**ШИНЖЛЭХ УХААН ТЕХНОЛОГИЙН ИХ СУРГУУЛЬ**

**Мэдээлэл холбооны технологийн сургууль**



БИЕ ДААЛТЫН АЖЛЫН ТАЙЛАН

# Эх Хэлний Боловсруулалт (F.CSC336) 2024-2025 оны хичээлийн жилийн намар

**Бие даалтын ажлын нэр:** **Корпус үүсгэн хэлний загвар байгуулах**

# Хичээл заасан багш: Э.Батцэцэг

# **Бие даалтын ажил гүйцэтгэсэн:** С.Тэмүүжин B221960002

**Улаанбаатар хот. 2024 он**

import PyPDF2

import re

import nltk

from nltk import word\_tokenize

from nltk.util import ngrams

from collections import Counter

import random

def read\_pdf\_and\_clean\_text():

# Open the PDF file in binary read mode

with open('/content/Анне Франкийн өдрийн тэмдэглэл.pdf', 'rb') as file:

# Create a PdfReader object to read the PDF file

pdf = PyPDF2.PdfReader(file)

# Initialize an empty string to store the extracted text

text\_content = ''

# Loop through each page in the PDF

for page\_num in range(len(pdf.pages)):

# Access the current page

page = pdf.pages[page\_num]

# Extract the text content of the current page

page\_text = page.extract\_text()

# If the page contains text, process it

if page\_text:

# Remove specific characters (e.g., '.', ',') using regex

cleaned\_text = re.sub(r'[.,]', '', page\_text)

# Append the cleaned text to the overall text content

text\_content += cleaned\_text

# Return the cleaned text content extracted from the PDF

return text\_content

def read\_pdf\_and\_clean\_text():

# Open the PDF file in binary read mode

with open('/content/Анне Франкийн өдрийн тэмдэглэл.pdf', 'rb') as file:

# Create a PdfReader object to read the PDF file

pdf = PyPDF2.PdfReader(file)

# Initialize an empty string to store the extracted text

text\_content = ''

# Loop through each page in the PDF

for page\_num in range(len(pdf.pages)):

# Access the current page

page = pdf.pages[page\_num]

# Extract the text content of the current page

page\_text = page.extract\_text()

# If the page contains text, process it

if page\_text:

# Remove specific characters (e.g., '.', ',') using regex

cleaned\_text = re.sub(r'[.,]', '', page\_text)

# Append the cleaned text to the overall text content

text\_content += cleaned\_text

# Return the cleaned text content extracted from the PDF

return text\_content

extracted\_text = read\_pdf\_and\_clean\_text()

words\_counted, sentences\_counted = count\_words\_and\_sentences(extracted\_text)

print(f"Number of words: {words\_counted}")

print(f"Number of sentences: {sentences\_counted}")

**Output:**

Number of words: 52175

Number of sentences: 649

# Download the NLTK 'punkt' package for tokenization

nltk.download('punkt')

# Tokenize the extracted text into individual words

tokenized\_words = word\_tokenize(extracted\_text)

# Unigrams: Each word as a single unit (essentially the tokenized words)

unigrams = tokenized\_words

# Bigrams: Pairs of consecutive words from the tokenized text

bigrams = list(ngrams(tokenized\_words, 2))

# Trigrams: Groups of three consecutive words from the tokenized text

trigrams = list(ngrams(tokenized\_words, 3))

def generate\_text(ngram\_list, word\_count=500):

# Create a dictionary to map n-gram keys to possible next words

ngram\_dict = {}

for ngram in ngram\_list:

# Split each n-gram into a key (all but the last word) and value (the last word)

key, value = ngram[:-1], ngram[-1]

# Add the value to the list of possible continuations for the key

ngram\_dict.setdefault(key, []).append(value)

# Select a random starting key from the n-gram dictionary

current = random.choice(list(ngram\_dict.keys()))

# Initialize the generated words with the selected starting key

generated\_words = list(current)

# Generate the desired number of words

for \_ in range(word\_count - len(current)):

# Get the list of possible next words for the current key

next\_words = ngram\_dict.get(current)

if not next\_words:

# If no continuation exists, randomly select a new starting key

current = random.choice(list(ngram\_dict.keys()))

generated\_words.extend(list(current))

else:

# Randomly choose the next word from the possible options

next\_word = random.choice(next\_words)

# Add the chosen word to the generated words

generated\_words.append(next\_word)

# Update the current key by sliding over (drop the first word, add the new one)

current = (\*current[1:], next\_word) if len(current) > 1 else (next\_word,)

# Join the generated words into a single string and return it

return ' '.join(generated\_words)

print("Unigram Text:\n", generate\_text([(w,) for w in unigrams]))

print("\nBigram Text:\n", generate\_text(bigrams))

print("\nTrigram Text:\n", generate\_text(trigrams))

**Unigram Text:**

хүрч инээж Германы чөлөөтэй лагалдсан ажээ нэг дуугарах хийж -Аннегийн дуусаж унасан ньэрчүүд Хайрт юм өдрийн Тэрхоёр нөхрөө сонор эмч орох дандаа их баярлав мөрөөднөм Хайрт даа ирвэгнэх ньгомдчих хүчтэй чадах цагт ньном Өглөөний ньөвдсөн юм болгов арилж их бүх дураараа үзсэнгүй гэж ! : хивсээ нээлттэй илүү өглөө хүртэл улам шүү талын хүн балчирхан гай усандоогуур минь Бид үзүүлж нь мөргөснөө таваг ” хөтлөгдөж зөвшөөрөхгүй үлгэр саналаа битгий гурав бодож байна эхэлтэл хураачихвал дан чюу Харрийг маань өгнө өглөө улам болж члбайсан олз чинь би хэн чүгүй яалаа Тэгээд ньхаах миний гуай азжаргалд үүбайх сурагчдын Саун Аавын надтай тийм арав Мийп юмбилээ Санаснаас нь түчигнэж зовлонг ньтүүний дараа баясгалант байнга Аав гэхбайх Есд дээгүүр хоол ажил чүгүй чхүртэл тэндээс боллоо найртай Хөөрхий ихюмбэлэглэв цоргуудаа байв Китти үерхэхээ Мэдээж Христос цөөн өдрөөр минь зургадугаар гэхэд өвөл байгаа амьдралд хэрэг очих Анне1944 хоёр оны ? асааж тухай өөлөх таминь ньхэлсэн сэтгэлийн гарч гэрийн зугтахын хөвгүүнтэй гэдгийг оны 15 хэрэв сэтгэлээ ашигтайгаа дээр Бусад оройг хоёрдугаар гэсэн баригдан алдсан Нүүрээ чадна болох үзэж Ээжийнхээр гуай кофе зөвхөн Тэгэхлээр идэж юм чадахгүй минь хэлэхэд ньавсан Тэрнь аваад үнсэлцдэг хүртэл олж хэрүүлийг Дон цагбайв нь хутга бид юу цайлган сургуульдаа уйлан явдал явахыг Харин “ цаг туульс ” дайнд болж Намайг уншив Тэдтэгээд одоогийн бай өрөөнд чихэнд чадалтай боддог мэдэж еврейчүүдийг нэвтрүүлэгт байсан чин гуай эрэн юмуугэдгийг ээжээсээ дутагдлыг байтал дугхийгээд гэдэг гэв сайн Харамсалтай ньминий байх биүдээс үдэд үздэг санаа Аав билээ хулгайч Даан хаалгаа бас очин болохоор өгдөг ах болов байжээ Дуссел

**Bigram Text:**

балгүй сэдэв рүүоръё Хоргодох байрныхны аммурийлцдаггүй тэр зовлонг мэдрэв Би чмөчөөгөөөгөхгүй ноён Коофуйс ходоодондоо мэсзасал хийлгэх боллоо Петер маань найм дөчин тав хүрэв үүүгүй юусалаад уначих вий гээд баснэлээд хэдэн мянган хүнийг үргэлж лэндтэнд хань болов Мань хүн чашгүй тэгээд яагаад чөөрчилж болохгүй Машины тухайд хөнгөн байх шиг хамаг нууцыг хадгалж байгаад өгсөн байв Гадны хүмүүст байхгүй юу хийх ёстой юмболов уудаа Гайхалтайяа түүнээс хойш манайх чинь харин нэгхоног хүлээ гэвГэвч биаавын зөвлөснийг чимэдэж байгаа Затийм байж болно шүү Өдөр ньтэнд Мийп шашны баяраар “ Ээж бид шинэ мэдээ тарсны учир ньВангийн ордны өмнөх том нүдээрээ ээжийг надтай танилцсанаасаа хойш мөн л ухаарна Дараа нь байдал муудах ньюубайхав Өөрийг ньхэн нэгэн хэмээр хангинана Би энд байгааг мэдэхгүй сургуулийн баясгалант үе минь ухаан суухад Даан гуайнхан миний ажлын хуучин хэвэндээ л чухам яахгэж юуны өмнө ньМийпийн авчирсан дуусашгүй энэихзовлон дайны дараа Петерийг өрөөнд хүлээсэн байна билээ Харин битгий л баттай ойлгож байна уу ? Чухам Маргот ньАннегаасгурван насаар эгчмэд байв Даарахаднус яарахад шээс байж санасандаа хүрч чадахгүй бизээ Нүдэнд минь ! Бигэдэг тийм чичмээр санагдахгүй байна Юугэхгээд байгааг минь ! Өчигдөр аав орж гурван долоо хоног тоолон суухаас өөр хөвгүүнтэй хамт байж вэбиХөөрхий минь чибиднийг хэрэлдэж муудалцан цаг хагасын өмнө чзаримдаа ганцаардах үебайсан лдаа Биээжийгээ өрөвдөж басэх үршиг байх даа Гэтэл Даан гуай дэргэд ирж очдог нь “ Элдорадо ” гэхзэргээр гэж Францтай байлдсан тавдугаар сарын хугацаанд чутга төгөлдөр хуурч Агнес Кингво София Монте төгөлдөр хуурч Агнес Кингво София Монте Черкессийн сурвалжит хатагтай Франк хэмээх азтай охин буюу ээж тэнгэр газар руухөтлөв Тэгтэл юуч болсон гэнэ Тэгтэл Даан гуайн өдөр шөнөгүй тачигнавч англид олигтой хохирол учруулж чадахгүй юмаа Бияасхийж Петерийг давгүй харж суусан бөгөөд бусдыг ньном зохиол буюу хатагтай үүрийн гэгээнээс үдшийн бүрий тасраагүй байхад нь Германчууд ухрахдаа оршин суугчдыг нүүлгээд эхэлбэл бичтэсвэр тэвчээрээ барна даагэж бодно Гэхдээ аарсаар юммундахбиш Аавын ширээний чийдэн гэнэт залуусын эрхийг хамгаалагч болчихлоо Гэрт зүгээр сайн өдөр бүрномын сангаас “ Би мөн байх Чибас бид хоёр мөнгөө бүгдий ньайлуудад тавьчихсанХэрэв юм амсвал яахболгэж дотроо “ Бурхан минь эрхлэн угтав Долоо хоногийн өмнө хаалга хаагдав Нэг өдөр хагас цаг ирээсэй тэгвэл нь хайлчихсан байна Ноён Кралер ээжид баглаа цэцэг ирэв Хүн ардыг зөөлөө гээд эмөгчээ Эмээ минь ! Бичинь толгой өндийхгүй ихажилтай сууна Дөрөв хагаст завсарлаж дахиад лхонх жингэнэхэд бидтүргэн жагсаал авцгаав Бинөгөө тавиуртай нууц ” хааны угсаа залгамжлах гүнж хэнгэгч ханхүүтэй гэрлэх бол арай гэж “ алив цохилтонд нэг дэвэргэж нэгдоордуулна Арай лшинэ гутлыг минь баяртай байна Хажуу хаалгаар ньМаргот бид хоёр бараг харанхуйд баймаар байсан биз гэв Түүнийг хаалга цоожлоод явахад ньбиш Эртхэн шиг харагдсан байх шиг ярьдаг байгаа түүнд баяр тэмдэглэхтэй манатай байх шүү дээ -Тиймээ гэхдээ биүдээс хойш хэдхоног өнгөрсний дараа үнэр саран ээлжлэн солигддог дөрвөн эмэгтэй хүнбиеэ барьж чадахааргүй байсан юмбол доо Тэгэхэд таүдээс өмнө биүдээс хойш намайг хэн мэдэх хийж гарав Харвал ардусгоожиж байв Жорлонд чорж болохгүй гээгүй гэнэ Нэг мэдэхэд номын жагсаалтанд оруулж цэвэрхэн дуудаж ччадахгүй ! Ашгүй бүгд лминий өдрийн тэмдэглэлОршил Асар богинохон даашинз өмсөөд битөдөлгүй хөнжилдөө шургах амжвал зөгийн балтай сүү гашилж ядахдаа салхивч нээж олсон Бид ямар

**Trigram Text:**

читэгэж бодоогүй байж чболох юм Тэгвэл ярилцахыг хүсдэг мөртөө үгээр илэрхийлэхэд хэцүү зүйлийг минь тэр мэдэж авна шүү дээ Биийм юман дээр аав ээжийн бодлыг өөрчлөхийг бодно Ерньлбиэнэ бүхний учрыг олохыг хүссэн охины чин хүслийн үгийг сая сая хүүхдүүд уншсаар байна Аннелис Марие буюу Анне Франк хэмээх еврей охин юм Түүний бичиж үлдээсэн өдрийн тэмдэглэлийг чинь уншиж болох уу ? Биллав тэднийг зөвшөөрнө гэж бодохгүй байна Гэхдээ юугэж боддогий ньасуух болно Лав лмуугаар санадаггүй байх Чибас бид хоёрын хэн хэн ньнэг юмнаас болгоомжлоод байх шиг байна Веермахт германы цэргийн албанд ажилладаггүй бүхгерман эмэгтэйчүүд хүүхдээ эрэг орчмын германы бэхлэлтэн дээр таван сая килограмм бөмбөг хаясан онгоцуудын түрүүч ньбуудалдаа буцаж иржээ Өнөөдөр 20000 нисэх онгоц германы хотыг бөмбөгдөхөөр шөнө бүр Холландын дээгүүр нисэн өнгөрч орос африкт хэдэн мянган хүн цусанд живж байна Дэлхий ертөнц бүхэлдээ дайнд нэрвэгдэн газар сайгүй алаан хядаан дуулдан түүнд мэргэжлийн хүмүүс заримдаа цагдаа манаач нар чөөрсдөө оролцох болжээ Хэн хүнгүй л жаахан ахиухан юм бусдаасаа өрсөөд долоочихсон гэдэг болж Авдаг ньнэмэгдэхгүй болохоор эцэстээ бүххүнхуурч мэхлэх болсон байна Залуучуудын хэрэг эрхэлсэн цагдаагийн газар ажлаа дийлэхгүй байгаа гэнэ Бие маань нэглбиш гэррүүгээ явахаас даа гэхэд ньбид түүнийг ихдордсоныг анх удаагаа гурилын түмпэн цэвэрлэж бөөн инээдэм болов Ээж чгэсэн ааваас өмнө өөр хүнтэй ойртоогүй ньлав Петерийн гарыг дэрлэн зүрхээ цээжинд ньнаан толгойгоо нийлүүлээд суудгийг минь миний найз охид мэдвэл юугэх бол ? ” гэлцэн дууалдав Шүүгээн дотор эрээн мяраан цаасаар хээлж хуарлан харПетерийн багаар бэлэгдэн чимсэн том сагс байж байлаа Бидтэр сагсаа дээш авчирангуут ерөөлийн бадаг шүлэг бичиж хийжээ Христос мэндэлсний баярын анхны өдөр үргэлжид лнэгийг боддогоороо Пимын тухай бодмоор санагдана Тэгэхдээ бүр аавынхаа ноднин өөрөө ярьсан залуугийнх нь хайр дурлалын тухай бодоод байх ньээ дээ ? гэсэн асуулт гарч байнаТиймээ “ Христос шашинт хүн хийсэн гэм нүглээ ганц өөрийнхөө хар толгойгоор хүлээх ёстой Харин еврей хүний гэмбурууг бүхеврей үндэстэн үүрэх ёстой ! ” гээд үсрэн босов Аав тэгэхэд муу юмболж мэднэ гэж түүний гараас бариад хамтдаа дээш гарав Намайг “ Ашгүй оноод хэлчихжээ Хэлэх гэсэн юмаа лхэлж ” гэхчлэн тунгаан боддог юм Үнэнээ хэлэхэд биЛизийг сар тиймээ бараг жилээр шахам мартжээ ЗаЛиз минь дайн дуусаж би чөлөөлөгдөж эрх чөлөөт амьдралдаа эргэн орно гэдгийн мэдэж байгаа атал ягбичээд эхлэхээр нэгл болж өгөхгүй юмаа Тэр маань олигтой юм болохгүй болхогийн хувин юмуу зуух рууорж чмэдэх юм Энэ чбас лгайхмаар байгаа биз ? Энэ долоо хоногийн сүүлчийн өдрөөс эхлэн бид цагхугацааг баримжаалах юмгүй болов Яагаад гэвэл нууцыг минь хэзээ чхэнд чхэлэхгүй дуугүй байдгийг чинь бимэдэх юм Эхнийх ньээжийн тухай яримаар байдгийг минь аав ээж гуравт ч халдав Петер лбитгий авчихаасай Тэрнадтай үнсэлцэхийг хүсч байгаа даа Намайг тэр “ Яагаад болохгүй гэж хатуу шийдэв Ширээнд суухаасаа өмнө би : -Чи наад даравгараа татна байгаа Миний бодлоор бид хоёрын ажил унтлагын өрөө рүү “ нүүж ” залуу хос бидний орон дээр надтай хамт хэвтэв Нарийхан орон дээр ньхэвтэв Мань эрдургүйцсэн чби тоосонгүй Урьд орой ньхүн алим бэлэглээдбайхад надад арай эелдэг байж болдоггүй юм бол байшингийн дээвэрт гарав Даан гуай хүүгийнхээ шохоорхлыг ерюман чинээ санасангүй Харин уншиж дуусаагүй номыг нь аав ньхурааж авсан ньПетерт дэндүү шударга