Wikipedia Summarizer

Samantha Fields-Samotowka

CISB-63

CRN 21061

October 29, 2023

Explanation of Project

I will be creating a basic summarizer for Wikipedia pages

Techniques Employed

- Web scraping
- Regular expressions
- Tokenizing words
- Removing stop words
- Stemming
- TF-IDF Vectorization

Pages tested

Found using the "random article" feature on Wikipedia, with the exception of the sixth, which holds personal meaning to me.

- Smooth Toadfish
- Nomen Dubium
- Fortuna
- The Snow Hole
- Garfield
- LGBT Community

Preparatory Code

Import Statements

In [1]:
```
#Web scraping libraries
import requests
from bs4 import BeautifulSoup

#General libraries
import re
import numpy as np
import warnings
warnings.filterwarnings("ignore")
import os
from os import path

#Visualisation
import matplotlib.pyplot as plt
from wordcloud import WordCloud

#NLTK NLP libraries
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize

#Vectorization library
from sklearn.feature_extraction.text import TfidfVectorizer
```

Downloading dependencies

In [2]:
```
nltk.download("punkt")
nltk.download("stopwords")
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Samantha\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Samantha\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[2]:
```
True
```

In [3]:
```
stop_words = set(stopwords.words("english"))
```

Text file creation

In [4]:
```
def write_text_to_file(directory, title, content, file_type):
    file_name = f"{file_type}_{title}.txt"
    file_path = os.path.join(directory, file_name)

    if not os.path.exists(directory):
        os.mkdir(directory)

    with open(file_path, "w", encoding="utf-8") as file:
        file.write(content)
```

Preprocessing text

In [5]:
```
def preprocess_text(text):
    words = word_tokenize(text)

    cleaned_sentence = re.sub(r"[^a-zA-Z]", " ", text)
    cleaned_words = word_tokenize(cleaned_sentence)

    cleaned_words = [word.lower() for word in cleaned_words if word.isalnum() and word

    return " ".join(cleaned_words)
```

Data Exploration and Preparation

Web page input and web scraping

In [6]:
```
#Enter a URL from any Wikipedia page
url = input("Please enter the Wikipedia URL you would like to summarize from: ")
```

```
Please enter the Wikipedia URL you would like to summarize from: https://en.wikipedi
a.org/wiki/LGBT_community
```

In [7]:
```
#Scraping the page for the content
response = requests.get(url)
html_content = response.content
soup = BeautifulSoup(html_content, "html.parser")
```

In [8]:
```
#Get the page title
title = soup.find("h1", {"id": "firstHeading"}).text

print(f"Summarizing the page: {title}.")
```

```
Summarizing the page: LGBT community.
```

In [9]:
```
#Write a text file of the page content
#This is more for future versions
directory = "./data/"
original_file_content = str(soup)

write_text_to_file(directory, title, original_file_content, "original_text")
```

In [10]:
```
#Extract only the paragraph content
paragraphs = soup.find_all("p")
paragraph_texts = [paragraph.get_text() for paragraph in paragraphs]
```

Data cleaning

This consists of:

1. Converting the text to lowercase
2. Using a regex to remove all punctuation and non-alphanumeric characters
3. Tokenizing the text into individual words
4. Removing the stop words
5. Stemming the tokens

6. Stripping the tokens words
7. Rebuilding the words into sentences
8. Rebuilding the sentences into paragraphs

In [11]:
```
lowercase_text = [text.lower() for text in paragraph_texts]

cleaned_text = [re.sub(r"[^a-zA-Z0-9\s]", "", text) for text in lowercase_text]

tokenized_text = [word_tokenize(text) for text in cleaned_text]

filtered_text = [[word for word in tokens if word not in stop_words] for tokens in tok

stemmer = PorterStemmer()
stemmed_text = [[stemmer.stem(word) for word in tokens] for tokens in filtered_text]
```

In [12]:
```
final_text = [[word for word in tokens if word.strip()] for tokens in stemmed_text]

sentences = [" ".join(tokens) for tokens in final_text]
```

In [13]:
```
processed_paragraphs = "\n\n".join(sentences)
```

In [14]:
```
#Write a text file of the processed text
#This is more for comparisons in future versions
directory = "./data/"
processed_file_content = processed_paragraphs

write_text_to_file(directory, title, processed_file_content, "processed_text")
```

In [15]:
```
#Display the original sentences
original_sentences = processed_paragraphs.strip().split(".")

print(original_sentences)
```

['lgbt commun also known lgbtq commun lgbtqia commun gay commun queer commun loos def in group lesbian gay bisexu transgend individu unit common cultur social movement com mun gener celebr pride divers individu sexualitynot verifi bodi lgbt activist sociolo gist see lgbt communitybuild counterweight heterosex homophobia biphobia transphobia sexual conformist pressur exist larger societi term pride sometim gay pride express l gbt commun ident collect strength pride parad provid prime exampl use demonstr gener mean termnot verifi bodi lgbt commun divers polit affili peopl lesbian gay bisexu tra nsgend consid part lgbt commun\n\ngroup may consid part lgbt commun includ gay villag lgbt right organ lgbt employe group compani lgbt student group school univers lgbtaff irm religi group\n\nlgbt commun may organ support movement civil right promot lgbt ri ght variou place around world time highprofil celebr broader societi may offer strong support organ certain locat exampl lgbt advoc entertain madonna state ask perform man i pride event around world would never ever turn new york city4\n\nlgbt glbt initi st and lesbian gay bisexu transgend use sinc 1990 term adapt initi lgb use replac term g ay refer commun whole begin variou form larg earli 1990s5citat need\n\nmovement alway includ lgbt peopl oneword unifi term 1950 earli 1980 gay see gay liber throughout 197 0 80 number group lesbian member profeminist polit prefer repres lesbian gay6 earli n ineti group shift name base lesbian gay bisexu transgend lgbt queer also increasingli reclaim oneword altern everlengthen string initi especi use radic polit group use que er sinc 80s6\n\niniti well common variant lgbtq adopt mainstream 1990s7 umbrella term use label topic sexual gender ident exampl lgbt movement advanc project term commun c enter servic specif member lgbt commun lgbt commun center comprehens studi center aro und unit states8\n\niniti lgbt intend emphas divers sexual gender identitybas cultur may refer anyon nonheterosexu noncisgend instead exclus peopl lesbian gay bisexu tran sgender9 recogn inclus popular variant add letter q identifi queer question sexual id ent lgbtq record sinc 19961011\n\ndisagr precis word best still present 2023 propos a d letter make particip group explicit12 detractor approach argu ad letter implicitli exclud other make wors brandingcit need\n\ngay commun frequent associ certain symbol especi rainbow rainbow flag greek lambda symbol l liber triangl ribbon gender symbol also use gay accept symbol mani type flag repres subdivis gay commun commonli recogn one rainbow flag accord gilbert baker creator commonli known rainbow flag color repre s valu commun\n\nlater pink indigo remov flag result presentday flag first present 19 79 pride parad flag includ victori aid flag leather pride flag bear pride flag13\n\nl ambda symbol origin adopt gay activist allianc new york 1970 broke away larger gay li ber front lambda chosen peopl might confus colleg symbol recogn gay commun symbol unl ess one actual involv commun back decemb 1974 lambda offici declar intern symbol gay lesbian right intern gay right congress edinburgh scotland13\n\ntriangl becam symbol gay commun holocaust repres jew homosexu kill german law holocaust homosexu label pin k triangl distinguish jew regular prison polit prison black triangl similarli symbol femal repres lesbian sisterhood\n\npink yellow triangl use label jewish homosexu gend er symbol much longer list variat homosexu bisexu relationship clearli recogniz may p opularli seen symbol symbol relat gay commun gay pride includ gayteen suicid awar rib bon aid awar ribbon labri purpl rhinoceros1415\n\nfall 1995 human right campaign adop t logo yellow equal sign deep blue squar becom one recogniz symbol lesbian gay bisexu transgend commun logo spot world becom synonym fight equal right lgbt people16\n\none notabl recent chang made philadelphia pennsylvania june 8 2017 ad two new stripe rain bow flag one black one brown intend highlight member color within lgbt community17\n \nlgbt commun repres social compon global commun believ mani includ heterosexu alli u nderrepres area civil right current struggl gay commun larg brought global unit state world war ii brought togeth mani closet rural men around nation expos progress attitu d part europ upon return home war mani men decid band togeth citi rather return small town fledgl commun would soon becom polit begin gay right movement includ monument in cid place like stonewal today mani larg citi gay lesbian commun center mani univers c olleg across world support center lgbt student human right campaign18 lambda legal em pow spirit foundation19 glaad20 advoc lgbt peopl wide rang issu unit state also inter n lesbian gay associ 1947 unit kingdom adopt univers declar human right udhr lgbt act ivist clung concept equal inalien right peopl regardless race gender sexual orient de clar specif mention gay right discuss equal freedom discrimination21 1962 clark polak join janu societi philadelphia pennsylvania22 year becam presid 1968 announc societi would chang name homosexu law reform societi homosexu will fli color stewart 1968\n\n

part world partnership right marriag extend samesex coupl advoc samesex marriag cite rang benefit deni peopl marri includ immigr health care inherit properti right famili oblig protect reason marriag extend samesex coupl oppon samesex marriag within gay co mmun argu fight achiev benefit mean extend marriag right samesex coupl privat benefit eg health care made avail peopl regardless relationship statu argu samesex marriag mo vement within gay commun discrimin famili compos three intim partner opposit samesex marriag movement within gay commun confus opposit outsid communitycit need\n\ncontemp orari lesbian gay commun grow complex place american western european media lesbian g ay men often portray inaccur televis film media gay commun often portray mani stereot yp gay men portray flamboy bold like minor group caricatur intend ridicul margin grou p23\n\ncurrent widespread ban refer childrel entertain refer occur almost invari gene r controversi 1997 american comedian ellen degener came closet popular sitcom mani sp onsor wendi fast food chain pull advertising24 also portion media attempt make gay co mmun includ publicli accept televis show grace queer eye straight guy increas public reflect come movement lgbt commun celebr came show develop 2004 show l word depict lg bt commun controversi benefici commun increas visibl lgbt peopl allow lgbt commun uni t organ demand chang also inspir mani lgbt peopl come out25\n\nunit state gay peopl f requent use symbol social decad celebr evangelist organ focu famili mani lgbt organ e xist repres defend gay commun exampl gay lesbian allianc defam unit state stonewal uk work media help portray fair accur imag gay community2627\n\ncompani advertis gay com mun lgbt activist use ad slogan promot gay commun view subaru market forest outback s logan choic way built later use eight us citi street gay right events28\n\nsocial med ia often use platform lgbt commun congreg share resourc search engin social network s ite provid numer opportun lgbt peopl connect one anoth addit play key role ident crea tion selfpresentation293031 social network site allow commun build well anonym allow peopl engag much littl would like32 varieti social media platform includ facebook tik tok tumblr twitter youtub differ associ audienc afford norms31 vari platform allow in clus member lgbt commun agenc decid engag selfpres themselves31 exist lgbt commun dis cours social media platform essenti disrupt reproduct hegemon cisheteronorm repres wi de varieti ident exist32\n\nban adult content 2018 tumblr platform uniqu suit share t ran stori build community33 mainstream social media platform like tiktok also benefic i tran commun creat space folk share resourc transit stori normal tran identity34 fou nd access lgbt content peer commun search engin social network site allow ident accep t pride within lgbt individuals35\n\nalgorithm evalu criteria control content recomme nd user search engin social network site30 reproduc stigmat discours domin within soc ieti result neg impact lgbt selfperception30 social media algorithm signific impact f ormat lgbt commun culture36 algorithm exclus occur exclusionari practic reinforc algo rithm across technolog landscap directli result exclud margin identities34 exclus ide nt represent caus ident insecur lgbt peopl perpetu cisheteronorm ident discourse34 lg bt user alli found method subvert algorithm may suppress content order continu build onlin communities34\n\naccord witeckcomb commun inc marketresearchcom 2006 buy power unit state gay lesbian approxim 660 billion expect exceed 835 billion 201137 gay cons um loyal specif brand wish support compani support gay commun also provid equal right lgbt worker uk buy power sometim abbrevi pink pound38\n\naccord articl jame hipp lgbt american like seek compani advertis will pay higher price premium product servic attr ibut median household incom compar samesex coupl oppositesex coupl studi show glbt am erican twice like graduat colleg twice like individu incom 60000 twice like household incom 250000 more39\n\nalthough mani claim lgbt commun affluent compar heterosexu con sum research proven false40 howev lgbt commun still import segment consum demograph s pend power loyalti brand have41 witeckcomb commun calcul adult lgbt buy power 830 bil lion 201340 samesex partner household spend slightli averag home given shop trip42 al so make shop trip compar nonlgbt households42 averag differ spend samesex partner hom e 25 percent higher averag unit state household42 accord univers maryland gay male pa rtner earn 10000 less averag compar heterosexu men40 howev partner lesbian receiv 700 0 year heterosexu marri women40 henc samesex partner heterosexu partner equal concern consum affluence40\n\nlgbt commun recogn one largest consum travel travel includ annu al trip sometim even multipl annual trip annual lgbt commun spend around 65 billion t ravel total 10 percent unit state travel market40 mani common travel factor play lgbt travel decis destin especi tailor lgbt commun like travel places40\n\nsurvey conduct 2012 younger american like identifi gay42 statist continu decreas age adult age 1829

three time like identifi lgbt senior older 6542 statist lgbt commun taken account dem ograph find trend pattern specif products40 consum identifi lgbt like regularli engag variou activ oppos identifi heterosexual40 accord commun market inc 90 percent lesbia n 88 percent gay men dine friend regularli similarli 31 percent lesbian 50 percent ga y men visit club bar40\n\nhome likelihood lgbt women children home nonlgbt women equa l42 howev lgbt men half like compar nonlgbt men children home42 household incom sixte en percent lgbt american rang 90000 per year comparison 21 percent overal adult popul ation42 howev key differ identifi lgbt fewer children collect comparison heterosexu p artners40 anoth factor hand lgbt popul color continu face incom barrier along rest ra ce issu expectedli earn less affluent predicted40\n\nanalysi gallup survey show detai l estim year 2012 2014 metropolitan area highest percentag lgbt commun san francisco california next highest portland oregon austin texas43\n\n2019 survey twospirit lgbtq popul canadian citi hamilton ontario call map void twospirit lgbtq experi hamilton sh ow 906 respond came sexual orient 489 identifi bisexualpansexu 216 identifi gay 183 i dentifi lesbian 49 identifi queer 63 identifi categori consist indic asexu heterosexu question gave respons sexual orientation44\n\n2019 survey tran nonbinari peopl canada call tran puls canada show 2873 respond came sexual orient 13 identifi asexu 28 ident ifi bisexu 13 identifi gay 15 identifi lesbian 31 identifi pansexu 8 identifi straigh t heterosexu 4 identifi twospirit 9 identifi unsur questioning45\n\nsurvey carri 2021 gallup found 71 us adult identifi lesbian gay bisexu transgend someth straight hetero sexual46\n\nmarket toward lgbt commun alway strategi among advertis last three four d ecad corpor america creat market nich lgbt commun three distinct phase defin market t urnov 1 shun 1980 2 curios fear 1990 3 pursuit 2000s47\n\nrecent market pick lgbt dem ograph spike samesex marriag 2014 market figur new way tie person sexual orient produ ct sold42 effort attract member lgbt commun product market research develop market me thod reach new families42 advertis histori shown market famili alway wife husband chi ldren42 today necessarili case could famili two father two mother one child six child ren break away tradit famili set market research notic need recogn differ famili conf igurations42\n\none area market subject fall stereotyp lgbt commun market toward comm un may corner target audienc altern lifestyl categori ultim other lgbt community42 se nsit import market toward commun market toward lgbt commun advertis respect boundari \n\nmarket also refer lgbt singl characterist make individual42 area target along lgb t segment race age cultur incom levels42 know consum give market power41\n\nalong att empt engag lgbt commun research found gender disagr among product respect consumers47 instanc gay male may want feminin product wherea lesbian femal may interest masculin product hold entir lgbt commun possibl differ far greater47 past gender seen fix cong ruent represent individu sex understood sex gender fluid separ research also note eva lu product person biolog sex equal determin selfconcept47 custom respons advertis dir ect toward gay men women like interest product41 import factor goal market indic futu r loyalti product brand\n\n2001 studi examin possibl root caus mental disord lesbian gay bisexu peopl cochran psychologist vicki may univers california explor whether ong o discrimin fuel anxieti depress stressrel mental health problem among lgb people48 a uthor found strong evid relationship two48 team compar 74 lgb 2844 heterosexu respond rate lifetim daili experi discrimin hire job deni bank loan well feel perceiv discrim ination48 lgb respond report higher rate perceiv discrimin heterosexu everi categori relat discrimin team found48 howev gay youth consid higher risk suicid literatur revi ew publish journal adolesc state gay inandofitself caus increas suicid rather review note find previou studi suggest thesuicid attempt significantli associ psychosoci str essor includ gender nonconform earli awar gay victim lack support school dropout fami li problem acquaint suicid attempt homeless substanc abus psychiatr disord stressor a lso experienc heterosexu adolesc shown preval among gay adolescents49 despit recent p rogress lgbt right gay men continu experi high rate loneli depress come out50\n\nlgbt multicultur divers within lgbt commun represent differ sexual orient gender identitie sa well differ ethnic languag religi group within lgbt commun time lgbt multicultur r elat may consid inclus lgbt commun larger multicultur model exampl universities51 mul ticultur model includ lgbt commun togeth equal represent larg minor group african ame rican unit statescit need\n\ntwo movement much common polit concern toler real differ divers minor statu invalid valu judgment appli differ way life5253\n\nresearch identi fi emerg gay lesbian commun sever progress time period across world includ renaiss en lighten modern westernization54 depend geograph locat commun experienc opposit exist

other nonetheless began permeat societi social politically54\n\nciti space earli mode rn europ host wealth gay activ howev scene remain semisecret long period time54 date back 1500 citi condit apprenticeship labor relat live arrang abund student artist act iv hegemon norm surround femal societ statu typic venic florenc italy54 circumst mani open mind young peopl attract citi settings54 consequ abund samesex interact began ta ke place54 mani connect form often led occurr casual romant sexual relationship preva l increas quit rapidli time point becam subcultur commun own54 literatur ballroom cul tur gradual made way onto scene becam integr despit transgress societ views54 perhap wellknown ball magicc amsterdam london also recogn lead locat lgbt commun establishme nt54 1950 urban space boom gay venu bar public sauna commun member could come togethe r54 pari london particularli attract lesbian popul platform social educ well54 urban occas import lgbt commun includ carniv rio de janeiro brazil mardi gra sydney austral ia well variou pride parad host bigger citi around world54\n\nway lgbt peopl use citi backdrop join social abl join forc polit well new sens collect provid somewhat safeti net individu voic demand equal rights55 unit state specif sever key polit event taken place urban context includ limit\n\nfollow event lgbt commun subcultur began grow sta bil nationwid phenomenon59 gay bar becam popular larg cities59 gay particularli incre as number cruis area public bath hous ymca urban space continu welcom experi liber wa y living59 lesbian led format literari societi privat social club samesex housing59 c ore communitybuild took place new york citi san francisco citi like st loui lafayett park wa chicago quickli follow suit59\n\nciti afford host prime condit allow better i ndividu develop well collect movement otherwis avail rural spaces55 first foremost ur ban landscap offer lgbt better prospect meet lgbt form network relationships55 one id eal platform within framework free labor market mani capitalist societi entic peopl b reak away often damag tradit nuclear famili order pursu employ bigger cities59 make m ove space afford new liberti realm sexual ident also kinship55 research describ phase resist confin expect normativity55 urban lgbt demonstr push back variou outlet includ style dress way talk carri chose build community55 social scienc perspect relationshi p citi lgbt commun oneway street lgbt give back much term econom contribut ie pink mo ney activ polit too54\n\ncompar white lgbt individu lgbt peopl color often experi pre judic stereotyp discrimin basi sexual orient gender ident also basi race60 nadal coll eagu discuss lgbtq peopl color experi intersect microaggress target variou aspect soc ial identities6061 neg experi microaggress come cisgend heterosexu white individu cis gend heterosexu individu race60 lgbt commun usual domin white people60\n\nlgbt peopl color feel comfort repres within lgbt spaces60 comprehens systemat review exist publi sh research literatur around experi lgbt individu color find common theme exclus larg white lgbt spaces60 space typic domin white lgbt individu promot white western valu o ften leav lgbt individu color feel though must choos racial commun gender sexual orie nt community60 gener western societi often subtli code gay white white lgbt folk ofte n seen face lgbt cultur values60\n\ntopic come reveal one sexual orient gender ident public associ white valu expect mainstream discussions60 white western cultur place v alu abil speak openli one ident famili one particular studi found lgbt particip color view famili silenc ident support accepting60 exampl collectivist cultur view come pro cess famili affair rather individu one furthermor annual nation come day center white perspect event meant help lgbt person feel liber comfort skin60 howev lgbt peopl colo r nation come day view neg light6062 commun color come publicli advers consequ risk p erson sens safeti well famili commun relationships60 white lgbt peopl tend collect re ject differ perspect come result possibl isol lgbt sibl color60']

TF-IDF Vectorization

Compressing the Code

Instead of having multiple lines of code to process all the text, we can use a single, compressed function to greatly reduce the complexity.

In [16]:
```
preprocessed_sentences = []
for sentence in sentences:
    preprocessed_sentence = preprocess_text(sentence)
    preprocessed_sentences.append(preprocessed_sentence)
```

In [17]:
```
print(sentences)
```

['', 'lgbt commun also known lgbtq commun lgbtqia commun gay commun queer commun loos defin group lesbian gay bisexu transgend individu unit common cultur social movement commun gener celebr pride divers individu sexualitynot verifi bodi lgbt activist soci ologist see lgbt communitybuild counterweight heterosex homophobia biphobia transphob ia sexual conformist pressur exist larger societi term pride sometim gay pride expres s lgbt commun ident collect strength pride parad provid prime exampl use demonstr gen er mean termnot verifi bodi lgbt commun divers polit affili peopl lesbian gay bisexu transgend consid part lgbt commun', 'group may consid part lgbt commun includ gay vil lag lgbt right organ lgbt employe group compani lgbt student group school univers lgb taffirm religi group', 'lgbt commun may organ support movement civil right promot lgb t right variou place around world time highprofil celebr broader societi may offer st rong support organ certain locat exampl lgbt advoc entertain madonna state ask perfor m mani pride event around world would never ever turn new york city4', 'lgbt glbt ini ti stand lesbian gay bisexu transgend use sinc 1990 term adapt initi lgb use replac t erm gay refer commun whole begin variou form larg earli 1990s5citat need', 'movement alway includ lgbt peopl oneword unifi term 1950 earli 1980 gay see gay liber througho ut 1970 80 number group lesbian member profeminist polit prefer repres lesbian gay6 e arli nineti group shift name base lesbian gay bisexu transgend lgbt queer also increa singli reclaim oneword altern everlengthen string initi especi use radic polit group use queer sinc 80s6', 'initi well common variant lgbtq adopt mainstream 1990s7 umbrel la term use label topic sexual gender ident exampl lgbt movement advanc project term commun center servic specif member lgbt commun lgbt commun center comprehens studi ce nter around unit states8', 'initi lgbt intend emphas divers sexual gender identitybas cultur may refer anyon nonheterosexu noncisgend instead exclus peopl lesbian gay bise xu transgender9 recogn inclus popular variant add letter q identifi queer question se xual ident lgbtq record sinc 19961011', 'disagr precis word best still present 2023 p ropos ad letter make particip group explicit12 detractor approach argu ad letter impl icitli exclud other make wors brandingcit need', 'gay commun frequent associ certain symbol especi rainbow rainbow flag greek lambda symbol l liber triangl ribbon gender symbol also use gay accept symbol mani type flag repres subdivis gay commun commonli recogn one rainbow flag accord gilbert baker creator commonli known rainbow flag colo r repres valu commun', 'later pink indigo remov flag result presentday flag first pre sent 1979 pride parad flag includ victori aid flag leather pride flag bear pride flag 13', 'lambda symbol origin adopt gay activist allianc new york 1970 broke away larger gay liber front lambda chosen peopl might confus colleg symbol recogn gay commun symb ol unless one actual involv commun back decemb 1974 lambda offici declar intern symbo l gay lesbian right intern gay right congress edinburgh scotland13', 'triangl becam s ymbol gay commun holocaust repres jew homosexu kill german law holocaust homosexu lab el pink triangl distinguish jew regular prison polit prison black triangl similarli s ymbol femal repres lesbian sisterhood', 'pink yellow triangl use label jewish homosex u gender symbol much longer list variat homosexu bisexu relationship clearli recogniz may popularli seen symbol symbol relat gay commun gay pride includ gayteen suicid awa r ribbon aid awar ribbon labri purpl rhinoceros1415', 'fall 1995 human right campaign adopt logo yellow equal sign deep blue squar becom one recogniz symbol lesbian gay bi sexu transgend commun logo spot world becom synonym fight equal right lgbt people16', 'one notabl recent chang made philadelphia pennsylvania june 8 2017 ad two new stripe rainbow flag one black one brown intend highlight member color within lgbt community1 7', 'lgbt commun repres social compon global commun believ mani includ heterosexu all i underrepres area civil right current struggl gay commun larg brought global unit st ate world war ii brought togeth mani closet rural men around nation expos progress at titud part europ upon return home war mani men decid band togeth citi rather return s mall town fledgl commun would soon becom polit begin gay right movement includ monume nt incid place like stonewal today mani larg citi gay lesbian commun center mani univ ers colleg across world support center lgbt student human right campaign18 lambda leg al empow spirit foundation19 glaad20 advoc lgbt peopl wide rang issu unit state also intern lesbian gay associ 1947 unit kingdom adopt univers declar human right udhr lgb t activist clung concept equal inalien right peopl regardless race gender sexual orie nt declar specif mention gay right discuss equal freedom discrimination21 1962 clark polak join janu societi philadelphia pennsylvania22 year becam presid 1968 announc so cieti would chang name homosexu law reform societi homosexu will fli color stewart 19

68', 'part world partnership right marriag extend samesex coupl advoc samesex marriag cite rang benefit deni peopl marri includ immigr health care inherit properti right f amili oblig protect reason marriag extend samesex coupl oppon samesex marriag within gay commun argu fight achiev benefit mean extend marriag right samesex coupl privat b enefit eg health care made avail peopl regardless relationship statu argu samesex mar riag movement within gay commun discrimin famili compos three intim partner opposit s amesex marriag movement within gay commun confus opposit outsid communitycit need', 'contemporari lesbian gay commun grow complex place american western european media l esbian gay men often portray inaccur televis film media gay commun often portray mani stereotyp gay men portray flamboy bold like minor group caricatur intend ridicul marg in group23', 'current widespread ban refer childrel entertain refer occur almost inva ri gener controversi 1997 american comedian ellen degener came closet popular sitcom mani sponsor wendi fast food chain pull advertising24 also portion media attempt make gay commun includ publicli accept televis show grace queer eye straight guy increas p ublic reflect come movement lgbt commun celebr came show develop 2004 show l word dep ict lgbt commun controversi benefici commun increas visibl lgbt peopl allow lgbt comm un unit organ demand chang also inspir mani lgbt peopl come out25', 'unit state gay p eopl frequent use symbol social decad celebr evangelist organ focu famili mani lgbt o rgan exist repres defend gay commun exampl gay lesbian allianc defam unit state stone wal uk work media help portray fair accur imag gay community2627', 'compani advertis gay commun lgbt activist use ad slogan promot gay commun view subaru market forest ou tback slogan choic way built later use eight us citi street gay right events28', 'soc ial media often use platform lgbt commun congreg share resourc search engin social ne twork site provid numer opportun lgbt peopl connect one anoth addit play key role ide nt creation selfpresentation293031 social network site allow commun build well anonym allow peopl engag much littl would like32 varieti social media platform includ facebo ok tiktok tumblr twitter youtub differ associ audienc afford norms31 vari platform al low inclus member lgbt commun agenc decid engag selfpres themselves31 exist lgbt comm un discours social media platform essenti disrupt reproduct hegemon cisheteronorm rep res wide varieti ident exist32', 'ban adult content 2018 tumblr platform uniqu suit s hare tran stori build community33 mainstream social media platform like tiktok also b enefici tran commun creat space folk share resourc transit stori normal tran identity 34 found access lgbt content peer commun search engin social network site allow ident accept pride within lgbt individuals35', 'algorithm evalu criteria control content re commend user search engin social network site30 reproduc stigmat discours domin withi n societi result neg impact lgbt selfperception30 social media algorithm signific imp act format lgbt commun culture36 algorithm exclus occur exclusionari practic reinforc algorithm across technolog landscap directli result exclud margin identities34 exclus ident represent caus ident insecur lgbt peopl perpetu cisheteronorm ident discourse34 lgbt user alli found method subvert algorithm may suppress content order continu buil d onlin communities34', 'accord witeckcomb commun inc marketresearchcom 2006 buy powe r unit state gay lesbian approxim 660 billion expect exceed 835 billion 201137 gay co nsum loyal specif brand wish support compani support gay commun also provid equal rig ht lgbt worker uk buy power sometim abbrevi pink pound38', 'accord articl jame hipp l gbt american like seek compani advertis will pay higher price premium product servic attribut median household incom compar samesex coupl oppositesex coupl studi show glb t american twice like graduat colleg twice like individu incom 60000 twice like house hold incom 250000 more39', 'although mani claim lgbt commun affluent compar heterosex u consum research proven false40 howev lgbt commun still import segment consum demogr aph spend power loyalti brand have41 witeckcomb commun calcul adult lgbt buy power 83 0 billion 201340 samesex partner household spend slightli averag home given shop trip 42 also make shop trip compar nonlgbt households42 averag differ spend samesex partne r home 25 percent higher averag unit state household42 accord univers maryland gay ma le partner earn 10000 less averag compar heterosexu men40 howev partner lesbian recei v 7000 year heterosexu marri women40 henc samesex partner heterosexu partner equal co ncern consum affluence40', 'lgbt commun recogn one largest consum travel travel inclu d annual trip sometim even multipl annual trip annual lgbt commun spend around 65 bil lion travel total 10 percent unit state travel market40 mani common travel factor pla y lgbt travel decis destin especi tailor lgbt commun like travel places40', 'survey c onduct 2012 younger american like identifi gay42 statist continu decreas age adult ag

```
e 1829 three time like identifi lgbt senior older 6542 statist lgbt commun taken acco
unt demograph find trend pattern specif products40 consum identifi lgbt like regularl
i engag variou activ oppos identifi heterosexual40 accord commun market inc 90 percen
t lesbian 88 percent gay men dine friend regularli similarli 31 percent lesbian 50 pe
rcent gay men visit club bar40', 'home likelihood lgbt women children home nonlgbt wo
men equal42 howev lgbt men half like compar nonlgbt men children home42 household inc
om sixteen percent lgbt american rang 90000 per year comparison 21 percent overal adu
lt population42 howev key differ identifi lgbt fewer children collect comparison hete
rosexu partners40 anoth factor hand lgbt popul color continu face incom barrier along
rest race issu expectedli earn less affluent predicted40', 'analysi gallup survey sho
w detail estim year 2012 2014 metropolitan area highest percentag lgbt commun san fra
ncisco california next highest portland oregon austin texas43', '2019 survey twospiri
t lgbtq popul canadian citi hamilton ontario call map void twospirit lgbtq experi ham
ilton show 906 respond came sexual orient 489 identifi bisexualpansexu 216 identifi g
ay 183 identifi lesbian 49 identifi queer 63 identifi categori consist indic asexu he
terosexu question gave respons sexual orientation44', '2019 survey tran nonbinari peo
pl canada call tran puls canada show 2873 respond came sexual orient 13 identifi asex
u 28 identifi bisexu 13 identifi gay 15 identifi lesbian 31 identifi pansexu 8 identi
fi straight heterosexu 4 identifi twospirit 9 identifi unsur questioning45', 'survey
carri 2021 gallup found 71 us adult identifi lesbian gay bisexu transgend someth stra
ight heterosexual46', 'market toward lgbt commun alway strategi among advertis last t
hree four decad corpor america creat market nich lgbt commun three distinct phase def
in market turnov 1 shun 1980 2 curios fear 1990 3 pursuit 2000s47', 'recent market pi
ck lgbt demograph spike samesex marriag 2014 market figur new way tie person sexual o
rient product sold42 effort attract member lgbt commun product market research develo
p market method reach new families42 advertis histori shown market famili alway wife
husband children42 today necessarili case could famili two father two mother one chil
d six children break away tradit famili set market research notic need recogn differ
famili configurations42', 'one area market subject fall stereotyp lgbt commun market
toward commun may corner target audienc altern lifestyl categori ultim other lgbt com
munity42 sensit import market toward commun market toward lgbt commun advertis respec
t boundari', 'market also refer lgbt singl characterist make individual42 area target
along lgbt segment race age cultur incom levels42 know consum give market power41',
'along attempt engag lgbt commun research found gender disagr among product respect c
onsumers47 instanc gay male may want feminin product wherea lesbian femal may interes
t masculin product hold entir lgbt commun possibl differ far greater47 past gender se
en fix congruent represent individu sex understood sex gender fluid separ research al
so note evalu product person biolog sex equal determin selfconcept47 custom respons a
dvertis direct toward gay men women like interest product41 import factor goal market
indic futur loyalti product brand', '2001 studi examin possibl root caus mental disor
d lesbian gay bisexu peopl cochran psychologist vicki may univers california explor w
hether ongo discrimin fuel anxieti depress stressrel mental health problem among lgb
people48 author found strong evid relationship two48 team compar 74 lgb 2844 heterose
xu respond rate lifetim daili experi discrimin hire job deni bank loan well feel perc
eiv discrimination48 lgb respond report higher rate perceiv discrimin heterosexu ever
i categori relat discrimin team found48 howev gay youth consid higher risk suicid lit
eratur review publish journal adolesc state gay inandofitself caus increas suicid rat
her review note find previou studi suggest thesuicid attempt significantli associ psy
chosoci stressor includ gender nonconform earli awar gay victim lack support school d
ropout famili problem acquaint suicid attempt homeless substanc abus psychiatr disord
stressor also experienc heterosexu adolesc shown preval among gay adolescents49 despi
t recent progress lgbt right gay men continu experi high rate loneli depress come out
50', 'lgbt multicultur divers within lgbt commun represent differ sexual orient gende
r identitiesa well differ ethnic languag religi group within lgbt commun time lgbt mu
lticultur relat may consid inclus lgbt commun larger multicultur model exampl univers
ities51 multicultur model includ lgbt commun togeth equal represent larg minor group
african american unit statescit need', 'two movement much common polit concern toler
real differ divers minor statu invalid valu judgment appli differ way life5253', 'res
earch identifi emerg gay lesbian commun sever progress time period across world inclu
d renaiss enlighten modern westernization54 depend geograph locat commun experienc op
```

```
posit exist other nonetheless began permeat societi social politically54', 'citi spac
e earli modern europ host wealth gay activ howev scene remain semisecret long period
time54 date back 1500 citi condit apprenticeship labor relat live arrang abund studen
t artist activ hegemon norm surround femal societ statu typic venic florenc italy54 c
ircumst mani open mind young peopl attract citi settings54 consequ abund samesex inte
ract began take place54 mani connect form often led occurr casual romant sexual relat
ionship preval increas quit rapidli time point becam subcultur commun own54 literatur
ballroom cultur gradual made way onto scene becam integr despit transgress societ vie
ws54 perhap wellknown ball magicc amsterdam london also recogn lead locat lgbt commun
establishment54 1950 urban space boom gay venu bar public sauna commun member could c
ome together54 pari london particularli attract lesbian popul platform social educ we
ll54 urban occas import lgbt commun includ carniv rio de janeiro brazil mardi gra syd
ney australia well variou pride parad host bigger citi around world54', 'way lgbt peo
pl use citi backdrop join social abl join forc polit well new sens collect provid som
ewhat safeti net individu voic demand equal rights55 unit state specif sever key poli
t event taken place urban context includ limit', 'follow event lgbt commun subcultur
began grow stabil nationwid phenomenon59 gay bar becam popular larg cities59 gay part
icularli increas number cruis area public bath hous ymca urban space continu welcom e
xperi liber way living59 lesbian led format literari societi privat social club sames
ex housing59 core communitybuild took place new york citi san francisco citi like st
loui lafayett park wa chicago quickli follow suit59', 'citi afford host prime condit
allow better individu develop well collect movement otherwis avail rural spaces55 fir
st foremost urban landscap offer lgbt better prospect meet lgbt form network relation
ships55 one ideal platform within framework free labor market mani capitalist societi
entic peopl break away often damag tradit nuclear famili order pursu employ bigger ci
ties59 make move space afford new liberti realm sexual ident also kinship55 research
describ phase resist confin expect normativity55 urban lgbt demonstr push back variou
outlet includ style dress way talk carri chose build community55 social scienc perspe
ct relationship citi lgbt commun oneway street lgbt give back much term econom contri
but ie pink money activ polit too54', 'compar white lgbt individu lgbt peopl color of
ten experi prejudic stereotyp discrimin basi sexual orient gender ident also basi rac
e60 nadal colleagu discuss lgbtq peopl color experi intersect microaggress target var
iou aspect social identities6061 neg experi microaggress come cisgend heterosexu whit
e individu cisgend heterosexu individu race60 lgbt commun usual domin white people6
0', 'lgbt peopl color feel comfort repres within lgbt spaces60 comprehens systemat re
view exist publish research literatur around experi lgbt individu color find common t
heme exclus larg white lgbt spaces60 space typic domin white lgbt individu promot whi
te western valu often leav lgbt individu color feel though must choos racial commun g
ender sexual orient community60 gener western societi often subtli code gay white whi
te lgbt folk often seen face lgbt cultur values60', 'topic come reveal one sexual ori
ent gender ident public associ white valu expect mainstream discussions60 white weste
rn cultur place valu abil speak openli one ident famili one particular studi found lg
bt particip color view famili silenc ident support accepting60 exampl collectivist cu
ltur view come process famili affair rather individu one furthermor annual nation com
e day center white perspect event meant help lgbt person feel liber comfort skin60 ho
wev lgbt peopl color nation come day view neg light6062 commun color come publicli ad
vers consequ risk person sens safeti well famili commun relationships60 white lgbt pe
opl tend collect reject differ perspect come result possibl isol lgbt sibl color60']
```

TF-IDF Vectorization

In [18]:

```
#Instantiate the vectorizer
tfidf = TfidfVectorizer()

#Feed the preprocessed sentences into the vectorizer
tfidf_matrix = tfidf.fit_transform(sentences)
tfidf_scores = tfidf_matrix.sum(axis=1)
```

In [19]:
```
#Get the highest scoring sentences based on the vectorization
#The negative index number is how many sentences you want
#This can be adjusted, depending on your needs
top_sentence_indices = np.argsort(tfidf_scores, axis=0)[-10:]
```

In [20]:
```
#Display the top sentences
top_sentence_indices
```

Out[20]:
```
matrix([[ 1],
        [39],
        [22],
        [50],
        [46],
        [19],
        [47],
        [40],
        [16],
        [44]], dtype=int64)
```

In [21]:
```
#Selecting a single sentence from the matrix
#Note: The index is the index of the matrix, not the value within the matrix
sentences[3]
```

Out[21]:
```
'lgbt commun may organ support movement civil right promot lgbt right variou place ar
ound world time highprofil celebr broader societi may offer strong support organ cert
ain locat exampl lgbt advoc entertain madonna state ask perform mani pride event arou
nd world would never ever turn new york city4'
```

In [22]:
```
#Create the summary
top_sentences = []

for index in range(len(top_sentence_indices)):
    val = sentences[index]
    #print(index, val)
    top_sentences.append(sentence[index])

    summary = "".join(val) + "."

    print(summary)
```

```
.
lgbt commun also known lgbtq commun lgbtqia commun gay commun queer commun loos defin
group lesbian gay bisexu transgend individu unit common cultur social movement commun
gener celebr pride divers individu sexualitynot verifi bodi lgbt activist sociologist
see lgbt communitybuild counterweight heterosex homophobia biphobia transphobia sexua
l conformist pressur exist larger societi term pride sometim gay pride express lgbt c
ommun ident collect strength pride parad provid prime exampl use demonstr gener mean
termnot verifi bodi lgbt commun divers polit affili peopl lesbian gay bisexu transgen
d consid part lgbt commun.
group may consid part lgbt commun includ gay villag lgbt right organ lgbt employe gro
up compani lgbt student group school univers lgbtaffirm religi group.
lgbt commun may organ support movement civil right promot lgbt right variou place aro
und world time highprofil celebr broader societi may offer strong support organ certa
in locat exampl lgbt advoc entertain madonna state ask perform mani pride event aroun
d world would never ever turn new york city4.
lgbt glbt initi stand lesbian gay bisexu transgend use sinc 1990 term adapt initi lgb
use replac term gay refer commun whole begin variou form larg earli 1990s5citat need.
movement alway includ lgbt peopl oneword unifi term 1950 earli 1980 gay see gay liber
throughout 1970 80 number group lesbian member profeminist polit prefer repres lesbia
n gay6 earli nineti group shift name base lesbian gay bisexu transgend lgbt queer als
o increasingli reclaim oneword altern everlengthen string initi especi use radic poli
t group use queer sinc 80s6.
initi well common variant lgbtq adopt mainstream 1990s7 umbrella term use label topic
sexual gender ident exampl lgbt movement advanc project term commun center servic spe
cif member lgbt commun lgbt commun center comprehens studi center around unit states
8.
initi lgbt intend emphas divers sexual gender identitybas cultur may refer anyon nonh
eterosexu noncisgend instead exclus peopl lesbian gay bisexu transgender9 recogn incl
us popular variant add letter q identifi queer question sexual ident lgbtq record sin
c 19961011.
disagr precis word best still present 2023 propos ad letter make particip group expli
cit12 detractor approach argu ad letter implicitli exclud other make wors brandingcit
need.
gay commun frequent associ certain symbol especi rainbow rainbow flag greek lambda sy
mbol l liber triangl ribbon gender symbol also use gay accept symbol mani type flag r
epres subdivis gay commun commonli recogn one rainbow flag accord gilbert baker creat
or commonli known rainbow flag color repres valu commun.
```

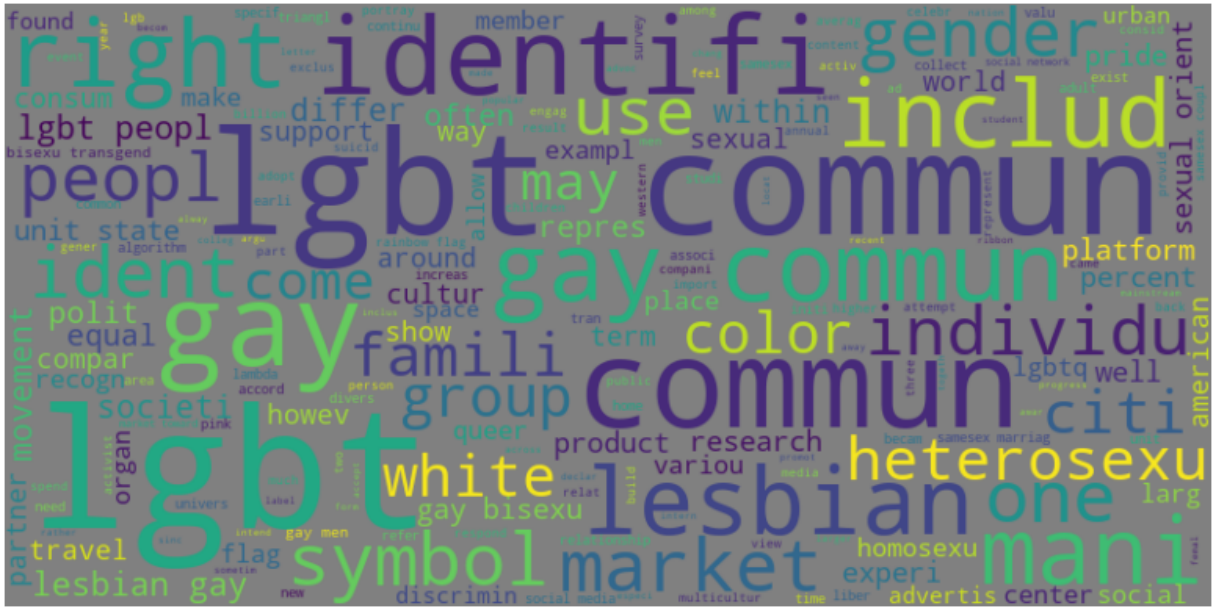
In [23]:
```
#Write the summary to a text file
#This is more for future versions
directory = "./data/"
content_summary = summary

write_text_to_file(directory, title, content_summary, "summary")
```

Visualization

In [24]:
```
#Creating a word cloud of the most important words
wordcloud = WordCloud(width=800, height=400, background_color="grey").generate(process

plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



Summary and Conclusion

Summary

This project used web scraping to pull the content from a provided Wikipedia page, then processed it down into a summary. It used various text processing techniques to prepare the text for Text Vectorization and Term Frequency-Inverse Document Frequency analysis. The TF-IDF vectors were used to select the most important sentences from the text, then combine them into a summary.

Conclusion

Unfortunately, my code was unable to construct a proper summary of the page, instead only using the stemmed words in the summary. This does cause some issues in the functionality of the project as a whole. However, the link below is for a thread on Stack Overflow, and has some interesting ideas on "undoing" stemming. I would very much like to look more into this, and possibly apply it to a future version of the project.

Reversing Stemming