DATASETS AND READERS

Datasets

We use a number of datasets in this book. Some of them are freely available, some are freely downloadable but have licences or terms and conditions that restrict what you can do with them, some you have to pay for. We provide tools for unpacking most of the freely available datasets used in the book and putting them where they need to be, and for a few we provide tools for downloading them as well. Some of them have restrictions on what you can do with them, and most of them say that while we can write programs that do things using them we cannot directly redistribute them (but we can tell you where to download them from). You should always look at the licences before downloading them. We have tried to stick to corpora without restrictive licences where we can, and will always point to the licence conditions when we tell you where to get them from, **but it is up to you to obey the terms and conditions.**

The programs we will be using expect the corpora to be inside a directory called CORPORA. You can choose where that should be kept, but you should specify it inside the program file basics/corpora.py. I keep everything inside a folder called /Library/WebServer/CGI-Executables/SENTIMENTS, which is perhaps a slightly odd place to keep it but it works for me, so my basics/corpora.py contains the line

CORPORA = "/Library/WebServer/CGI-Executables/SENTIMENTS/CORPORA"

You need to set this to wherever you're going to want to keep the corpora.

In order to put corpora in the appropriate place inside CORPORA, they need to be downloaded. Some of the sources we use allow people to download the data and then make use of programs that manipulate it in various ways, but does not allow us to download and distribute it directly. We therefore have a directory inside CORPORA called DOWNLOADS which is where you should put the downloaded data. Our scripts will then unpack it and convert it into the form we need it to be and put it it in the correct place within CORPORA, but you will have to do the actual downloads. We will specify where to get it from and where you should put it in the relevant scripts. The general structure of this directory is as shown below (files that you will have to download are highlighted in red):

--|CORPORA--|DOWNLOADS--|TEXTS--|BNC--|2554.zip

|

|TREEBANKS--|UDT--|ud-treebanks-v2.12.tgz

|

|TWEETS--|CARER--|EN--|data.jsonl.gz

|

|dataset\_infos.json

|

|IMDB--|EN--|aclImdb\_v1.tar.gz

|

|SEM11--|AR--|E-c-Ar-dev.zip

|

|E-c-Ar-train.zip

|

|EN--|E-c-En-dev.zip

|

|E-c-En-train.zip

|

|ES--|E-c-Es-dev.zip

|

|E-c-Es-train.zip

|

|SEM4--|AR--|EI-oc-Ar-dev.zip

|

|EI-oc-Ar-train.zip

|

|EN--|EI-oc-En-dev.zip

|

|EI-oc-En-train.zip

|

|ES--|EI-oc-Es-dev.zip

|

|EI-oc-Es-train.zip

|

|WASSA--|EN--|anger-ratings-0to1.dev.target.txt

|

|anger-ratings-0to1.train.txt

|

|fear-ratings-0to1.dev.target.txt

|

|fear-ratings-0to1.train.txt

|

|joy-ratings-0to1.dev.target.txt

|

|joy-ratings-0to1.train.txt

|

|sadness-ratings-0to1.dev.target.txt

|

|sadness-ratings-0to1.train.txt

|

|TEXTS--|BNC--|download--|Doc--|HTML--|BNCdes.html

|

|bibliog.html

...

|

|Src

|

|Texts--|A--|A0--|A00.xml

|

|A01.xml

...

|

|A1--|A10.xml

|

|A11.xml

...

...

|

|B--|B0--|B01.xml

|

|B02.xml

...

|

|B1--|B10.xml

|

|B11.xml

...

...

...

|

|

|TREEBANKS--|UDT--|ud-treebanks-v2.12--|UD\_Abaza-ATB--|LICENSE.txt

|

|wholething

|

|UD\_Afrikaans-AfriBooms--|LICENSE.txt

|

|wholething

...

|

|TWEETS--|CARER--|EN--|wholething.csv

|

|IMDB--|EN--|wholething.csv

|

|KWT--|KWT.M--|AR--|wholething.csv

|

|KWT.U--|AR--|wholething.csv

|

|QATAR--|AR--|wholething.csv

|

|SEM11--|AR--|wholething.csv

|

|EN--|wholething.csv

|

|ES--|wholething.csv

|

|SEM4--|AR--|wholething.csv

|

|EN--|wholething.csv

|

|ES--|wholething.csv

|

|WASSA--|EN--|wholething.csv

Texts and Treebanks

We will be using two sorts of datasets. We use various textual resources – some just plain texts, some annotated texts, some treebanks – for developing and testing preprocessing steps like morphological analyis, identification of of compound terms and so on; and we use collections of annotated tweets for the main business of the book, i.e. developing and testing emotion identification algorithms. We start by looking at how to obtain and unpack the main textual resources.

British National Corpus (BNC)

To download and install the BNC, do the following.

Read the licence at <http://www.natcorp.ox.ac.uk/docs/licence.html>: you must obey the terms and conditions of this licence if you are going to use this resource. Then download it from

https://ota.bodleian.ox.ac.uk/repository/xmlui/bitstream/handle/20.500.12024/2554/2554.zip?sequence=3&isAllowed=y

and put the file 2554.zip that you get from there in <CORPORA>/DOWNLOADS/TEXTS/BNC (i.e. inside a directory called DOWNLOADS inside the one where you're putting the corpora, which you should set in basics/corpora.py as described above). Then, in the main directory where you are keeping the program code do:

>>> from basics import datasets

>>> bnc = datasets.BNC()

If there is a file called 2554.zip inside <CORPORA>/DOWNLOADS/TEXTS/BNC this will unzip it and put it in a directory with the structure shown above inside <CORPORA>/TEXTS/BNC, where the actual data is in files inside A0, A1, ... Once you've done this you can delete 2554.zip to save space – it's quite big so you may not want to keep it lying around once you've unpacked it.

If you call datasets.BNC() without having downloaded 2554.zip it will remind you of where to get it and where to put it.

>>> datasets.BNC()

You need to download the file 2554.zip from https://ota.bodleian.ox.ac.uk/repository/xmlui/bitstream/handle/20.500.12024/2554/2554.zip?sequence=3&isAllowed=y to <CORPORA>/DOWNLOADS/TEXTS/BNC before you can install the BNC. Please check the licence details at http://www.natcorp.ox.ac.uk/docs/licence.html before downloading it

Universal Dependency Treebank (UDT)

Again you should download the source for the UDT manually from

<https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-5150/ud-treebanks-v2.12.tgz?sequence=1&isAllowed=y.>

to

<CORPORA>/DOWNLOADS/TREEBANKS/UDT/ud-treebanks-v2.12.tgz

and then once you've done that you can unpack it and put it where the programs expect it to be by doing

>>> udt = datasets.UDT()

UNZIPPING THE MAIN SOURCE FILE

CONVERTING INDIVIDUAL TREEBANKS

REMOVING NON-CCAS FILES

This will put the UDT in a directory with the following structure inside CORPORA/TREEBANKS (it's inside TREEBANKS rather than TEXTS because the data is annotated so that it can be interpreted as a set of trees), where the actual data is in files in the wholething files inside the per language subdirectories. There is a lot more material in the UDT itself, but since we do not use it for anything we collect all the trees from files within a subdirectory into a single file called wholething and then remove everything else (except for the licence details for the subdirectory) to save space.

However, since some of the files have restrictive licences we read through all the licences and delete all the datasets which do not have Creative-Commons-Share-Alike (CCSA) licenses. CCSA licenses allow you to do most things that you would want to do (though you should check at <https://creativecommons.org/licenses/> to be sure), so we keep those and delete the others.

--|UDT--|ud-treebanks-v2.12--|UD\_Abaza-ATB--|LICENSE.txt

|

|wholething

|

|UD\_Afrikaans-AfriBooms--|LICENSE.txt

|

|wholething

|

|UD\_Akkadian-PISANDUB--|LICENSE.txt

|

|wholething

|

|...

As with the BNC, if you try to install the UDT without having downloaded the required file you'll get a message telling you where to get it from.

>>> udt = datasets.UDT()

UNZIPPING THE MAIN SOURCE FILE

Before you can install the UDT, you have to download the file

https://lindat.mff.cuni.cz/repository/xmlui/bitstream/handle/11234/1-5150/ud-treebanks-v2.12.tgz?sequence=1&isAllowed=y

to

<CORPORA>/DOWNLOADS/TREEBANKS/UDT/ud-treebanks-v2.12.tgz

(<CORPORA> is the place where you have chosen to store the datasets: you should set it in corpora.py).

The UDT consists of a collection of treebanks supplied by different people, with a variety of licences. Different treebanks from within the overall collection have different licences. When you download the main .tgz file you get all of them. This downloader will remove all except the ones with Creative-Commons-Attribution-ShareAlike licences (see https://creativecommons.org/licenses/? for moredetails) and the .tgz file itself when it finishes.

DO NOT USE THESE TREEBANKS UNLESS YOU AGREE TO THE TERMS OF C-C-A-S LICENCES.

Tweets

We also need datasets containing tweets annotated with emotion labels for developing and evaluating the actual emotion classification algorithms. Again these have a range of licences, so in some cases you need to download the data yourself before using our scripts to unpack and reorganise and in others the downloader itself downloads the data as well as unpacking it. **If you want to use one the downloadable datasets, you must read the licence/terms-and-conditions and you must abide by these.**

We keep the tweet data inside CORPORA/TWEETS in directories which are in turn split into language specific directories each of which contains a single file called wholething.csv. Wholething.csv will be a tab-separated file of the following form, i.e. the first two columns are the tweet ID and the tweet itself and the remainder specify what emotions it expresses – 1 denotes that this tweet does express this emotion and 0 that it does not. This format makes it easy to allow a single tweet to denote 0 emotions (if all the emotion columns are 0) or one emotion (if exactly one emotion column is 1) or more than one (if more than one column is 1).

|  |
| --- |
|  |
| |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | ID | tweet | anger | fear | joy | sadness | | 40000 | Depression sucks! #depression | 0 | 0 | 0 | 1 | | 40001 | Feeling worthless as always #depression | 0 | 0 | 0 | 1 | |
|  |

WASSA

The WASSA data comes from .<https://saifmohammad.com/WebPages/EmotionIntensity-SharedTask.html> The data itself is stored as a collection of files with names which indicate the emotion expressed by the tweets they contain, e.g. anger-ratings-0to1.train.txt and anger-ratings-0to1.dev.target.txt both contain tweets that express anger, and joy-ratings-0to1.train.txt and joy-ratings-0to1.dev.target.txt contain tweets that express joy. The WASSA data is split into training and development sets, but we merge these into a single file called wholething.csv because we want to be in control of which data is used for training, which is used for development and which for testing.

As usual, you have to download the datasets from the original website before you can start. In this case you need to download a set of files called

[http://saifmohammad.com/WebDocs/EmoInt%20Train%20Data/anger-ratings-0to1.train.txt](http://saifmohammad.com/WebDocs/EmoInt Train Data/anger-ratings-0to1.train.txt)

[http://saifmohammad.com/WebDocs/EmoInt%20Dev%20Data/anger-ratings-0to1.dev.target.txt](http://saifmohammad.com/WebDocs/EmoInt Dev Data/anger-ratings-0to1.dev.target.txt)

[http://saifmohammad.com/WebDocs/EmoInt%20Train%20Data/fear-ratings-0to1.train.txt](http://saifmohammad.com/WebDocs/EmoInt Train Data/fear-ratings-0to1.train.txt)

[http://saifmohammad.com/WebDocs/EmoInt%20Dev%20Data/fear-ratings-0to1.dev.target.txt](http://saifmohammad.com/WebDocs/EmoInt Dev Data/fear-ratings-0to1.dev.target.txt)

[http://saifmohammad.com/WebDocs/EmoInt%20Train%20Data/joy-ratings-0to1.train.txt](http://saifmohammad.com/WebDocs/EmoInt Train Data/joy-ratings-0to1.train.txt)

[http://saifmohammad.com/WebDocs/EmoInt%20Dev%20Data/joy-ratings-0to1.dev.target.txt](http://saifmohammad.com/WebDocs/EmoInt Dev Data/joy-ratings-0to1.dev.target.txt)

[http://saifmohammad.com/WebDocs/EmoInt%20Train%20Data/sadness-ratings-0to1.train.txt](http://saifmohammad.com/WebDocs/EmoInt Train Data/sadness-ratings-0to1.train.txt)

http://saifmohammad.com/WebDocs/EmoInt%20Dev%20Data/sadness-ratings-0to1.dev.target.txt

to

<CORPORA>/DOWNLOADS/TWEETS/WASSA/EN/anger-ratings-0to1.train.txt

<CORPORA>/DOWNLOADS/TWEETS/WASSA/EN/anger-ratings-0to1.dev.target.txt

<CORPORA>/DOWNLOADS/TWEETS/WASSA/EN/fear-ratings-0to1.train.txt

<CORPORA>/DOWNLOADS/TWEETS/WASSA/EN/fear-ratings-0to1.dev.target.txt

<CORPORA>/DOWNLOADS/TWEETS/WASSA/EN/joy-ratings-0to1.train.txt

<CORPORA>/DOWNLOADS/TWEETS/WASSA/EN/joy-ratings-0to1.dev.target.txt

<CORPORA>/DOWNLOADS/TWEETS/WASSA/EN/sadness-ratings-0to1.train.txt

<CORPORA>/DOWNLOADS/TWEETS/WASSA/EN/sadness-ratings-0to1.dev.target.txt

on your machine (remember, <CORPORA> is where you have set CORPORA to be in basics/corpora.py).

Once you've downloaded these files then

>>> wassa = datasets.WASSA()

will create <CORPORA>/TWEETS/WASSA/EN/wholething.csv, which is what we need.

SEM4/SEM11

To get the SEM4/SEM11 datasets, you have to download the source files. For SEM4 these are

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/EI-oc/English/2018-EI-oc-En-dev.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/EI-oc/English/EI-oc-En-train.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/EI-oc/Arabic/2018-EI-oc-Ar-dev.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/EI-oc/Arabic/2018-EI-oc-Ar-train.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/EI-oc/Spanish/2018-EI-oc-Es-dev.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/EI-oc/Spanish/2018-EI-oc-Es-train.zip>

which you have to download to

<CORPORA>/DOWNLOADS/TWEETS/SEM4/EN/2018-EI-oc-En-dev.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM4/EN/EI-oc-En-train.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM4/AR/2018-EI-oc-Ar-dev.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM4/AR/2018-EI-oc-Ar-train.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM4/ES/2018-EI-oc-Es-dev.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM4/ES/2018-EI-oc-Es-train.zip

and for SEM11 they are

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/E-c/English/2018-E-c-En-dev.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/E-c/English/2018-E-c-En-train.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/E-c/Arabic/2018-E-c-Ar-dev.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/E-c/Arabic/2018-E-c-Ar-train.zip>

<http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/E-c/Spanish/2018-E-c-Es-dev.zip>

http://saifmohammad.com/WebDocs/AIT-2018/AIT2018-DATA/E-c/Spanish/2018-E-c-Es-train.zip

which have to be downloaded to

<CORPORA>/DOWNLOADS/TWEETS/SEM11/EN/2018-E-c-En-dev.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM11/EN/2018-E-c-En-train.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM11/AR/2018-E-c-Ar-dev.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM11/AR/2018-E-c-Ar-train.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM11/ES/2018-E-c-Es-dev.zip

<CORPORA>/DOWNLOADS/TWEETS/SEM11/ES/2018-E-c-Es-train.zip

**Note the name of the English training set for SEM4, which does not fit the pattern of the names for the other SEM4 files.**

Once you have these files, doing

>>> sem4 = datasets.SEM4()

Converting data in /Library/WebServer/CGI-Executables/SENTIMENTS/CORPORA/TWEETS/SEM4

>>> sem11 = datasets.SEM11()

Converting data in /Library/WebServer/CGI-Executables/SENTIMENTS/CORPORA/DOWNLOADS/TWEETS/SEM11

will make

--|TWEETS--|SEM11--|AR--|wholething.csv

|

|EN--|wholething.csv

|

|ES--|wholething.csv

|

|SEM4--|AR--|wholething.csv

|

|EN--|wholething.csv

|

|ES--|wholething.csv

As always, if you haven't download the files to the right place you will get a message telling you where to get them and where to put them.

CARER

To get the CARER dataset, download

<https://huggingface.co/datasets/dair-ai/emotion/resolve/main/data/dataset_infos.json>

to

<CORPORA>/DOWNLOADS/TWEETS/CARER/EN/dataset\_infos.json

and

<https://huggingface.co/datasets/dair-ai/emotion/resolve/main/data/data.jsonl.gz>

to

<CORPORA>/DOWNLOADS/TWEETS/CARER/EN/data.jsonl.gz

Doing

carer = datasets.CARER()

will then create the required <CORPORA>/TWEETS/CARER/EN/wholething.csv

IMDB

To get the IMDB dataset, download

<https://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz>

to

<CORPORA>/DOWNLOADS/TWEETS/IMDB/EN/aclImdb\_v1.tar.gz

(this one is quite big and will take a few minutes to download) and do

>>> imdb = datasets.IMDB()

UNZIPPING MAIN FILE

CONVERTING DATA FILES

KWT

The KWT.M and KWT.U datasets are our own and hence we can distribute them with the code, so there is nothing to download. Just do

>>> kwtm = datasets.KWTM()

>>> kwtu = datasets.KWTU()