Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working

Introduction

In this project, we developed a program that calculates the conditional probabilities of the bigrams in our test file using a text file. This method is used in speech and language processing.

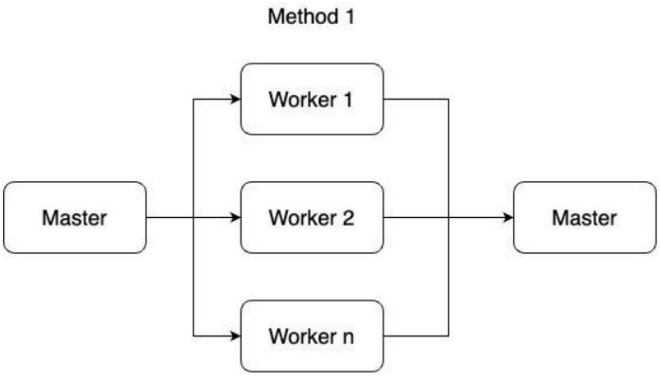
What is bigram?

A term In the topic of natural language processing. consecutive sequences of n words are called n-grams. Consecutive sequences of 2 words is called bigram.

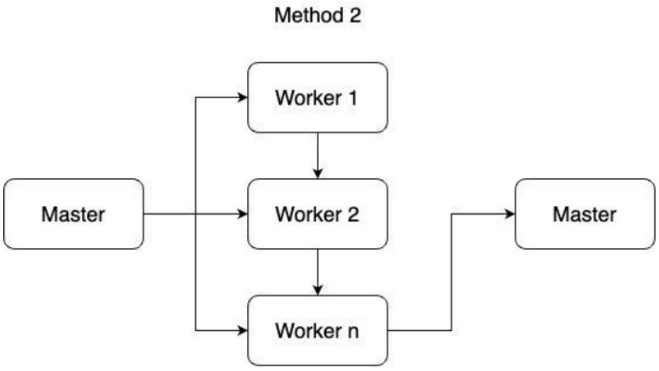
What is the main purpose of this project?

The main purpose is to ensure that the bigram conditional probabilities are calculated in parallel by communicating between parallel programs using MPI (Message Passing Interface). We use two different flow to do this(same result different technic).

In first method: process(which index is 0) split text data and send other processes and other process calculate bigrams and unigrams. And after that first process(0 index) calculate conditional probabilities of bigrams and print results.



In second method: process (which index is 0) split text data and send other processes and other process get before calculated result and add this data to their calculated bigram and unigram data and send them next process. And after that first process (0 index) print result of conditional probabilities of bigrams.



Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working

Program Interface

- The user can run the program via a terminal session in the program directory with Python installed in it.
- The user also needs mpi4py library for Python installed and open mpi installed.
- The command template is: “mpiexec -n 5 python main.py --input_file data/sample_text.txt --merge_method MASTER --test_file data/test.txt”
- The user can specify the number of processes by changing the number argument after “-n”. The number cannot be greater than processor size of the computer.
- The user can specify the input file directory by changing the argument comes after –input_file.
- The user can specify the test file (the bigram probabilities s/he wants in that file) directory by changing the argument comes after –test_file.
- The user can specify the merge method of the program by changing the argument comes after –merge_method. There are 2 options which are ‘MASTER’ and ‘WORKERS’.
- The program should print the bigrams and their probabilities to the terminal.
- The user may interrupt the program by pressing ctrl + c combination.
- The program terminates itself after the execution.
- The program works case sensitive.

Program Execution

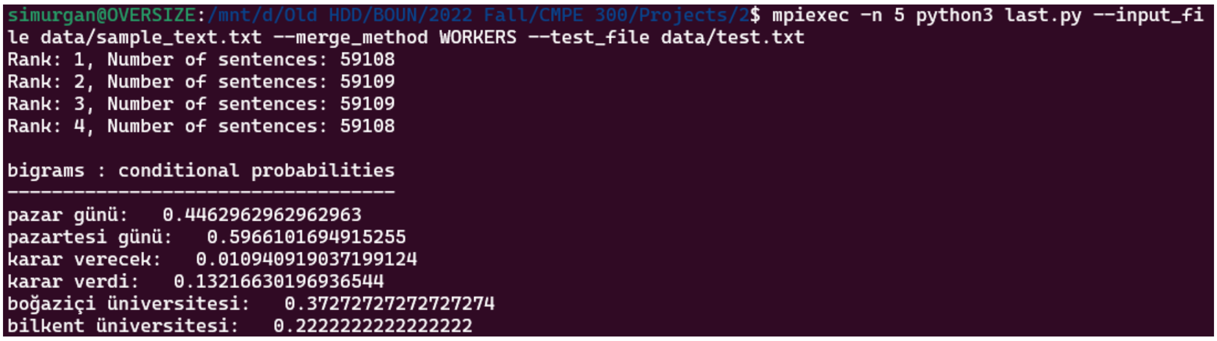
The program takes two txt files as input, input_file and test_file. It calculates the probabilities of the bigrams (pairs of words) given in the test_file and calculates their conditional probabilities by counting in the input_file.

The program may runs on multiple processes. This means the program divides its workload to the specified number of processes and finishes the job faster.

The program calculates the conditional probabilities as the number of the occurrence of bigrams in the input_file and divides it to the number of occurrences of the first word of the bigrams.

Run example:

Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working



```
simurgan@OVERSIZE:/mnt/d/Old HDD/BOUN/2022 Fall/CMPE 300/Projects/2$ mpiexec -n 5 python3 last.py --input_fi
le data/sample_text.txt --merge_method WORKERS --test_file data/test.txt
Rank: 1, Number of sentences: 59108
Rank: 2, Number of sentences: 59109
Rank: 3, Number of sentences: 59109
Rank: 4, Number of sentences: 59108

bigrams : conditional probabilities
-----------------------------------
pazar günü:   0.4462962962962963
pazartesi günü:   0.5966101694915255
karar verecek:   0.010940919037199124
karar verdi:   0.13216630196936544
boğaziçi üniversitesi:   0.37272727272727274
bilkent üniversitesi:   0.2222222222222222
```

Input and Output

The output of the program is printed in the format:

- "rank:<rank number>, number of sentences: <sentence number for each proccess > "
- "<bigram>: <probability>".

As the number of processes increase, the speed of the program increases too.

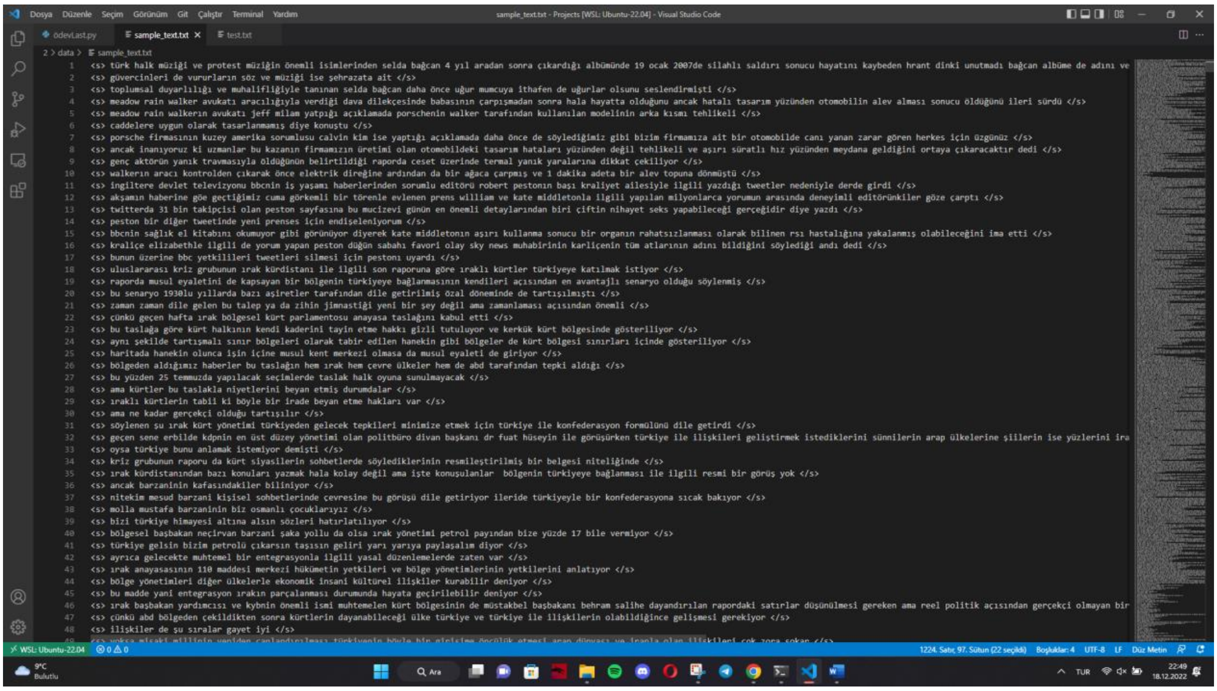
The input_file can be any text. For the sake of the program execution and to run the program fast, the input_file should be given as lines because the program divides the lines to the processes evenly. The test_file must be given as one pair of words in each line.

The user may use any kind of flags to indicate anything in the input_file. For example, the user my want the probability of occurance of a specific word, i.e. bogazici, in the first place of the lines and adds a flag to indicate it, i.e. '<s>'. To see the conditional probability, the user adds a line '<s> bogazici' to the test_file.

Example:

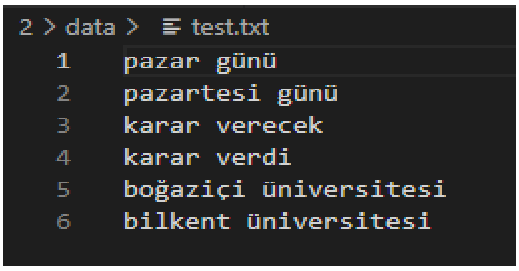
input_file example:

Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working


Dosya Düzenle Seçim Görünüm Git Çalıştır Terminal Yardım
sample_text.txt - Projects [WSL: Ubuntu-22.04] - Visual Studio Code
ödevLast.py
sample_text.txt
test.txt
2 > data > sample_text.txt
WSL: Ubuntu-22.04
0 0
1224. Satır, 97. Sütun (22 seçildi) Boşluklar: 4 UTF-8 LF Düz Metin
9°C Bulutlu
Ara
TUR
22:49 18.12.2022

```
<s> türk halk müziği ve protest müziğin önemli isimlerinden selda bağcan 4 yıl aradan sonra çıkardığı albümünde 19 ocak 2007de silahlı saldırı sonucu hayatını kaybeden hrant dinki unutmadı bağcan albüme de adını ve
<s> güvercinleri de vururların söz ve müziği ise şehrazata ait </s>
<s> toplumsal duyarlılığı ve muhalifliğiyle tanınan selda bağcan daha önce uğur mumcuya ithafen de uğurlar olsunu seslendirmişti </s>
<s> meadow rain walker avukatı aracılığıyla verdiği dava dilekçesinde babasının çarpışmadan sonra hala hayatta olduğunu ancak hatalı tasarım yüzünden otomobilin alev alması sonucu öldüğünü ileri sürdü </s>
<s> meadow rain walkerın avukatı jeff milam yatpığı açıklamada porschenin walker tarafından kullanılan modelinin arka kısmı tehlikeli </s>
<s> caddelere uygun olarak tasarlanmamış diye konuştu </s>
<s> porsche firmasının kuzey amerika sorumlusu calvin kim ise yaptığı açıklamada daha önce de söylediğimiz gibi bizim firmamıza ait bir otomobilde canı yanan zarar gören herkes için üzgünüz </s>
<s> ancak inanıyoruz ki uzmanlar bu kazanın firmamızın üretimi olan otomobildeki tasarım hataları yüzünden değil tehlikeli ve aşırı süratlı hız yüzünden meydana geldiğini ortaya çıkaracaktır dedi </s>
<s> genç aktörün yanık travmasıyla öldüğünün belirtildiği raporda ceset üzerinde termal yanık yaralarına dikkat çekiliyor </s>
<s> walkerın aracı kontrolden çıkarak önce elektrik direğine ardından da bir ağaca çarpmış ve 1 dakika adeta bir alev topuna dönmüştü </s>
<s> ingiltere devlet televizyonu bbcnin iş yaşamı haberlerinden sorumlu editörü robert pestonın başı kraliyet ailesiyle ilgili yazdığı tweetler nedeniyle derde girdi </s>
<s> akşamın haberine göe geçtiğimiz cuma görkemli bir törenle evlenen prens william ve kate middletonla ilgili yapılan milyonlarca yorumun arasında deneyimli editörünkiler göze çarptı </s>
<s> twitterda 31 bin takipçisi olan peston sayfasına bu mucizevi günün en önemli detaylarından biri çiftin nihayet seks yapabileceği gerçeğidir diye yazdı </s>
<s> peston bir diğer tweetinde yeni prenses için endişeleniyorum </s>
<s> bbcnin sağlık el kitabını okumuyor gibi görünüyor diyerek kate middletonın aşırı kullanma sonucu bir organın rahatsızlanması olarak bilinen rsı hastalığına yakalanmış olabileceğini ima etti </s>
<s> kraliçe elizabethle ilgili de yorum yapan peston düğün sabahı favori olay sky news muhabirinin karliçenin tüm atlarının adını bildiğini söylediği andı dedi </s>
<s> bunun üzerine bbc yetkilileri tweetleri silmesi için pestonı uyardı </s>
<s> uluslararası kriz grubunun ırak kürdistanı ile ilgili son raporuna göre ıraklı kürtler türkiyeye katılmak istiyor </s>
<s> raporda musul eyaletini de kapsayan bir bölgenin türkiyeye bağlanmasının kendileri açısından en avantajlı senaryo olduğu söylenmiş </s>
<s> bu senaryo 1930lu yıllarda bazı aşiretler tarafından dile getirilmiş özal döneminde de tartışılmıştı </s>
<s> zaman zaman dile gelen bu talep ya da zihin jimnastiği yeni bir şey değil ama zamanlaması açısından önemli </s>
<s> çünkü geçen hafta ırak bölgesel kürt parlamentosu anayasa taslağını kabul etti </s>
<s> bu taslağa göre kürt halkının kendi kaderini tayin etme hakkı gizli tutuluyor ve kerkük kürt bölgesinde gösteriliyor </s>
<s> aynı şekilde tartışmalı sınır bölgeleri olarak tabir edilen hanekin gibi bölgeler de kürt bölgesi sınırları içinde gösteriliyor </s>
<s> haritada hanekin olunca işin içine musul kent merkezi olmasa da musul eyaleti de giriyor </s>
<s> bölgeden aldığımız haberler bu taslağın hem ırak hem çevre ülkeler hem de abd tarafından tepki aldığı </s>
<s> bu yüzden 25 temmuzda yapılacak seçimlerde taslak halk oyuna sunulmayacak </s>
<s> ama kürtler bu taslakla niyetlerini beyan etmiş durumdalar </s>
<s> ıraklı kürtlerin tabii ki böyle bir irade beyan etme hakları var </s>
<s> ama ne kadar gerçekçi olduğu tartışılır </s>
<s> söylenen şu ırak kürt yönetimi türkiyeden gelecek tepkileri minimize etmek için türkiye ile konfederasyon formülünü dile getirdi </s>
<s> geçen sene erbilde kdpnin en üst düzey yönetimi olan politbüro divan başkanı dr fuat hüseyin ile görüşürken türkiye ile ilişkileri geliştirmek istediklerini sünnilerin arap ülkelerine şiilerin ise yüzlerini ira
<s> oysa türkiye bunu anlamak istemiyor demişti </s>
<s> kriz grubunun raporu da kürt siyasilerin sohbetlerde söylediklerinin resmileştirilmiş bir belgesi niteliğinde </s>
<s> ırak kürdistanından bazı konuları yazmak hala kolay değil ama işte konuşulanlar  bölgenin türkiyeye bağlanması ile ilgili resmi bir görüş yok </s>
<s> ancak barzaninin kafasındakiler biliniyor </s>
<s> nitekim mesud barzani kişisel sohbetlerinde çevresine bu görüşü dile getiriyor ileride türkiyeyle bir konfederasyona sıcak bakıyor </s>
<s> molla mustafa barzaninin biz osmanlı çocuklarıyız </s>
<s> bizi türkiye himayesi altına alsın sözleri hatırlatılıyor </s>
<s> bölgesel başbakan neçirvan barzani şaka yollu da olsa ırak yönetimi petrol payından bize yüzde 17 bile vermiyor </s>
<s> türkiye gelsin bizim petrolü çıkarsın taşısın geliri yarı yarıya paylaşalım diyor </s>
<s> ayrıca gelecekte muhtemel bir entegrasyonla ilgili yasal düzenlemelerde zaten var </s>
<s> ırak anayasasının 110 maddesi merkezi hükümetin yetkileri ve bölge yönetimlerinin yetkilerini anlatıyor </s>
<s> bölge yönetimleri diğer ülkelerle ekonomik insani kültürel ilişkiler kurabilir deniyor </s>
<s> bu madde yani entegrasyon ırakın parçalanması durumunda hayata geçirilebilir deniyor </s>
<s> ırak başbakan yardımcısı ve kybnin önemli ismi muhtemelen kürt bölgesinin de müstakbel başbakanı behram salihe dayandırılan rapordaki satırlar düşünülmesi gereken ama reel politik açısından gerçekçi olmayan bir
<s> çünkü abd bölgeden çekildikten sonra kürtlerin dayanabileceği ülke türkiye ve türkiye ile ilişkilerin olabildiğince gelişmesi gerekiyor </s>
<s> ilişkiler de şu sıralar gayet iyi </s>
<s> yoksa misakı millinin yeniden canlandırılması türkiyenin böyle bir girişime öncülük etmesi arap dünyası ve iranla olan ilişkileri çok zora sokan </s>
```

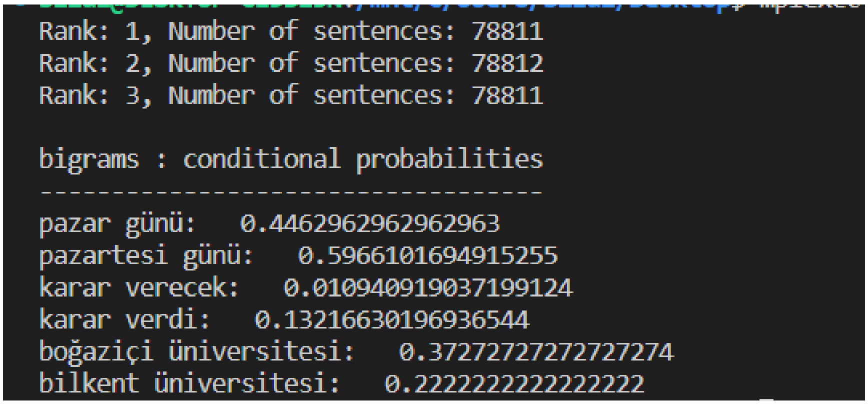
test_file example:



```
pazar günü
pazartesi günü
karar verecek
karar verdi
boğaziçi üniversitesi
bilkent üniversitesi
```

output example:

Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working



```
Rank: 1, Number of sentences: 78811
Rank: 2, Number of sentences: 78812
Rank: 3, Number of sentences: 78811

bigrams : conditional probabilities
-----------------------------------
pazar günü:   0.4462962962962963
pazartesi günü:   0.5966101694915255
karar verecek:   0.010940919037199124
karar verdi:   0.13216630196936544
boğaziçi üniversitesi:   0.37272727272727274
bilkent üniversitesi:   0.2222222222222222
```

Program Structure

First of all, the program will be executed by all of the processes. This means that, all of the lines will be executed by all of the processes. Therefore, we have to arrange them accordingly via conditional structures.

Firstly, we imported mpi4py and sys libraries. With sys, we take given program arguments. With mpi4py, we use multiple processes.

We get the number of processes, number of workers, and the rank of the processes. We will use the rank of the processes in the conditional structures that we decide which process do what.

After that, we initialized the test_uni and test_bi dictionaries and we store the bigrams and unigrams given in the test_file in these dictionaries. All processes do that because we will use them with all of the processes.

If the rank of the process is 0, the program gets in the first if block, so only the first process executes the lines in that block. We call that process MASTER.

If the rank is not 0, the program gets in the else block, so the lines in that block are executed by all of the processes except the first one. We call these processes WORKERS.

From the beginning of this if block until the end, the program will be divided, so there is no other line out of these if and else blocks.

In the if block (which the MASTER process executes only) the program takes the lines of the input_file and divide them to the workers evenly. To do that, the program iterates in the range of 1 and the number of processes and sends some of the lines to the workers via the send method of the mpi4py library.

After that, the program checks the merge method and do the tasks accordingly. If the merge method is MASTER, the program here (the MASTER process) receives the counted data via the recv method of the mpi4py library, test_bi and test_uni dictionaries in an array which comes from all of the other processes, and sum all of them in its own test_bi and test_uni dictionary.

If the merge method is WORKERS, the program receives that data from only the process with the highest rank via the recv method of the mpi4py library. With this merge method, the data is summed through the workers while coming to the MASTER process, so it doesn't need any additional summation.

After that, the MASTER process iterates over the bigrams in the test_bi dictionary and prints the conditional probabilities calculated as dividing the number of the occurance of the first word to the number of the occurrence of the pair.

In the else block (which only the WORKER processes execute), the program prints the rank of the process and the number of sentences that process got firstly. Then, in the input file, it counts the occurrences of the first words of the bigrams and bigrams which given in the test_file and stores the values in its own test_bi and test_uni dictionaries.

If the merge method is specified as WORKERS, the test_bi and test_uni dictionaries of the former process is received via the recv method of the mpi4py library and summed to the new ones, except that if the rank is 1 (the lowest WORKER).

An array of test_bi and test_uni is assigned to the result variable and sent to another process via send method of the mpi4py library. If the merge method is WORKERS, the result is sent to the next worker process (to the master process if the rank is the highest) and if the merge method is MASTER, it is sent to the MASTER process.

When a process is executing a line which the recv method is called, the program stops and waits until the data comes for that process. Therefore, the program works accordingly.

Examples

Input:

<s> türk halk müziği ve protest müziğin önemli isimlerinden selda bağcan 4 yıl aradan sonra çıkardığı albümünde 19 ocak 2007de silahlı saldırı sonucu hayatını kaybeden hrant dinki unutmadı bağcan albüme de adını veren güvercinleri de vururlar şarkısını hrant dinke adadı </s>

<s> güvercinleri de vururların söz ve müziği ise şehrazata ait </s>

<s> toplumsal duyarlılığı ve muhalifliğiyle tanınan selda bağcan daha önce uğur mumcuya ithafen de uğurlar olsunu seslendirmişti </s>

<s> caddelere uygun olarak tasarlanmamış diye konuştu </s>

Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working

<s> porsche firmasının kuzey amerika sorumlusu calvin kim ise yaptığı açıklamada daha önce de söylediğimiz gibi bizim firmamıza ait bir otomobilde canı yanan zarar gören herkes için üzgünüz </s>

<s> ancak inanıyoruz ki uzmanlar bu kazanın firmamızın üretimi olan otomobildeki tasarım hataları yüzünden değil tehlikeli ve aşırı süratlı hız yüzünden meydana geldiğini ortaya çıkaracaktır dedi </s>

<s> genç aktörün yanık travmasıyla öldüğünün belirtildiği raporda ceset üzerinde termal yanık yaralarına dikkat çekiliyor </s>

<s> ingiltere devlet televizyonu bbcnin iş yaşamı haberlerinden sorumlu editörü robert pestonın başı kraliyet ailesiyle ilgili yazdığı tweetler nedeniyle derde girdi </s>

<s> uluslararası kriz grubunun ırak kürdistanı ile ilgili son raporuna göre ıraklı kürtler türkiyeye katılmak istiyor </s>

<s> raporda musul eyaletini de kapsayan bir bölgenin türkiyeye bağlanmasının kendileri açısından en avantajlı senaryo olduğu söylenmiş </s>

Test:

<s> türk

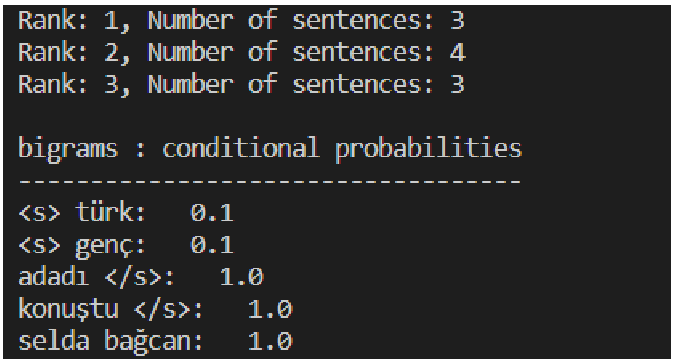
<s> genç

adadı </s>

konuştu </s>

selda bağcan

Output:



```
Rank: 1, Number of sentences: 3
Rank: 2, Number of sentences: 4
Rank: 3, Number of sentences: 3

bigrams : conditional probabilities
-----------------------------------
<s> türk:   0.1
<s> genç:   0.1
adadı </s>:   1.0
konuştu </s>:   1.0
selda bağcan:   1.0
```

In the first part of the output, we can easily see the number of sentences dropped by each worker, and then the bigrams in the test.txt part and the possibilities of these bigrams in this example.

By using “<s> türk”, we can see the conditional probability of starting sentences with the word “türk” in our input text.

By using “adadı </s>”, we can see the conditional probability of ending a line after the word “adadı” in our input text.

By using selda bağcan, we cen see the conditional probability of that the word ‘bağcan’ coming after the word ‘selda’

Improvements and Extensions

Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working

The program can be made modular and made usable by being accessed by another script. By pulling data from different file types, the desired bigrams can be obtained in these data. The program can also be modified to calculate trigram or any n-gram probabilities.

The weak point of the program is that it is slower than other programming languages because the language used is python. Another weak point is dependency for the open-mpi used for the program to work because installing open-mpi is difficult for windows. Actually it doesn't support windows, the user has to use another OS or tools like WSL or VMBOX to use the program.

The strong point of the program is that it mainly consists of a code block written in a short, easy to understand way. The input in the text file does not have to be line by line for the program to work. Also, not all sentences have to start with <s> and end with </s>. In addition, the user can edit the input file and the test part as desired.

Difficulties Encountered

- We had a hard time grasping the logic of the assignment, so we first thought about the working logic of MPI.
- We started to write the code on ubuntu installed in VirtualBox. But sharing and writing the code among us is difficult because VirtualBox is quite slow.
- When we wanted to write the code and run VirtualBox on Windows, we could not test the program because VirtualBox uses 2 cores.
- When we tried to increase the number of cores, we recognized that we have an upper limit for this. Therefore we couldn't test the program with high number of processes.
- We had a hard time writing the code in windows and testing it in VirtualBox.
- For these reasons, we installed WSL on windows and wanted to add the mpi4py library, this was a very long process.
- Finally, after the installations, we developed the program and did our tests on windows.

Conclusion

MPI programming is a really different and new experience for us. We saw that it fastens the programs with the same objectives by being able to dividing the job to each processor the computer has. However, the development and debugging was so different and harder than normal programming. The advantages it has can be taken when they are needed. Also we saw that not all tasks can be divided into processes. There are some we can do and some we cannot. To implement the programs better, the logic and usage of the MPI should be learnt deeply.

Reference

https://www.open-mpi.org/

https://indico.desy.de/event/12535/contributions/9450/attachments/6466/7452/CL5_MPI_Lecture.pdf

https://web.stanford.edu/~jurafsky/slp3/

https://web.stanford.edu/~jurafsky/slp3/3.pdf

Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working

https://www.techtarget.com/searchenterprisedesktop/definition/message-passing-interface-MPI

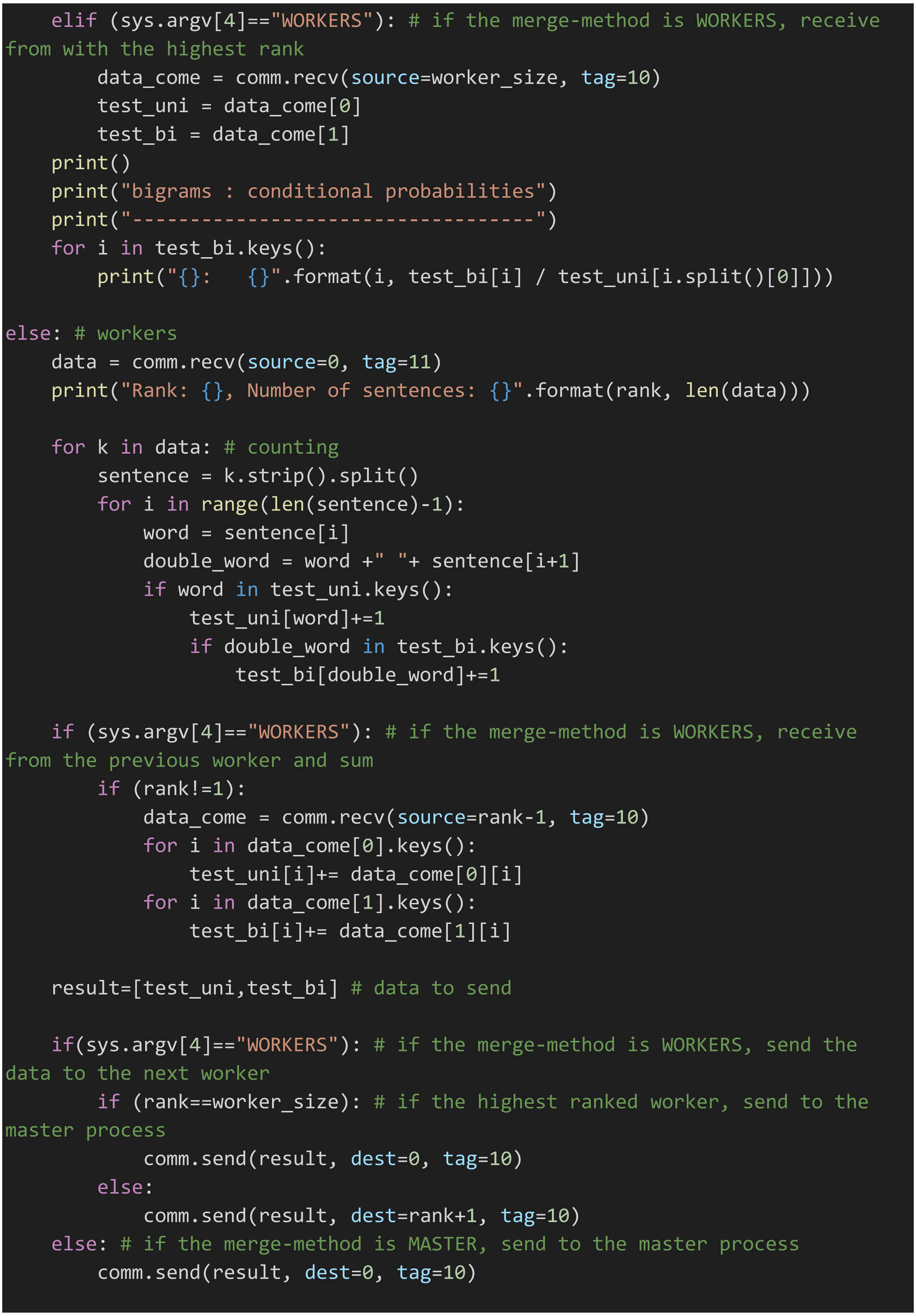
https://learn.microsoft.com/en-us/windows/wsl/install

Appendix

The source code of the program:

```
from mpi4py import MPI
import sys

comm = MPI.COMM_WORLD
rank = comm.Get_rank()
world_size = comm.Get_size()
worker_size = world_size - 1

test_uni=dict()
test_bi=dict()
test = open(sys.argv[6], "r")
test_lines = test.readlines()

for test_line in test_lines:
    test_words = test_line.strip().split()
    test_uni[test_words[0]] = 0
    test_bi[test_words[0]+" "+test_words[1]] = 0

if rank == 0: # master process
    file = open(sys.argv[2], "r")
    lines = file.readlines()
    size = len(lines)
    line_per_rank = size / worker_size

    worker_first=0
    for i in range(1,world_size): # split and send data equally to the workers
        worker_last = round(i * line_per_rank)
        comm.send(lines[worker_first:worker_last], dest=i, tag=11)
        worker_first=worker_last

    if (sys.argv[4]=="MASTER"): # if the merge-method is MASTER, receive from
all
        for j in range(1,world_size):
            data_come = comm.recv(source=j, tag=10)
            for i in data_come[0].keys():
                test_uni[i]+= data_come[0][i]
            for i in data_come[1].keys():
                test_bi[i]+= data_come[1][i]
```

Student Name: Bilal Atım – Ömer Şükrü Uyduran
Student Number: 2019400168 - 2018400234
Compile Status: Compiling
Program Status: Working



```
    elif (sys.argv[4]=="WORKERS"): # if the merge-method is WORKERS, receive
from with the highest rank
        data_come = comm.recv(source=worker_size, tag=10)
        test_uni = data_come[0]
        test_bi = data_come[1]
    print()
    print("bigrams : conditional probabilities")
    print("-----------------------------------")
    for i in test_bi.keys():
        print("{}:   {}".format(i, test_bi[i] / test_uni[i.split()[0]]))

else: # workers
    data = comm.recv(source=0, tag=11)
    print("Rank: {}, Number of sentences: {}".format(rank, len(data)))

    for k in data: # counting
        sentence = k.strip().split()
        for i in range(len(sentence)-1):
            word = sentence[i]
            double_word = word +" "+ sentence[i+1]
            if word in test_uni.keys():
                test_uni[word]+=1
                if double_word in test_bi.keys():
                    test_bi[double_word]+=1

    if (sys.argv[4]=="WORKERS"): # if the merge-method is WORKERS, receive
from the previous worker and sum
        if (rank!=1):
            data_come = comm.recv(source=rank-1, tag=10)
            for i in data_come[0].keys():
                test_uni[i]+= data_come[0][i]
            for i in data_come[1].keys():
                test_bi[i]+= data_come[1][i]

    result=[test_uni,test_bi] # data to send

    if(sys.argv[4]=="WORKERS"): # if the merge-method is WORKERS, send the
data to the next worker
        if (rank==worker_size): # if the highest ranked worker, send to the
master process
            comm.send(result, dest=0, tag=10)
        else:
            comm.send(result, dest=rank+1, tag=10)
    else: # if the merge-method is MASTER, send to the master process
        comm.send(result, dest=0, tag=10)
```