Automated Newsletter Creation

DORINA KARAMETI 01528726
ELISE LANDMAN 01551237
WIOLETA SACHA 01446367

WU Vienna University of Economics and Business
DATA SCIENCE LAB – WINTER TERM
20 January 2020

TABLE OF CONTENTS

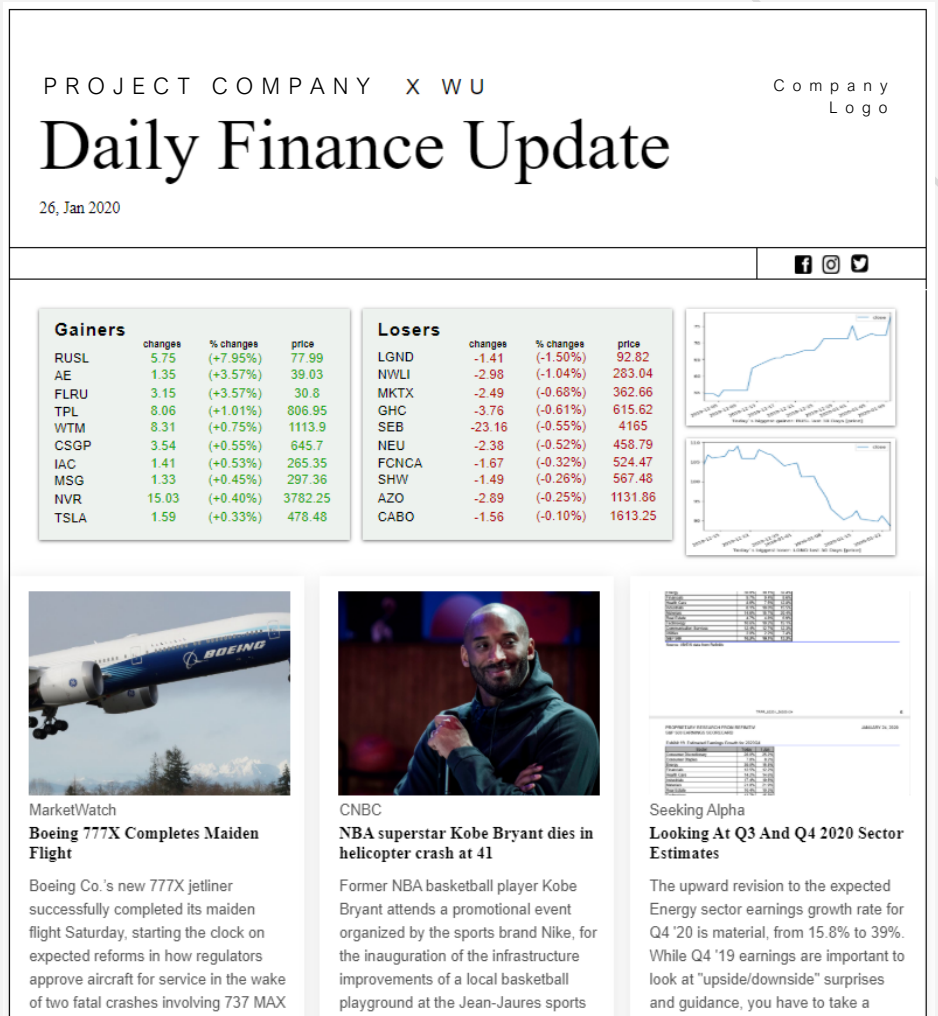
PROJECT OVERVIEW

GOAL

- extracting **trending financial news articles** and **stocks**
- displaying them **automatically** on a **daily created newsletter**
- for **portfolio trading customers**

WHAT WE ACHIEVED

- fully **automated** creation of **HTML newsletter**
- in approx. **5min**
- including most trending **financial news articles** and **stocks**



PROJECT COMPANY X WU

Company Logo

Daily Finance Update

26, Jan 2020

Gainers

	changes	% changes	price
RUSL	5.75	(+7.95%)	77.99
AE	1.35	(+3.57%)	39.03
FLRU	3.15	(+3.57%)	30.8
TPL	8.06	(+1.01%)	806.95
WTM	8.31	(+0.75%)	1113.9
CSGP	3.54	(+0.55%)	645.7
IAC	1.41	(+0.53%)	265.35
MSG	1.33	(+0.45%)	297.36
NVR	15.03	(+0.40%)	3782.25
TSLA	1.59	(+0.33%)	478.48

Losers

	changes	% changes	price
LGND	-1.41	(-1.50%)	92.82
NWLI	-2.98	(-1.04%)	283.04
MKTX	-2.49	(-0.68%)	362.66
GHC	-3.76	(-0.61%)	615.62
SEB	-23.16	(-0.55%)	4165
NEU	-2.38	(-0.52%)	458.79
FCNCA	-1.67	(-0.32%)	524.47
SHW	-1.49	(-0.26%)	567.48
AZO	-2.89	(-0.25%)	1131.86
CABO	-1.56	(-0.10%)	1613.25

MarketWatch

Boeing 777X Completes Maiden Flight

Boeing Co.'s new 777X jetliner successfully completed its maiden flight Saturday, starting the clock on expected reforms in how regulators approve aircraft for service in the wake of two fatal crashes involving 737 MAX

CNBC

NBA superstar Kobe Bryant dies in helicopter crash at 41

Former NBA basketball player Kobe Bryant attends a promotional event organized by the sports brand Nike, for the inauguration of the infrastructure improvements of a local basketball playground at the Jean-Jaures sports

Seeking Alpha

Looking At Q3 And Q4 2020 Sector Estimates

The upward revision to the expected Energy sector earnings growth rate for Q4 '20 is material, from 15.8% to 39%. While Q4 '19 earnings are important to look at "upside/downside" surprises and guidance, you have to take a

WHAT WE ACHIEVED

Extraction of news articles from RSS feeds

Python Libraries:
NewsPaper, NLTK, SKlearn

Extraction of title, content, link, keywords, image

Scraping stocks

API:
Financial Modeling Prep

Article comparison and analysis:

Euclidean Distance Measure

Cosine Similarity / Soft Cosine Measure

Newsletter creation

Automatic creation of HTML Newsletter with extracted content

APPROACH:

1. Extraction of News Articles

- Several **top rated news sources**:

CNN, CNBC, Financial Times, Wall Street Journal,

Yahoo Finance, ...

- Daily between **600 and 900 articles** scraped
- **Cleaned** and turned into a Pandas **dataframe**

YAHOO! FINANCE

FT FINANCIAL TIMES Business

CNN

Business Standard

WSJ

APPROACH:

1. Extraction of News Articles

NewsPapers_finance.json - Notepad

File Edit Format View Help

```
{"CNN":
        {"rss": ["http://rss.cnn.com/rss/money_latest.rss",
                 "http://rss.cnn.com/rss/money_news_economy.rss",
                 "http://rss.cnn.com/rss/money_news_companies.rss" ]},
"CNBC":
        {"rss": ["https://www.cnbc.com/id/10000664/device/rss/rss.html",
                 "https://www.cnbc.com/id/10001147/device/rss/rss.html",
                 "https://www.cnbc.com/id/15839135/device/rss/rss.html",
                 "https://www.cnbc.com/id/20910258/device/rss/rss.html",
                 "https://www.cnbc.com/id/15839069/device/rss/rss.html",
                 "http://www.cnbc.com/id/20409666/device/rss/rss.html?x=1"]},
"CBN":
        {"rss": ["https://www1.cbn.com/rss-cbn-news-finance.xml"]},

"MarketWatch":
        {"rss": ["http://feeds.marketwatch.com/marketwatch/topstories/",
                 "http://feeds.marketwatch.com/marketwatch/marketpulse/"]},
```

```
57  articles downloaded from  Wall Street Journal
1  articles downloaded from  CNN
37  articles downloaded from  The Guardian
41  articles downloaded from  MarketWatch
270  articles downloaded from  Business Standard
31  articles downloaded from  Fortune
11  articles downloaded from  Skynews
2  articles downloaded from  Wall Street Survivor
57  articles downloaded from  Nasdaq
53  articles downloaded from  New York Times
1  articles downloaded from  Daily Telegraph
1  articles downloaded from  Reddit
1  articles downloaded from  Financial Times
97  articles downloaded from  Yahoo Finance
73  articles downloaded from  CNBC
51  articles downloaded from  Investing.com
31  articles downloaded from  Seeking Alpha
2  articles downloaded from  CBN

817  total articles downloaded
```

	source	link	published_date	published_time	title	text	keywords	image	summary	clean_title	clean_text
0	CNN	https://www.cnbc...	2020-01-17	15:07:13 UTC	Wall Street sees...	Trailing closely...	market street tr...	https://image.cn...	The e-commerce g...	wall street see ...	trailing closely...
1	CNBC	https://www.cnbc...	2020-01-17	15:41:09 UTC	David Tepper say...	David Tepper, bi...	hes love horse m...	https://image.cn...	David Tepper, bi...	david tepper say...	david tepper bil...
2	CNBC	https://www.cnbc...	2020-01-17	12:39:27 UTC	Stocks making th...	Check out the co...	alibaba quarter ...	https://image.cn...	Regions Financia...	stock making big...	check company ma...
3	CNBC	https://www.cnbc...	2020-01-16	22:33:30 UTC	Here's what happ...	A trader works o...	earnings quarter...	https://image.cn...	Solid economic d...	happened stock m...	trader work floo...
4	CNBC	https://www.cnbc...	2020-01-16	18:48:35 UTC	Criminals are us...	The Indiana cred...	steal credit uni...	https://image.cn...	The credit union...	criminal using f...	indiana credit u...

APPROACH:

2. Extraction of Stocks

Which stocks to show:

- Biggest winners and biggest losers of the day
- API FinancialModelingPrep
- Licensing: „The FinancialModelingprep API is a set of services designed for developers and engineers. It can be used to build high-quality apps and services. We’re always working to improve the FinancialModelingPrep API. “ *

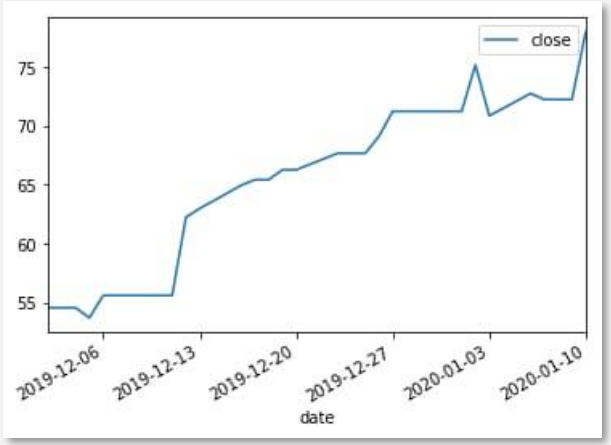
* https://www.financialmodelingprep.com/developer/docs/

APPROACH:

2. Extraction of Stocks

	ticker	changes	price	changesPercentage	companyName
0	NVR	15.03	3782.25	(+0.40%)	NVR Inc.
1	WTM	8.31	1113.9	(+0.75%)	White Mountains Insurance Group Ltd.
2	TPL	8.06	806.95	(+1.01%)	Texas Pacific Land Trust
3	RUSL	5.75	77.99	(+7.95%)	Direxion Daily Russia Bull 3x Shares
4	CSGP	3.54	645.7	(+0.55%)	CoStar Group Inc.

Gainers dataframe



Biggest gainer for 11.01.2020

	date	open	high	low	close	volume	unadjustedVolume	change	changePercent	vwap	label	changeOverTime
0	2020-01-16	3900.10	3935.00	3865.00	3870.52	19284.0	19284.0	29.58	0.758	3890.17333	January 16, 20	0.00758
1	2020-01-15	3829.00	3910.18	3829.00	3903.99	31400.0	31400.0	-74.99	-1.958	3881.05667	January 15, 20	-0.01958
2	2020-01-14	3804.47	3828.28	3800.00	3819.04	20100.0	20100.0	-14.57	-0.383	3815.77333	January 14, 20	-0.00383
3	2020-01-13	3780.89	3821.45	3765.13	3809.51	24500.0	24500.0	-28.62	-0.757	3798.69667	January 13, 20	-0.00757
4	2020-01-10	3752.48	3799.29	3735.00	3781.31	24600.0	24600.0	-28.83	-0.768	3771.86667	January 10, 20	-0.00768
5	2020-01-09	3786.00	3818.00	3696.54	3741.70	60500.0	60500.0	44.30	1.170	3752.08000	January 09, 20	0.01170
6	2020-01-08	3818.63	3849.69	3775.01	3785.81	35100.0	35100.0	32.82	0.859	3803.50333	January 08, 20	0.00859
7	2020-01-07	3775.94	3810.00	3774.00	3806.12	23700.0	23700.0	-30.18	-0.799	3796.70667	January 07, 20	-0.00799

Dataframe of historical data

APPROACH:

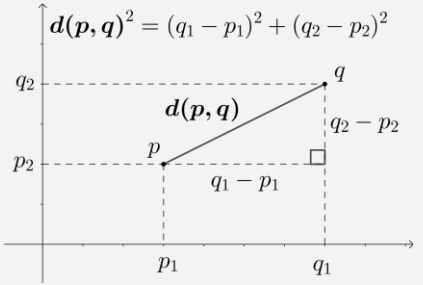
3. Similarity Analysis

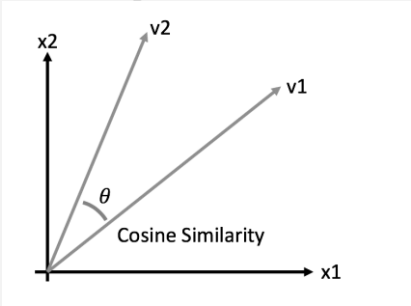
Selecting the articles by most trending topics:

- Creating word count vectors of all articles
- Computing **Euclidean Distance** between all articles
- **Ranking and clustering** by least distance = most similar
- Computing **Cosine Similarity** between articles in clusters
- Extracting articles in cluster with **highest similarity value**

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_i-q_i)^2+\cdots+(p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

Euclidean Distance between p and q in 2-dim space





Cosine Similarity between vector v1 and v2 in 2-dim space

APPROACH:

3. Article Ranking

- Extracting articles with **highest similarity values**

	source	link	published_date	title	text	clean_title	clean_text
117	Business Standard	https://www.business-standard.com/ar...	2020-01-19	Shin Kyuk-ho, founder of South Korea...	The founder of South Korea's sprawli...	shin kyukho founder south korean ret...	founder south korea sprawling retail...
277	Investing.com	https://www.investing.com/news/stock...	2020-01-19	Founder of South Korean retail giant...	By Hyunjoo Jin and Joyce Lee\n\nSEOU...	founder south korean retail giant lo...	hyunjoo jin joyce lee seoul reuters ...

Example of a cluster

	source	link	published_date	published_time	title	text	keywords	image	summary	clean_title	clean_text
99	Business Standard	https://www.busi...	2020-01-19	20:59:00	Answers to the S...	1. Connect Zarat...	648 zarathustra ...	https://bsmedia....	Connect Zarathus...	answer strategis...	connect zarathus...
144	Business Standard	https://www.busi...	2020-01-18	22:31:00	ED attaches asse...	Assets worth ove...	london agency fu...	https://bsmedia....	Assets worth ove...	ed attache asset...	asset worth r cr...
214	Seeking Alpha	https://www.nyti...	2020-01-19	10:00:29 UTC	A Look at Davos ...	1976\n\nIn an ef...	wider wrote week...	https://static01...	1976In an effort...	look davos year	effort engage so...
152	Business Standard	https://www.busi...	2020-01-18	17:05:00	ED attaches asse...	Assets worth ove...	sanjay london ag...	https://bsmedia....	Assets worth ove...	ed attache asset...	asset worth r cr...
216	Nasdaq	https://www.nyti...	2020-01-19	10:00:08 UTC	Davos by the Num...	$4,000\n\nThe en...	raised world dav...	https://static01...	$4,000The entry ...	davos number	entry fee world ...

APPROACH:

4. Creation of Newsletter

Importing chosen articles to the html template

Automatic creation with just one click

html file includes:

- 6 articles
- images for the articles, news sources and links
- the stocks (gainers / losers)
- date, social media links, etc.

Business Standard

ED attaches assets worth Rs 204 cr of ex-Bhushan Power CMD Singhal

Assets worth over Rs 204 crore, including houses in Delhi and London, of former Bhushan Power and Steel (BPSL) CMD Sanjay Singal have been attached under the anti-money laundering law, the (ED) said on Saturday. "The attached assets consist of movab...

Gainers

	changes	% changes	price
NVR	15.03	(+0.40%)	3782.25
WTM	8.31	(+0.75%)	1113.9
TPL	8.06	(+0.75%)	806.95
RUSL	5.75	(+7.95%)	77.99

Losers

	changes	% changes	price
SEB	-23.16	(-0.55%)	4165
GHC	-3.76	(-0.61%)	615.62
NWLI	-2.98	(-1.04%)	283.04
AZO	-2.89	(-0.25%)	1131.86

CHALLENGES:

ARTICLES:

- Getting the right content from the RSS feed (description, title etc.)
- How to find trending articles?
- Comparison of similarity measures
- Excluding the correct articles from the similarity clusters

STOCKS:

- Exporting the graphs
- What stocks to include?

HTML:

- Inserting Python variables in HTML
- Right layout

FUTURE WORK:

NEWSLETTER

- Personalize according to interest
- Split articles into predetermined categories
- Cluster by customer chosen topics

ARTICLE ANALYSIS

other measures for filtering out trending articles:

- e.g. based on articles featuring the most trending stocks
- filtered by current “trending” keywords, etc.

THANK YOU !

BUT WAIT

…. NOW WE WILL SHOW YOU TODAYS NEWSLETTER IN A LIVE DEMO