The Hebrew University Of Jerusalem

2009-2010

Text Mining Project Final Paper

Course Name

Text Mining For Business Applications

Submission Date

Thursday, April, 25th, 2010

Table of Contents

Abstract:

Our project performs sentiment analysis on financial news articles. The input is a company ticker, and the result is a score from the range [-1, 1]. The score reflects what kind of sentiment the articles have for the company**.**

The program, once it receives a specific ticker, will search for articles of that ticker in top financial news sites, such as Google Finance, Yahoo! Finance, and download articles of the past 30 days from the current date.

The downloaded articles will be cleaned and converted from HTML format to a text format, eliminating all advertisements, links, and information that are not directly related to the article itself.

Once the files are cleaned and converted from HTML to Text files, we will perform sentence splitting, part of speech tagging, and named entity recognition on all the articles downloaded.

Then, the program goes over all the sentences that refer to the ticker and checks if it has positive or negative impact on the article, these sentences or part of sentences will be marked as dark red for very bad sentiment, red for bad sentiment, green for good sentiment, and light green for very good sentiment. Based on these sentiments we will provide a score to each article.

Anaphora resolution is partially done. We noticed that most of the articles don't use "they" or "it" to refer to the company, but there is some use of "the company" to refer to the company being checked; so, we checked only if "the company" phrases refer to the company being checked. If so, then all of the sentences that contain this phrase in it are checked.

Once everything is done for all articles, the results are presented as HTML pages. The Main HTML file will include a table consisting of all companies we've worked on and the calculated score of the sentiment analysis of all articles harvested and analyzed. Each ticker will link to another file that includes all articles related to that specific ticker, with their scores. Finally, each article name will link back to the article itself, where the article's related sentences are tagged by the proper color if deemed to provide negative or positive sentiment for the company.

We have decided to implement 5 different levels of analysis instead of the regular three:

- Very Negative: score of -2
- Negative: score of -1
- Neutral: score of 0; will not affect the final score
- Positive: score of 1
- Very Positive: score of 2

The sentiment analyzer looks for patterns as well as events.

System Architecture:

Crawling:

- Contacting main financial news article hubs, such as Google Finance, or Yahoo! Finance, and downloading relevant article URLs, and their dates of the last 30 days, for each ticker, from every website, depending on input.
- For each ticker, reading the URL list from these sites and creating a list saving the URL and date of each URL as well as its title for further use by the cleaner (aka pre-processing stage).

Pre-Processing:

- Reading each HTML file, deleting irrelevant parts, such as menus, comments section, Java Script functions, CSS blocks, and advertisements.
- Converting the cleaned HTML to regular text, by removing all the HTML tags from it, and fixing spacing and paragraph splitting.

Linguistic Analysis:

- Parsing each text file, to identify relevant sentences for each company, and tagging each sentence related to the company, with appropriate score, ranging from very negative to very positive.

Scoring:

- Each document, after being analyzed and each relevant sentence tagged, a final score of all articles is calculated based on these tags. The score is from [-1, 1]. Where +1 is very good sentiment for the company in that article, and -1 is a very bad sentiment for the company in that article.
- Once each article is scanned for negative and positive sentiment, we calculate the final score of the company using these values, where events get a double effect than regular positive or negative sentences. The score is calculated using the formula taught in class.

Presentation:

- Once everything is done, the results will be viewed in an easy to see way with proper links backtracking to the articles themselves.
- Main page includes company name, company ticker, and the calculated score for the company.
- Each line will link to all articles of that specific company, where a list of article and score can be seen. There, each article name will link to the tagged article itself.

Project Modules:

- **Crawler.py**: given command line input, downloads the links to articles of ticker(s) from a specific site.
- **Cleaner.py**: links downloaded by the crawler, are downloaded, cleaned, and saved in text format.
- **SentimentAnalyzer.py**: performs sentiment analysis on the articles downloaded by the cleaner, according to patterns and events. Sentences that have sentiment for the examined ticker are tagged, and statistics of how many sentences were found of each of the 4 levels are saved. Anaphora resolution is performed for “the company” phrases.
- **Presenter.py**: calculated the final score of each ticker based on the results given by the previous section, and creates HTML files to present the results (tickers and their scores, and HTML files of the tagged articles).

Requirements:

- Python 2.6.4: http://www.python.org/ftp/python/2.6.4/python-2.6.4.msi
- BeautifulSoup v3.0.8: http://www.crummy.com/software/BeautifulSoup/download/3.x/BeautifulSoup-3.0.8.tar.gz
- Nltk package, stemmer, part of speech tagger, sentence splitter, tokenizer packages.
- companyList.txt: contains lines, in each line a pair of the format: “Ticker:CompanyName” for the 10 companies in our list:
  - MRK:Merck & Co Inc
  - C:Citigroup Inc
  - DIS:Walt Disney Co
  - NOK:Nokia
  - GS:Goldman Sachs Group Inc
  - POT:Potash Corp of Saskatchewan Inc
  - KMB:Kimberly-Clark Corp
  - LVS:Las Vegas Sands Corp
  - INTC:Intel Corp
  - AA:Alcoa Inc

NOTE: The snapshots in this file are not always of our final results. Final results can be viewed through the index.html file found under: project/present/. A screen dump of running the project is found in project.log

Crawler.py:

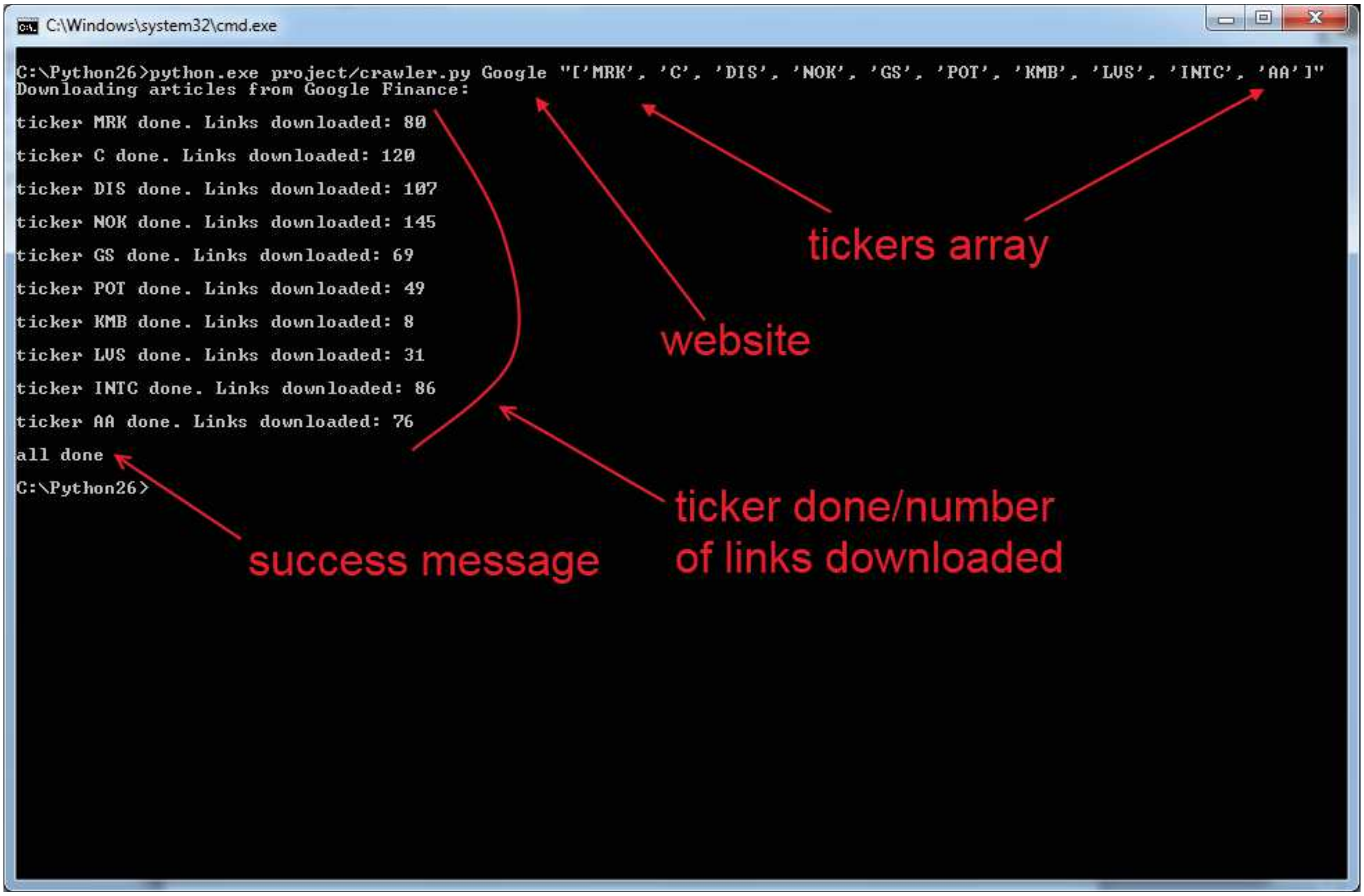
- Command line: from outside project folder
    - Python project/crawler.py [search engine] [“[‘ticker1’, ‘ticker2’, …, ‘tickerN’]”]
- Where:
    - search engine = Google Finance, Yahoo! Finance
    - ticker = MRK, C,
- The crawler depending on input, will contact the requested site, and download the ticker(s) links to articles of the last 30 days, starting from the current date.
- The crawler will harvest the article links as well as their dates.

Usage:

- python project/crawler.py <financeSite> <Ticker list>.
- financeSite is either Google or Yahoo.
- Ticker list is of the format: “[‘Ticker, ‘Ticker, ‘Ticker’]”.

Example:

- Downloading article list from Google Finance, for given tickers.

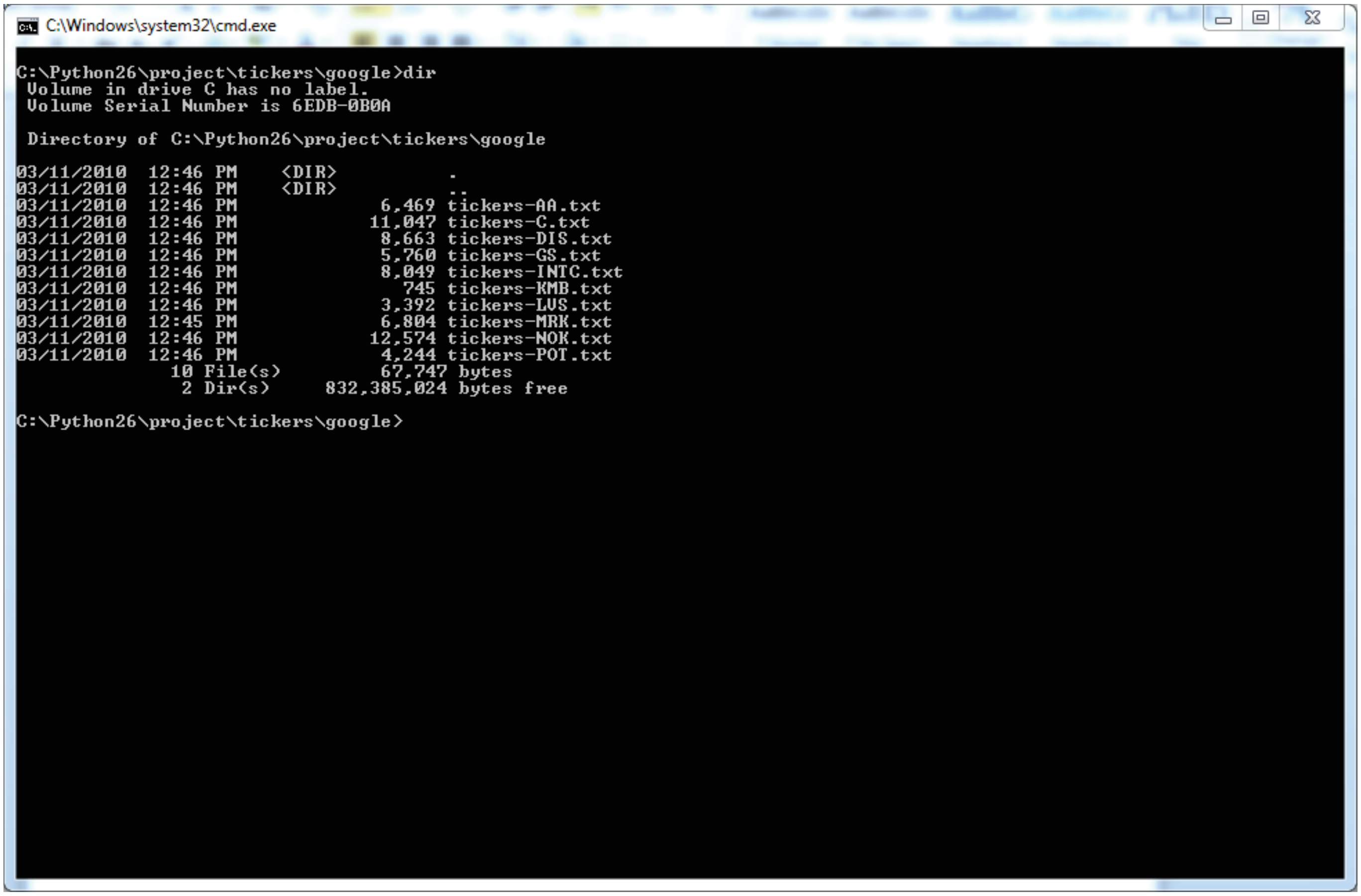


C:\Windows\system32\cmd.exe

```
C:\Python26>python.exe project/crawler.py Google "['MRK', 'C', 'DIS', 'NOK', 'GS', 'POT', 'KMB', 'LVS', 'INTC', 'AA']"
Downloading articles from Google Finance:

ticker MRK done. Links downloaded: 80

ticker C done. Links downloaded: 120

ticker DIS done. Links downloaded: 107

ticker NOK done. Links downloaded: 145

ticker GS done. Links downloaded: 69

ticker POT done. Links downloaded: 49

ticker KMB done. Links downloaded: 8

ticker LVS done. Links downloaded: 31

ticker INTC done. Links downloaded: 86

ticker AA done. Links downloaded: 76

all done

C:\Python26>
```

Result:

- A special folder named Google; inside it each ticker has its own file where the links of the articles are stored.

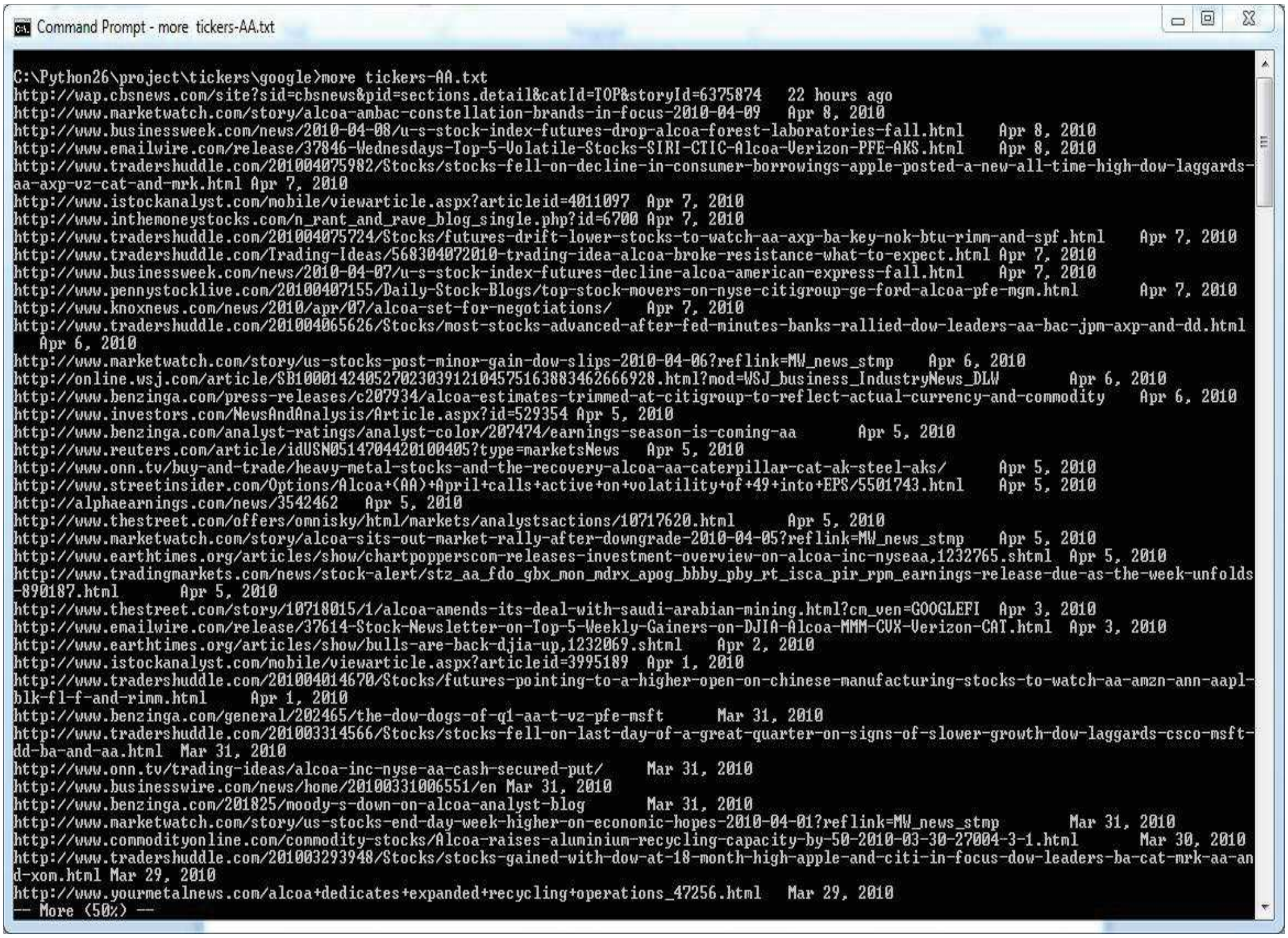


```
C:\Python26\project\tickers\google>dir
 Volume in drive C has no label.
 Volume Serial Number is 6EDB-0B0A

 Directory of C:\Python26\project\tickers\google

03/11/2010  12:46 PM    <DIR>          .
03/11/2010  12:46 PM    <DIR>          ..
03/11/2010  12:46 PM             6,469 tickers-AA.txt
03/11/2010  12:46 PM            11,047 tickers-C.txt
03/11/2010  12:46 PM             8,663 tickers-DIS.txt
03/11/2010  12:46 PM             5,760 tickers-GS.txt
03/11/2010  12:46 PM             8,049 tickers-INTC.txt
03/11/2010  12:46 PM               745 tickers-KMB.txt
03/11/2010  12:46 PM             3,392 tickers-LVS.txt
03/11/2010  12:45 PM             6,804 tickers-MRK.txt
03/11/2010  12:46 PM            12,574 tickers-NOK.txt
03/11/2010  12:46 PM             4,244 tickers-POT.txt
              10 File(s)         67,747 bytes
               2 Dir(s)     832,385,024 bytes free

C:\Python26\project\tickers\google>
```

- Each file consists of lines, where each line has a link to an article, and its date.

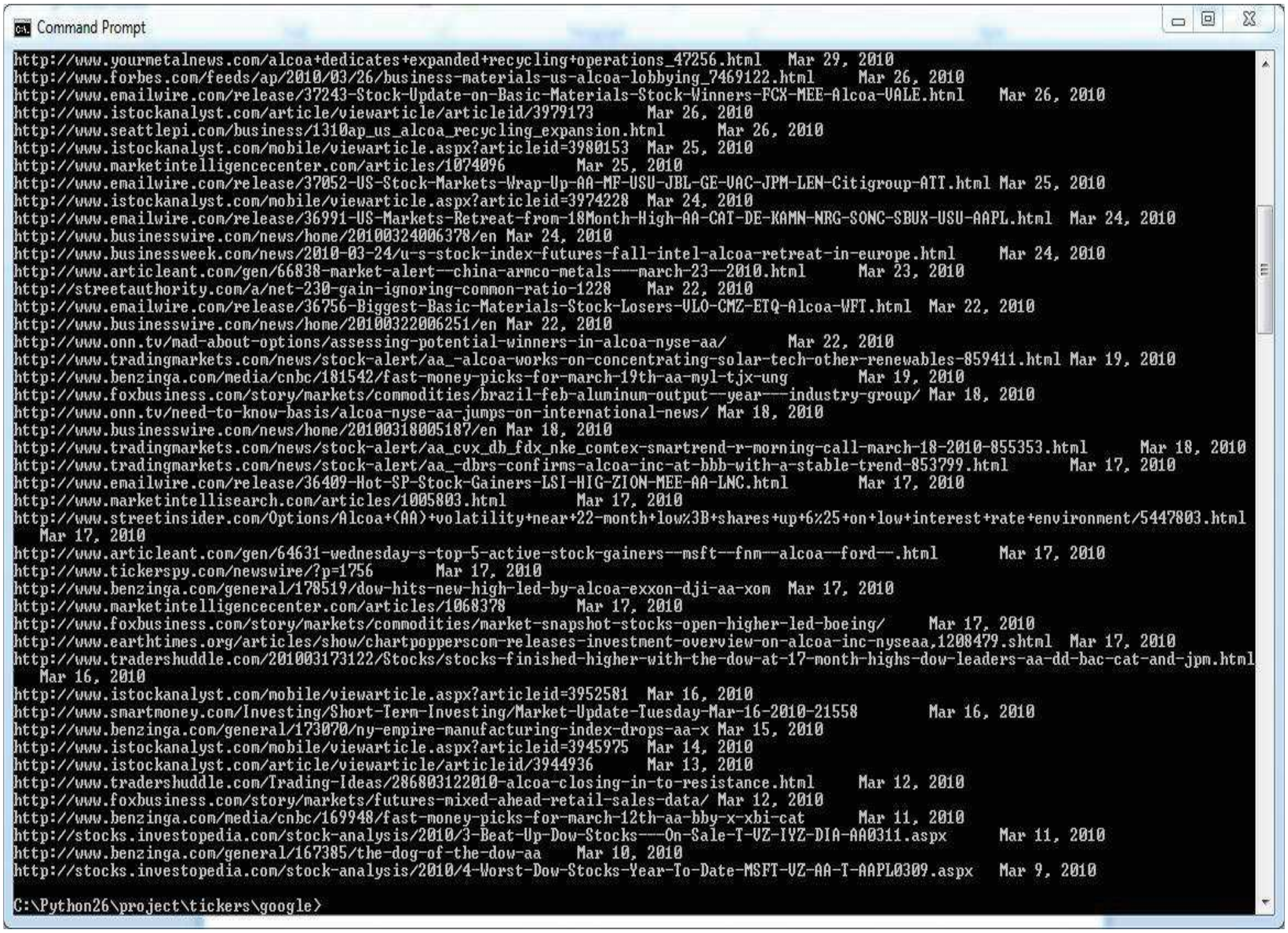
- Example File Content:
  - Let’s have a look at the AA file.

File Part 1:



```
C:\Python26\project\tickers\google>more tickers-AA.txt
http://wap.cbsnews.com/site?sid=cbsnews&pid=sections.detail&catId=TOP&storyId=6375874   22 hours ago
http://www.marketwatch.com/story/alcoa-ambac-constellation-brands-in-focus-2010-04-09   Apr 8, 2010
http://www.businessweek.com/news/2010-04-08/u-s-stock-index-futures-drop-alcoa-forest-laboratories-fall.html    Apr 8, 2010
http://www.emailwire.com/release/37846-Wednesdays-Top-5-Volatile-Stocks-SIRI-CTIC-Alcoa-Verizon-PFE-AKS.html    Apr 8, 2010
http://www.tradershuddle.com/201004075982/Stocks/stocks-fell-on-decline-in-consumer-borrowings-apple-posted-a-new-all-time-high-dow-laggards-
aa-axp-vz-cat-and-mrk.html Apr 7, 2010
http://www.istockanalyst.com/mobile/viewarticle.aspx?articleid=4011097  Apr 7, 2010
http://www.inthemoneystocks.com/n_rant_and_rave_blog_single.php?id=6700 Apr 7, 2010
http://www.tradershuddle.com/201004075724/Stocks/futures-drift-lower-stocks-to-watch-aa-axp-ba-key-nok-btu-rimm-and-spf.html    Apr 7, 2010
http://www.tradershuddle.com/Trading-Ideas/568304072010-trading-idea-alcoa-broke-resistance-what-to-expect.html Apr 7, 2010
http://www.businessweek.com/news/2010-04-07/u-s-stock-index-futures-decline-alcoa-american-express-fall.html    Apr 7, 2010
http://www.pennystocklive.com/20100407155/Daily-Stock-Blogs/top-stock-movers-on-nyse-citigroup-ge-ford-alcoa-pfe-mgm.html       Apr 7, 2010
http://www.knoxnews.com/news/2010/apr/07/alcoa-set-for-negotiations/    Apr 7, 2010
http://www.tradershuddle.com/201004065626/Stocks/most-stocks-advanced-after-fed-minutes-banks-rallied-dow-leaders-aa-bac-jpm-axp-and-dd.html
   Apr 6, 2010
http://www.marketwatch.com/story/us-stocks-post-minor-gain-dow-slips-2010-04-06?reflink=MW_news_stmp    Apr 6, 2010
http://online.wsj.com/article/SB10001424052702303912104575163883462666928.html?mod=WSJ_business_IndustryNews_DLW        Apr 6, 2010
http://www.benzinga.com/press-releases/c207934/alcoa-estimates-trimmed-at-citigroup-to-reflect-actual-currency-and-commodity    Apr 6, 2010
http://www.investors.com/NewsAndAnalysis/Article.aspx?id=529354 Apr 5, 2010
http://www.benzinga.com/analyst-ratings/analyst-color/207474/earnings-season-is-coming-aa       Apr 5, 2010
http://www.reuters.com/article/idUSN0514704420100405?type=marketsNews   Apr 5, 2010
http://www.onn.tv/buy-and-trade/heavy-metal-stocks-and-the-recovery-alcoa-aa-caterpillar-cat-ak-steel-aks/      Apr 5, 2010
http://www.streetinsider.com/Options/Alcoa+(AA)+April+calls+active+on+volatility+of+49+into+EPS/5501743.html    Apr 5, 2010
http://alphaearnings.com/news/3542462   Apr 5, 2010
http://www.thestreet.com/offers/omnisky/html/markets/analystsactions/10717620.html      Apr 5, 2010
http://www.marketwatch.com/story/alcoa-sits-out-market-rally-after-downgrade-2010-04-05?reflink=MW_news_stmp    Apr 5, 2010
http://www.earthtimes.org/articles/show/chartpopperscom-releases-investment-overview-on-alcoa-inc-nyseaa,1232765.shtml  Apr 5, 2010
http://www.tradingmarkets.com/news/stock-alert/stz_aa_fdo_gbx_mon_mdrx_apog_bbby_pby_rt_isca_pir_rpm_earnings-release-due-as-the-week-unfolds
-890187.html       Apr 5, 2010
http://www.thestreet.com/story/10718015/1/alcoa-amends-its-deal-with-saudi-arabian-mining.html?cm_ven=GOOGLEFI  Apr 3, 2010
http://www.emailwire.com/release/37614-Stock-Newsletter-on-Top-5-Weekly-Gainers-on-DJIA-Alcoa-MMM-CVX-Verizon-CAT.html  Apr 3, 2010
http://www.earthtimes.org/articles/show/bulls-are-back-djia-up,1232069.shtml    Apr 2, 2010
http://www.istockanalyst.com/mobile/viewarticle.aspx?articleid=3995189  Apr 1, 2010
http://www.tradershuddle.com/201004014670/Stocks/futures-pointing-to-a-higher-open-on-chinese-manufacturing-stocks-to-watch-aa-amzn-ann-aapl-
blk-fl-f-and-rimm.html     Apr 1, 2010
http://www.benzinga.com/general/202465/the-dow-dogs-of-q1-aa-t-vz-pfe-msft      Mar 31, 2010
http://www.tradershuddle.com/201003314566/Stocks/stocks-fell-on-last-day-of-a-great-quarter-on-signs-of-slower-growth-dow-laggards-csco-msft-
dd-ba-and-aa.html  Mar 31, 2010
http://www.onn.tv/trading-ideas/alcoa-inc-nyse-aa-cash-secured-put/     Mar 31, 2010
http://www.businesswire.com/news/home/20100331006551/en Mar 31, 2010
http://www.benzinga.com/201825/moody-s-down-on-alcoa-analyst-blog       Mar 31, 2010
http://www.marketwatch.com/story/us-stocks-end-day-week-higher-on-economic-hopes-2010-04-01?reflink=MW_news_stmp        Mar 31, 2010
http://www.commodityonline.com/commodity-stocks/Alcoa-raises-aluminium-recycling-capacity-by-50-2010-03-30-27004-3-1.html       Mar 30, 2010
http://www.tradershuddle.com/201003293948/Stocks/stocks-gained-with-dow-at-18-month-high-apple-and-citi-in-focus-dow-leaders-ba-cat-mrk-aa-an
d-xom.html Mar 29, 2010
http://www.yourmetalnews.com/alcoa+dedicates+expanded+recycling+operations_47256.html   Mar 29, 2010
-- More (50%) --
```

File Part 2:

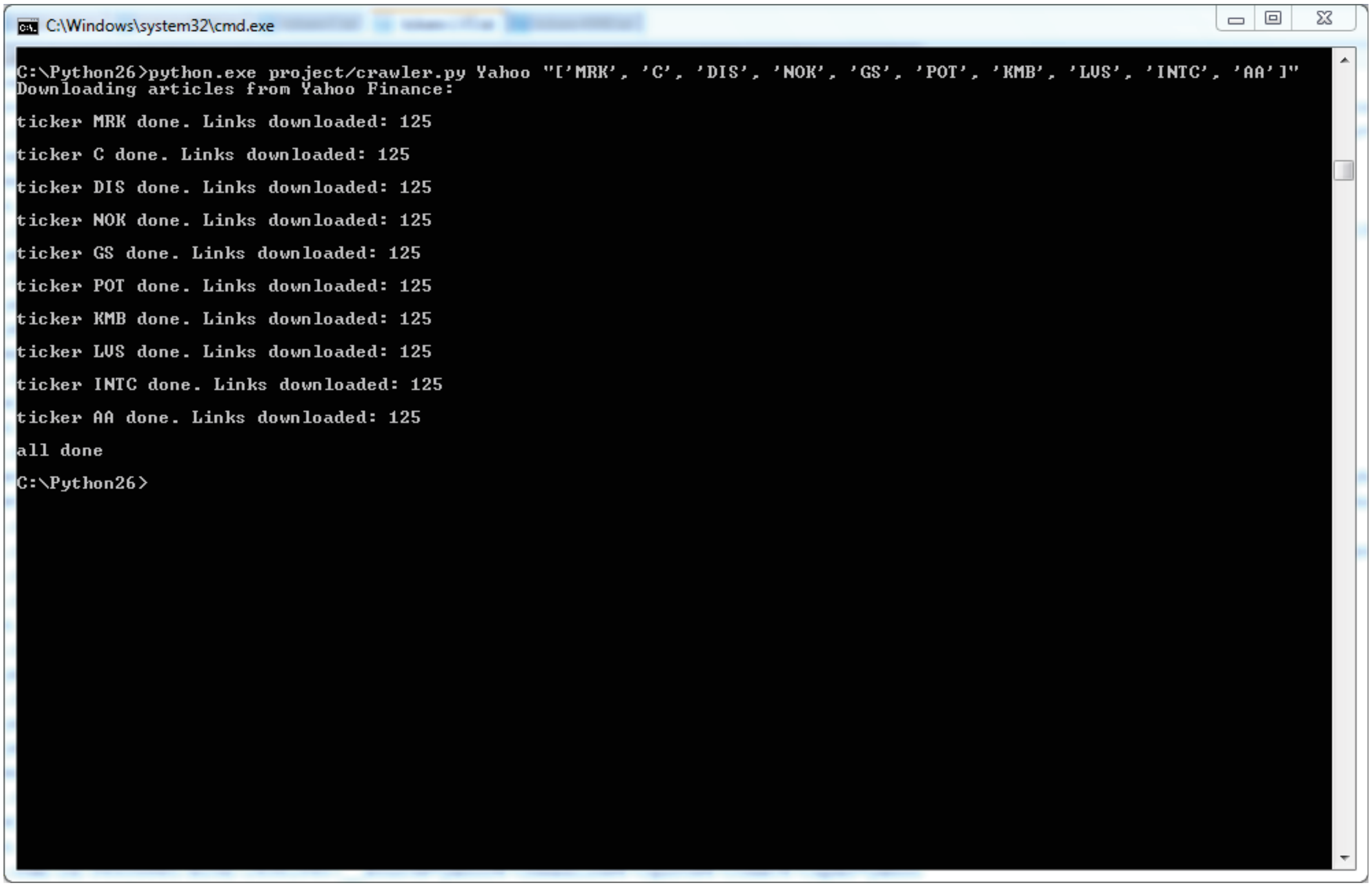


```
http://www.yourmetalnews.com/alcoa+dedicates+expanded+recycling+operations_47256.html   Mar 29, 2010
http://www.forbes.com/feeds/ap/2010/03/26/business-materials-us-alcoa-lobbying_7469122.html     Mar 26, 2010
http://www.emailwire.com/release/37243-Stock-Update-on-Basic-Materials-Stock-Winners-FCX-MEE-Alcoa-VALE.html    Mar 26, 2010
http://www.istockanalyst.com/article/viewarticle/articleid/3979173      Mar 26, 2010
http://www.seattlepi.com/business/1310ap_us_alcoa_recycling_expansion.html      Mar 26, 2010
http://www.istockanalyst.com/mobile/viewarticle.aspx?articleid=3980153  Mar 25, 2010
http://www.marketintelligencecenter.com/articles/1074096        Mar 25, 2010
http://www.emailwire.com/release/37052-US-Stock-Markets-Wrap-Up-AA-MF-USU-JBL-GE-VAC-JPM-LEN-Citigroup-ATT.html Mar 25, 2010
http://www.istockanalyst.com/mobile/viewarticle.aspx?articleid=3974228  Mar 24, 2010
http://www.emailwire.com/release/36991-US-Markets-Retreat-from-18Month-High-AA-CAT-DE-KAMN-NRG-SONC-SBUX-USU-AAPL.html  Mar 24, 2010
http://www.businesswire.com/news/home/20100324006378/en Mar 24, 2010
http://www.businessweek.com/news/2010-03-24/u-s-stock-index-futures-fall-intel-alcoa-retreat-in-europe.html     Mar 24, 2010
http://www.articleant.com/gen/66838-market-alert--china-armco-metals---march-23--2010.html      Mar 23, 2010
http://streetauthority.com/a/net-230-gain-ignoring-common-ratio-1228    Mar 22, 2010
http://www.emailwire.com/release/36756-Biggest-Basic-Materials-Stock-Losers-VLO-CMZ-ETQ-Alcoa-WFT.html  Mar 22, 2010
http://www.businesswire.com/news/home/20100322006251/en Mar 22, 2010
http://www.onn.tv/mad-about-options/assessing-potential-winners-in-alcoa-nyse-aa/       Mar 22, 2010
http://www.tradingmarkets.com/news/stock-alert/aa_-alcoa-works-on-concentrating-solar-tech-other-renewables-859411.html Mar 19, 2010
http://www.benzinga.com/media/cnbc/181542/fast-money-picks-for-march-19th-aa-myl-tjx-ung        Mar 19, 2010
http://www.foxbusiness.com/story/markets/commodities/brazil-feb-aluminum-output--year---industry-group/ Mar 18, 2010
http://www.onn.tv/need-to-know-basis/alcoa-nyse-aa-jumps-on-international-news/ Mar 18, 2010
http://www.businesswire.com/news/home/20100318005187/en Mar 18, 2010
http://www.tradingmarkets.com/news/stock-alert/aa_cvx_db_fdx_nke_comtex-smartrend-r-morning-call-march-18-2010-855353.html      Mar 18, 2010
http://www.tradingmarkets.com/news/stock-alert/aa_-dbrs-confirms-alcoa-inc-at-bbb-with-a-stable-trend-853799.html       Mar 17, 2010
http://www.emailwire.com/release/36409-Hot-SP-Stock-Gainers-LSI-HIG-ZION-MEE-AA-LNC.html        Mar 17, 2010
http://www.marketintellisearch.com/articles/1005803.html        Mar 17, 2010
http://www.streetinsider.com/Options/Alcoa+(AA)+volatility+near+22-month+low%3B+shares+up+6%25+on+low+interest+rate+environment/5447803.html
   Mar 17, 2010
http://www.articleant.com/gen/64631-wednesday-s-top-5-active-stock-gainers--msft--fnm--alcoa--ford--.html       Mar 17, 2010
http://www.tickerspy.com/newswire/?p=1756       Mar 17, 2010
http://www.benzinga.com/general/178519/dow-hits-new-high-led-by-alcoa-exxon-dji-aa-xom  Mar 17, 2010
http://www.marketintelligencecenter.com/articles/1068378        Mar 17, 2010
http://www.foxbusiness.com/story/markets/commodities/market-snapshot-stocks-open-higher-led-boeing/     Mar 17, 2010
http://www.earthtimes.org/articles/show/chartpopperscom-releases-investment-overview-on-alcoa-inc-nyseaa,1208479.shtml  Mar 17, 2010
http://www.tradershuddle.com/201003173122/Stocks/stocks-finished-higher-with-the-dow-at-17-month-highs-dow-leaders-aa-dd-bac-cat-and-jpm.html
   Mar 16, 2010
http://www.istockanalyst.com/mobile/viewarticle.aspx?articleid=3952581  Mar 16, 2010
http://www.smartmoney.com/Investing/Short-Term-Investing/Market-Update-Tuesday-Mar-16-2010-21558        Mar 16, 2010
http://www.benzinga.com/general/173070/ny-empire-manufacturing-index-drops-aa-x Mar 15, 2010
http://www.istockanalyst.com/mobile/viewarticle.aspx?articleid=3945975  Mar 14, 2010
http://www.istockanalyst.com/article/viewarticle/articleid/3944936      Mar 13, 2010
http://www.tradershuddle.com/Trading-Ideas/286803122010-alcoa-closing-in-to-resistance.html     Mar 12, 2010
http://www.foxbusiness.com/story/markets/futures-mixed-ahead-retail-sales-data/ Mar 12, 2010
http://www.benzinga.com/media/cnbc/169948/fast-money-picks-for-march-12th-aa-bby-x-xbi-cat      Mar 11, 2010
http://stocks.investopedia.com/stock-analysis/2010/3-Beat-Up-Dow-Stocks---On-Sale-T-VZ-IYZ-DIA-AA0311.aspx      Mar 11, 2010
http://www.benzinga.com/general/167385/the-dog-of-the-dow-aa    Mar 10, 2010
http://stocks.investopedia.com/stock-analysis/2010/4-Worst-Dow-Stocks-Year-To-Date-MSFT-VZ-AA-T-AAPL0309.aspx   Mar 9, 2010

C:\Python26\project\tickers\google>
```

- Total links counted: 84.

Example 2:

- Downloading from Yahoo! Finance:



```
C:\Python26>python.exe project/crawler.py Yahoo "['MRK', 'C', 'DIS', 'NOK', 'GS', 'POT', 'KMB', 'LVS', 'INTC', 'AA']"
Downloading articles from Yahoo Finance:

ticker MRK done. Links downloaded: 125

ticker C done. Links downloaded: 125

ticker DIS done. Links downloaded: 125

ticker NOK done. Links downloaded: 125

ticker GS done. Links downloaded: 125

ticker POT done. Links downloaded: 125

ticker KMB done. Links downloaded: 125

ticker LVS done. Links downloaded: 125

ticker INTC done. Links downloaded: 125

ticker AA done. Links downloaded: 125

all done

C:\Python26>
```

- Same as in Google Finance example, each ticker will have its own file, in Yahoo directory, with the links included.
- Note: 125 links per ticker is not a mistake. We've double checked them; each file contains 125 links exactly.

Cleaner.py:

- **Pre-Cleaning stage**: The files created by the crawler are read; each link is downloaded, and saved into an HTML file for processing.
- **First** thing we have done is preprocessing the source document, simplifying the DOM tree; this is done to reduce the number of places where the cleaning process can go wrong.
- **Then**, the process of cleaning links, menus, advertisements and other garbage is done on the result.
- **Third**, all the HTML tags are stripped and the result is copied to a text file.
- **The cleanup process** takes each ticker, and for each ticker folder contains the articles list of that ticker, reads each line – the line contains a link, and then downloads it, cleans it, and strips it from all HTML tags, and saves it in a text file.
- This process is done for all tickers, and for all the links saved by the crawler.

Simplification Process:

- If there is a link (“a href”) tag inside a paragraph (“p”) tag, they are merged together, removing the link, but keeping its caption.
- If there are formatting tags, such as font (“font”), and span (“span”) inside a table data cell (“td”) tag, they are merged together. Removing the formatting inside the table data cell tag.
- If there are div (“div”) or span (“span”) tags inside a paragraph (“p”) tag, these tags are removed to increase simplicity.
- Tags such as “strong”, “em”, “i”, and “b” are removed from the file, since they are merely formatting tags that do not affect the HTML file structure.

Cleanup Process: (links/menus/advertisements/other garbage)

- All JavaScript (“script”) tags, comments tags (“<!—“), and CSS (“style”) tags are removed.
- Any text tag consists of at least 20 words, is kept, while the rest are deleted, this will keep the main article intact while deleting the scattered unrelated words from the page around the article, such as add a comment section, and copyright sentences.
- Since the article title is normally less than 20 words, and normally contains the company name in it, we added an exception to this case: Since we are given the company tag, we harvested the company name from MSN Money website, and checked if the article title either contains the ticker or the company name, if so, the paragraph is kept – this will also allow small paragraphs that does include the company name but are too short, from getting removed.
- Any table row (‘tr’) and table data cell (‘td’) tag that contains 4 or more links – as direct children, is removed. Since this is most likely a menu, an advertisement, or just a bunch of unneeded links.

- Any unordered list ('ul') tag containing 4 or more links – direct children or just a descendant, is removed.
- Tags with id as "footer.*", ".*footer", "comment.*" or ".*comment" are removed completely as well.
- After deleting the garbage around the article, the HTML tags are removed and spacing fixed. The result is saved in a text file.

Logging:

- For each ticker a log file is created, named:< Ticker>_logFile.txt that contains the following information:
    - Stored article file name.
    - Article link.
    - Article title.
    - Article date.
- These log files are later used to create the presentation files with the results.

Note:

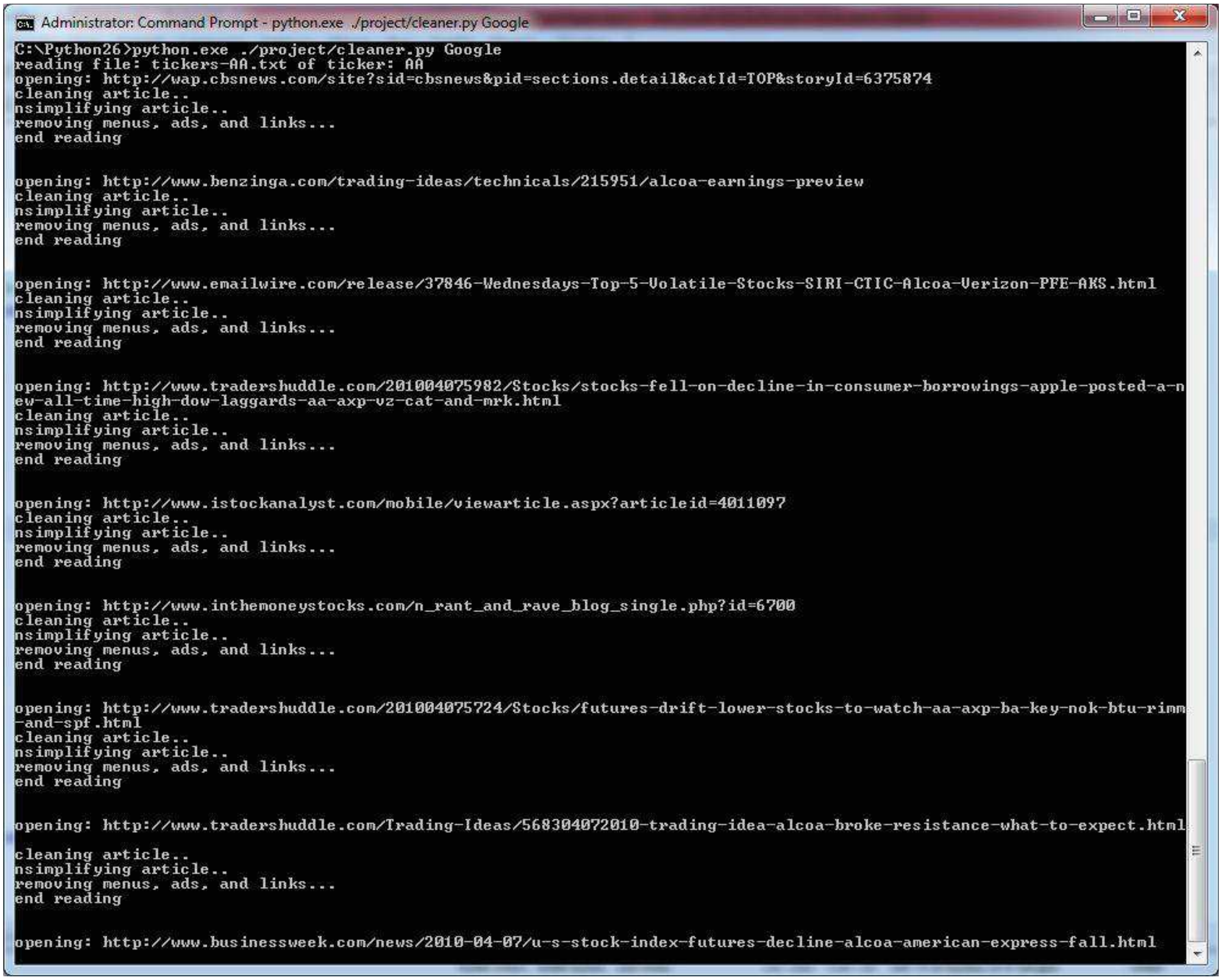
- There was a lot of thought and research time done into creating the cleaner, we did not really find any useful tools to serve our purpose; so we've built our cleaner from scratch. The results are quite good – but not perfect, but they still allow us to do the sentiment analysis without interference – even with some garbage text stayed in the text file, the sentiment analysis process will almost always ignore these sentences, since they have no relation to the company at all.

Usage:

- python project/cleaner.py <financeSite>
- financeSite is either Google or Yahoo
- The cleaner will download the articles harvested by the crawler and cleans them.

Example:

- All articles downloaded from Google finance in the example above are to be cleaned:



Administrator: Command Prompt - python.exe ./project/cleaner.py Google

```
C:\Python26>python.exe ./project/cleaner.py Google
reading file: tickers-AA.txt of ticker: AA
opening: http://wap.cbsnews.com/site?sid=cbsnews&pid=sections.detail&catId=TOP&storyId=6375874
cleaning article..
nsimplifying article..
removing menus, ads, and links...
end reading


opening: http://www.benzinga.com/trading-ideas/technicals/215951/alcoa-earnings-preview
cleaning article..
nsimplifying article..
removing menus, ads, and links...
end reading


opening: http://www.emailwire.com/release/37846-Wednesdays-Top-5-Volatile-Stocks-SIRI-CTIC-Alcoa-Verizon-PFE-AKS.html
cleaning article..
nsimplifying article..
removing menus, ads, and links...
end reading


opening: http://www.tradershuddle.com/201004075982/Stocks/stocks-fell-on-decline-in-consumer-borrowings-apple-posted-a-n
ew-all-time-high-dow-laggards-aa-axp-vz-cat-and-mrk.html
cleaning article..
nsimplifying article..
removing menus, ads, and links...
end reading


opening: http://www.istockanalyst.com/mobile/viewarticle.aspx?articleid=4011097
cleaning article..
nsimplifying article..
removing menus, ads, and links...
end reading


opening: http://www.inthemoneystocks.com/n_rant_and_rave_blog_single.php?id=6700
cleaning article..
nsimplifying article..
removing menus, ads, and links...
end reading


opening: http://www.tradershuddle.com/201004075724/Stocks/futures-drift-lower-stocks-to-watch-aa-axp-ba-key-nok-btu-rimm
-and-spf.html
cleaning article..
nsimplifying article..
removing menus, ads, and links...
end reading


opening: http://www.tradershuddle.com/Trading-Ideas/568304072010-trading-idea-alcoa-broke-resistance-what-to-expect.html

cleaning article..
nsimplifying article..
removing menus, ads, and links...
end reading


opening: http://www.businessweek.com/news/2010-04-07/u-s-stock-index-futures-decline-alcoa-american-express-fall.html
```

- Until all articles are cleaned and transformed to text format.

sentimentAnalyzer.py:

- This stage is done after the cleaner. Read company ticker and name from companyList.txt, found under project/; for each ticker, read its cleaned articles found under: project/articles/<ticker name>_articles, perform sentiment analysis for each article, and write the results in <ticker name>_scores.txt.
- **Sentiment Analysis:**
    - Load dictionaries into memory. See tools below for the dictionaries we used.
    - Read article and split it into sentences.
    - Tag words with part of speech tagging.
    - Tag entities with named entity recognition tagging, such as: ORGANIZATION, PERSON etc.
    - For each sentence that contains the name of the company/ticker, contains a "the company" phrase which refers to the ticker:
        - Extract phrases to evaluate them:
            - "up" and "down" phrases, using the following regular expression: {<RB><TO>?<CD>}'.
            - Verb phrases together with their predicate: {<JJR>?<V.*><CD>?(<RB>|<TO>)*<V.*>*(<IN>?(<PRP.*>|<DT>)?<RB>?<CD>?<JJ.*>*<N.*>?<N.*>?<N.*>?<JJ.*>*)*}.
            - Noun phrases that were not extracted with a verb, in the previous section.
              {<DT>?<RB>*<N.*>?<N.*>?<N.*>?<CD>?(<JJ.*>+<CD>?<N.*>?<N.*>?<N.*>?<N.*>?)|(<JJ.*>*<CD>?<N.*><N.*>?<N.*>?<N.*>?)}
              We restricted it to maximum 5 nouns since some sentence included a huge list of nouns, and they did not form a noun phrase.
        - For each of the extracted phrases evaluate it, and compute a score:
            - The score is an integer in the range [-2,2]. If the score!=0,:
                - Check if the phrase refers to the company, since the sentence could talk about more than one company, and the sentence splitter did not always work well. We look for the last entity that appears before the phrase in the sentence, if it or the extracted phrase contain the company name/ticker then:
                    - Tag phrase with a color tag: dark red = -2, red = -1, green = 1, light green= 2.
                - The text with the tagged phrases is saved in a file called: article<number>_tagged.txt, found under: project/articles/<ticker>_articles

- **Evaluating Phrases:**
  - If the current word in the phrase is one of the **intensify words** (“ALWAYS”, “AWFULLY”, “BEST”,”CLEARLY”, etc.) then return: 2 *value of the phrase after it. If it’s not followed by anything, like: “BEST”,”HIGHEST”,”LOWEST”,”LOWER”,”UNDISPUTED”,”UNEQUIVOCAL”,”UNSURPASSED”, then return it’s value (2).
  - “**Up**”, ”**above**”**,** “ **down**”: “up” and “above”, have the value of 1. “Down’ has the value of -1. (“up” and “down” are checked again here, since in the extraction, they’re extracted only if they are followed by a number)
  - If the current word is “**never**”, or “**no**”: return the value of the phrase that follows it multiplied by -1.
  - If the current word is “**not**”:
    - If it’s followed by “**only**”: then it functions as an intensifier; return: 2 * value of the phrase that follows it.
    - Otherwise, return: -1 * value of the phrase that follows it
  - If the current word is a **verb**:
    - Check if the verb is an **event:** check in the event dictionary, if the verb is there; if so return its value (2 or -2).
    - Check if it’s one of the verbs that **switch polarity** of the phrase that comes after them. Verbs like: “prevent”, and “avoid”. If so, return: -1 * value of the phrase that follows it.
    - If it’s a **positive verb**: return of the value of the phrase that follows it, unless there’s none or its value is 0, then return 1.
    - If it’s a **negative verb**: we noticed that negative verbs usually switch the polarity of the phrase that comes after them; so, the default behavior is to return: -1 * value of the following phrase. We added a couple of rules for exceptions.
  - If the current word is a **noun**:
    - If it’s an **event**: like in the verb section, return its value (2 or -2).
    - If it’s a **positive noun**: return its value (1).
    - If it’s a **negative noun**: return its value (-1).
    - Otherwise, remove the first word and evaluate the word that comes after it.
  - If the current word is an **adjective**(JJ):
    - If the adjective is **neutral**, then remove it, and proceed to evaluate the next word in the phrase.
    - Otherwise, add all of the values of the consecutive adjectives together then multiply that value with the value of the noun that follows them, or return their value if its value is 0 or there’s no noun. The returned value is then normalized to be in the range of [-1,2]; So “horrid winnings”, would have the value of -1. We didn’t exactly know how to deal with expressions

like, “beautiful horrid poem”, since it is not exactly negative nor positive; It’s both! So we regarded it as neutral.

- **Logging**: For each article that we evaluated, we created a log file called “article<number>_dump.txt, found under: project/articles/<ticker>_articles/
  The log file contains each sentence that appears in the article, followed by its POS tagging, followed by its Named Entity Recognition tagging, then the phrases that were extracted from it, if it contains the company name/ticker, and the tagged phrases, if any. See example, for more information.

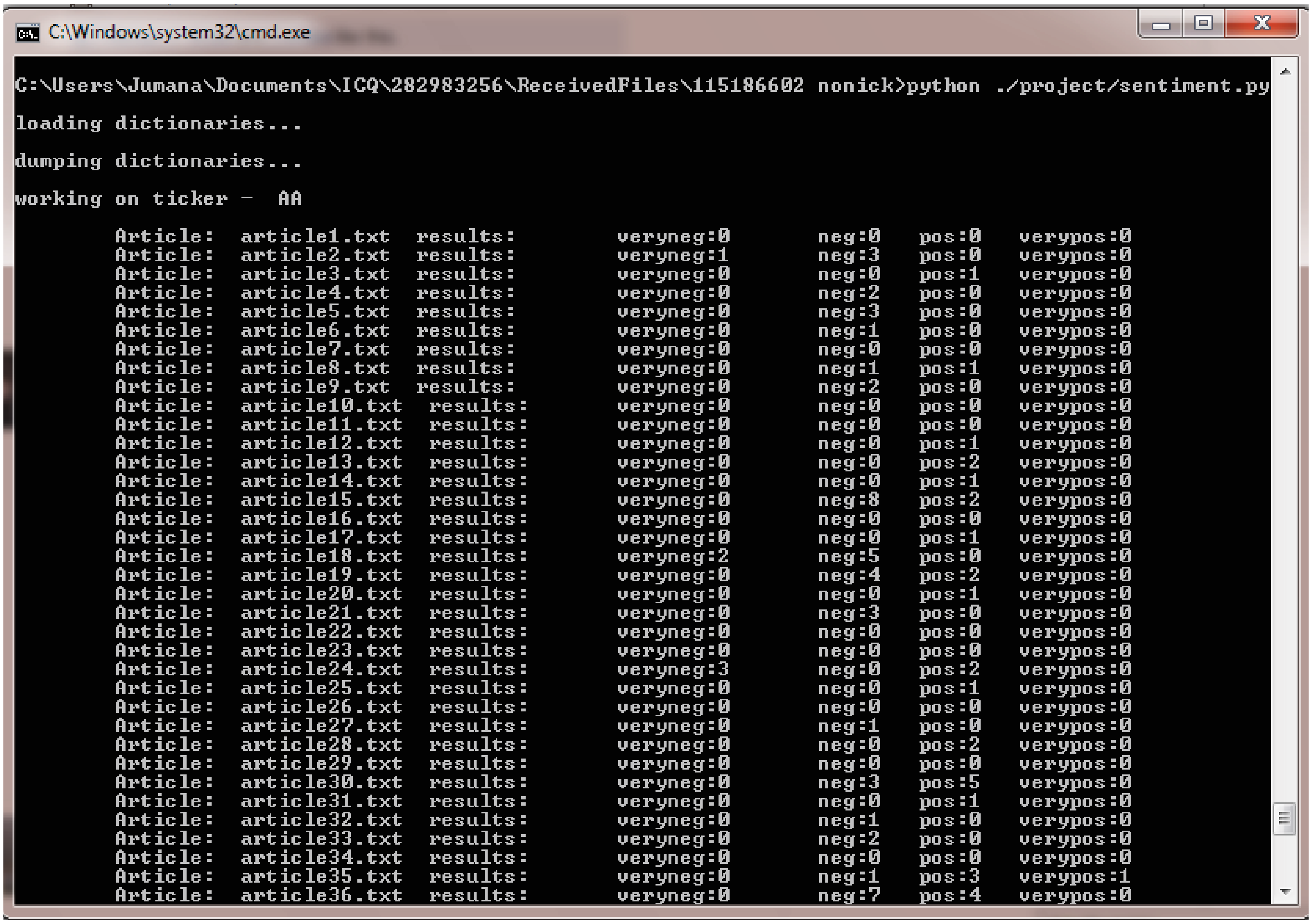
- **Tools**:
  - **Dictionaries**:
    We used the following dictionaries with some modification:
      - http://www.nd.edu/~mcdonald/Data/Finance_Word_Lists/ND_FinTerms_Negative_v2.txt for the negative dictionary.
      - http://www.nd.edu/~mcdonald/Data/Finance_Word_Lists/ND_FinTerms_Positive_v2.txt for the positive dictionary.
      - http://www.nd.edu/~mcdonald/Data/Finance_Word_Lists/ND_FinTerms_ModalStrong_v2.txt for the intensifying words dictionary.
    - The negative dictionary consists of words that give bad sentiment for the sentence and a score of -1.
    - The positive dictionary consists of words that give good sentiment for the sentence and a score of +1.

    Where the strong modal words, if found before a positive or negative phrase – the score will be multiplied by 2, giving it a very bad sentiment or a very good sentiment
    - In addition, we created 2 other dictionaries. One for good events and the second for bad events. The good evens had a score of 2, whereas the bad ones had a score of -2.

  - **Other tools**:
    - We used the NLTK tools found in: http://www.nltk.org/download together with its data, which can be downloaded from the same link. The site contains information on how to install the NLTK tools. The NLTK tools were used for:
      - Sentence Splitting.
      - POS (Part of Speech) Tagging.
      - Named Entity Recognition.
      - Chunking – creating chunks using regular expressions on POS tags.
      - Stemming – stemming words before looking them up in the dictionaries.
- **Notes**:

All words are stemmed before they are looked up in the dictionaries; with the exception of intensify words.

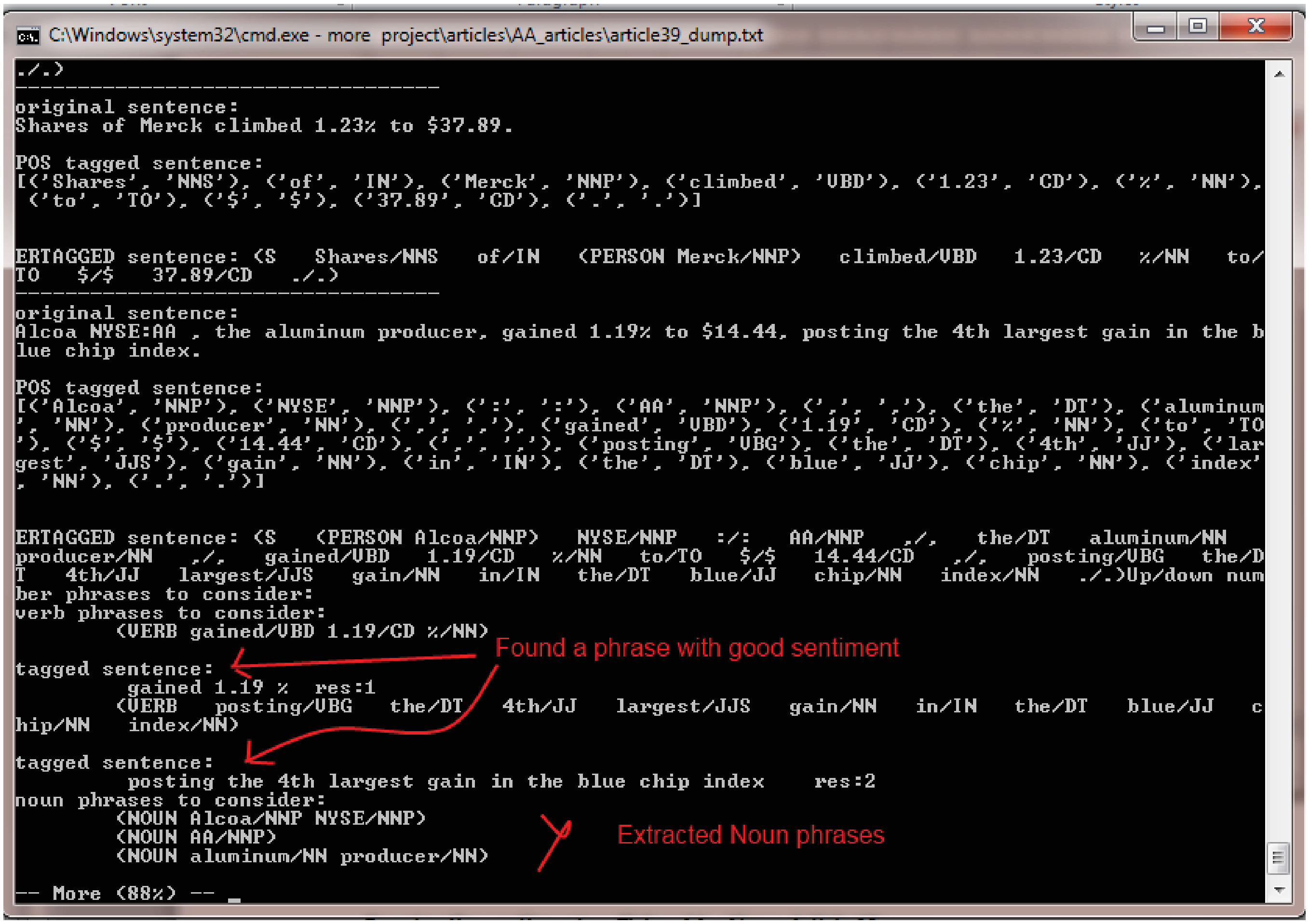
Example:



```
C:\Users\Jumana\Documents\ICQ\282983256\ReceivedFiles\115186602 nonick>python ./project/sentiment.py

loading dictionaries...

dumping dictionaries...

working on ticker -  AA

        Article:  article1.txt  results:        veryneg:0       neg:0   pos:0   verypos:0
        Article:  article2.txt  results:        veryneg:1       neg:3   pos:0   verypos:0
        Article:  article3.txt  results:        veryneg:0       neg:0   pos:1   verypos:0
        Article:  article4.txt  results:        veryneg:0       neg:2   pos:0   verypos:0
        Article:  article5.txt  results:        veryneg:0       neg:3   pos:0   verypos:0
        Article:  article6.txt  results:        veryneg:0       neg:1   pos:0   verypos:0
        Article:  article7.txt  results:        veryneg:0       neg:0   pos:0   verypos:0
        Article:  article8.txt  results:        veryneg:0       neg:1   pos:1   verypos:0
        Article:  article9.txt  results:        veryneg:0       neg:2   pos:0   verypos:0
        Article:  article10.txt  results:       veryneg:0       neg:0   pos:0   verypos:0
        Article:  article11.txt  results:       veryneg:0       neg:0   pos:0   verypos:0
        Article:  article12.txt  results:       veryneg:0       neg:0   pos:1   verypos:0
        Article:  article13.txt  results:       veryneg:0       neg:0   pos:2   verypos:0
        Article:  article14.txt  results:       veryneg:0       neg:0   pos:1   verypos:0
        Article:  article15.txt  results:       veryneg:0       neg:8   pos:2   verypos:0
        Article:  article16.txt  results:       veryneg:0       neg:0   pos:0   verypos:0
        Article:  article17.txt  results:       veryneg:0       neg:0   pos:1   verypos:0
        Article:  article18.txt  results:       veryneg:2       neg:5   pos:0   verypos:0
        Article:  article19.txt  results:       veryneg:0       neg:4   pos:2   verypos:0
        Article:  article20.txt  results:       veryneg:0       neg:0   pos:1   verypos:0
        Article:  article21.txt  results:       veryneg:0       neg:3   pos:0   verypos:0
        Article:  article22.txt  results:       veryneg:0       neg:0   pos:0   verypos:0
        Article:  article23.txt  results:       veryneg:0       neg:0   pos:0   verypos:0
        Article:  article24.txt  results:       veryneg:3       neg:0   pos:2   verypos:0
        Article:  article25.txt  results:       veryneg:0       neg:0   pos:1   verypos:0
        Article:  article26.txt  results:       veryneg:0       neg:0   pos:0   verypos:0
        Article:  article27.txt  results:       veryneg:0       neg:1   pos:0   verypos:0
        Article:  article28.txt  results:       veryneg:0       neg:0   pos:2   verypos:0
        Article:  article29.txt  results:       veryneg:0       neg:0   pos:0   verypos:0
        Article:  article30.txt  results:       veryneg:0       neg:3   pos:5   verypos:0
        Article:  article31.txt  results:       veryneg:0       neg:0   pos:1   verypos:0
        Article:  article32.txt  results:       veryneg:0       neg:1   pos:0   verypos:0
        Article:  article33.txt  results:       veryneg:0       neg:2   pos:0   verypos:0
        Article:  article34.txt  results:       veryneg:0       neg:0   pos:0   verypos:0
        Article:  article35.txt  results:       veryneg:0       neg:1   pos:3   verypos:1
        Article:  article36.txt  results:       veryneg:0       neg:7   pos:4   verypos:0
```

Snapshot of running the program; evaluating articles for the ticker: AA. It shows the name of the article together with its result.

Running the sentiment on Ticker AA - Alcoa, Article 39:



C:\Windows\system32\cmd.exe - more project\articles\AA_articles\article39_dump.txt

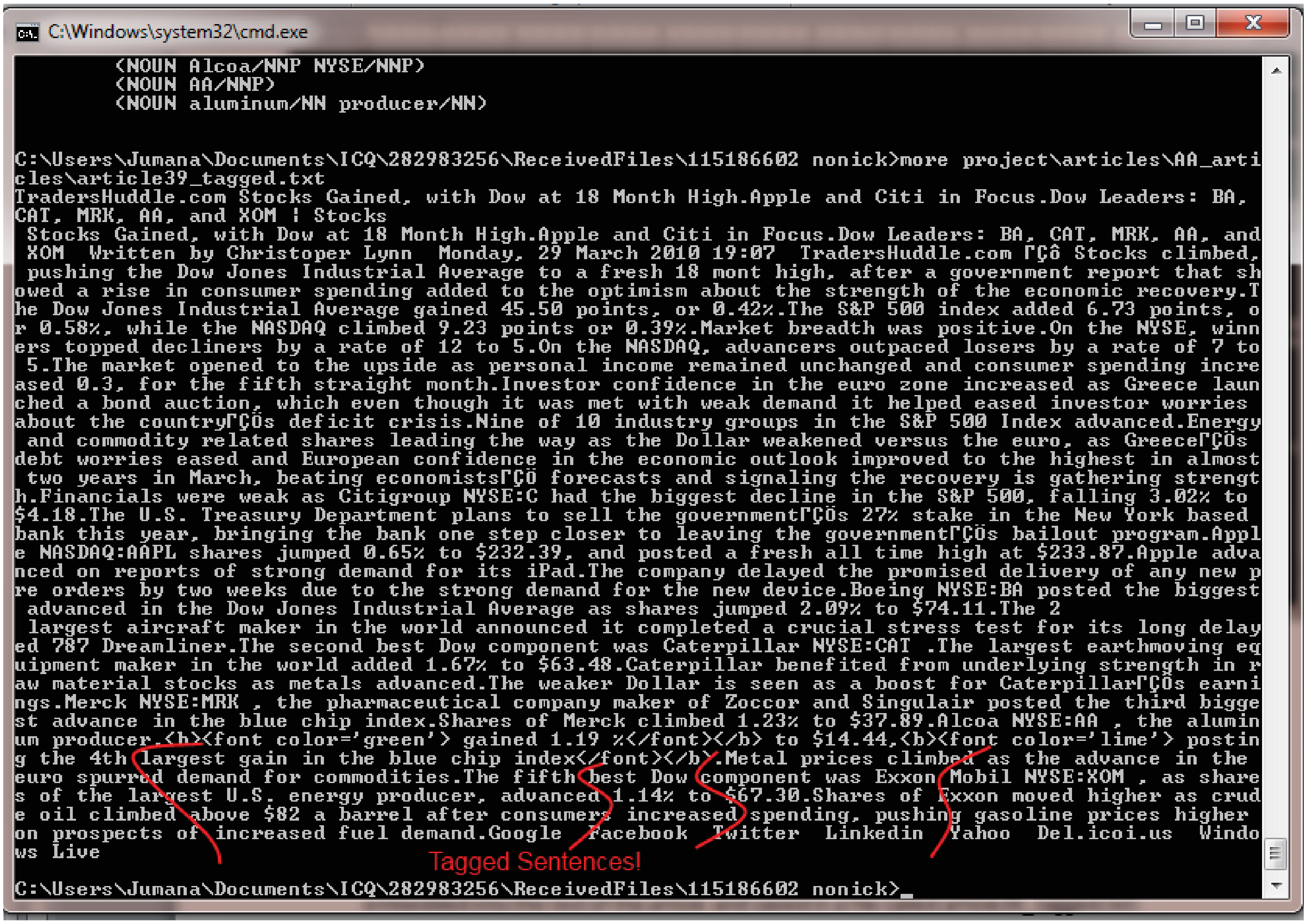
```
./.)
----------------------------------
original sentence:
Shares of Merck climbed 1.23% to $37.89.

POS tagged sentence:
[('Shares', 'NNS'), ('of', 'IN'), ('Merck', 'NNP'), ('climbed', 'VBD'), ('1.23', 'CD'), ('%', 'NN'),
 ('to', 'TO'), ('$', '$'), ('37.89', 'CD'), ('.', '.')]


ERTAGGED sentence: (S   Shares/NNS   of/IN   (PERSON Merck/NNP)   climbed/VBD   1.23/CD   %/NN   to/
TO   $/$   37.89/CD   ./.)
----------------------------------
original sentence:
Alcoa NYSE:AA , the aluminum producer, gained 1.19% to $14.44, posting the 4th largest gain in the b
lue chip index.

POS tagged sentence:
[('Alcoa', 'NNP'), ('NYSE', 'NNP'), (':', ':'), ('AA', 'NNP'), (',', ','), ('the', 'DT'), ('aluminum
', 'NN'), ('producer', 'NN'), (',', ','), ('gained', 'VBD'), ('1.19', 'CD'), ('%', 'NN'), ('to', 'TO
'), ('$', '$'), ('14.44', 'CD'), (',', ','), ('posting', 'VBG'), ('the', 'DT'), ('4th', 'JJ'), ('lar
gest', 'JJS'), ('gain', 'NN'), ('in', 'IN'), ('the', 'DT'), ('blue', 'JJ'), ('chip', 'NN'), ('index'
, 'NN'), ('.', '.')]


ERTAGGED sentence: (S   (PERSON Alcoa/NNP)   NYSE/NNP   :/:   AA/NNP   ,/,   the/DT   aluminum/NN
producer/NN   ,/,   gained/VBD   1.19/CD   %/NN   to/TO   $/$   14.44/CD   ,/,   posting/VBG   the/D
T   4th/JJ   largest/JJS   gain/NN   in/IN   the/DT   blue/JJ   chip/NN   index/NN   ./.)Up/down num
ber phrases to consider:
verb phrases to consider:
        (VERB gained/VBD 1.19/CD %/NN)

tagged sentence:
         gained 1.19 %   res:1
        (VERB   posting/VBG   the/DT   4th/JJ   largest/JJS   gain/NN   in/IN   the/DT   blue/JJ   c
hip/NN   index/NN)

tagged sentence:
         posting the 4th largest gain in the blue chip index    res:2
noun phrases to consider:
        (NOUN Alcoa/NNP NYSE/NNP)
        (NOUN AA/NNP)
        (NOUN aluminum/NN producer/NN)

-- More (88%) -- _
```

Found a phrase with good sentiment

Extracted Noun phrases

Snapshot from the file article39_dump.txt; contains each sentence found in the article with its POs tags, and its Entity Recognition tags. If the sentence contains “Alcoa” or “AA”, as in the case of the last sentence of the page, then phrases that might contain sentiment for the company are extracted (Up/down + number phrases, verb phrases, and noun phrases). If one of them contains good/bad sentiment then it’s printed together with its result, like in the case of “gained 1.19%”.

When a sentence with good/bad sentiment is found then it's tagged with color tags embedded into the rest of the article and saved in a file called article39_tagged.txt:



```
        (NOUN Alcoa/NNP NYSE/NNP)
        (NOUN AA/NNP)
        (NOUN aluminum/NN producer/NN)


C:\Users\Jumana\Documents\ICQ\282983256\ReceivedFiles\115186602 nonick>more project\articles\AA_arti
cles\article39_tagged.txt
TradersHuddle.com Stocks Gained, with Dow at 18 Month High.Apple and Citi in Focus.Dow Leaders: BA,
CAT, MRK, AA, and XOM | Stocks
 Stocks Gained, with Dow at 18 Month High.Apple and Citi in Focus.Dow Leaders: BA, CAT, MRK, AA, and
 XOM  Written by Christoper Lynn  Monday, 29 March 2010 19:07  TradersHuddle.com ΓÇô Stocks climbed,
 pushing the Dow Jones Industrial Average to a fresh 18 mont high, after a government report that sh
owed a rise in consumer spending added to the optimism about the strength of the economic recovery.T
he Dow Jones Industrial Average gained 45.50 points, or 0.42%.The S&P 500 index added 6.73 points, o
r 0.58%, while the NASDAQ climbed 9.23 points or 0.39%.Market breadth was positive.On the NYSE, winn
ers topped decliners by a rate of 12 to 5.On the NASDAQ, advancers outpaced losers by a rate of 7 to
 5.The market opened to the upside as personal income remained unchanged and consumer spending incre
ased 0.3, for the fifth straight month.Investor confidence in the euro zone increased as Greece laun
ched a bond auction, which even though it was met with weak demand it helped eased investor worries
about the countryΓÇÖs deficit crisis.Nine of 10 industry groups in the S&P 500 Index advanced.Energy
 and commodity related shares leading the way as the Dollar weakened versus the euro, as GreeceΓÇÖs
debt worries eased and European confidence in the economic outlook improved to the highest in almost
 two years in March, beating economistsΓÇÖ forecasts and signaling the recovery is gathering strengt
h.Financials were weak as Citigroup NYSE:C had the biggest decline in the S&P 500, falling 3.02% to
$4.18.The U.S. Treasury Department plans to sell the governmentΓÇÖs 27% stake in the New York based
bank this year, bringing the bank one step closer to leaving the governmentΓÇÖs bailout program.Appl
e NASDAQ:AAPL shares jumped 0.65% to $232.39, and posted a fresh all time high at $233.87.Apple adva
nced on reports of strong demand for its iPad.The company delayed the promised delivery of any new p
re orders by two weeks due to the strong demand for the new device.Boeing NYSE:BA posted the biggest
 advanced in the Dow Jones Industrial Average as shares jumped 2.09% to $74.11.The 2
 largest aircraft maker in the world announced it completed a crucial stress test for its long delay
ed 787 Dreamliner.The second best Dow component was Caterpillar NYSE:CAT .The largest earthmoving eq
uipment maker in the world added 1.67% to $63.48.Caterpillar benefited from underlying strength in r
aw material stocks as metals advanced.The weaker Dollar is seen as a boost for CaterpillarΓÇÖs earni
ngs.Merck NYSE:MRK , the pharmaceutical company maker of Zoccor and Singulair posted the third bigge
st advance in the blue chip index.Shares of Merck climbed 1.23% to $37.89.Alcoa NYSE:AA , the alumin
um producer.<b><font color='green'> gained 1.19 %</font></b> to $14.44,<b><font color='lime'> postin
g the 4th largest gain in the blue chip index</font></b>.Metal prices climbed as the advance in the
euro spurred demand for commodities.The fifth best Dow component was Exxon Mobil NYSE:XOM , as share
s of the largest U.S. energy producer, advanced 1.14% to $67.30.Shares of Exxon moved higher as crud
e oil climbed above $82 a barrel after consumers increased spending, pushing gasoline prices higher
on prospects of increased fuel demand.Google  Facebook  Twitter  Linkedin  Yahoo  Del.icoi.us  Windo
ws Live

C:\Users\Jumana\Documents\ICQ\282983256\ReceivedFiles\115186602 nonick>_
```

- "Gained 1.19%" is tagged with a green color tag. "Gain" is a positive verb and none of its following phrases has any sentiment. "AA" is the last entity found before it.
- "Posting the 4th largest gain in the blue chip index" is tagged with a light green color tag. "Largest" is one of the intensify words, and "gain" is a positive word; therefore, the result is 2.

When all of the sentences have been checked, the result is added to article39_dump.txt, and to the file: AA_scores.txt:

veryneg:0   neg:0 pos:1 verypos:1

Presenter.py:

The files are stored under '/project/present/'. The main file is called index.html and it contains a table with this information for each company:

- Number of times a very negative sentiment appeared in the articles related to the company, in the last 30 days.
- Number of times a negative sentiment appeared in the articles related to the company, in the last 30 days.
- Number of times a positive sentiment appeared in the articles related to the company, in the last 30 days.
- Number of times a very positive sentiment appeared in the articles related to the company, in the last 30 days.
- Average sentiment score for the company depending on the sentiment analysis of all articles in the last 30 days.
- Company ticker
- Company name.

The company ticker symbol is also a link to another file especially made for the company, an HTML file named by the company ticker. These files are located under the "/present/data/<Ticker>.html". Each file contains a table with the following information for each article:

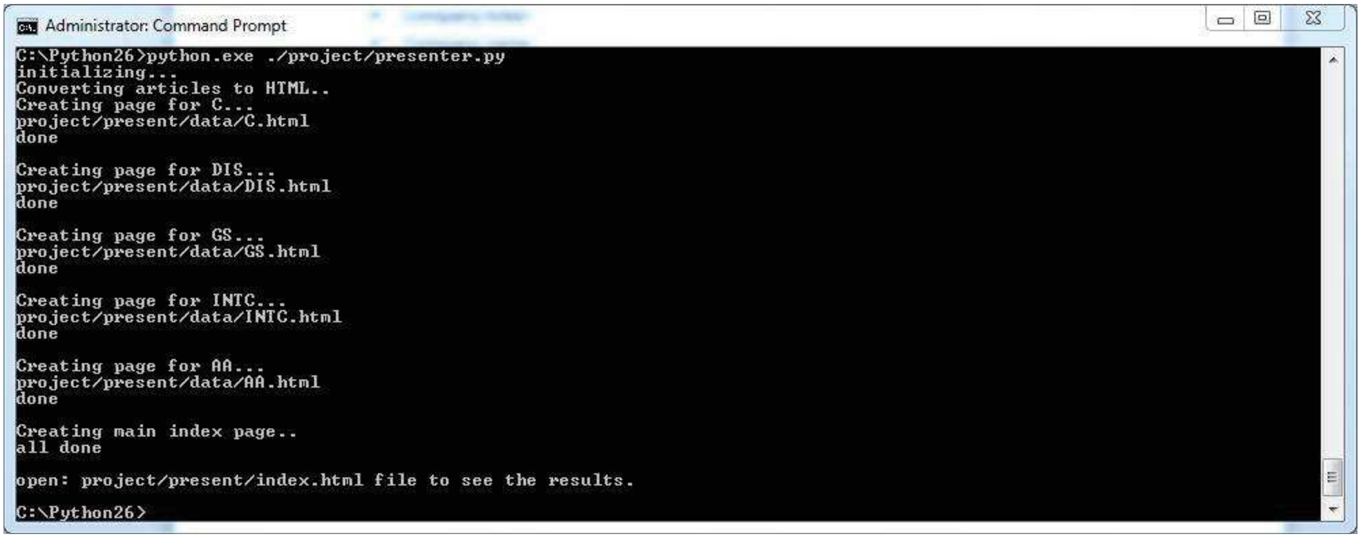
- Number of times a very negative sentiment related to the company appeared in the article.
- Number of times a negative sentiment related to the company appeared in the article.
- Number of times a positive sentiment related to the company appeared in the article.
- Number of times a very positive sentiment related to the company appeared in the article.
- A "Tagged Article" field which contains a link to the tagged article itself, stored under the "/present/data/ticker/" path, presented in an HTML format, where related sentences are marked by a color depending on their sentiment analysis score.
- The article title, which is a link to the article URL itself.
- The date on which the article appeared on the internet.

To see the final results open the following file:

Project/present/index.html

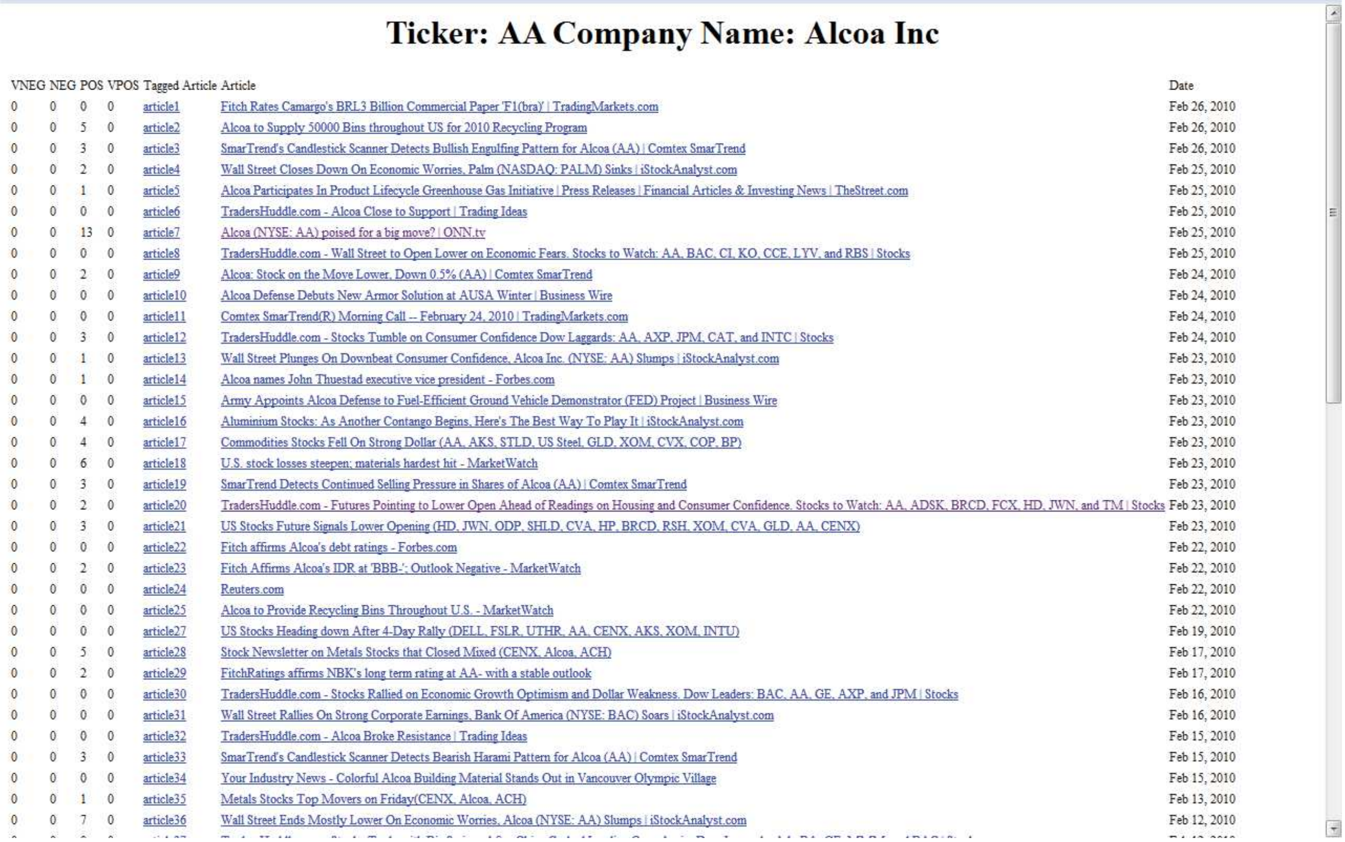
Example:

- Creates the pages with statistics for all articles found, the tickers are: C, DIS, GS, INTC, and AA.



```
C:\Python26>python.exe ./project/presenter.py
initializing...
Converting articles to HTML..
Creating page for C...
project/present/data/C.html
done

Creating page for DIS...
project/present/data/DIS.html
done

Creating page for GS...
project/present/data/GS.html
done

Creating page for INTC...
project/present/data/INTC.html
done

Creating page for AA...
project/present/data/AA.html
done

Creating main index page..
all done

open: project/present/index.html file to see the results.

C:\Python26>
```

Result Example: AA

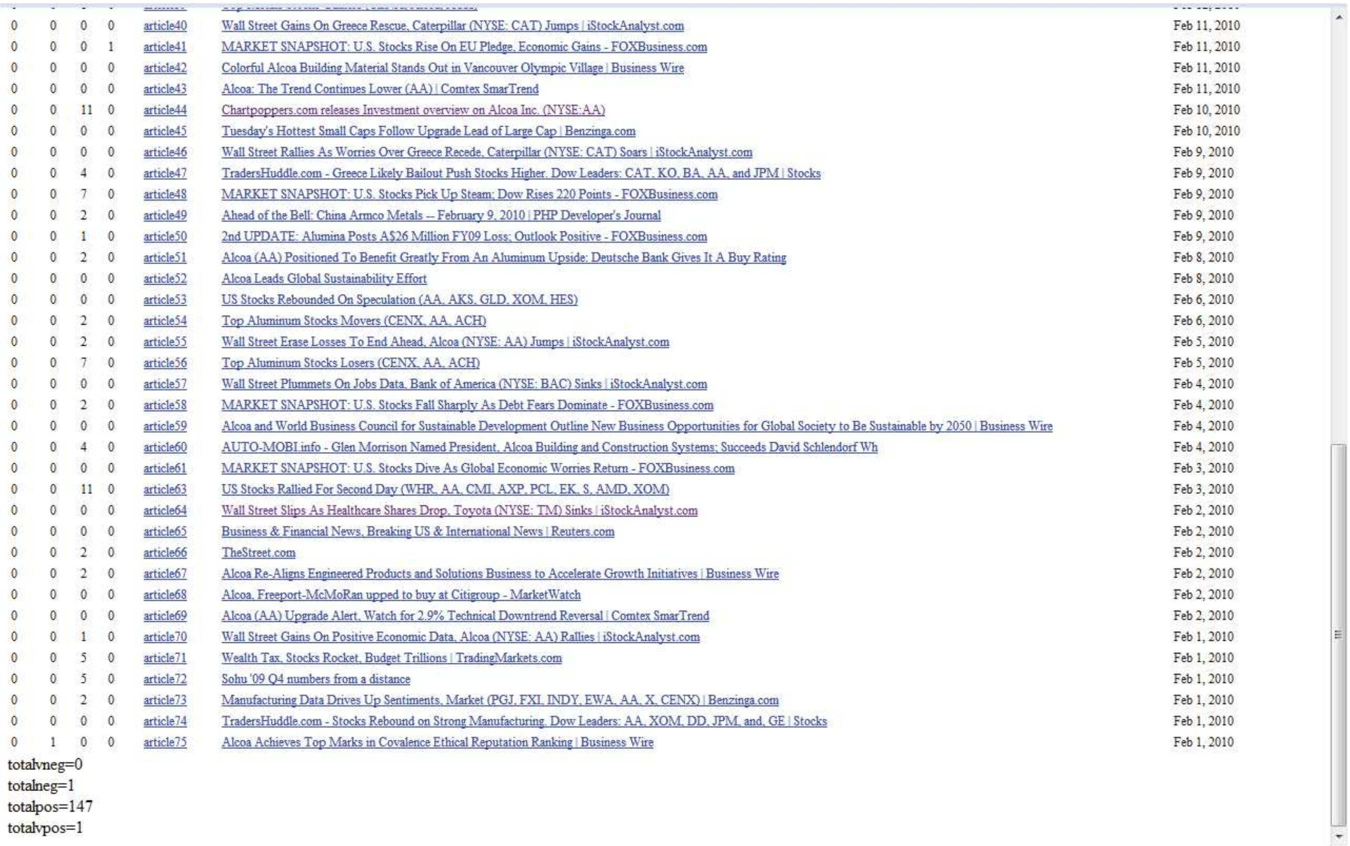


Ticker: AA Company Name: Alcoa Inc

VNEG	NEG	POS	VPOS	Tagged Article	Article	Date
0	0	0	0	article1	Fitch Rates Camargo's BRL3 Billion Commercial Paper 'F1(bra)' | TradingMarkets.com	Feb 26, 2010
0	0	5	0	article2	Alcoa to Supply 50000 Bins throughout US for 2010 Recycling Program	Feb 26, 2010
0	0	3	0	article3	SmarTrend's Candlestick Scanner Detects Bullish Engulfing Pattern for Alcoa (AA) | Comtex SmarTrend	Feb 26, 2010
0	0	2	0	article4	Wall Street Closes Down On Economic Worries, Palm (NASDAQ: PALM) Sinks | iStockAnalyst.com	Feb 25, 2010
0	0	1	0	article5	Alcoa Participates In Product Lifecycle Greenhouse Gas Initiative | Press Releases | Financial Articles & Investing News | TheStreet.com	Feb 25, 2010
0	0	0	0	article6	TradersHuddle.com - Alcoa Close to Support | Trading Ideas	Feb 25, 2010
0	0	13	0	article7	Alcoa (NYSE: AA) poised for a big move? | ONN.tv	Feb 25, 2010
0	0	0	0	article8	TradersHuddle.com - Wall Street to Open Lower on Economic Fears. Stocks to Watch: AA, BAC, CI, KO, CCE, LYV, and RBS | Stocks	Feb 25, 2010
0	0	2	0	article9	Alcoa: Stock on the Move Lower, Down 0.5% (AA) | Comtex SmarTrend	Feb 24, 2010
0	0	0	0	article10	Alcoa Defense Debuts New Armor Solution at AUSA Winter | Business Wire	Feb 24, 2010
0	0	0	0	article11	Comtex SmarTrend(R) Morning Call -- February 24, 2010 | TradingMarkets.com	Feb 24, 2010
0	0	3	0	article12	TradersHuddle.com - Stocks Tumble on Consumer Confidence Dow Laggards: AA, AXP, JPM, CAT, and INTC | Stocks	Feb 24, 2010
0	0	1	0	article13	Wall Street Plunges On Downbeat Consumer Confidence, Alcoa Inc. (NYSE: AA) Slumps | iStockAnalyst.com	Feb 23, 2010
0	0	1	0	article14	Alcoa names John Thuestad executive vice president - Forbes.com	Feb 23, 2010
0	0	0	0	article15	Army Appoints Alcoa Defense to Fuel-Efficient Ground Vehicle Demonstrator (FED) Project | Business Wire	Feb 23, 2010
0	0	4	0	article16	Aluminium Stocks: As Another Contango Begins, Here's The Best Way To Play It | iStockAnalyst.com	Feb 23, 2010
0	0	4	0	article17	Commodities Stocks Fell On Strong Dollar (AA, AKS, STLD, US Steel, GLD, XOM, CVX, COP, BP)	Feb 23, 2010
0	0	6	0	article18	U.S. stock losses steepen; materials hardest hit - MarketWatch	Feb 23, 2010
0	0	3	0	article19	SmarTrend Detects Continued Selling Pressure in Shares of Alcoa (AA) | Comtex SmarTrend	Feb 23, 2010
0	0	2	0	article20	TradersHuddle.com - Futures Pointing to Lower Open Ahead of Readings on Housing and Consumer Confidence. Stocks to Watch: AA, ADSK, BRCD, FCX, HD, JWN, and TM | Stocks	Feb 23, 2010
0	0	3	0	article21	US Stocks Future Signals Lower Opening (HD, JWN, ODP, SHLD, CVA, HP, BRCD, RSH, XOM, CVA, GLD, AA, CENX)	Feb 23, 2010
0	0	0	0	article22	Fitch affirms Alcoa's debt ratings - Forbes.com	Feb 22, 2010
0	0	2	0	article23	Fitch Affirms Alcoa's IDR at 'BBB-'; Outlook Negative - MarketWatch	Feb 22, 2010
0	0	0	0	article24	Reuters.com	Feb 22, 2010
0	0	0	0	article25	Alcoa to Provide Recycling Bins Throughout U.S. - MarketWatch	Feb 22, 2010
0	0	0	0	article27	US Stocks Heading down After 4-Day Rally (DELL, FSLR, UTHR, AA, CENX, AKS, XOM, INTU)	Feb 19, 2010
0	0	5	0	article28	Stock Newsletter on Metals Stocks that Closed Mixed (CENX, Alcoa, ACH)	Feb 17, 2010
0	0	2	0	article29	FitchRatings affirms NBK's long term rating at AA- with a stable outlook	Feb 17, 2010
0	0	0	0	article30	TradersHuddle.com - Stocks Rallied on Economic Growth Optimism and Dollar Weakness. Dow Leaders: BAC, AA, GE, AXP, and JPM | Stocks	Feb 16, 2010
0	0	0	0	article31	Wall Street Rallies On Strong Corporate Earnings, Bank Of America (NYSE: BAC) Soars | iStockAnalyst.com	Feb 16, 2010
0	0	0	0	article32	TradersHuddle.com - Alcoa Broke Resistance | Trading Ideas	Feb 15, 2010
0	0	3	0	article33	SmarTrend's Candlestick Scanner Detects Bearish Harami Pattern for Alcoa (AA) | Comtex SmarTrend	Feb 15, 2010
0	0	0	0	article34	Your Industry News - Colorful Alcoa Building Material Stands Out in Vancouver Olympic Village	Feb 15, 2010
0	0	1	0	article35	Metals Stocks Top Movers on Friday(CENX, Alcoa, ACH)	Feb 13, 2010
0	0	7	0	article36	Wall Street Ends Mostly Lower On Economic Worries, Alcoa (NYSE: AA) Slumps | iStockAnalyst.com	Feb 12, 2010

In this page, we see the results for Alcoa Inc. ticker: AA.

- First coloumn shows the number of times very negative remarks were found in the article, second coloumn shows number of very negative remarks. Third coloumn, positive remarks and forth coloumn shows the very positive remarks about the company in the article.
- Each line consists of an article, where articleX contains a link to the tagged article, where the hyper link links to the original article on the web.
- The last coloumn shows the date the article has been released.

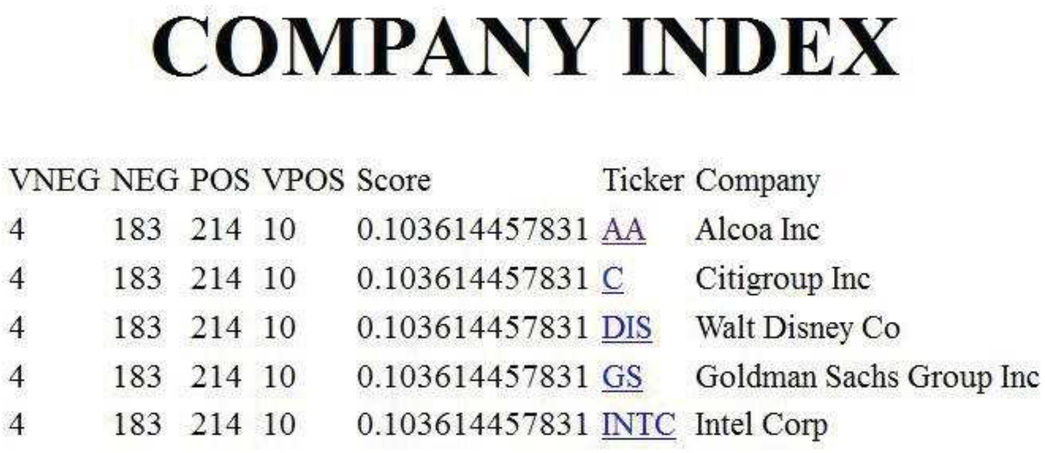


0	0	0	0	article40	Wall Street Gains On Greece Rescue, Caterpillar (NYSE: CAT) Jumps | iStockAnalyst.com	Feb 11, 2010
0	0	0	1	article41	MARKET SNAPSHOT: U.S. Stocks Rise On EU Pledge, Economic Gains - FOXBusiness.com	Feb 11, 2010
0	0	0	0	article42	Colorful Alcoa Building Material Stands Out in Vancouver Olympic Village | Business Wire	Feb 11, 2010
0	0	0	0	article43	Alcoa: The Trend Continues Lower (AA) | Comtex SmarTrend	Feb 11, 2010
0	0	11	0	article44	Chartpoppers.com releases Investment overview on Alcoa Inc. (NYSE:AA)	Feb 10, 2010
0	0	0	0	article45	Tuesday's Hottest Small Caps Follow Upgrade Lead of Large Cap | Benzinga.com	Feb 10, 2010
0	0	0	0	article46	Wall Street Rallies As Worries Over Greece Recede, Caterpillar (NYSE: CAT) Soars | iStockAnalyst.com	Feb 9, 2010
0	0	4	0	article47	TradersHuddle.com - Greece Likely Bailout Push Stocks Higher. Dow Leaders: CAT, KO, BA, AA, and JPM | Stocks	Feb 9, 2010
0	0	7	0	article48	MARKET SNAPSHOT: U.S. Stocks Pick Up Steam; Dow Rises 220 Points - FOXBusiness.com	Feb 9, 2010
0	0	2	0	article49	Ahead of the Bell: China Armco Metals -- February 9, 2010 | PHP Developer's Journal	Feb 9, 2010
0	0	1	0	article50	2nd UPDATE: Alumina Posts A$26 Million FY09 Loss; Outlook Positive - FOXBusiness.com	Feb 9, 2010
0	0	2	0	article51	Alcoa (AA) Positioned To Benefit Greatly From An Aluminum Upside: Deutsche Bank Gives It A Buy Rating	Feb 8, 2010
0	0	0	0	article52	Alcoa Leads Global Sustainability Effort	Feb 8, 2010
0	0	0	0	article53	US Stocks Rebounded On Speculation (AA, AKS, GLD, XOM, HES)	Feb 6, 2010
0	0	2	0	article54	Top Aluminum Stocks Movers (CENX, AA, ACH)	Feb 6, 2010
0	0	2	0	article55	Wall Street Erase Losses To End Ahead, Alcoa (NYSE: AA) Jumps | iStockAnalyst.com	Feb 5, 2010
0	0	7	0	article56	Top Aluminum Stocks Losers (CENX, AA, ACH)	Feb 5, 2010
0	0	0	0	article57	Wall Street Plummets On Jobs Data, Bank of America (NYSE: BAC) Sinks | iStockAnalyst.com	Feb 4, 2010
0	0	2	0	article58	MARKET SNAPSHOT: U.S. Stocks Fall Sharply As Debt Fears Dominate - FOXBusiness.com	Feb 4, 2010
0	0	0	0	article59	Alcoa and World Business Council for Sustainable Development Outline New Business Opportunities for Global Society to Be Sustainable by 2050 | Business Wire	Feb 4, 2010
0	0	4	0	article60	AUTO-MOBI.info - Glen Morrison Named President, Alcoa Building and Construction Systems; Succeeds David Schlendorf Wh	Feb 4, 2010
0	0	0	0	article61	MARKET SNAPSHOT: U.S. Stocks Dive As Global Economic Worries Return - FOXBusiness.com	Feb 3, 2010
0	0	11	0	article63	US Stocks Rallied For Second Day (WHR, AA, CMI, AXP, PCL, EK, S, AMD, XOM)	Feb 3, 2010
0	0	0	0	article64	Wall Street Slips As Healthcare Shares Drop, Toyota (NYSE: TM) Sinks | iStockAnalyst.com	Feb 2, 2010
0	0	0	0	article65	Business & Financial News, Breaking US & International News | Reuters.com	Feb 2, 2010
0	0	2	0	article66	TheStreet.com	Feb 2, 2010
0	0	2	0	article67	Alcoa Re-Aligns Engineered Products and Solutions Business to Accelerate Growth Initiatives | Business Wire	Feb 2, 2010
0	0	0	0	article68	Alcoa, Freeport-McMoRan upped to buy at Citigroup - MarketWatch	Feb 2, 2010
0	0	0	0	article69	Alcoa (AA) Upgrade Alert, Watch for 2.9% Technical Downtrend Reversal | Comtex SmarTrend	Feb 2, 2010
0	0	1	0	article70	Wall Street Gains On Positive Economic Data, Alcoa (NYSE: AA) Rallies | iStockAnalyst.com	Feb 1, 2010
0	0	5	0	article71	Wealth Tax, Stocks Rocket, Budget Trillions | TradingMarkets.com	Feb 1, 2010
0	0	5	0	article72	Sohu '09 Q4 numbers from a distance	Feb 1, 2010
0	0	2	0	article73	Manufacturing Data Drives Up Sentiments, Market (PGJ, FXI, INDY, EWA, AA, X, CENX) | Benzinga.com	Feb 1, 2010
0	0	0	0	article74	TradersHuddle.com - Stocks Rebound on Strong Manufacturing. Dow Leaders: AA, XOM, DD, JPM, and, GE | Stocks	Feb 1, 2010
0	1	0	0	article75	Alcoa Achieves Top Marks in Covalence Ethical Reputation Ranking | Business Wire	Feb 1, 2010

totalvneg=0
totalneg=1
totalpos=147
totalvpos=1

At the end of the page, we show the number of times each sentiment has appeared in the articles above for this company. In our example 0 times for very negative sentiment, once for negative sentiment, 147 times for positive sentiments, and once for very positive sentiment.

Result: Main Index



COMPANY INDEX

VNEG	NEG	POS	VPOS	Score	Ticker	Company
4	183	214	10	0.103614457831	AA	Alcoa Inc
4	183	214	10	0.103614457831	C	Citigroup Inc
4	183	214	10	0.103614457831	DIS	Walt Disney Co
4	183	214	10	0.103614457831	GS	Goldman Sachs Group Inc
4	183	214	10	0.103614457831	INTC	Intel Corp

- **VNEG** represents the number of times a very negative sentiment is found in the article.
- **NEG** represents the number of times a negative sentiment is found in the article.
- **POS** represents the number of times a positive sentiment is found in the article.
- **VPOS** represents the number of times a very positive sentiment is found in the article.
- **Score** is the score of the sentiment analysis for the company in all articles, following the formula given in class.
- **Ticker**, the company ticker.
- **Company**, the company name.

Installation Instructions:

1. Extract all the contest of the archive file into any directory.
2. Download Python 2.6.4:
   http://www.python.org/ftp/python/2.6.4/python-2.6.4.msi
3. Download BeautifulSoup v3.0.8:
   http://www.crummy.com/software/BeautifulSoup/download/3.x/BeautifulSoup-3.0.8.tar.gz
   And extract it under python's Lib/site-packages folder
4. Download the Nltk tools and data:
   http://www.nltk.org/download
   The site contains information on how to install the NLTK tools & data. Again, the NLTk tools must be extracted in the site-packages folder.
5. The project is built from several parts; they must be run in the following sequence from outside "project" folder:
   a. Crawler:
      python project/crawler.py [search engine] ["['ticker1', 'ticker2', …, 'tickerN']"]
   b. Cleaner:
      python project/cleaner.py [search engine]
   c. Sentiment analysis:
      python project/sentimentAnalyzer.py
   d. presentation:
      python project/presenter.py

   [search engine] is either Goolge or Yahoo.
6. Another option is to run runme.bat file if you are under windows OS. It will run all the steps on all 10 tickers searching articles from the last 30 days in Google; performing all processes: crawling, cleaning, sentiment analysis and presentation. The result will be saved under /project/present.