RFA 2019 - Introduction to web scraping and natural language processing in Java

Loïc Maréchal

November 21, 2019

Objectives

- Understand the principles of natural language processing (NLP), i.e. quantitative text analysis
- Understand the principles of web browsing emulation and web scraping on SEC forms, web page content and twitter
- Conduct a basic sentiment extraction with dictionaries

Outline

- Introduction to NLP
- Introduction to Java
- How to download SEC forms from Edgar in bulk
- Read and parse a text (a 10-K form)
- Twitter connection and live NLP

Introduction to NLP - My own experience

- Sugar production data of the Thai Sugar Millers Association

* รายงานแยกตามประเภท (ภาค) *

ชื่อโรงงาน	เปิดหีบ	ณ วันนี้ (*ปิดหีบ)	รวมวันเดินเครื่อง	ปริมาณอ้อยเข้าหีบ (ตัน)		รวม ปริมาณอ้อย (ตัน)	เฉลี่ย C.C.S ถึงวันนี้	น้ำตาลทรายขาว (กส.)			น้ำตาลทรายดิบ			* ชนิดอื่นๆ	รวมทั้งสิ้น	น้ำตาล/ ตันอ้อย กก.	ปริมาณการผลิต กากน้ำตาล (ตัน)	เฉลี่ย/ ตันอ้อย กก.
				อ้อยสด	อ้อยไฟไหม้			ขาวธรรมดา	ขาวบริสุทธิ์	รวม	เทกอง (ตัน)	กระสอบ	รวม (กส.)					
ภาคตะวันออกเฉียงเหนือ																		
อีสาน	1/12/49	1/12/49	1	605.350	276.300	881.650	9.88	0.00	0.00	0.00	0.000	0.00	0.00	0.00	0.00	0.00	0.000	0.00
มิตรผล (กาฬสินธุ์)	1/12/49	1/12/49	1	3,286.790	3,928.170	7,214.960	10.47	0.00	0.00	0.00	0.000	0.00	0.00	0.00	0.00	0.00	0.000	0.00
เกษตรผล	1/12/49	1/12/49	1	2,239.200	6,891.930	9,131.130	9.57	0.00	0.00	0.00	122.450	0.00	1,224.50	0.00	1,224.50	13.41	0.000	0.00
มิตรภูเวียง	1/12/49	1/12/49	1	8,027.840	477.810	8,505.650	9.82	0.00	0.00	0.00	0.000	0.00	0.00	0.00	0.00	0.00	0.000	0.00
กุมภวาปี	1/12/49	1/12/49	1	1,286.390	2,932.280	4,218.670	9.62	0.00	0.00	0.00	0.000	0.00	0.00	0.00	0.00	0.00	0.000	0.00
ขอนแก่น	1/12/49	1/12/49	1	0.000	0.000	0.000	0.00	0.00	0.00	0.00	0.000	0.00	0.00	0.00	0.00	0.00	0.000	0.00
รวม				**15,445.570**	**14,506.490**	**29,952.060**	**9.87**	**0.00**	**0.00**	**0.00**	**122.450**	**0.00**	**1,224.50**	**0.00**	**1,224.50**	**4.09**	**0.000**	**0.00**
รวมทั้งสิ้น				**15,445.570**	**14,506.490**	**29,952.060**	**9.87**	**0.00**	**0.00**	**0.00**	**122.450**	**0.00**	**1,224.50**	**0.00**	**1,224.50**	**4.09**	**0.000**	**0.00**

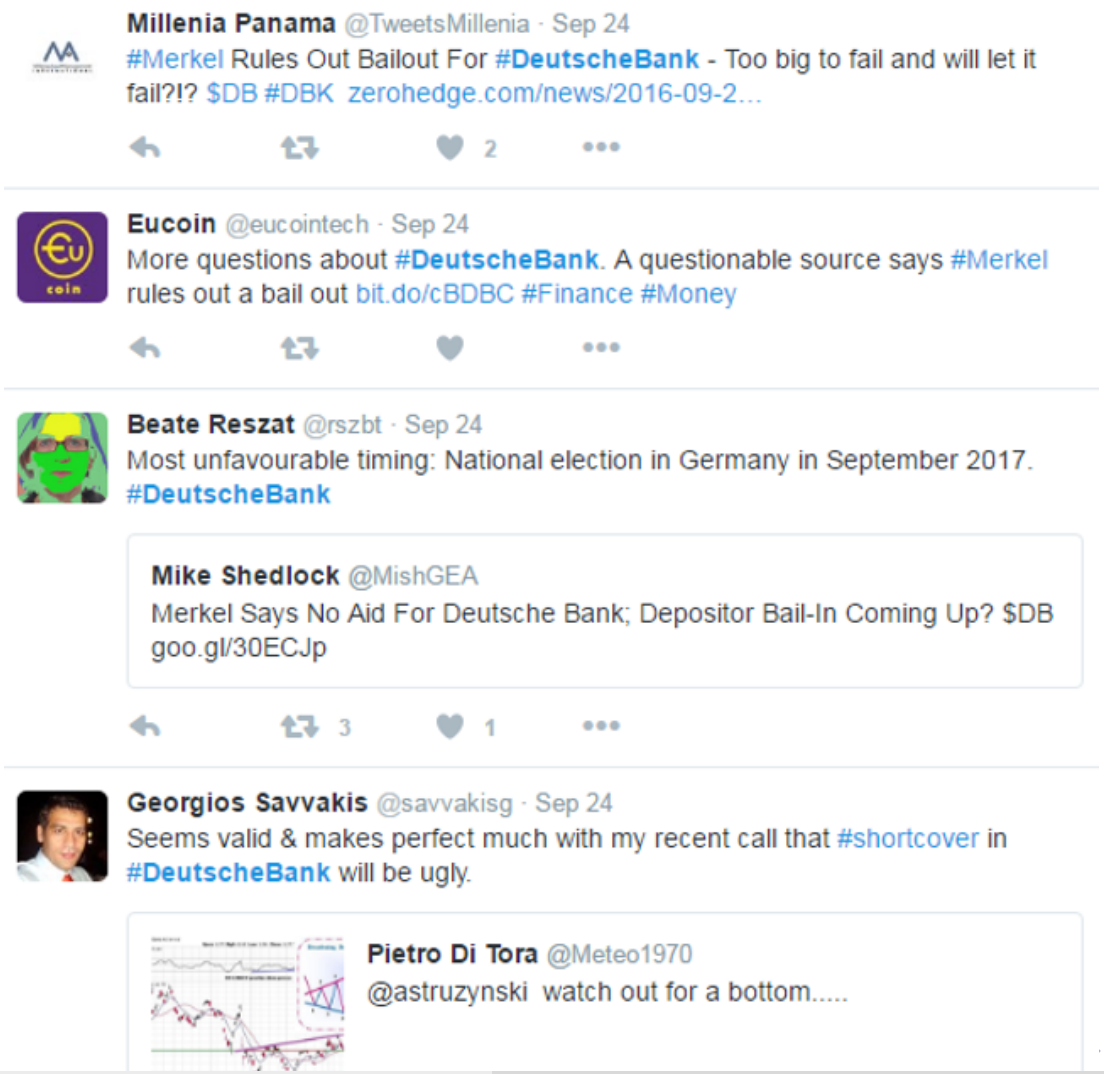
Introduction to NLP - Regular expression example with a 10-K filing

- 10-K including a negative tone message : Abercombie & Fitch Q1-2014 :

  *Our **inability** to obtain commercial insurance at acceptable prices or our **failure** to adequately reserve for self-insured **exposures** might increase our expenses and **adversely** impact our financial results*

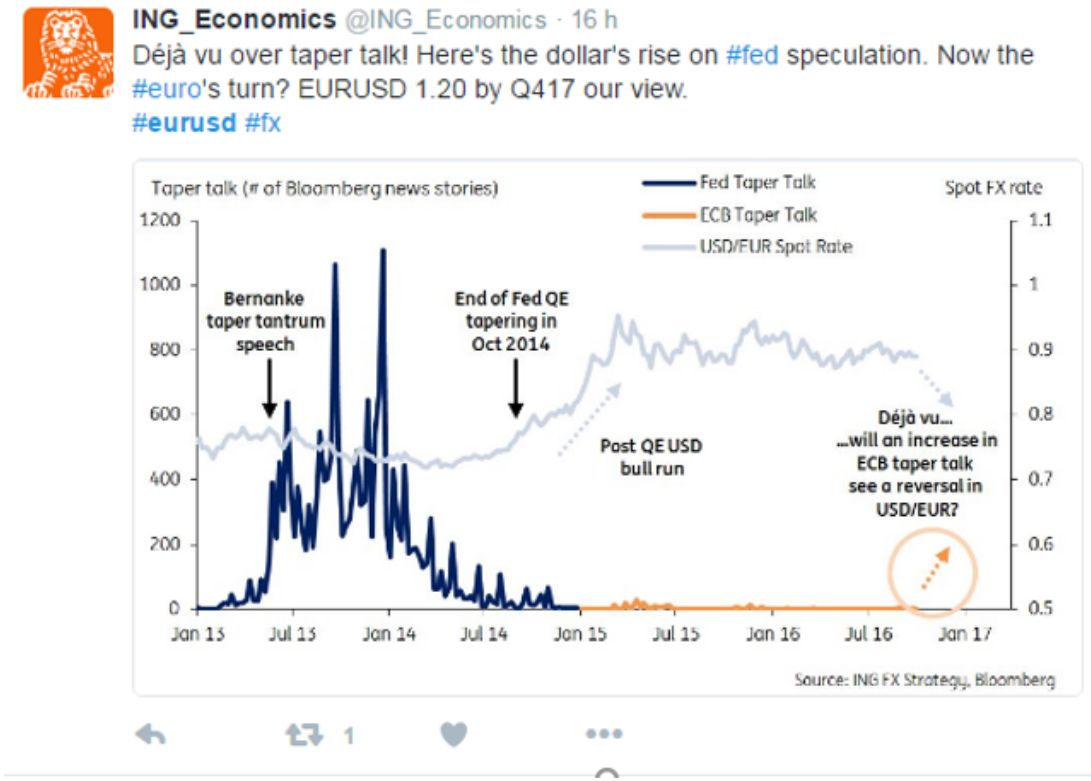
Introduction to NLP - Regular expression example with Twitter

- We focus on Deutsche Bank hashtags
- We recognize the tweets tone with regular expression
- For instance : “fail”, “bailout”, “unfavourable”, “bail-in”, “ugly”



Introduction to NLP - Issue 1 : Identifying the topic of interest

- Difficult to associate the topic with the tone
- Association of "rise" with "dollar", "euro" or "eurusd" implies different interpretations



Introduction to NLP - Issue 2 : Negative sentence with positive lexicon

- 10-K : Abercombie & Fitch Q1-2014 :

  [. . .] it is an important goal of ours that ***we return to positive growth****, particularly in our core U.S. business, and we believe the steps we are taking as we execute against our long-range strategic plan can put us in position to achieve this goal.*

Introduction to NLP - Issue 3 : Sarcasm

- A simple NLP algorithm would catch “fine” three times
- How can we deal with it ?

Introduction to NLP - Advantages and drawbacks

- Advantages
  - Deliver more refined measures than only quantitative analysis
  - Big amount of data acquired and read fast (vs hand collection)
  - Is already powerful, but full potential is not yet unleashed (deep learning)

- Drawbacks
  - Difficult to keep track of what is inside the black box
  - Difficult to calibrate as text format may vary from one file to the next
  - It would take an infinite amount of time to get success rates up to 99%
  - Need to accept 90-95% and live with some measurement error (otherwise the optimization time converges to the hand collection time)

Introduction to Java - The Java language

- Java developped since 1982 – Available to the public in 1995
- Fully object oriented programming (everything is an object / different from e.g., C++)
- Portable, ran in Windows, Unix systems, smartphones, watches, cars or fridges
- Huge online support, huge amount of packages
- Drawbacks ? Might be slightly slower than C/C++ (only for perfectly well optimized code), is not originally designed for textual analysis (unlike Perl). Less flexible than R or Matlab for statistics.

Introduction to Java - Application

- A_introduction.java : entry point and printing in console

  public static void main(String[] args){
  System.out.println("hello world");
  }

- B_instantiate.java : how to instantiate a class
- C_variables.java : how to declare variables
- D_conditional_statements.java
- E_loops.java
- F_methods.java

How to download SEC forms from Edgar in bulk - Before

- Files used to be obtained from the SEC's FTP site : ftp.sec.gov
- Files are available on the FTP 24 hours after they are filed
- Do large downloads early in the morning (Swiss time)
- Some networks (e.g. University/Eduroam) used to be blocked by the FTP
- Now we need to work around this with HTTPS access and web browsing emulation
- The end of the Edgar/SEC FTP - Workaround with HTTPS access and web browsing emulation

How to download SEC forms from Edgar in bulk

- Download index file : G_getIdx.java - download index files and identify filings from index files to download

Description: Master Index of EDGAR Dissemination Feed by Company Name
Last Data Received: September 6, 2006
Comments: webmaster@sec.gov
Anonymous FTP: ftp://ftp.sec.gov/edgar/

Company Name	Form Type	CIK	Date Filed	File Name
033 ASSET MANAGEMENT LLC /	13F-HR	1114831	2006-08-11	edgar/data/1114831/0001110550-06-000042.txt
1 800 CONTACTS INC	10-Q	1050122	2006-08-10	edgar/data/1050122/0001104659-06-053544.txt

- Download filings : H_getFiles.java - creates a file containing the names of all files to be downloaded and then downloads the filings.
- Process Filings (e.g., extract data, etc.) : use I_parseFiles - get a raw sentiment score for the 10-K report

Twitter connection and live NLP

- Connect to a Twitter feed of US economic hashtags, stock tickers, economic and market keywords within the tweet body
- Feed an array with tweet content
- Load two arrays of positive and negative tone dictionaries
- Add counters of positive and negative words match
- Substract the negative from the positive counter, and get a live rough estimate of the market sentiment
- Compare with the live S&P 500 or record for longer term analysis
- Let's check the results !

loic.marechal@unine.ch