261072449_Assignment_5

March 20, 2024

NL2DS - Winter 2024

Assignment 5 – Language Phylogeny, Clustering

Name: **Sam Zhang**

Student ID: **261072449**

This homework consists of 38 points.

There are two types of exercise:

- "Problems" require writing code.
  - Replace `# your code here` with your answer.
  - The code block should run when all code above it in this file has also been run.
  - If you skip some problems, it's your responsibility to make sure that all code blocks which you filled out still run.
- "Questions" require writing text. Replace "**put your answer here**" with your answer.

For "Problems": * **You may find code from the course CoLab notebooks useful for this assignment.**
* Every `# Put your answer here` can be solved by a few lines of code, often 1-2 lines.
* **Do not reimplement any major functionality, such as calculating edit distance, linkage methods, etc.** * Following the contents of these CoLab notebooks, you should: * Use `sklearn` functionality as much as possible for machine learning tools. (For example, do not compute clusters in Problem 4 using a different library.) * Use `pandas` functionalty as much as possible for basic data manipulation and analysis. * Do not delete any code, unless it is marked as `# some code to get you started`.

Please make sure to follow directions carefully, including maximum lengths for "Question" answers. Failure to follow directions may result in partial or no credit for the relevant problem/question.

In this assignment, we will look at some cross-linguistic word form data and use some of the tools we saw in class to build family trees of languages based on the sound forms of words—otherwise known as "optimal phylogenies."

We will use data from the following recent paper.

Dellert, Johannes, Daneyko, T., Muench, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., Muehlenbernd, R., Wahle, J., and Jaeger, G. (2020). Northeuralex: A wide-coverage lexical database of northern eurasia. Language Resources & Evaluation, 54(273–301).

This data can be found here as well.

Copy the data to your drive folder from: here, here, and here.

1 Part 1

1.1 Question 1 (3 points)

Question 1: Read the paper and/or Northeuralex's website as much as necessary to answer this question.

What is the Northeuralex dataset? Give a brief overview, including: * What kind of data is it? * What is the purpose of this data (what sorts of scientific questions or practical applications will it be used to address)? * How was it constructed?

Your answer should not refer to low-level details, such as file names or what columns are present in different files. Just give an overview of no more than one paragraph that gives the gist for someone unfamiliar with the dataset.

Answer 1: The NorthEuraLex dataset is a wide-coverage lexical database focusing on Northern Eurasia, including Europe. It's a collection of data designed to support computational studies in historical linguistics by providing machine-readable information that linguists typically gather from dictionaries or other sources. The dataset encompasses a large list of 1,016 concepts across 107 languages from diverse language families, totaling over 121,614 dictionary forms uniformly transcribed into the International Phonetic Alphabet (IPA). The dataset was constructed using a systematic procedure to extract lexical data from existing resources, supplemented by input from experts and native speakers for revisions and quality improvements. It aims to offer a reliable benchmark for testing automated methods in historical linguistics, especially for under-researched areas, by providing a uniform IPA transcription across all included word lists, which can be converted into other transcription formats as needed

1.2 Question 2 (3 points)

Now, let's read in the wordforms in this dataset.

[98]:
```
from google.colab import drive
drive.mount('/content/drive/')

import pandas as pd
wordforms=pd.read_csv("/content/drive/My Drive/northeuralex.csv")
display(wordforms)
```

```
Drive already mounted at /content/drive/; to attempt to forcibly remount, call
drive.mount("/content/drive/", force_remount=True).
```

	Language_ID	Glottocode	Concept_ID	Word_Form	rawIPA	\
0	fin	finn1318	Auge::N	silmä	silmæ	
1	fin	finn1318	Ohr::N	korva	k r	
2	fin	finn1318	Nase::N	nenä	n næ	
3	fin	finn1318	Mund::N	suu	su	
4	fin	finn1318	Zahn::N	hammas	h m s	

```
…            …        …              …        …            …
121608         cmn   mand1415    verkaufen::V                  mâ͡
121609         cmn   mand1415     bezahlen::V           fû t j͡n
121610         cmn   mand1415       zahlen::V              t  fû
121611         cmn   mand1415  beherrschen::V            t͡ŋtˆ
121612         cmn   mand1415     ertragen::V              ə͡nnâ͡

                      IPA    ASJP      List     Dolgo Next_Step
0               s i l m æ   silmE     SILME     SVRMV  validate
1               k   r      korwa     KURWA     KVRWV  validate
2                 n   n æ    nEnE      NENE      NVNV  validate
3                   s u u      su        SY        SV  validate
4             h   m m   s   hamas     HAMAS     HVMVS  validate
…                   …     …       …        …       …
121608              m a       mai       MAI        MV  validate
121609  f u _ t   j   n  fuCyEn  BY_CJE2N  PV_KJV1N  validate
121610          t    f u    C3fu     CI1BY     KV1PV  validate
121611       t   ŋ t      tuNC3    TY2NCI    TV1NKV  validate
121612         ə n n a    L3nai   RE2NNAI    RV1NNV  validate

[121613 rows x 10 columns]
```

Question 2: Describe the meaning of the `Language_ID`, `Concept_ID`, `rawIPA` and `IPA` columns of the data. Why are there separate `rawIPA` and `IPA` columns?

A2:

- Language_ID: The identifier of a specific language within the dataset, each Language_ID is unique to a language.
- Concept_ID: The identifier to a specific concept or lexical entry within the dataset, used to uniquely identify the meanings or concepts for which words are provided across different languages.
- rawIPA: The rawIPA column contains the phonetic transcriptions of words as they were initially generated or extracted from the source materials, prior to any standardization or normalization.
- IPA: The standardized or normalized International Phonetic Alphabet (IPA) transcriptions of the lexical items.

The existence of separate rawIPA and IPA columns allows for a distinction between the unedited phonetic transcriptions directly taken or automatically generated from the original resources (rawIPA) and the standardized or normalized phonetic transcriptions (IPA) that have been processed for consistency and comparability between datapoints.

1.3 Question 3 (2 points)

Now let's read in some metadata about the languages.

[99]:
```
languages=pd.read_csv("/content/drive/My Drive/northeuralex-languages.csv")
display(languages)
```

```
                name glotto_code iso_code             family     subfamily  \
0            Finnish    finn1318      fin             Uralic        Finnic
1      North Karelian    kare1335      krl             Uralic        Finnic
2    Olonets Karelian    livv1243      olo             Uralic        Finnic
3               Veps    veps1250      vep             Uralic        Finnic
4           Estonian    esto1258      ekk             Uralic        Finnic
..               …         …      …                 …             …
102           Dargwa    darg1241      dar  Nakh-Daghestanian  Daghestanian
103          Chechen    chec1245      che  Nakh-Daghestanian          Nakh
104   Standard Arabic    stan1318      arb       Afro-Asiatic       Semitic
105     Modern Hebrew    hebr1245      heb       Afro-Asiatic       Semitic
106  Mandarin Chinese    mand1415      cmn       Sino-Tibetan       Sinitic

     latitude  longitude
0     61.0000    24.4500
1     65.1691    30.8655
2     61.0000    33.0000
3     60.3353    34.7865
4     59.2500    24.7500
..        …          …
102   42.4257    47.4388
103   43.5000    45.5000
104   27.9625    43.8525
105   31.1056    35.0179
106   40.0209   116.2280

[107 rows x 7 columns]
```

Question 3: Describe the meaning of the `family`, `iso_code`, and `subfamily` columns of the data.

A3: - family: The language family (group of languages derived from a common ancestral language) to which a particular language belongs. - iso_code: Unique identifier of a language based on the ISO 639-3 codes, which are the International Standard for language codes. - subfamily: Further classification of the language within a particular language family into smaller groups based on more specific shared linguistic features.

1.4 Question 4 (2 points)

Now let's read in some further data about the concepts.

[100]:
```
concepts=pd.read_csv("/content/drive/My Drive/northeuralex-concepts.csv")
display(concepts)
```

```
     number  position_in_ranking ranking_value        id_nelex gloss_en  \
0         1                   44     -2,539237         Auge::N      eye
1         2                   34     -2,649194          Ohr::N      ear
```

```
2          3                  149     -1,995463         Nase::N     nose
3          4                   25     -2,762589         Mund::N    mouth
4          5                   31     -2,670705         Zahn::N    tooth
…      …                   …           …               …      …
1011    1012                  140     -2,029052    verkaufen::V     sell
1012    1013                  198     -1,822012     bezahlen::V  pay for
1013    1014                  235     -1,715766       zahlen::V      pay
1014    1015                  899      0,118183  beherrschen::V     rule
1015    1016                  751     -0,491453     ertragen::V   endure

       gloss_ru                    annotation_en       annotation_en.1  \
0                                 [[Anatomie]]           [[anatomy]]
1                                 [[Anatomie]]           [[anatomy]]
2                                 [[Anatomie]]           [[anatomy]]
3                                 [[Anatomie]]           [[anatomy]]
4              [BSP:menschlicher Schneidezahn]    [EX:human incisor]
…         …                               …                 …
1011                             [BSP:Ware]            [EX:goods]
1012                            [BSP:Ware]            [EX:goods]
1013                    [BSP:im Restaurant]  [EX:in a restaurant]
1014                             [BSP:Land]          [EX:country]
1015                          [BSP:Schmerz]            [EX:pain]

              annotation_ru concepticon  concepticon_id concepticon_proposed  \
0               [[    ]]         EYE            1248                  EYE
1               [[    ]]         EAR            1247                  EAR
2               [[    ]]        NOSE            1221                 NOSE
3               [[    ]]       MOUTH             674                MOUTH
4            [  :    ]       TOOTH            1380                TOOTH
…                     …         …               …                    …
1011            [  :   ]        SELL            1571                 SELL
1012         [  :     ]         NaN               0              PAY_FOR
1013      [  :       ]         PAY             718                  PAY
1014          [  :    ]        RULE            1846                 RULE
1015  [     ,   :  ]      ENDURE             833               ENDURE

     comments
0         NaN
1         NaN
2         NaN
3         NaN
4         NaN
…       …
1011      NaN
1012      NaN
1013      NaN
1014      NaN
1015      NaN
```

[1016 rows x 13 columns]

Question 4: Describe the meaning of the `id_nelex`, `gloss_en`, and `position_in_ranking` columns of the data.

A4: - id_nelex: The unique identifier of the concept - gloss_en: The English translation of the concept represented by the lexical item - position_in_ranking: The represents the ranking of concepts based on a basicness score developed by Dellert and Buch (2018).

2 Part 2

2.1 Problem 1 (2 points)

It will be useful to merge all of the meta-information into the main wordforms dataframe.

[101]:
```
# Problem 1a: rename the appropriate columns in the languages and concepts␣
 ↪dataframes to make this merge possible.
languages = languages.rename(columns={
  'iso_code':'Language_ID'
})
concepts = concepts.rename(columns={
  'id_nelex':'Concept_ID'
})

# Problem 1b: Use the merge function to merge the three dataframes into one.
wordforms = wordforms.merge(languages).merge(concepts)

display(wordforms)
```

```
       Language_ID Glottocode   Concept_ID Word_Form      rawIPA  \
0              fin   finn1318      Auge::N     silmä       silmæ
1              krl   kare1335      Auge::N     silmä       silmæ
2              olo   livv1243      Auge::N     silmy        silm
3              vep   veps1250      Auge::N     sil'm       sil m
4              ekk   esto1258      Auge::N      silm       s ilm
…              …         …          …       …          …
121608         che   chec1245  ertragen::V            sədet ə
121609         arb   stan1318  ertragen::V              at a qa
121610         arb   stan1318  ertragen::V             taħamala
121611         heb   hebr1245  ertragen::V                saval
121612         cmn   mand1415  ertragen::V            ə̃nnə̂

                    IPA      ASJP      List     Dolgo Next_Step  …  \
0             s i l m æ     silmE     SILME     SVRMV  validate  …
1             s i l m æ     silmE     SILME     SVRMV  validate  …
2             s i l m       silmi     SILMY     SVRMV  validate  …
3              s i l  m      silm      SILM      SVRM  validate  …
4              s  i l m      silm      SILM      SVRM  validate  …
```

```
…                  …       …       …       …       …  …
121608    s ə d e t t ə    s3det3    SETETE    SVTVTV    review  …
121609     a t  a a q a     ataqa     ATAKA     VTVKV  validate  …
121610  t a ħ a m a l a  taGamala  TAHAMALA  TVHVMVRV  validate  …
121611        s a v a l     saval     SABAL     SVWVR  validate  …
121612       ə n n a       L3nai   RE2NNAI    RV1NNV  validate  …

       ranking_value gloss_en   gloss_ru  annotation_en  annotation_en.1  \
0          -2,539237      eye            [[Anatomie]]      [[anatomy]]
1          -2,539237      eye            [[Anatomie]]      [[anatomy]]
2          -2,539237      eye            [[Anatomie]]      [[anatomy]]
3          -2,539237      eye            [[Anatomie]]      [[anatomy]]
4          -2,539237      eye            [[Anatomie]]      [[anatomy]]
…                …      …         …            …                 …
121608     -0,491453   endure         [BSP:Schmerz]        [EX:pain]
121609     -0,491453   endure         [BSP:Schmerz]        [EX:pain]
121610     -0,491453   endure         [BSP:Schmerz]        [EX:pain]
121611     -0,491453   endure         [BSP:Schmerz]        [EX:pain]
121612     -0,491453   endure         [BSP:Schmerz]        [EX:pain]

                annotation_ru  concepticon  concepticon_id  \
0                  [[    ]]          EYE            1248
1                  [[    ]]          EYE            1248
2                  [[    ]]          EYE            1248
3                  [[    ]]          EYE            1248
4                  [[    ]]          EYE            1248
…                        …          …              …
121608  [     ,    :  ]       ENDURE            833
121609  [     ,    :  ]       ENDURE            833
121610  [     ,    :  ]       ENDURE            833
121611  [     ,    :  ]       ENDURE            833
121612  [     ,    :  ]       ENDURE            833

       concepticon_proposed comments
0                       EYE      NaN
1                       EYE      NaN
2                       EYE      NaN
3                       EYE      NaN
4                       EYE      NaN
…                        …      …
121608               ENDURE      NaN
121609               ENDURE      NaN
121610               ENDURE      NaN
121611               ENDURE      NaN
121612               ENDURE      NaN

[121613 rows x 28 columns]
```

2.2 Problem 2 (2 points)

In this problem set, we will make use of the `lingpy` package of tools for historical linguistics. You can find more information on this here. We'll start by installing the package.

[ ]:
```
!pip install lingpy
```

In order to make our computations below more manageable, we will focus on the Indo-european languages which you can read more about here. We will also focus just on the top 20 concepts as determined by their rank.

[103]:
```
#Problem 2a: Filter out the non-Indo-European languages from the dataframes
wordforms = wordforms[wordforms['family']== 'Indo-European']

#Problem 2b: Filter the concepts to include those less than or equal to rank 20␣
 ↪in the dataframe.
wordforms = wordforms[wordforms['position_in_ranking'] <= 20]

display(wordforms)
```

```
       Language_ID Glottocode Concept_ID Word_Form  rawIPA          IPA  ASJP  \
7163           ben   beng1280  Wasser::N            d l      d   l   jol
7164           hin   hind1269  Wasser::N            d əl      d  ə l   j3l
7165           hin   hind1269  Wasser::N          pa ni   p a a n i i  pani
7166           pbu   nort2646  Wasser::N            o bə        o b ə   ob3
7167           pes   west2369  Wasser::N               b           b    ob
…              …         …        …       …     …           …    …
107539         cat   stan1289   geben::V     donar   duna      d u n a  duna
107540         spa   stan1288   geben::V       dar     da        d a    dar
107541         por   port1283   geben::V       dar     dar       d a r   dar
107542         ita   ital1282   geben::V      dare    dare      d a r e  dare
107543         ron   roma1327   geben::V        da      da          d a    da

        List Dolgo Next_Step  … ranking_value gloss_en gloss_ru  \
7163     CUL   KVR  validate  …      -2,92383    water
7164     CEL   KVR  validate  …      -2,92383    water
7165    PANI  PVNV  validate  …      -2,92383    water
7166     UPE   VPV  validate  …      -2,92383    water
7167      OP    VP  validate  …      -2,92383    water
…      …   …       …  …           …       …      …
107539  TYNA  TVNV  validate  …     -3,482883     give
107540   TAR   TVR  validate  …     -3,482883     give
107541   TAR   TVR  validate  …     -3,482883     give
107542  TARE  TVRV  validate  …     -3,482883     give
107543    TA    TV  validate  …     -3,482883     give

                    annotation_en  annotation_en.1    annotation_ru  \
7163               [kaltes Wasser]     [cold water]  [       ]
7164               [kaltes Wasser]     [cold water]  [       ]
```

```
7165                [kaltes Wasser]     [cold water]  [       ]
7166                [kaltes Wasser]     [cold water]  [       ]
7167                [kaltes Wasser]     [cold water]  [       ]
…                               …                …             …
107539  [allgemein, BSP:Gegenstand]               []            []
107540  [allgemein, BSP:Gegenstand]               []            []
107541  [allgemein, BSP:Gegenstand]               []            []
107542  [allgemein, BSP:Gegenstand]               []            []
107543  [allgemein, BSP:Gegenstand]               []            []

        concepticon  concepticon_id concepticon_proposed comments
7163          WATER             948                WATER      NaN
7164          WATER             948                WATER      NaN
7165          WATER             948                WATER      NaN
7166          WATER             948                WATER      NaN
7167          WATER             948                WATER      NaN
…               …               …                    …        …
107539         GIVE            1447                 GIVE      NaN
107540         GIVE            1447                 GIVE      NaN
107541         GIVE            1447                 GIVE      NaN
107542         GIVE            1447                 GIVE      NaN
107543         GIVE            1447                 GIVE      NaN

[817 rows x 28 columns]
```

3 Part 3

3.1 Problem 3 (6 points)

Our goal is to use agglomerative clustering to try to reconstruct the tree for the indoeuropean languages. You can find a reference tree (for families) here.

In order to do this, we will need to construct a matrix of similarities between the languages, called a confusion matrix.

We will compute the (normalized) levenshtein distance between the strings for each concept for each pair of languages. For instance, we will compute the normalized levenshtein distance between the words for Wasser::N (water in English) for German and English and then similarily for all 19 other concepts. If there are multiple words for the same concept, take the average across all pair possibilities. We will then average these values (i.e., average across all concepts) to find the similarity between German and English. We will do this for all pairs of languages to create a list of lists representing the confusion matrix.

Note that running your code will take a few minutes.

Hint: Make use of the `lp.align.pairwise.edit_dist` function from lingpy.

[ ]:
```
import lingpy as lp
import numpy as np
from tqdm import tqdm
```

```
#Problem 3: fill the confusion matrix on
#the "IPA" fields for each language.

#initialize confusion matrix
language_list = list(wordforms['Language_ID'].unique())
concept_list = list(wordforms['Concept_ID'].unique())
confusion = [[0 for j in range(len(language_list))] for i in␣
 ↪range(len(language_list))]

def normalized_levenshtein(words1, words2):
    distances = []
    for word1 in words1:
        for word2 in words2:
            distance = lp.align.pairwise.edit_dist(word1, word2,␣
 ↪normalized=True)
            distances.append(distance)
    return np.mean(distances)

with tqdm(total=len(language_list)**2) as pbar:
  for i, language1 in enumerate(language_list):
      for j, language2 in enumerate(language_list):
          pbar.update(1)
          if i >= j:
              # the matrix is symmetric, so we don't need to calculate twice
              continue

          language_distances = []
          for concept in concept_list:
              words_language1 = wordforms[(wordforms['Language_ID'] ==␣
 ↪language1) & (wordforms['Concept_ID'] == concept)]['IPA'].tolist()
              words_language2 = wordforms[(wordforms['Language_ID'] ==␣
 ↪language2) & (wordforms['Concept_ID'] == concept)]['IPA'].tolist()
              distance = normalized_levenshtein(words_language1,␣
 ↪words_language2)
              language_distances.append(distance)

          avg_distance = np.mean(language_distances)
          confusion[i][j] = avg_distance
          confusion[j][i] = avg_distance
```

After running it, clear the *output* of the above cell (by clicking on the cross at top left of the output part) so that it doesn't clutter the pdf.

3.2 Question 5 (2 points)

Now that we have computed a matrix of similarities, we can use clustering algorithms to try to build phylogenetic trees representing the languages historical relationships. First, let's use the `lp.algorithm.clustering.flat_cluster` function from `lingpy` to derive a flat clustering of languages.

[105]:
```
# let's add a lihne of code here to make them readable
language_map = dict(zip(languages['Language_ID'], languages['name']))
language_list = list([language_map[lang] for lang in language_list])

lp.algorithm.clustering.flat_cluster('upgma', 0.6, confusion, language_list)
```

[105]:
```
{0: ['Bengali', 'Hindi'],
 2: ['Northern Pashto'],
 3: ['Western Farsi', 'Northern Kurdish'],
 5: ['Ossetian'],
 6: ['Armenian'],
 7: ['Modern Greek'],
 8: ['Standard Albanian'],
 10: ['Croatian',
  'Slovene',
  'Bulgarian',
  'Slovak',
  'Polish',
  'Czech',
  'Belarusian',
  'Russian',
  'Ukrainian',
  'Lithuanian',
  'Latvian'],
 20: ['Icelandic',
  'Norwegian (Bokmål)',
  'Swedish',
  'Danish',
  'German',
  'Dutch',
  'English'],
 27: ['Irish'],
 28: ['Welsh', 'Breton'],
 30: ['Latin'],
 31: ['French'],
 32: ['Catalan', 'Spanish', 'Portuguese', 'Italian', 'Romanian']}
```

Question 5: Do you recognize any of the clusters of languages? Are there any noteworthy errors in this clustering? (You may first need to learn a bit about Indo-European languages.)

A5: We observe a few obvious clusters, such as Slavic languages in number 10, Bengali and Hindi in number 1 (two very close countries geographically), and germanic languages in number 20.

There are a few questionable cluster placements: The individual placcement of Latin and French are interesting, as they may definitely a place in the Romance language cluster (32), particularly, Latin is generally considered an ancestral language of modern romance languages, and French is unequivocally a romance language with similar features as Spanish and Italian.

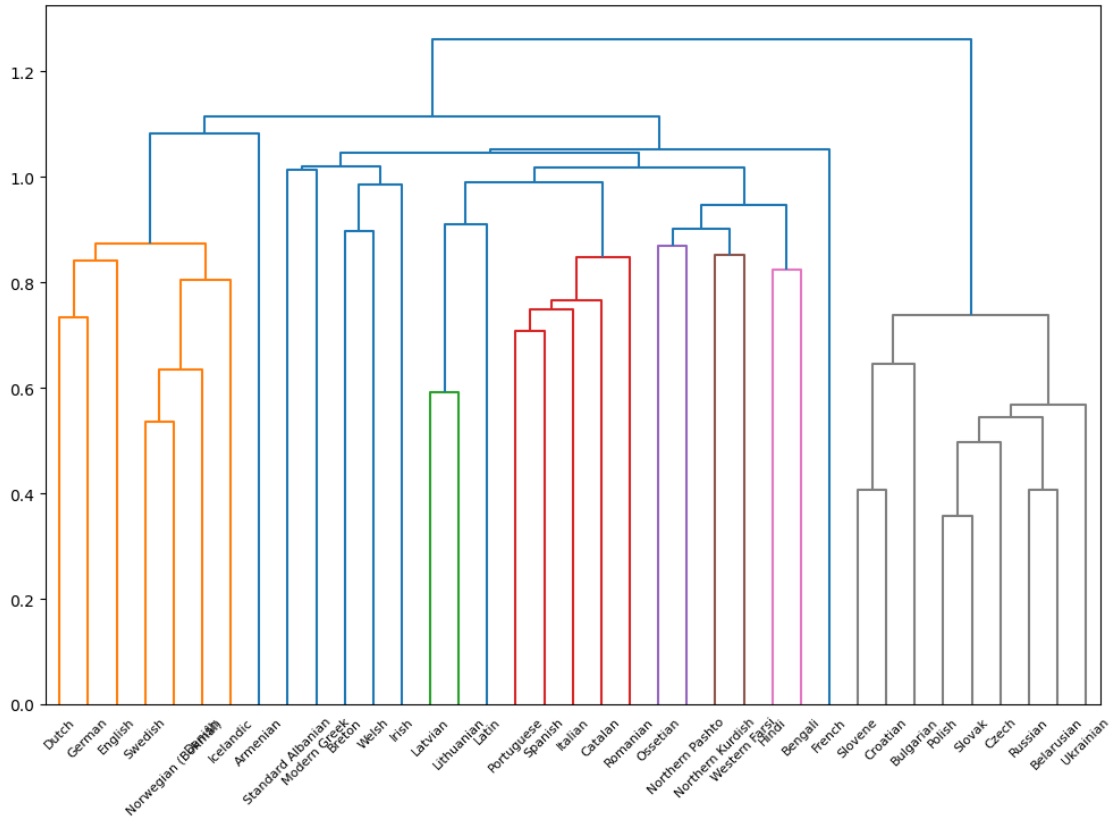
3.3 Part 4

3.4 Problem 4 (2 points)

Now we will build our own dendrogram using the clustering algorithms available in `scipy`. You can read in particular about the `linkage` function and the `dendrogram` function.

[106]:
```
from scipy.cluster.hierarchy import dendrogram, linkage
from sklearn.metrics import v_measure_score
import matplotlib.pyplot as plt

#Problem 4: use the linkage function with the average linkage method to compute␣
 ↪the clustering.

linked = linkage(confusion, method='average')

#plot the results using dendrogram
def llf(id):
  return language_list[id]

plt.figure(figsize=(12, 8))
dendrogram(linked,
           p=100,
           truncate_mode="level",
           orientation='top',
           distance_sort='descending',
           show_leaf_counts=False,
           leaf_label_func=llf)

plt.show()
```

```
<ipython-input-106-ae75ce65eb5e>:7: ClusterWarning: scipy.cluster: The symmetric
non-negative hollow observation matrix looks suspiciously like an uncondensed
distance matrix
  linked = linkage(confusion, method='average')
```



3.5 Question 6 (2 points)

Question 6: Do you recognize any of the clusters of languages at any of the levels? Are there any noteworthy errors in this clustering?

A6: We observe some similar groupings as Q5, such as Germanic languages, Romance languages and Slavic languages. A few pairings are unusual, such as the two errors mentioned in Q5, but also, Latin with Latvian and Lithuanian, who are generally not associated together. Albanian appears to be grouped with Modern Greek, which also not that share many similarities linguistically.

4 Part 5

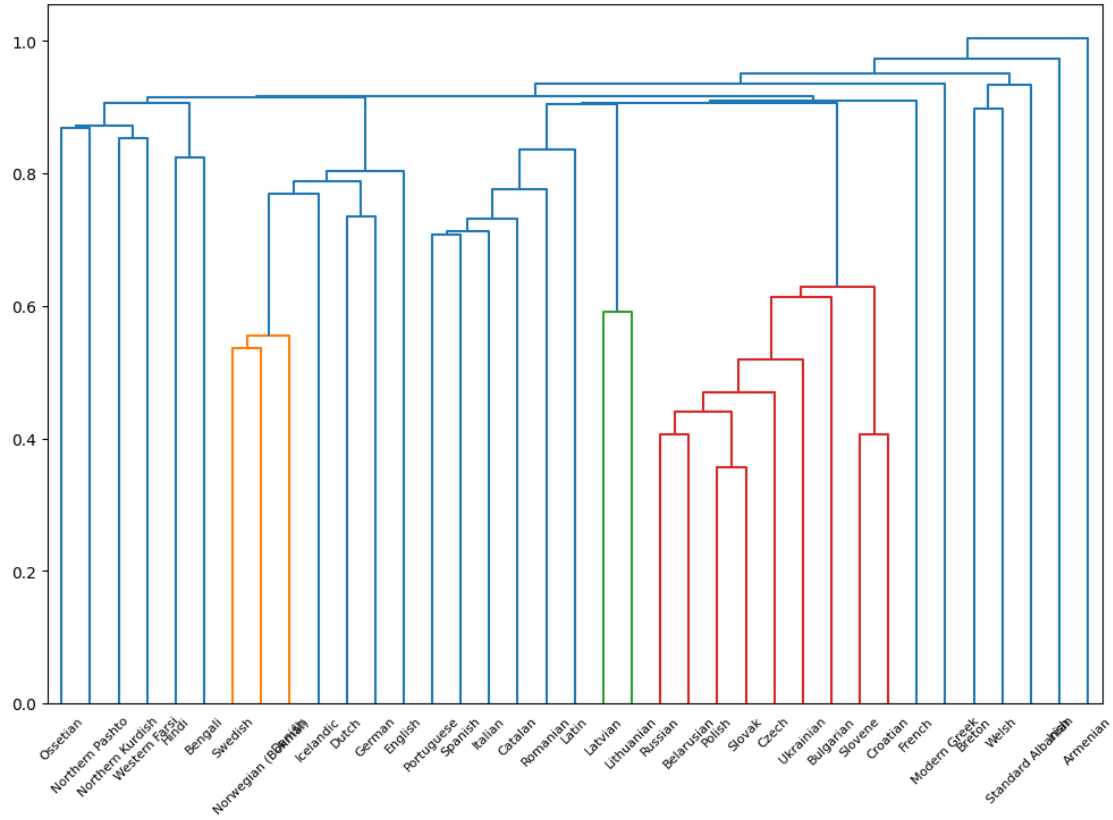
4.1 Problem 5 / Question 7 (4 points)

(4 points: 2 points for code, 2 points for answer)

[107]:
```
linked = linkage(confusion, method='single')

plt.figure(figsize=(12, 8))
dendrogram(linked,
           p=100,
           truncate_mode="level",
```

```
            orientation='top',
            distance_sort='descending',
            show_leaf_counts=False,
            leaf_label_func=llf)

plt.show()
```

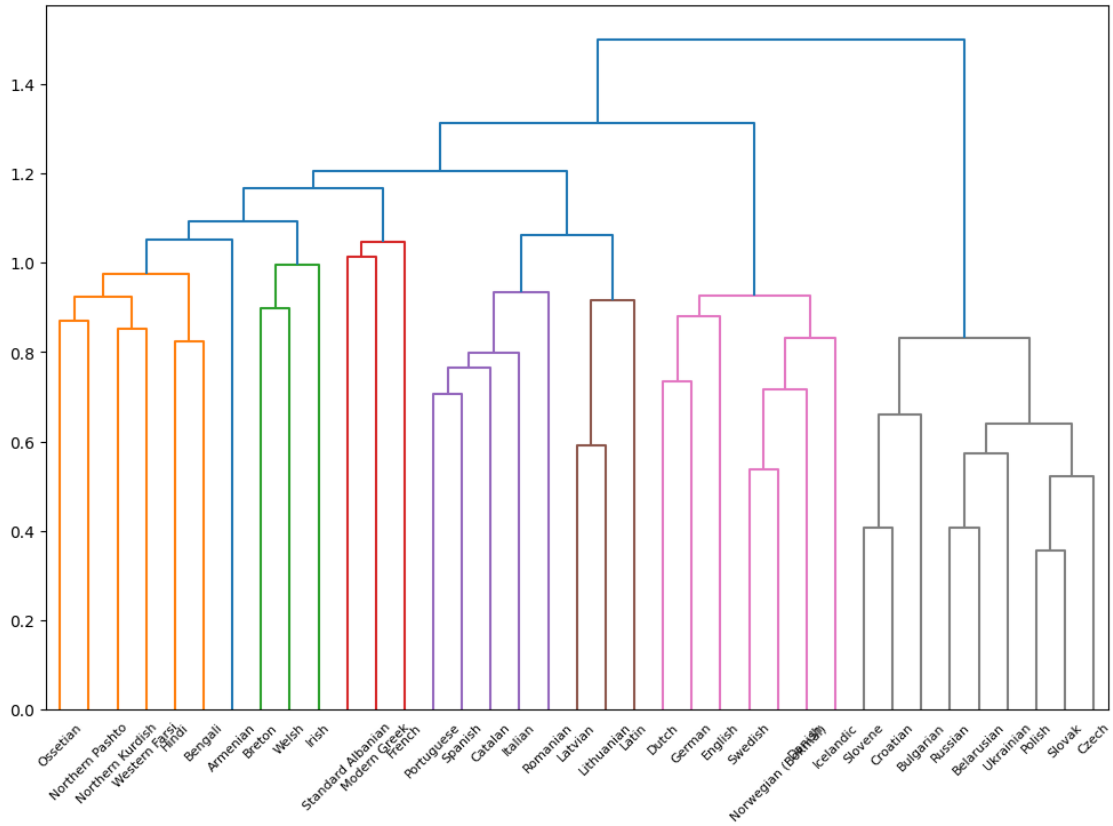
```
<ipython-input-107-eb5f51be08f9>:1: ClusterWarning: scipy.cluster: The symmetric
non-negative hollow observation matrix looks suspiciously like an uncondensed
distance matrix
  linked = linkage(confusion, method='single')
```



[108]:
```
linked = linkage(confusion, method='complete')

plt.figure(figsize=(12, 8))
dendrogram(linked,
            p=100,
            truncate_mode="level",
            orientation='top',
            distance_sort='descending',
            show_leaf_counts=False,
```

```
            leaf_label_func=llf)

plt.show()
```

```
<ipython-input-108-9b673b1df503>:1: ClusterWarning: scipy.cluster: The symmetric
non-negative hollow observation matrix looks suspiciously like an uncondensed
distance matrix
  linked = linkage(confusion, method='complete')
```



Question 7: Try two of the other linkage methods and describe how they change the results.

A7: In the `single` method, where languages are joined one at a time, we observe less compact clusters, with more elonged and less compact branches than with the `average` linkage method, implyinga greater distance between clusters. The `complete` method that minimizes the maximumintra-cluster distance shows the opposite, with essentially 7 large clusters of languages, in some ways performing better as it included French in the Romance languages, but have also mistakenly over-grouped some languages, some as Persian language Farsi with Hindi.

4.2 Problem 6 / Question 8 (4 points)

(4 points: 2 points for code, 2 points for answer)

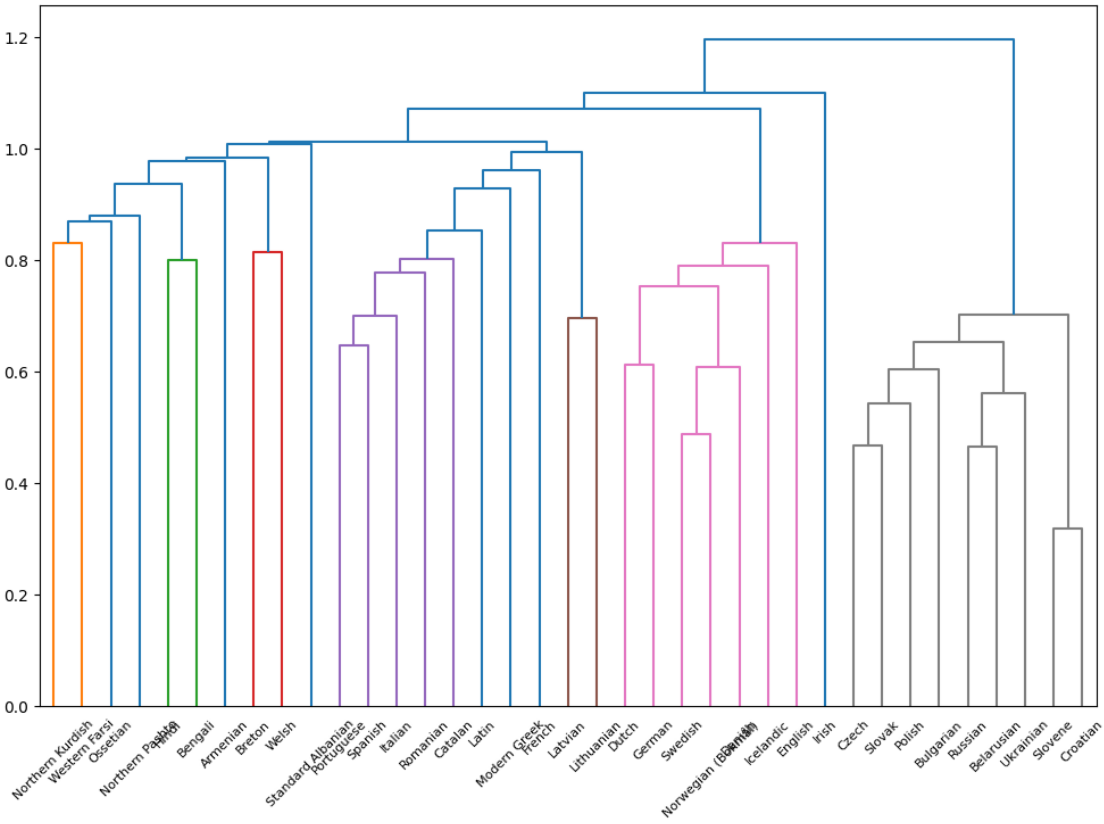
[123]:
```
# re-read wordforms for the complete dataframe
wordforms=pd.read_csv("/content/drive/My Drive/northeuralex.csv")
wordforms = wordforms.merge(languages).merge(concepts)
```

[ ]:
```
wordforms = wordforms[wordforms['family']== 'Indo-European']
wordforms = wordforms[wordforms['position_in_ranking'] <= 70]
display(wordforms)

language_list = list(wordforms['Language_ID'].unique())
concept_list = list(wordforms['Concept_ID'].unique())
confusion_more_concepts = [[0 for j in range(len(language_list))] for i in␣
 ↪range(len(language_list))]

with tqdm(total=len(language_list)**2) as pbar:
  for i, language1 in enumerate(language_list):
      for j, language2 in enumerate(language_list):
          pbar.update(1)
          if i >= j:
              # the matrix is symmetric, so we don't need to calculate twice
              continue

          language_distances = []
          for concept in concept_list:
              words_language1 = wordforms[(wordforms['Language_ID'] ==␣
 ↪language1) & (wordforms['Concept_ID'] == concept)]['IPA'].tolist()
              words_language2 = wordforms[(wordforms['Language_ID'] ==␣
 ↪language2) & (wordforms['Concept_ID'] == concept)]['IPA'].tolist()
              distance = normalized_levenshtein(words_language1,␣
 ↪words_language2)
              language_distances.append(distance)

          avg_distance = np.mean(language_distances)
          confusion_more_concepts[i][j] = avg_distance
          confusion_more_concepts[j][i] = avg_distance

language_list = list([language_map[lang] for lang in language_list])
```

[128]:
```
linked = linkage(confusion_more_concepts, method='average')

plt.figure(figsize=(12, 8))
dendrogram(linked,
           p=100,
           truncate_mode="level",
           orientation='top',
           distance_sort='descending',
           show_leaf_counts=False,
           leaf_label_func=llf)
```

```
plt.show()
```

```
<ipython-input-128-931fbf67af4c>:1: ClusterWarning: scipy.cluster: The symmetric
non-negative hollow observation matrix looks suspiciously like an uncondensed
distance matrix
  linked = linkage(confusion_more_concepts, method='average')
```



Question 8: Try increasing the number of concepts we use to compute our confusion matrix to be higher than 20. Does it change the results?

A8: With an increased number of concepts to 70, we notice overrall very similar clusters, with only mior changes such as a new cluster consisting of Breton and Welsh. The results are pretty much unchanged, suggesting that the results for the experiments prior to this one is representative of the entire dataset.

5 Part 6

Let's evaluate the quality of the clustering from Question 5, relative to the two class labels we have: `family` and `subfamily`.

5.1 Problem 7 (4 points)

Write code to compute the V measure scores for this clustering relative to `family` and `subfamily`, then print them.

[146]:
```
## Problem 7
from sklearn.metrics import v_measure_score
from scipy.cluster.hierarchy import fcluster


predicted_labels = fcluster(linked, t=0.6 , criterion='distance')

# build the true labels
lang_to_family= dict(zip(languages['Language_ID'], languages['family']))
lang_to_subfamily= dict(zip(languages['Language_ID'], languages['subfamily']))

family_map = dict(zip(languages['family'], range(len(languages['family']))))
subfamily_map = dict(zip(languages['subfamily'],␣
 ↪range(len(languages['subfamily']))))

language_list = list(wordforms['Language_ID'].unique())

family_labels = [family_map[lang_to_family[lang]] for lang in language_list]
subfamily_labels = [subfamily_map[lang_to_subfamily[lang]] for lang in␣
 ↪language_list]

## save the two V measure scores as v_measure_family and v_measure_subfamily
v_measure_family = v_measure_score(family_labels, predicted_labels)
v_measure_subfamily = v_measure_score(subfamily_labels, predicted_labels)

print(v_measure_family)
print(v_measure_subfamily)
```

```
0.0
0.7191569844205944
```

6 To Submit

To submit: * Name this notebook `YOUR_STUDENT_ID_Assignment_5.ipynb` and download it. * Convert this `.ipynb` file to a `.pdf` (e.g., using the following instructions).
* Upload the PDF to the Gradescope assignment "Assignment 5".
* Submit the `.ipynb` file on myCourses under Assignment 5.

(Note: `Print > Save as PDF` **will not work** because it will not display your figures correctly.)

You can convert the notebook to a PDF using the following instructions.

7 Converting this notebook to a PDF

1. Make sure you have renamed the notebook, e.g. `000000000_Assignment_5.ipynb` where `000000000` is your student ID.
2. Make sure to save the notebook (`ctrl/cmd + s`).

Make sure Google Drive is mounted (it likely already is from the first question).

[ ]:
```
from google.colab import drive
drive.mount('/content/drive/')
!ls "/content/drive/My Drive/Colab Notebooks/"
```

3. Install packages for converting .ipynb to .pdf

[148]:
```
!apt-get -q install texlive-xetex texlive-fonts-recommended␣
↪texlive-plain-generic
```

```
Reading package lists…
Building dependency tree…
Reading state information…
The following additional packages will be installed:
  dvisvgm fonts-droid-fallback fonts-lato fonts-lmodern fonts-noto-mono fonts-
texgyre
  fonts-urw-base35 libapache-pom-java libcommons-logging-java libcommons-parent-
java
  libfontbox-java libfontenc1 libgs9 libgs9-common libidn12 libijs-0.35
libjbig2dec0 libkpathsea6
  libpdfbox-java libptexenc1 libruby3.0 libsynctex2 libteckit0 libtexlua53
libtexluajit2 libwoff1
  libzzip-0-13 lmodern poppler-data preview-latex-style rake ruby ruby-net-
telnet ruby-rubygems
  ruby-webrick ruby-xmlrpc ruby3.0 rubygems-integration t1utils teckit tex-
common tex-gyre
  texlive-base texlive-binaries texlive-latex-base texlive-latex-extra texlive-
latex-recommended
  texlive-pictures tipa xfonts-encodings xfonts-utils
Suggested packages:
  fonts-noto fonts-freefont-otf | fonts-freefont-ttf libavalon-framework-java
  libcommons-logging-java-doc libexcalibur-logkit-java liblog4j1.2-java poppler-
utils ghostscript
  fonts-japanese-mincho | fonts-ipafont-mincho fonts-japanese-gothic | fonts-
ipafont-gothic
  fonts-arphic-ukai fonts-arphic-uming fonts-nanum ri ruby-dev bundler debhelper
gv
  | postscript-viewer perl-tk xpdf | pdf-viewer xzdec texlive-fonts-recommended-
doc
  texlive-latex-base-doc python3-pygments icc-profiles libfile-which-perl
  libspreadsheet-parseexcel-perl texlive-latex-extra-doc texlive-latex-
recommended-doc
```

```
  texlive-luatex texlive-pstricks dot2tex prerex texlive-pictures-doc vprerex
default-jre-headless
  tipa-doc
The following NEW packages will be installed:
  dvisvgm fonts-droid-fallback fonts-lato fonts-lmodern fonts-noto-mono fonts-
texgyre
  fonts-urw-base35 libapache-pom-java libcommons-logging-java libcommons-parent-
java
  libfontbox-java libfontenc1 libgs9 libgs9-common libidn12 libijs-0.35
libjbig2dec0 libkpathsea6
  libpdfbox-java libptexenc1 libruby3.0 libsynctex2 libteckit0 libtexlua53
libtexluajit2 libwoff1
  libzzip-0-13 lmodern poppler-data preview-latex-style rake ruby ruby-net-
telnet ruby-rubygems
  ruby-webrick ruby-xmlrpc ruby3.0 rubygems-integration t1utils teckit tex-
common tex-gyre
  texlive-base texlive-binaries texlive-fonts-recommended texlive-latex-base
texlive-latex-extra
  texlive-latex-recommended texlive-pictures texlive-plain-generic texlive-xetex
tipa
  xfonts-encodings xfonts-utils
0 upgraded, 54 newly installed, 0 to remove and 38 not upgraded.
Need to get 182 MB of archives.
After this operation, 571 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/main amd64 fonts-droid-fallback all
1:6.0.1r16-1.1build1 [1,805 kB]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 fonts-lato all 2.0-2.1
[2,696 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy/main amd64 poppler-data all
0.4.11-1 [2,171 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tex-common all 6.17
[33.7 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy/main amd64 fonts-urw-base35 all
20200910-1 [6,367 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libgs9-common
all 9.55.0~dfsg1-0ubuntu5.6 [751 kB]
Get:7 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libidn12 amd64
1.38-4ubuntu1 [60.0 kB]
Get:8 http://archive.ubuntu.com/ubuntu jammy/main amd64 libijs-0.35 amd64
0.35-15build2 [16.5 kB]
Get:9 http://archive.ubuntu.com/ubuntu jammy/main amd64 libjbig2dec0 amd64
0.19-3build2 [64.7 kB]
Get:10 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libgs9 amd64
9.55.0~dfsg1-0ubuntu5.6 [5,031 kB]
Get:11 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libkpathsea6
amd64 2021.20210626.59705-1ubuntu0.2 [60.4 kB]
Get:12 http://archive.ubuntu.com/ubuntu jammy/main amd64 libwoff1 amd64
1.0.2-1build4 [45.2 kB]
```

```
Get:13 http://archive.ubuntu.com/ubuntu jammy/universe amd64 dvisvgm amd64
2.13.1-1 [1,221 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy/universe amd64 fonts-lmodern all
2.004.5-6.1 [4,532 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy/main amd64 fonts-noto-mono all
20201225-1build1 [397 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy/universe amd64 fonts-texgyre all
20180621-3.1 [10.2 MB]
Get:17 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libapache-pom-java
all 18-1 [4,720 B]
Get:18 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libcommons-parent-
java all 43-1 [10.8 kB]
Get:19 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libcommons-logging-
java all 1.2-2 [60.3 kB]
Get:20 http://archive.ubuntu.com/ubuntu jammy/main amd64 libfontenc1 amd64
1:1.1.4-1build3 [14.7 kB]
Get:21 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libptexenc1
amd64 2021.20210626.59705-1ubuntu0.2 [39.1 kB]
Get:22 http://archive.ubuntu.com/ubuntu jammy/main amd64 rubygems-integration
all 1.18 [5,336 B]
Get:23 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 ruby3.0 amd64
3.0.2-7ubuntu2.4 [50.1 kB]
Get:24 http://archive.ubuntu.com/ubuntu jammy/main amd64 ruby-rubygems all
3.3.5-2 [228 kB]
Get:25 http://archive.ubuntu.com/ubuntu jammy/main amd64 ruby amd64 1:3.0~exp1
[5,100 B]
Get:26 http://archive.ubuntu.com/ubuntu jammy/main amd64 rake all 13.0.6-2 [61.7
kB]
Get:27 http://archive.ubuntu.com/ubuntu jammy/main amd64 ruby-net-telnet all
0.1.1-2 [12.6 kB]
Get:28 http://archive.ubuntu.com/ubuntu jammy/universe amd64 ruby-webrick all
1.7.0-3 [51.8 kB]
Get:29 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 ruby-xmlrpc all
0.3.2-1ubuntu0.1 [24.9 kB]
Get:30 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libruby3.0
amd64 3.0.2-7ubuntu2.4 [5,113 kB]
Get:31 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libsynctex2
amd64 2021.20210626.59705-1ubuntu0.2 [55.6 kB]
Get:32 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libteckit0 amd64
2.5.11+ds1-1 [421 kB]
Get:33 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libtexlua53
amd64 2021.20210626.59705-1ubuntu0.2 [120 kB]
Get:34 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 libtexluajit2
amd64 2021.20210626.59705-1ubuntu0.2 [267 kB]
Get:35 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libzzip-0-13 amd64
0.13.72+dfsg.1-1.1 [27.0 kB]
Get:36 http://archive.ubuntu.com/ubuntu jammy/main amd64 xfonts-encodings all
1:1.0.5-0ubuntu2 [578 kB]
```

```
Get:37 http://archive.ubuntu.com/ubuntu jammy/main amd64 xfonts-utils amd64
1:7.7+6build2 [94.6 kB]
Get:38 http://archive.ubuntu.com/ubuntu jammy/universe amd64 lmodern all
2.004.5-6.1 [9,471 kB]
Get:39 http://archive.ubuntu.com/ubuntu jammy/universe amd64 preview-latex-style
all 12.2-1ubuntu1 [185 kB]
Get:40 http://archive.ubuntu.com/ubuntu jammy/main amd64 t1utils amd64
1.41-4build2 [61.3 kB]
Get:41 http://archive.ubuntu.com/ubuntu jammy/universe amd64 teckit amd64
2.5.11+ds1-1 [699 kB]
Get:42 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tex-gyre all
20180621-3.1 [6,209 kB]
Get:43 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 texlive-
binaries amd64 2021.20210626.59705-1ubuntu0.2 [9,860 kB]
Get:44 http://archive.ubuntu.com/ubuntu jammy/universe amd64 texlive-base all
2021.20220204-1 [21.0 MB]
Get:45 http://archive.ubuntu.com/ubuntu jammy/universe amd64 texlive-fonts-
recommended all 2021.20220204-1 [4,972 kB]
Get:46 http://archive.ubuntu.com/ubuntu jammy/universe amd64 texlive-latex-base
all 2021.20220204-1 [1,128 kB]
Get:47 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libfontbox-java all
1:1.8.16-2 [207 kB]
Get:48 http://archive.ubuntu.com/ubuntu jammy/universe amd64 libpdfbox-java all
1:1.8.16-2 [5,199 kB]
Get:49 http://archive.ubuntu.com/ubuntu jammy/universe amd64 texlive-latex-
recommended all 2021.20220204-1 [14.4 MB]
Get:50 http://archive.ubuntu.com/ubuntu jammy/universe amd64 texlive-pictures
all 2021.20220204-1 [8,720 kB]
Get:51 http://archive.ubuntu.com/ubuntu jammy/universe amd64 texlive-latex-extra
all 2021.20220204-1 [13.9 MB]
Get:52 http://archive.ubuntu.com/ubuntu jammy/universe amd64 texlive-plain-
generic all 2021.20220204-1 [27.5 MB]
Get:53 http://archive.ubuntu.com/ubuntu jammy/universe amd64 tipa all 2:1.3-21
[2,967 kB]
Get:54 http://archive.ubuntu.com/ubuntu jammy/universe amd64 texlive-xetex all
2021.20220204-1 [12.4 MB]
Fetched 182 MB in 3s (58.3 MB/s)
Extracting templates from packages: 100%
Preconfiguring packages …
Selecting previously unselected package fonts-droid-fallback.
(Reading database … 121752 files and directories currently installed.)
Preparing to unpack …/00-fonts-droid-fallback_1%3a6.0.1r16-1.1build1_all.deb
…
Unpacking fonts-droid-fallback (1:6.0.1r16-1.1build1) …
Selecting previously unselected package fonts-lato.
Preparing to unpack …/01-fonts-lato_2.0-2.1_all.deb …
Unpacking fonts-lato (2.0-2.1) …
Selecting previously unselected package poppler-data.
```

```
Preparing to unpack …/02-poppler-data_0.4.11-1_all.deb …
Unpacking poppler-data (0.4.11-1) …
Selecting previously unselected package tex-common.
Preparing to unpack …/03-tex-common_6.17_all.deb …
Unpacking tex-common (6.17) …
Selecting previously unselected package fonts-urw-base35.
Preparing to unpack …/04-fonts-urw-base35_20200910-1_all.deb …
Unpacking fonts-urw-base35 (20200910-1) …
Selecting previously unselected package libgs9-common.
Preparing to unpack …/05-libgs9-common_9.55.0~dfsg1-0ubuntu5.6_all.deb …
Unpacking libgs9-common (9.55.0~dfsg1-0ubuntu5.6) …
Selecting previously unselected package libidn12:amd64.
Preparing to unpack …/06-libidn12_1.38-4ubuntu1_amd64.deb …
Unpacking libidn12:amd64 (1.38-4ubuntu1) …
Selecting previously unselected package libijs-0.35:amd64.
Preparing to unpack …/07-libijs-0.35_0.35-15build2_amd64.deb …
Unpacking libijs-0.35:amd64 (0.35-15build2) …
Selecting previously unselected package libjbig2dec0:amd64.
Preparing to unpack …/08-libjbig2dec0_0.19-3build2_amd64.deb …
Unpacking libjbig2dec0:amd64 (0.19-3build2) …
Selecting previously unselected package libgs9:amd64.
Preparing to unpack …/09-libgs9_9.55.0~dfsg1-0ubuntu5.6_amd64.deb …
Unpacking libgs9:amd64 (9.55.0~dfsg1-0ubuntu5.6) …
Selecting previously unselected package libkpathsea6:amd64.
Preparing to unpack …/10-libkpathsea6_2021.20210626.59705-1ubuntu0.2_amd64.deb
…
Unpacking libkpathsea6:amd64 (2021.20210626.59705-1ubuntu0.2) …
Selecting previously unselected package libwoff1:amd64.
Preparing to unpack …/11-libwoff1_1.0.2-1build4_amd64.deb …
Unpacking libwoff1:amd64 (1.0.2-1build4) …
Selecting previously unselected package dvisvgm.
Preparing to unpack …/12-dvisvgm_2.13.1-1_amd64.deb …
Unpacking dvisvgm (2.13.1-1) …
Selecting previously unselected package fonts-lmodern.
Preparing to unpack …/13-fonts-lmodern_2.004.5-6.1_all.deb …
Unpacking fonts-lmodern (2.004.5-6.1) …
Selecting previously unselected package fonts-noto-mono.
Preparing to unpack …/14-fonts-noto-mono_20201225-1build1_all.deb …
Unpacking fonts-noto-mono (20201225-1build1) …
Selecting previously unselected package fonts-texgyre.
Preparing to unpack …/15-fonts-texgyre_20180621-3.1_all.deb …
Unpacking fonts-texgyre (20180621-3.1) …
Selecting previously unselected package libapache-pom-java.
Preparing to unpack …/16-libapache-pom-java_18-1_all.deb …
Unpacking libapache-pom-java (18-1) …
Selecting previously unselected package libcommons-parent-java.
Preparing to unpack …/17-libcommons-parent-java_43-1_all.deb …
Unpacking libcommons-parent-java (43-1) …
```

```
Selecting previously unselected package libcommons-logging-java.
Preparing to unpack …/18-libcommons-logging-java_1.2-2_all.deb …
Unpacking libcommons-logging-java (1.2-2) …
Selecting previously unselected package libfontenc1:amd64.
Preparing to unpack …/19-libfontenc1_1%3a1.1.4-1build3_amd64.deb …
Unpacking libfontenc1:amd64 (1:1.1.4-1build3) …
Selecting previously unselected package libptexenc1:amd64.
Preparing to unpack …/20-libptexenc1_2021.20210626.59705-1ubuntu0.2_amd64.deb
…
Unpacking libptexenc1:amd64 (2021.20210626.59705-1ubuntu0.2) …
Selecting previously unselected package rubygems-integration.
Preparing to unpack …/21-rubygems-integration_1.18_all.deb …
Unpacking rubygems-integration (1.18) …
Selecting previously unselected package ruby3.0.
Preparing to unpack …/22-ruby3.0_3.0.2-7ubuntu2.4_amd64.deb …
Unpacking ruby3.0 (3.0.2-7ubuntu2.4) …
Selecting previously unselected package ruby-rubygems.
Preparing to unpack …/23-ruby-rubygems_3.3.5-2_all.deb …
Unpacking ruby-rubygems (3.3.5-2) …
Selecting previously unselected package ruby.
Preparing to unpack …/24-ruby_1%3a3.0~exp1_amd64.deb …
Unpacking ruby (1:3.0~exp1) …
Selecting previously unselected package rake.
Preparing to unpack …/25-rake_13.0.6-2_all.deb …
Unpacking rake (13.0.6-2) …
Selecting previously unselected package ruby-net-telnet.
Preparing to unpack …/26-ruby-net-telnet_0.1.1-2_all.deb …
Unpacking ruby-net-telnet (0.1.1-2) …
Selecting previously unselected package ruby-webrick.
Preparing to unpack …/27-ruby-webrick_1.7.0-3_all.deb …
Unpacking ruby-webrick (1.7.0-3) …
Selecting previously unselected package ruby-xmlrpc.
Preparing to unpack …/28-ruby-xmlrpc_0.3.2-1ubuntu0.1_all.deb …
Unpacking ruby-xmlrpc (0.3.2-1ubuntu0.1) …
Selecting previously unselected package libruby3.0:amd64.
Preparing to unpack …/29-libruby3.0_3.0.2-7ubuntu2.4_amd64.deb …
Unpacking libruby3.0:amd64 (3.0.2-7ubuntu2.4) …
Selecting previously unselected package libsynctex2:amd64.
Preparing to unpack …/30-libsynctex2_2021.20210626.59705-1ubuntu0.2_amd64.deb
…
Unpacking libsynctex2:amd64 (2021.20210626.59705-1ubuntu0.2) …
Selecting previously unselected package libteckit0:amd64.
Preparing to unpack …/31-libteckit0_2.5.11+ds1-1_amd64.deb …
Unpacking libteckit0:amd64 (2.5.11+ds1-1) …
Selecting previously unselected package libtexlua53:amd64.
Preparing to unpack …/32-libtexlua53_2021.20210626.59705-1ubuntu0.2_amd64.deb
…
Unpacking libtexlua53:amd64 (2021.20210626.59705-1ubuntu0.2) …
```

```
Selecting previously unselected package libtexluajit2:amd64.
Preparing to unpack
…/33-libtexluajit2_2021.20210626.59705-1ubuntu0.2_amd64.deb …
Unpacking libtexluajit2:amd64 (2021.20210626.59705-1ubuntu0.2) …
Selecting previously unselected package libzzip-0-13:amd64.
Preparing to unpack …/34-libzzip-0-13_0.13.72+dfsg.1-1.1_amd64.deb …
Unpacking libzzip-0-13:amd64 (0.13.72+dfsg.1-1.1) …
Selecting previously unselected package xfonts-encodings.
Preparing to unpack …/35-xfonts-encodings_1%3a1.0.5-0ubuntu2_all.deb …
Unpacking xfonts-encodings (1:1.0.5-0ubuntu2) …
Selecting previously unselected package xfonts-utils.
Preparing to unpack …/36-xfonts-utils_1%3a7.7+6build2_amd64.deb …
Unpacking xfonts-utils (1:7.7+6build2) …
Selecting previously unselected package lmodern.
Preparing to unpack …/37-lmodern_2.004.5-6.1_all.deb …
Unpacking lmodern (2.004.5-6.1) …
Selecting previously unselected package preview-latex-style.
Preparing to unpack …/38-preview-latex-style_12.2-1ubuntu1_all.deb …
Unpacking preview-latex-style (12.2-1ubuntu1) …
Selecting previously unselected package t1utils.
Preparing to unpack …/39-t1utils_1.41-4build2_amd64.deb …
Unpacking t1utils (1.41-4build2) …
Selecting previously unselected package teckit.
Preparing to unpack …/40-teckit_2.5.11+ds1-1_amd64.deb …
Unpacking teckit (2.5.11+ds1-1) …
Selecting previously unselected package tex-gyre.
Preparing to unpack …/41-tex-gyre_20180621-3.1_all.deb …
Unpacking tex-gyre (20180621-3.1) …
Selecting previously unselected package texlive-binaries.
Preparing to unpack …/42-texlive-
binaries_2021.20210626.59705-1ubuntu0.2_amd64.deb …
Unpacking texlive-binaries (2021.20210626.59705-1ubuntu0.2) …
Selecting previously unselected package texlive-base.
Preparing to unpack …/43-texlive-base_2021.20220204-1_all.deb …
Unpacking texlive-base (2021.20220204-1) …
Selecting previously unselected package texlive-fonts-recommended.
Preparing to unpack …/44-texlive-fonts-recommended_2021.20220204-1_all.deb …
Unpacking texlive-fonts-recommended (2021.20220204-1) …
Selecting previously unselected package texlive-latex-base.
Preparing to unpack …/45-texlive-latex-base_2021.20220204-1_all.deb …
Unpacking texlive-latex-base (2021.20220204-1) …
Selecting previously unselected package libfontbox-java.
Preparing to unpack …/46-libfontbox-java_1%3a1.8.16-2_all.deb …
Unpacking libfontbox-java (1:1.8.16-2) …
Selecting previously unselected package libpdfbox-java.
Preparing to unpack …/47-libpdfbox-java_1%3a1.8.16-2_all.deb …
Unpacking libpdfbox-java (1:1.8.16-2) …
Selecting previously unselected package texlive-latex-recommended.
```

```
Preparing to unpack …/48-texlive-latex-recommended_2021.20220204-1_all.deb …
Unpacking texlive-latex-recommended (2021.20220204-1) …
Selecting previously unselected package texlive-pictures.
Preparing to unpack …/49-texlive-pictures_2021.20220204-1_all.deb …
Unpacking texlive-pictures (2021.20220204-1) …
Selecting previously unselected package texlive-latex-extra.
Preparing to unpack …/50-texlive-latex-extra_2021.20220204-1_all.deb …
Unpacking texlive-latex-extra (2021.20220204-1) …
Selecting previously unselected package texlive-plain-generic.
Preparing to unpack …/51-texlive-plain-generic_2021.20220204-1_all.deb …
Unpacking texlive-plain-generic (2021.20220204-1) …
Selecting previously unselected package tipa.
Preparing to unpack …/52-tipa_2%3a1.3-21_all.deb …
Unpacking tipa (2:1.3-21) …
Selecting previously unselected package texlive-xetex.
Preparing to unpack …/53-texlive-xetex_2021.20220204-1_all.deb …
Unpacking texlive-xetex (2021.20220204-1) …
Setting up fonts-lato (2.0-2.1) …
Setting up fonts-noto-mono (20201225-1build1) …
Setting up libwoff1:amd64 (1.0.2-1build4) …
Setting up libtexlua53:amd64 (2021.20210626.59705-1ubuntu0.2) …
Setting up libijs-0.35:amd64 (0.35-15build2) …
Setting up libtexluajit2:amd64 (2021.20210626.59705-1ubuntu0.2) …
Setting up libfontbox-java (1:1.8.16-2) …
Setting up rubygems-integration (1.18) …
Setting up libzzip-0-13:amd64 (0.13.72+dfsg.1-1.1) …
Setting up fonts-urw-base35 (20200910-1) …
Setting up poppler-data (0.4.11-1) …
Setting up tex-common (6.17) …
update-language: texlive-base not installed and configured, doing nothing!
Setting up libfontenc1:amd64 (1:1.1.4-1build3) …
Setting up libjbig2dec0:amd64 (0.19-3build2) …
Setting up libteckit0:amd64 (2.5.11+ds1-1) …
Setting up libapache-pom-java (18-1) …
Setting up ruby-net-telnet (0.1.1-2) …
Setting up xfonts-encodings (1:1.0.5-0ubuntu2) …
Setting up t1utils (1.41-4build2) …
Setting up libidn12:amd64 (1.38-4ubuntu1) …
Setting up fonts-texgyre (20180621-3.1) …
Setting up libkpathsea6:amd64 (2021.20210626.59705-1ubuntu0.2) …
Setting up ruby-webrick (1.7.0-3) …
Setting up fonts-lmodern (2.004.5-6.1) …
Setting up fonts-droid-fallback (1:6.0.1r16-1.1build1) …
Setting up ruby-xmlrpc (0.3.2-1ubuntu0.1) …
Setting up libsynctex2:amd64 (2021.20210626.59705-1ubuntu0.2) …
Setting up libgs9-common (9.55.0~dfsg1-0ubuntu5.6) …
Setting up teckit (2.5.11+ds1-1) …
Setting up libpdfbox-java (1:1.8.16-2) …
```

```
Setting up libgs9:amd64 (9.55.0~dfsg1-0ubuntu5.6) …
Setting up preview-latex-style (12.2-1ubuntu1) …
Setting up libcommons-parent-java (43-1) …
Setting up dvisvgm (2.13.1-1) …
Setting up libcommons-logging-java (1.2-2) …
Setting up xfonts-utils (1:7.7+6build2) …
Setting up libptexenc1:amd64 (2021.20210626.59705-1ubuntu0.2) …
Setting up texlive-binaries (2021.20210626.59705-1ubuntu0.2) …
update-alternatives: using /usr/bin/xdvi-xaw to provide /usr/bin/xdvi.bin
(xdvi.bin) in auto mode
update-alternatives: using /usr/bin/bibtex.original to provide /usr/bin/bibtex
(bibtex) in auto mode
Setting up lmodern (2.004.5-6.1) …
Setting up texlive-base (2021.20220204-1) …
/usr/bin/ucfr
/usr/bin/ucfr
/usr/bin/ucfr
/usr/bin/ucfr
mktexlsr: Updating /var/lib/texmf/ls-R-TEXLIVEDIST…
mktexlsr: Updating /var/lib/texmf/ls-R-TEXMFMAIN…
mktexlsr: Updating /var/lib/texmf/ls-R…
mktexlsr: Done.
tl-paper: setting paper size for dvips to a4:
/var/lib/texmf/dvips/config/config-paper.ps
tl-paper: setting paper size for dvipdfmx to a4:
/var/lib/texmf/dvipdfmx/dvipdfmx-paper.cfg
tl-paper: setting paper size for xdvi to a4: /var/lib/texmf/xdvi/XDvi-paper
tl-paper: setting paper size for pdftex to a4: /var/lib/texmf/tex/generic/tex-
ini-files/pdftexconfig.tex
Setting up tex-gyre (20180621-3.1) …
Setting up texlive-plain-generic (2021.20220204-1) …
Setting up texlive-latex-base (2021.20220204-1) …
Setting up texlive-latex-recommended (2021.20220204-1) …
Setting up texlive-pictures (2021.20220204-1) …
Setting up texlive-fonts-recommended (2021.20220204-1) …
Setting up tipa (2:1.3-21) …
Setting up texlive-latex-extra (2021.20220204-1) …
Setting up texlive-xetex (2021.20220204-1) …
Setting up rake (13.0.6-2) …
Setting up libruby3.0:amd64 (3.0.2-7ubuntu2.4) …
Setting up ruby3.0 (3.0.2-7ubuntu2.4) …
Setting up ruby (1:3.0~exp1) …
Setting up ruby-rubygems (3.3.5-2) …
Processing triggers for man-db (2.10.2-1) …
Processing triggers for fontconfig (2.13.1-4.2ubuntu5) …
Processing triggers for libc-bin (2.35-0ubuntu3.4) …
/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a symbolic link
```

```
/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a symbolic
link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link

Processing triggers for tex-common (6.17) …
Running updmap-sys. This may take some time… done.
Running mktexlsr /var/lib/texmf … done.
Building format(s) --all.
        This may take some time… done.
```

4. Convert to PDF (replace `000000000` with your student ID)

[ ]:
```
%env STUDENT_ID=261072449
!jupyter nbconvert --to pdf "/content/drive/My Drive/Colab Notebooks/
 ↪${STUDENT_ID}_Assignment_5.ipynb"
```

5. Download the resulting PDF file. If you are using Chrome, you can do so by running the following code. On other browsers, you can download the PDF using the file mananger on the left of the screen (Navigate to the file > Right Click > Download).

[ ]:
```
import os
from google.colab import files
files.download(f"/content/drive/My Drive/Colab Notebooks/{os.
 ↪environ['STUDENT_ID']}_Assignment_5.pdf")
```

6. Verify that your PDF correctly displays your figures and responses.