Practice Assignment 06

Download the ECB speeches dataset and the *daily EUR/USD* reference exchange rate from the ECB Statistical Data Warehouse. Next, save them as “speeches.csv” and “fx.csv”. For the speeches.csv, please keep only the “date” and “contents” columns.

Create a GitHub repository called “st2195_assignment_6” and include two scripts (one in Python and one in R) to perform the operations described below, and a README.md file.

1. Load and merge the datasets keeping all information available for the dates in which there is a measurement in “fx.csv”. [1 point]
2. Remove entries with obvious outliers or mistakes, if any. [1.5 points]
3. Handle missing observations for the exchange rate, if any. This should be done replacing any missing exchange rate with the latest information available. Whenever this cannot be done, the relevant entry should be removed entirely from the dataset. [1.5 points]
4. Calculate the exchange rate return. Extend the original dataset with the following variables: “good_news” (equal to 1 when the exchange rate return is larger than 0.5 percent, 0 otherwise) and “bad_news” (equal to 1 when the exchange rate return is lower than -0.5 percent, 0 otherwise). [1.5 points]
5. Remove the entries for which contents column has NA values. Generate and store in csv the following tables [1.5 points each]:
    a. “good_indicators” – with the 20 most common words (excluding articles, prepositions and similar connectors) associated with entries wherein “good_news” is equal to 1;

    b. “bad_indicators” – with the 20 most common words (excluding articles, prepositions and similar connectors) associated with entries wherein “bad_news” is equal to 1;

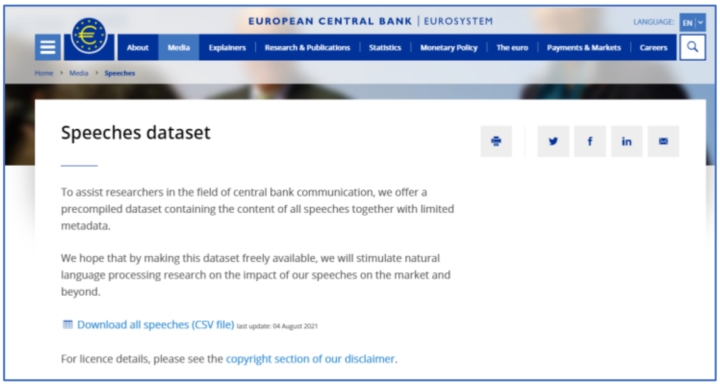
    Any observation from the common words found above?

Note that the original data should not be included in the GitHub repository, but only appropriately described and linked in the readme file.

Additional Notes:

- Challenging practice assignment
  - Discuss within your group
  - Do some research
  - Use the help functions, and search the documentation

- Download datasets
  - “speeches.csv” – Go to the given link ECB speeches dataset and download the speeches from there. Rename file as “speeches.csv”.
    - Sample R code for importing the CSV file is provided at the link



Speeches dataset

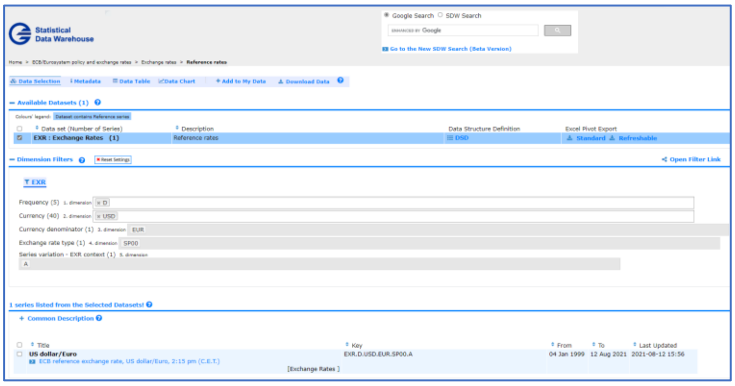
To assist researchers in the field of central bank communication, we offer a precompiled dataset containing the content of all speeches together with limited metadata.

We hope that by making this dataset freely available, we will stimulate natural language processing research on the impact of our speeches on the market and beyond.

Download all speeches (CSV file) last update: 04 August 2021

For licence details, please see the copyright section of our disclaimer.

  - “fx.csv” – Go to the given link and search for USD/EUR daily rates – Frequency=D, Currency=USD, Currency Denominator=EUR, Exchange Rate Type=SP00
    - Click on the link to the dataset (also shown below) and download the “CSV" version. Rename as “fx.csv”. ‘
    - https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=120.EXR.D.USD.EUR.SP00.A


Statistical Data Warehouse
○ Google Search ○ SDW Search
ENHANCED BY Google
Go to the New SDW Search (Beta Version)
Home > ECB/Eurosystem policy and exchange rates > Exchange rates > Reference rates
Data Selection
Metadata
Data Table
Data Chart
Add to My Data
Download Data
Reset Settings
Open Filter Link

Available Datasets (1)

Colours' legend: Dataset contains Reference series

	Data set (Number of Series)	Description	Data Structure Definition	Excel Pivot Export
☑	EXR : Exchange Rates (1)	Reference rates	DSD	Standard Refreshable

Dimension Filters

EXR

Frequency (5) 1. dimension × D

Currency (40) 2. dimension × USD

Currency denominator (1) 3. dimension EUR

Exchange rate type (1) 4. dimension SP00

Series variation - EXR context (1) 5. dimension A

1 series listed from the Selected Datasets!

+ Common Description

	Title	Key	From	To	Last Updated
□	US dollar/Euro ECB reference exchange rate, US dollar/Euro, 2:15 pm (C.E.T.) [Exchange Rates]	EXR.D.USD.EUR.SP00.A	04 Jan 1999	12 Aug 2021	2021-08-12 15:56

- Download list of stop words; one version can be found at the link below. You may also use any other sources of stop words.
  - https://countwordsfree.com/stopwords

- Hints
    - Load and merge datasets
        - Load “fx.csv” – You may need to skip a few lines to get to the data and also specify the indicator for missing data (NA)
        - Load “speeches.csv” – Transform this into just one row of content for each date (i.e., group contents by date and concatenate all contents per date as one row)
        - Merge the fx and speeches data into one
    - Remove entries with obvious outliers or mistakes
        - You may want to do a visual plot, and/or check the information/ structure/ summary
    - Handle missing observations for the exchange rate
        - Any missing data? Should we remove them?
        - How can we replace or fill them with appropriate data?
    - Calculate exchange rate return
    - Extend the original dataset with “good_news” and “bad_news” variables
    - Associate words with "good_news" and "bad_news", and output the top 20 for each category
        - Gather all the contents related to "good_news" and "bad_news" separately.
        - Write a function to get the most common words (excluding stop_words) related to "good_news" and "bad_news". You will need a word counter that stores the results in a dictionary.
    - Useful Libraries/Packages
        - R – dplyr, tidyr, zoo (for filling missing values), text2vec (for text analysis and NLP)
        - Python – pandas, string

- Several Useful References
    - https://cran.r-project.org/web/packages/text2vec/vignettes/text-vectorization.html
    - 9 functions that make natural language pre-processing a piece of cake | by Alejandra Vlerick | Towards Data Science