NLP Case Study

# Problem Statement

The *reuters 21578* corpus, containing 10,000 news articles, is a well-known text dataset published around 1996. In the past, it has been widely used by researchers for developing classification and other NLP methods. With the advent of deep learning, however, it is worth revisiting the corpus to see what new insights the corpus can yield.

The task for you is to apply modern NLP techniques to reuters 21578, and derive broad insights from the articles. We anticipate that candidates could use *topic modeling*, *weak supervision*, *single-shot and multi-shot* learners, and other *classification* techniques. However, we are deliberately leaving the problem statement open ended so you have the opportunity to show off your NLP chops.

At the end of your work, you will present your analysis to our team as a part of your (virtual) onsite interview process. Extra points for informative visualizations. We also expect to review your code, so please upload your code at a convenient (eg. GitHub) location and convey the URL to us.

# Dataset

Download the dataset (reuters21578.tar.gz) from: <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>

The data file reuters21578.tar.gz is around 9MB compressed, and 27MB uncompressed. Unzip the file, and you will see the following list of files:

**$ ls**

README.txt reut2-002.sgm reut2-013.sgm

all-exchanges-strings.lc.txt reut2-003.sgm reut2-014.sgm

all-orgs-strings.lc.txt reut2-004.sgm reut2-015.sgm

all-people-strings.lc.txt reut2-005.sgm reut2-016.sgm

all-places-strings.lc.txt reut2-006.sgm reut2-017.sgm

all-topics-strings.lc.txt reut2-007.sgm reut2-018.sgm

cat-descriptions\_120396.txt reut2-008.sgm reut2-019.sgm

feldman-cia-worldfactbook-data.txt reut2-009.sgm reut2-020.sgm

lewis.dtd reut2-010.sgm reut2-021.sgm

reut2-000.sgm reut2-011.sgm

reut2-001.sgm reut2-012.sgm

The README.txt file contains a detailed description of the dataset, and the other \*.txt. files contain corpus metadata. Each of the \*.sgm files contain the actual text of the articles in XML format as illustrated below:

<!DOCTYPE lewis SYSTEM "lewis.dtd">

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" NEWID="1">

<DATE>26-FEB-1987 15:01:01.79</DATE>

<TOPICS> <D>cocoa</D> </TOPICS>

<PLACES> <D>el-salvador</D> <D>usa</D> <D>uruguay</D> </PLACES>

<PEOPLE></PEOPLE>

<ORGS></ORGS>

<EXCHANGES></EXCHANGES>

<COMPANIES></COMPANIES>

<UNKNOWN>C T f0704reute u f BC-BAHIA-COCOA-REVIEW 02-26 0105</UNKNOWN>

<TEXT>

<TITLE>BAHIA COCOA REVIEW</TITLE>

<DATELINE>SALVADOR, Feb 26 -</DATELINE>

<BODY>Showers continued throughout the week in the Bahia cocoa zone,

alleviating the drought since early January and improving prospects for

…

after carnival which ends midday on February 27. Reuter</BODY>

</TEXT>

</REUTERS>

<REUTERS TOPICS="NO" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="5547" NEWID="4">

<DATE>26-FEB-1987 15:07:13.72</DATE>

<TOPICS></TOPICS>

<PLACES> <D>usa</D> <D>brazil</D> </PLACES>

<PEOPLE></PEOPLE>

<ORGS></ORGS>

<EXCHANGES></EXCHANGES>

<COMPANIES></COMPANIES>

<UNKNOWN>F f0725reute u f BC-TALKING-POINT/BANKAME 02-26 0105</UNKNOWN>

<TEXT>

<TITLE>TALKING POINT/BANKAMERICA &lt;BAC&gt; EQUITY OFFER</TITLE>

<AUTHOR>by Janie Gabbett, Reuters</AUTHOR>

<DATELINE>LOS ANGELES, Feb 26 -</DATELINE>

<BODY>BankAmerica Corp is not under pressure to act quickly on its proposed

equity offering and would do well to delay it because of the stock's recent

poor performance, banking analysts said. … </BODY>

</TEXT>

</REUTERS>

General Data Science Case Study

# Problem Statement

Data.gov is a U.S. government website launched in late May 2009 by the then Federal Chief Information Officer (CIO) of the United States. Data.gov aims to improve public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government. With the advent of modern Data Science, it is worth exploring it to check the value it can yield in the world of investing.

The task for you is to apply modern Data Science techniques to the [Allegheny County Property Assessments Dataset](https://data.wprdc.org/dataset/property-assessments) from Wprdc.org, with the purpose of identifying how this dataset could generate valuable insights for investing. We anticipate that candidates could use concepts from *probability theory*, *model selection*, *model validation and optimization techniques*, and other *Data Science* methods. Below are a few potential avenues for you to show off your Data Science skills, but the case study is open ended. Feel free to jump in and see where the data leads you!

* Is housing price in Allegheny county a martingale? Provide your answer using both a conceptual explanation and an empirical explanation using the case study dataset.
* Design and test a simple monthly "Allegheny County Home Value Index" using the case study dataset. As an example, see a methodology for a home value index described [here](https://www.zillow.com/research/zhvi-methodology-2019-deep-26226/).
* Using the case study dataset, design and test an investment strategy. Let the initial budget for your strategy be $5 million. Let the objective of your strategy be to maximize the value of your budgeted amount at investment time horizon by buying homes that appear on the market in Allegheny county starting January 1, 2016. Let the time horizon to check the resulting value of your investments be November 30, 2020. Use the case study dataset up until and including year 2015 for training and development, and test the developed strategy starting at year 2016. As needed, specify any additional assumptions for the analysis.

At the end of your work, you will present your analysis to our team as a part of your (virtual) onsite interview process. Extra points for informative visualizations. We also expect to review your code, so please upload your code at a convenient (eg. GitHub) location and convey the URL to us.

# Dataset

Download the dataset (Property Assessments Parcel Data) and other supporting files from this URL by clicking the corresponding “Download” buttons: <https://data.wprdc.org/dataset/property-assessments>

The main case study data file for Property Assessments Parcel Data (assessments.csv) is around 430MB uncompressed, and its timestamp varies depending on any data revisions. See the below screen capture for an example – the data file reference with the most recent time stamp of “FEB-2022” is shown in the green ellipse.

Graphical user interface, text, application

Description automatically generated

Please keep in mind that you need to download only one assessments.csv data file from the provided URL. **Use the file with the most recent time stamp that is available at the time of working on this case study. Please note, that for your case the time stamp may be more recent, such as MAR-2022, APR-2022 etc.**

After downloading the files at the URL above, you will see a list of files such as the below:

**$ ls**

alleghenycountypropertyassessmentdatauserguide-4.pdf

assessments.csv

property-assessment-data-dictionaryrev.pdf

property-assessments-data-dictionary.csv

The property-assessments-data-dictionary.csv, property-assessment-data-dictionaryrev.pdf

, and alleghenycountypropertyassessmentdatauserguide-4.pdf

files contain a detailed description of the dataset. The assessments.csv file contains the actual historical house attributes and sales records table in the standard comma-separated file format, as illustrated below:

**$ head -3 assessments.csv**

PARID,PROPERTYHOUSENUM,PROPERTYFRACTION,PROPERTYADDRESS,PROPERTYCITY,PROPERTYSTATE,PROPERTYUNIT,PROPERTYZIP,MUNICODE,MUNIDESC,SCHOOLCODE,SCHOOLDESC,LEGAL1,LEGAL2,LEGAL3,NEIGHCODE,NEIGHDESC,TAXCODE,TAXDESC,TAXSUBCODE,TAXSUBCODE\_DESC,OWNERCODE,OWNERDESC,CLASS,CLASSDESC,USECODE,USEDESC,LOTAREA,HOMESTEADFLAG,FARMSTEADFLAG,CLEANGREEN,ABATEMENTFLAG,RECORDDATE,SALEDATE,SALEPRICE,SALECODE,SALEDESC,DEEDBOOK,DEEDPAGE,PREVSALEDATE,PREVSALEPRICE,PREVSALEDATE2,PREVSALEPRICE2,CHANGENOTICEADDRESS1,CHANGENOTICEADDRESS2,CHANGENOTICEADDRESS3,CHANGENOTICEADDRESS4,COUNTYBUILDING,COUNTYLAND,COUNTYTOTAL,COUNTYEXEMPTBLDG,LOCALBUILDING,LOCALLAND,LOCALTOTAL,FAIRMARKETBUILDING,FAIRMARKETLAND,FAIRMARKETTOTAL,STYLE,STYLEDESC,STORIES,YEARBLT,EXTERIORFINISH,EXTFINISH\_DESC,ROOF,ROOFDESC,BASEMENT,BASEMENTDESC,GRADE,GRADEDESC,CONDITION,CONDITIONDESC,CDU,CDUDESC,TOTALROOMS,BEDROOMS,FULLBATHS,HALFBATHS,HEATINGCOOLING,HEATINGCOOLINGDESC,FIREPLACES,BSMTGARAGE,FINISHEDLIVINGAREA,CARDNUMBER,ALT\_ID,TAXYEAR,ASOFDATE

0001F00007000000,"601"," ","COMMONWEALTH PL","PITTSBURGH","PA"," ","15222","101","1st Ward - PITTSBURGH","47","Pittsburgh","LOT 363.86 X 280.14 X 110","","","51C01","PITTSBURGH URBAN","E","10 - Exempt","","","20","CORPORATION","G","GOVERNMENT","610","STATE GOVERNMENT","141961","","","","","","03-26-1948","0","3","LOVE&AFFECTION","","","","","","","300 LIBERTY AVE "," ","PITTSBURGH PA ","15222","1819400","8517700","10337100","0","1819400","8517700","10337100","1819400","8517700","10337100","","","","","","","","","","","","","","","","","","","","","","","","","","","","2020","01-DEC-20"

0001F00020000000,"601"," ","COMMONWEALTH PL","PITTSBURGH","PA"," ","15222","101","1st Ward - PITTSBURGH","47","Pittsburgh","LOT AVG705XAVG110","","","51C01","PITTSBURGH URBAN","E","10 - Exempt","","","20","CORPORATION","G","GOVERNMENT","610","STATE GOVERNMENT","155100","","","","","","12-22-1949","0","3","LOVE&AFFECTION","","","","","","","300 LIBERTY AVE "," ","PITTSBURGH PA ","15222","142700","9306000","9448700","0","142700","9306000","9448700","142700","9306000","9448700","","","","","","","","","","","","","","","","","","","","","","","","","","","","2020","01-DEC-20"