Name: Sina
Last name: Heidari
Student number: 400464201
Course: Natural Language Processing

Importing Libraries

In [1]:
```
from os import walk
import os
from parsivar import Normalizer
from parsivar import Tokenizer
from parsivar import FindStems
from hazm import Lemmatizer
import re
```

If you don't have hazm and parsivar installed, run the following command inside the working directory:
pip install -r requirements.txt

Reading the poem files into a dictionary

In [2]:
```
# derives current working folder path from getcwd function
poems_path = os.getcwd() + '/Poems'

# create a dic for storing the poems
poems = {}

# walk through all of the files inside Poems directory
for root, dirs, files in walk(poems_path):
    # iterate through all poem files
    for name in files:
        # join cwd path and file name and open the file
        to_read_file = open(os.path.join(root, name))
        # split file name with '.' and choose first field
        file_key = name.split('.')[0]
        # store in dic
        poems[file_key] = {'txt' : to_read_file.read(), 'tokenized': []}
        to_read_file.close()

# Ex: open file with key: 4844
print(poems['4844']['txt'])
```

هله ای شاه مپیچان سر و دستار مرو
هله ای ماه که نغزت رخ و رخسار مرو
در همه روی زمین چشم و دل باز که راست
مکن آزار مکن جانب اغیار مرو
مبر از یار مبر خانه اسرار مسوز
گل و گلزار مکن جانب هر خار مرو

مکن ای یار ستیزه دغل و جنگ مجوی

هله آن بار برفتی مکن این بار مرو

بنده و چاکر و پرورده و مولای توایم

ای دل و دین و حیات خوش ناچار مرو

هله سرنای توام مست نواهای توام

مشکن چنگ طرب را مسکل تار مرو

هله مخمور چه نالی بر مخمور دگر

پهلوی خم بنشین از بر خمار مرو

هله جان بخش بیا ای صدقات تو حیات

به از این خیر نباشد بجز این کار مرو

خاتم حسن و جمالی هله ای یوسف دهر

سوی مکاری اخوان ستمکار مرو

هله دیدار مهل برمگزین فکر و خیال

از عیان سر مکشان در پی آثار مرو

هله موسی زمان گرد برآر از دریا

دل فرعون مجو جانب انکار مرو

هله عیسی قران صحت رنجور گران

از برای دو سه ترسا سوی زنار مرو

هله ای شاهد جان خواجه جان‌های شهان

شیوه کن لب بگز و غبغبه افشار مرو

هله صدیق زمانی به تو ختم است وفا

جز سوی احمد بگزیده مختار مرو

جبرئیل کرمی سدره مقام و وطنت

همچو مرغان زمین بر سر شخسار مرو

تو یقین دار که بی‌تو نفسی جان نزید

در احسان بگشا و پس دیوار مرو

همه رندان و حریفان و بتان جمع شدند

وقت کار است بیا کار کن از کار مرو

هله باقی غزل را ز شهنشاه بجوی

همگی گوش شو اکنون سوی گفتار مرو

Normalize each poem document

In [3]:

```
normalizer = Normalizer()
for key in poems:
    poems[key] = normalizer.normalize(poems[key]['txt'])
    # replace half spaces and new line characters with single space
    poems[key] = {'txt': re.sub('|[\\n]' , ' ', poems[key]), 'tokenized': []}
```

In [4]:
```
print(poems['4844'])
```

{'txt': 'هله ای شاه میپچان سر و دستار مرو هله ای ماه که نغزت رخ و رخسار مرو در همه روی زمین چشم و دل باز که راست مکن
آزار مکن جانب اغیار مرو مبر از یار مبر خانه اسرار مسوز گل و گلزار مکن جانب هر خار مرو مکن ای یار ستیزه دغل و جنگ مجوی هله
آن بار برفتی مکن این بار مرو بنده و چاکر و پرورده و مولای توایم ای دل و دین و حیات خوش ناچار مرو هله سرنای توام مست نواهای توا
م مشکن چنگ طرب را مسکل تار مرو هله مخمور چه نالی بر مخمور دگر پهلوی خم بنشین از بر خمار مرو هله جان بخش بیا ای صدقات تو
حیات به از این خیر نباشد بجز این کار مرو خاتم حسن و جمالی هله ای یوسف دهر سوی مکاری اخوان ستمکار مرو هله دیدار مهل برمگزین ف
کر و خیال از عیان سر مکشان در پی آثار مرو هله موسی زمان گرد برآر از دریا دل فرعون مجو جانب انکار مرو هله عیسی قران صحت رن
جور گران از برای دو سه ترسا سوی زنار مرو هله ای شاهد جان خواجه جان های شهان شیوه کن لب بگز و غبغبه افشار مرو هله صدیق زما
نی به تو ختم است وفا جز سوی احمد بگزیده مختار مرو جبرئیل کرمی سدره مقام و وطنت همچو مرغان زمین بر سر شخسار مرو تو یقین دا
ر که بی تو نفسی جان نزید در احسان بگشا و پس دیوار مرو همه رندان و حریفان و بتان جمع شدند وقت کار است بیا کار کن از کار مرو
هله باقی غزل را ز شهنشاه بجوی همگی گوش شو اکنون سوی گفتار مرو', 'tokenized': []}

Tokenize each poem document

Tokenized poems are stored in 'tokenized' field of every dictionary element (which is a dictionary itself).

In [5]:
```
tokenizer = Tokenizer()
for key in poems:
    poems[key]['tokenized'] = tokenizer.tokenize_words(poems[key]['txt'])
```

In [6]:
```
print(poems['4844']['tokenized'])
```

['هله', 'ای', 'شاه', 'میپچان', 'سر', 'و', 'دستار', 'مرو', 'هله', 'ای', 'ماه', 'که', 'نغزت', 'رخ', 'و', 'رخسا
ر', 'مرو', 'در', 'همه', 'روی', 'زمین', 'چشم', 'و', 'دل', 'باز', 'که', 'راست', 'مکن', 'آزار', 'مکن', 'جانب',
'اغیار', 'مرو', 'مبر', 'از', 'یار', 'مبر', 'خانه', 'اسرار', 'مسوز', 'گل', 'و', 'گلزار', 'مکن', 'جانب', 'هر',
'خار', 'مرو', 'مکن', 'ای', 'یار', 'ستیزه', 'دغل', 'و', 'جنگ', 'مجوی', 'هله', 'آن', 'بار', 'برفتی', 'مکن', 'ای
ن', 'بار', 'مرو', 'بنده', 'و', 'چاکر', 'و', 'پرورده', 'و', 'مولای', 'توایم', 'ای', 'دل', 'و', 'دین', 'و', 'حیا
ت', 'خوش', 'ناچار', 'مرو', 'هله', 'سرنای', 'توام', 'مست', 'نواهای', 'توام', 'مشکن', 'چنگ', 'طرب', 'را', 'مسک
ل', 'تار', 'مرو', 'هله', 'مخمور', 'چه', 'نالی', 'بر', 'مخمور', 'دگر', 'پهلوی', 'خم', 'بنشین', 'از', 'بر', 'خما
ر', 'مرو', 'هله', 'جان', 'بخش', 'بیا', 'ای', 'صدقات', 'تو', 'حیات', 'به', 'از', 'این', 'خیر', 'نباشد', 'بجز',
'این', 'کار', 'مرو', 'خاتم', 'حسن', 'و', 'جمالی', 'هله', 'ای', 'یوسف', 'دهر', 'سوی', 'مکاری', 'اخوان', 'ستمکا
ر', 'مرو', 'هله', 'دیدار', 'مهل', 'برمگزین', 'فکر', 'و', 'خیال', 'از', 'عیان', 'سر', 'مکشان', 'در', 'پی', 'آثا
ر', 'مرو', 'هله', 'موسی', 'زمان', 'گرد', 'برآر', 'از', 'دریا', 'دل', 'فرعون', 'مجو', 'جانب', 'انکار', 'مرو',
'هله', 'عیسی', 'قران', 'صحت', 'رنجور', 'گران', 'از', 'برای', 'دو', 'سه', 'ترسا', 'سوی', 'زنار', 'مرو', 'هل
ه', 'ای', 'شاهد', 'جان', 'خواجه', 'جان', 'های', 'شهان', 'شیوه', 'کن', 'لب', 'بگز', 'و', 'غبغبه', 'افشار', 'مر
و', 'هله', 'صدیق', 'زمانی', 'به', 'تو', 'ختم', 'است', 'وفا', 'جز', 'سوی', 'احمد', 'بگزیده', 'مختار', 'مرو', 'ج
برئیل', 'کرمی', 'سدره', 'مقام', 'و', 'وطنت', 'همچو', 'مرغان', 'زمین', 'بر', 'سر', 'شخسار', 'مرو', 'تو', 'یقی

ن', 'دار', 'که', 'بی', 'تو', 'نفسی', 'جان', 'نزید', 'در', 'احسان', 'بگشا', 'و', 'پس', 'دیوار', 'مرو', 'همه',
'رندان', 'و', 'حریفان', 'و', 'بتان', 'جمع', 'شدند', 'وقت', 'کار', 'است', 'بیا', 'کار', 'کن', 'از', 'کار', 'مر
['و', 'هله', 'باقی', 'غزل', 'را', 'ز', 'شهنشاه', 'بجوی', 'همگی', 'گوش', 'شو', 'اکنون', 'سوی', 'گفتار', 'مرو

Removing stop-words

In [7]:
```
# read stop-words file
stop_file = open(os.getcwd() + '/Stopwords/Stopwords.txt', mode='r')
stop_file = stop_file.read()
# split them into a list of words
stop_file_split = stop_file.split('\n')
to_read_file.close()
tokenizer = Tokenizer()
# for every key inside poem dictionary
for key in poems:
    # a list comprehension to only choose words that are not included in our stop-words file (ig
    poems[key]['tokenized'] = [token for token in poems[key]['tokenized'] if token not in stop_f
```

In [8]:
```
print(poems['4844']['tokenized'])
```

['د', 'چشم', 'زمین', 'روی', 'مرو', 'هله', 'ماه', 'نغزت', 'رخ', 'رخسار', 'مرو', 'هله', 'شاه', 'میچان', 'دستار', 'مرو', 'هله',
ل', 'باز', 'راست', 'مکن', 'آزار', 'مکن', 'جانب', 'اغیار', 'مرو', 'مبر', 'یار', 'مبر', 'خانه', 'اسرار', 'مسوز',
'گل', 'گلزار', 'مکن', 'جانب', 'خار', 'مرو', 'مکن', 'یار', 'ستیزه', 'دغل', 'جنگ', 'مجوی', 'هله', 'بار', 'برفت
ی', 'مکن', 'بار', 'مرو', 'بنده', 'چاکر', 'پرورده', 'مولای', 'توایم', 'دل', 'دین', 'حیات', 'خوش', 'ناچار', 'مرو',
'هله', 'سرنای', 'توام', 'مست', 'نواهای', 'توام', 'مشکن', 'چنگ', 'طرب', 'مسکل', 'تار', 'مرو', 'هله', 'مخمور',
'نالی', 'مخمور', 'دگر', 'پهلوی', 'خم', 'بنشین', 'خمار', 'مرو', 'هله', 'جان', 'بخش', 'بیا', 'صدقات', 'حیات', 'خی
ر', 'نباشد', 'بجز', 'کار', 'مرو', 'خاتم', 'حسن', 'جمالی', 'هله', 'یوسف', 'دهر', 'مکاری', 'اخوان', 'ستمکار', 'مر
و', 'هله', 'دیدار', 'مهل', 'برمگزین', 'فکر', 'خیال', 'عیان', 'مکشان', 'پی', 'آثار', 'مرو', 'هله', 'موسی', 'زمان',
'گرد', 'برآر', 'دریا', 'دل', 'فرعون', 'مجو', 'جانب', 'انکار', 'مرو', 'هله', 'عیسی', 'قران', 'صحت', 'رنجور',
'گران', 'برای', 'دو', 'سه', 'ترسا', 'زنار', 'مرو', 'هله', 'شاهد', 'جان', 'خواجه', 'جان', 'های', 'شهان', 'شیو
ه', 'لب', 'بگز', 'غبغبه', 'افشار', 'مرو', 'هله', 'صدیق', 'زمانی', 'ختم', 'وفا', 'جز', 'احمد', 'بگزیده', 'مختار',
'مرو', 'جبرئیل', 'کرمی', 'سدره', 'مقام', 'وطنت', 'همچو', 'مرغان', 'زمین', 'شخسار', 'مرو', 'یقین', 'دار', 'بی',
'نفسی', 'جان', 'نزید', 'احسان', 'بگشا', 'دیوار', 'مرو', 'رندان', 'حریفان', 'بتان', 'جمع', 'شدند', 'وقت', 'کار', 'ب
یا', 'کار', 'کار', 'مرو', 'هله', 'باقی', 'غزل', 'شهنشاه', 'بجوی', 'همگی', 'گوش', 'شو', 'اکنون', 'گفتار', 'مر
و']

Lemmatizing tokenized words

In [9]:
```
lemmatizer = Lemmatizer()
# A list comprehension to lemmetize each word inside the tokenized list of words corresponding t
example_lemmetized = [lemmatizer.lemmatize(token) for token in poems['4844']['tokenized']]
```

In [10]:
```
print(example_lemmetized)
```

```
['د', 'چشم', 'زمین', 'روی', 'مرو', 'رخ', 'نغز', 'ماه', 'هله', 'مرو', 'دستار', 'میچان', 'شاه', 'هله',
'مسوز', 'اسرار', 'خانه', 'مبر', 'یار', 'مبر', 'مرو', 'اغیار', 'جانب', 'مکن', 'آزار', 'مکن', 'راست', 'باز', 'ل',
'برفت', 'بار', 'هله', 'مجوی', 'جنگ', 'دغل', 'ستیزه', 'یار', 'مکن', 'مرو', 'خار', 'جانب', 'مکن', 'گلزار', 'گل',
'مرو', 'ناچار', 'خوش', 'حیات', 'دین', 'دل', 'پرورده', 'مولا', 'توایم', 'چاکر', 'بنده', 'مرو', 'بار', 'مکن', 'ی',
'مخمور', 'ناال', 'مخور', 'هله', 'مرو', 'تار', 'مسکل', 'طرب', 'چنگ', 'مشکن', 'توام', 'نوا', 'مست', 'توام', 'سرنا', 'هله',
'حیات', 'صدقات', 'آمد#آ', 'بخش', 'جان', 'هله', 'مرو', 'خاتم', 'حسن', 'جمال', 'هله', 'یوسف', 'دهر', 'مکار', 'اخوان', 'ستمکار', 'م',
'موسی', 'هله', 'مرو', 'آثار', 'پی', 'عیان', 'مکشان', 'خیال', 'فکر', 'برمگزین', 'مهل', 'دیدار', 'هله', 'زما',
'رنجو', 'صحت', 'قران', 'عیسی', 'هله', 'مرو', 'انکار', 'جانب', 'مجو', 'فرعون', 'دل', 'دریا', 'برآر', 'گرد', 'ن',
'شی', 'های', 'جان', 'خواجه', 'جان', 'شاهد', 'هله', 'مرو', 'زنار', 'ترسا', 'سه', 'دو', 'برای', 'گران', 'ر',
'مختار', 'بگزیده', 'احمد', 'جز', 'وفا', 'ختم', 'زمان', 'صدیق', 'هله', 'مرو', 'افشار', 'غبغبه', 'بگز', 'لب', 'وه',
'دار', 'یقین', 'مرو', 'شخسار', 'زمین', 'مرغ', 'همچو', 'وطن', 'مقام', 'سدره', 'کرمی', 'جبرئیل', 'مرو',
'وقت', 'شد#شو', 'جمع', 'بتان', 'حریف', 'رند', 'مرو', 'دیوار', 'شهنشاه', 'جست#جو', 'همگی', 'گوش', 'شو', 'اکنون', 'گفتا',
'کار', 'آمد#آ',
'مرو', 'ر']
```

Stemming tokenized words

In [11]:
```
my_stemmer = FindStems()
# A list comprehension to find stems of each word inside the tokenized list of words correspondi
example_stems = [my_stemmer.convert_to_stem(token) for token in poems['4844']['tokenized']]
```

In [12]:
```
print(example_stems)
```

```
['د', 'چشم', 'زمین', 'روی', 'مرو', 'رخ', 'نغز', 'ماه', 'هله', 'مرو', 'دستار', 'میچان', 'شاه', 'هله',
'مسوز', 'اسرار', 'خانه', 'مبر', 'یار', 'مبر', 'مرو', 'اغیار', 'جانب', 'مکن', 'آزار', 'مکن', 'راست', 'باز', 'ل',
'برفت', 'بار', 'هله', 'مجوی', 'جنگ', 'دغل', 'ستیزه', 'یار', 'مکن', 'مرو', 'خار', 'جانب', 'مکن', 'گلزار', 'گل',
'مرو', 'ناچار', 'خوش', 'حیات', 'دین', 'دل', 'پرورده', 'مولا', 'توایم', 'چاکر', 'بنده', 'مرو', 'بار', 'مکن', 'ی',
'نا', 'مخور', 'هله', 'مرو', 'تار', 'مسکل', 'طرب', 'چنگ', 'مشکن', 'توام', 'نوا', 'مست', 'توام', 'سرنای', 'هله',
'حیات', 'صدقات', 'بیا', 'بخش', 'جان', 'هله', 'مرو', 'خاتم', 'حسن', 'جمال', 'هله', 'یوسف', 'دهر', 'مکاری', 'اخوان', 'ستمکا',
'موسی', 'هله', 'مرو', 'اثر', 'پی', 'مک', 'عیان', 'خیال', 'فکر', 'برمگزین', 'مهل', 'دیدار', 'هله', 'مرو',
'رن', 'صحت', 'قران', 'عیسی', 'هله', 'مرو', 'انکار', 'جانب', 'مجو', 'فرعون', 'دل', 'دریا', 'برآر', 'گرد', 'زمان',
'شهان', 'های', 'جان', 'خواجه', 'جان', 'شاهد', 'هله', 'مرو', 'زنار', 'ترسا', 'سه', 'دو', 'برای', 'گران', 'جور',
'بگزیده', 'احمد', 'جز', 'وفا', 'ختم', 'زمانی', 'صدیق', 'هله', 'مرو', 'افشار', 'غبغبه', 'گزید&گز', 'لب', 'شیوه',
'دار', 'یقین', 'مرو', 'شخسار', 'زمین', 'مرغ', 'همچو', 'وطن', 'مقام', 'سدره', 'کرمی', 'جبرئیل', 'مرو',
'مختار',
```

'بی', 'نفس', 'جان', 'نزید', 'احسان', 'گشود&گشا', 'دیوار', 'مرو', 'رند', 'حریف', 'بت', 'جمع', 'شد&شد', 'وقت', 'کار', 'بیا', 'کار', 'کار', 'مرو', 'هله', 'باقی', 'غزل', 'شهنشاه', 'جست&جو', 'همگی', 'گوش', 'شو', 'اکنون', 'گ
مرو', 'فتار']