Agenda for our Workshop

- Background on Textual Analysis in Accounting/Finance/Economics. How to measure tone?
  - Tetlock (2007)
  - Loughran and McDonald (2011)
- Implementing your first Textual Analysis in Python
  - Installing and starting Python
  - Transcript of earnings announcement call as example
  - Programming first steps in Python
  - Helpful software

Text used in our programming example

- Microsoft's 2020 Q2 earnings conference call transcript.
- You find Microsoft's transcripts on their webpage: https://www.microsoft.com/en-us/investor/events/events-recent.aspx
- Direct link to 2020 Q2 document: https://view.officeapps.live.com/op/view.aspx?src=https://c.s-microsoft.com/en-us/CMSFiles/TranscriptFY20Q2.docx?version=9674fe79-64c1-95db-10c0-1015c4c70d3c

Availability of earnings conference call transcripts

- Required by Regulation FD (Fair Disclosure).
- Thomson Reuters, Seekingalpha.com, and other (commercial) data providers offer conference call transcripts.
- Some companies release transcripts on their webpage.

Getting the transcript into Python

- txt files are the best input file type in Python.
- Additional Python packages allow to import Word documents.

First steps for our textual analysis in Python

- Open transcript in Word, manually copy text and insert it into an empty txt file.
- Start Spyder.
- Start writing the program code.

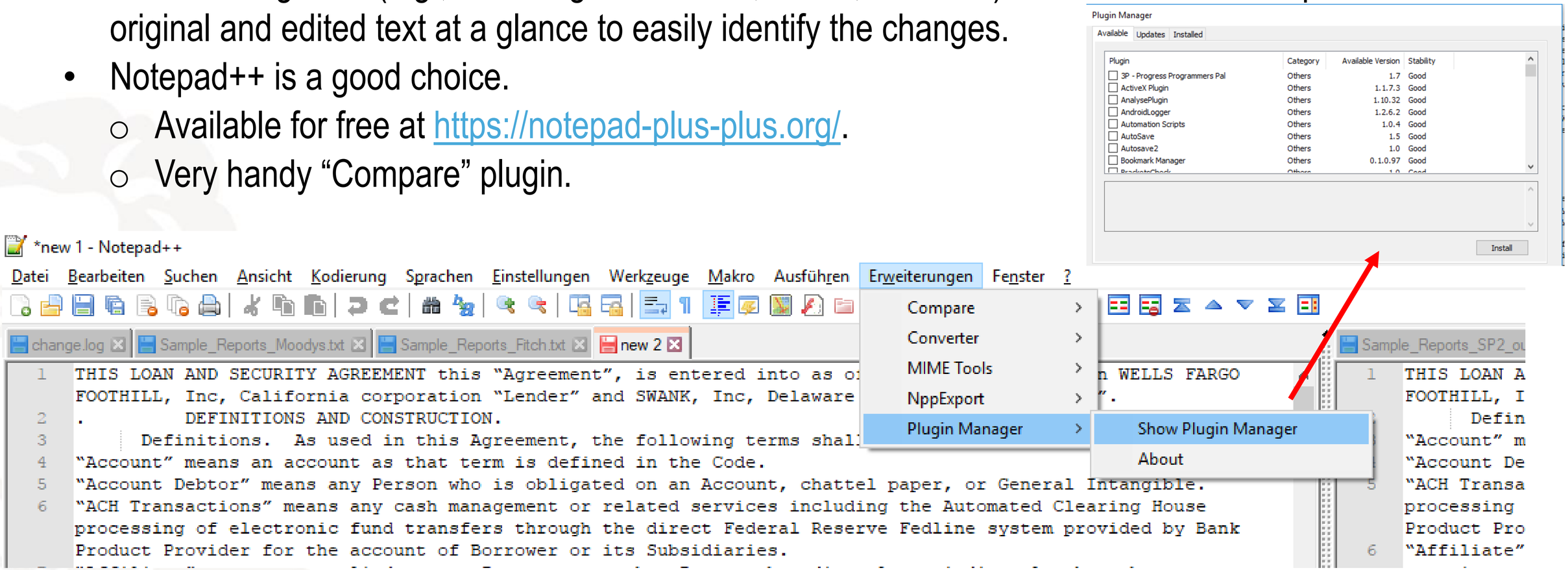
  → next section: Programming first steps in Python

Agenda for our Workshop

- Background on Textual Analysis in Accounting/Finance/Economics. How to measure tone?
  - Tetlock (2007)
  - Loughran and McDonald (2011)
- Implementing your first Textual Analysis in Python
  - Installing and starting Python
  - Transcript of earnings announcement call as example
  - Programming first steps in Python
  - Helpful software

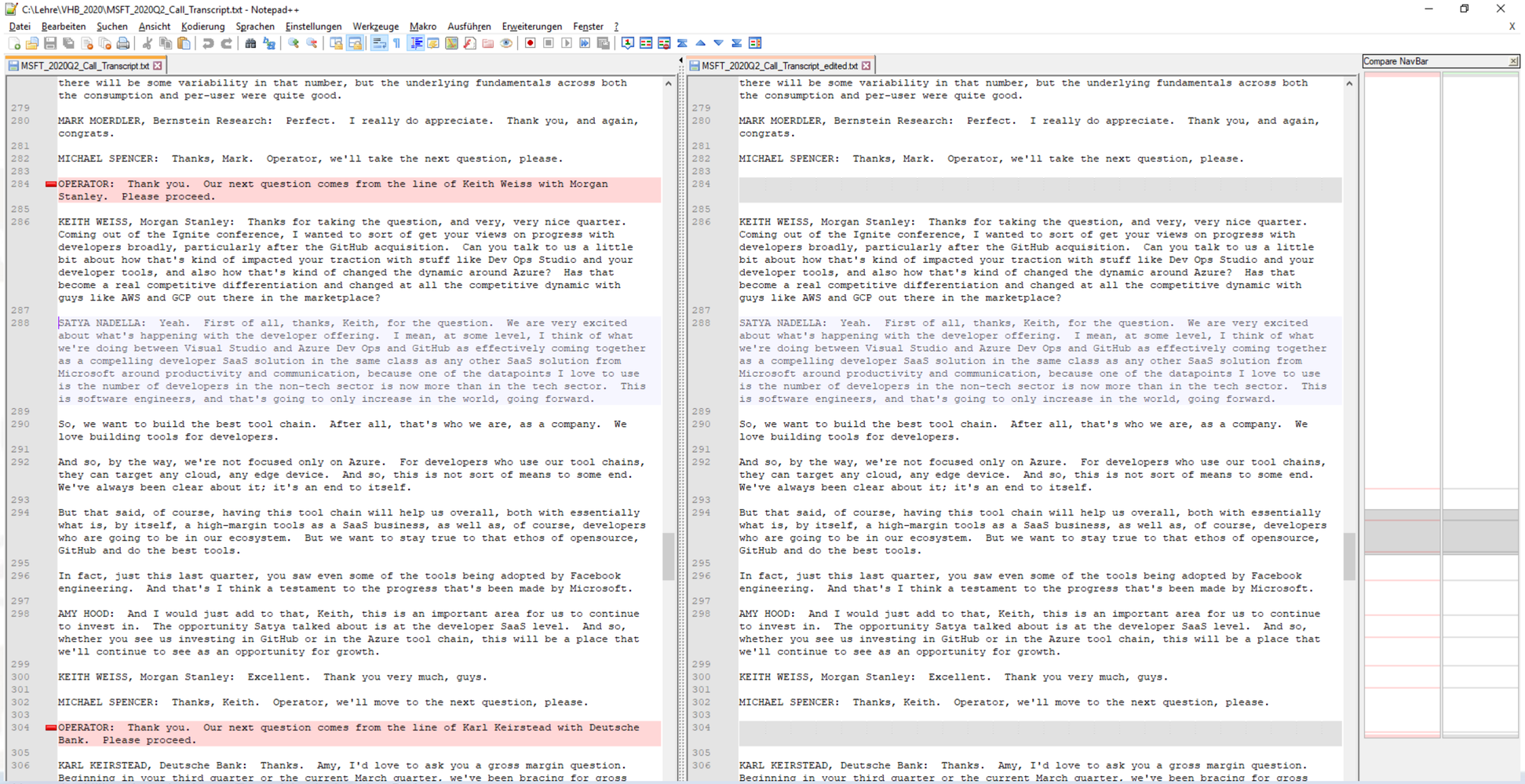
Software recommendation for text editor: Notepad++

- When editing texts (e.g., removing disclaimers, tables, numbers) we would like to compare the original and edited text at a glance to easily identify the changes.
- Notepad++ is a good choice.
  - Available for free at https://notepad-plus-plus.org/.
  - Very handy “Compare” plugin.



Helpful program – Notepad++ (2)

Compare plugin in Notepad++



GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Time for questions!

Please use the chat in the conference app.

Thank you very much for your attention!

Contact details

Prof. Dr. Alexander Hillert
Johann Wolfgang Goethe-University Frankfurt am Main
Theodor-W.-Adorno-Platz 3
60323 Frankfurt am Main
Phone: +49 (69) 798-33714
E-Mail: hillert@finance.uni-frankfurt.de