In [11]:
```
import requests
from bs4 import BeautifulSoup
from urllib import request
import nltk
import os
import argparse
import sys
from os import listdir
from os.path import isfile, join
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np
import matplotlib.pyplot as plt
from gensim.models import doc2vec
import gensim
from collections import namedtuple
import smart_open
import warnings
warnings.filterwarnings("ignore")
```

In [ ]:
```
#We first scrappe Gutenberg 100 most popular books page.
#https://katherinepully.com/project-gutenberg-scraper/

session = requests.Session()
books_directory="/Users/benjaminfell/gutenberg_data/text/"
gutenberg_url = "https://www.gutenberg.org/"

def get_links(url):
        r = session.get(url)
        soup = BeautifulSoup(r.content, features="lxml")
        data_links = soup.findAll("a", {"class": "link"})
        data_link = gutenberg_url + get_text_or_html_link(data_links)
        return data_link

def get_text_or_html_link(links):
        plaintext_link, html_link = "",""
        for link in links:
            if "Plain Text" in str(link):
                plaintext_link=link["href"]
            elif "HTML" in str(link):
                html_link=link["href"]

        data_link=""
        if (plaintext_link):
            data_link=plaintext_link
        elif (html_link):
            data_link=html_link
        return data_link

def download_book(booktitle,data_link):
        print("Downloading " % (booktitle))
        r=session.get(data_link)
        filename=books_directory+booktitle+".txt"
        file = open(filename, "w")
        file.write(r.text)
        file.close()
```

In [ ]:

```
r = session.get(URL)
soup = BeautifulSoup(r.content, features="lxml")
books = soup.findAll("li")

books_links = {}

for book in books[:100]:
    title = book.text
    link = book.find("a")["href"]
    if "ebooks" in str(book.find("a")["href"]):
        books_links[title]=link

for book in books_links.items():
    title=book[0]
    book_download_link=book[1]
    url = gutenberg_url + book_download_link[1:]
    data_link=get_links(url)
    download_book(title, data_link)
```

In [2]:

```
#Since the genre (Subject) directories are empty on the _Gutenberg project, we w
#We only care about the most representented genres so we don't annotate very spe

books_path = "/Users/benjaminfell/gutenberg_data/text/"
onlyfiles = [f for f in listdir(books_path) if isfile(join(books_path, f)) and f

books_genres = {}

books_genres["The Republic by Plato (241)"]=["Classical literature","Political s
books_genres["Moby Dick; Or, The Whale by Herman Melville (769)"]=["Psychologica
books_genres["Grimms' Fairy Tales by Jacob Grimm and Wilhelm Grimm (451)"]=["Fai
books_genres["Peter Pan by J. M.  Barrie (284)"]=["Fantasy literature","Fiction"
books_genres["The Hound of the Baskervilles by Arthur Conan Doyle (243)"]=["Dete
books_genres["The Scarlet Letter by Nathaniel Hawthorne (891)"]=["Adultery -- Fi
books_genres["The Awakening, and Selected Short Stories by Kate Chopin (306)"]=[
books_genres["The Strange Case of Dr. Jekyll and Mr. Hyde by Robert Louis Steven
books_genres["The Wonderful Wizard of Oz by L. Frank  Baum (246)"]=["Fantasy lit
books_genres["The Youngest Girl in the School by Evelyn Sharp (433)"]=["Juvenile
books_genres["Second Treatise of Government by John Locke (216)"]=["Political sc
books_genres["Pride and Prejudice by Jane Austen (1715)"]=["Young women -- Ficti
books_genres["Treasure Island by Robert Louis Stevenson (219)"]=["Sea stories","
books_genres["The Iliad by Homer (389)"]=["Epic poetry"]
books_genres["A Christmas Carol in Prose; Being a Ghost Story of Christmas by Ch
books_genres["The Translations of Beowulf: A Critical Bibliography by Chauncey B
books_genres["A Gamekeeper's Note-book by Owen Jones and Marcus Woodward (744)"]
books_genres["The Importance of Being Earnest: A Trivial Comedy for Serious Peop
books_genres["The Picture of Dorian Gray by Oscar Wilde (601)"]=["Didactic ficti
books_genres["Little Women by Louisa May Alcott (361)"]=["Young women -- Fiction
books_genres["Anne of Green Gables by L. M.  Montgomery (278)"]=["Orphans -- Fic
books_genres["Metamorphosis by Franz Kafka (423)"]=["Psychological fiction","Met
books_genres["The Prince by Niccolò Machiavelli (493)"]=["State, The -- Early wo
books_genres["The Adventures of Sherlock Holmes by Arthur Conan Doyle (669)"]=["
books_genres["Crime and Punishment by Fyodor Dostoyevsky (464)"]=["Detective and
books_genres["The Souls of Black Folk by W. E. B.  Du Bois (297)"]=["African Ame
books_genres["Joseph and his Brethren by W. K.  Tweedie (480)"]=[]
books_genres["Adventures of Huckleberry Finn by Mark Twain (355)"]=["Humorous st
books_genres["Dubliners by James Joyce (341)"]=["Short stories","Dublin (Ireland
books_genres["A Christmas Carol by Charles Dickens (263)"]=["Christmas stories",
```

```
books_genres["The Philippines a Century Hence by José Rizal (306)"]=["Philippine
books_genres["The Brothers Karamazov by Fyodor Dostoyevsky (235)"]=["Didactic fi
books_genres["The Romance of Lust: A classic Victorian erotic novel by Anonymous
books_genres["A Pickle for the Knowing Ones by Timothy Dexter (486)"]=["Biograph
books_genres["Märchen der Gebrüder Grimm 2 by Jacob Grimm and Wilhelm Grimm (220
books_genres["Anna Karenina by graf Leo Tolstoy (238)"]=["Adultery -- Fiction","
books_genres["Wuthering Heights by Emily Brontë (268)"]=["Psychological fiction"
books_genres["Ulysses by James Joyce (390)"]=["Psychological fiction","Domestic
books_genres["Tractatus Logico-Philosophicus by Ludwig Wittgenstein (320)"]=["Ph
books_genres["The American Diary of a Japanese Girl by Yoné Noguchi (623)"]=["Ja
books_genres["A Modest Proposal by Jonathan Swift (607)"]=["Political satire, En
books_genres["Heart of Darkness by Joseph Conrad (508)"]=["Psychological fiction
books_genres["A Warning to the Curious and Other Ghost Stories by M. R.  James (
books_genres["A Tale of Two Cities by Charles Dickens (556)"]=["War stories","Hi
books_genres["The Odyssey by Homer (248)"]=["Epic poetry"]
books_genres["Alice's Adventures in Wonderland by Lewis Carroll (748)"]=["Fantas
books_genres["Japanese Girls and Women by Alice Mabel Bacon (404)"]=["Women -- J
books_genres["Dracula by Bram Stoker (620)"]=["Horror tales","Gothic fiction","V
books_genres["War and Peace by graf Leo Tolstoy (412)"]=["Historical fiction","W
books_genres["Beyond Good and Evil by Friedrich Wilhelm Nietzsche (275)"]=["Ethi
books_genres["Les Misérables by Victor Hugo (217)"]=["Historical fiction","Epic
books_genres["Essays of Michel de Montaigne — Complete by Michel de Montaigne (2
books_genres["Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft She
books_genres["The Kama Sutra of Vatsyayana by Vatsyayana (235)"]=["Love","Sex"]
books_genres["Walden, and On The Duty Of Civil Disobedience by Henry David Thore
books_genres["The Great Gatsby by F. Scott  Fitzgerald (510)"]=["Psychological f
books_genres["Kwaidan: Stories and Studies of Strange Things by Lafcadio Hearn (
books_genres["The Prophet by Kahlil Gibran (282)"]=["Prose poems, American","Mys
books_genres["A Doll's House : a play by Henrik Ibsen (599)"]=["Drama"]
books_genres["Great Expectations by Charles Dickens (446)"]=["Orphans -- Fiction
books_genres["Narrative of the Life of Frederick Douglass, an American Slave by
books_genres["The Yellow Wallpaper by Charlotte Perkins Gilman (503)"]=["Psychol
books_genres["Siddhartha by Hermann Hesse (299)"]=["Spiritual life -- Fiction"]
books_genres["The Count of Monte Cristo, Illustrated by Alexandre Dumas (333)"]=
books_genres["The Adventures of Tom Sawyer, Complete by Mark Twain (238)"]=["Hum
books_genres["Anthem by Ayn Rand (218)"]=["Science fiction","Psychological ficti
books_genres["Don Quixote by Miguel de Cervantes Saavedra (249)"]=["Spain -- Soc
books_genres["Summer by Romain Rolland (361)"]=[]
books_genres["Jane Eyre: An Autobiography by Charlotte Brontë (500)"]=["Orphans
```

In [3]:

```
#Let's see what genres are more represented in the 100 most popular books:
genres_dicc = {}
for k,v in books_genres.items():
    for genres in v:
        if genres in genres_dicc.keys():
            genres_dicc[genres] += 1
        else:
            genres_dicc[genres] = 1
```

In [4]:

```
for w in sorted(genres_dicc, key=genres_dicc.get, reverse=True):
    print(w, genres_dicc[w])
```

```
Psychological fiction 11
England -- Fiction 6
Bildungsromans 6
Historical fiction 5
Love stories 5
```

```
Adventure stories 4
Domestic fiction 4
Biography 4
Orphans -- Fiction 4
Pirates -- Fiction 3
Adultery -- Fiction 3
Science fiction 3
Young women -- Fiction 3
Epic poetry 3
Ghost stories 3
Didactic fiction 3
Philosophy 3
Political science -- Early works to 1800 2
Sea stories 2
Fairy tales -- Germany 2
Fantasy literature 2
Fiction 2
Detective and mystery stories 2
Horror tales 2
Juvenile fiction 2
Christmas stories 2
Drama 2
Humorous stories 2
Boys -- Fiction 2
Epic literature 2
War stories 2
Revenge -- Fiction 2
Classical literature 1
Whaling -- Fiction 1
Ship captains -- Fiction 1
Fairies -- Fiction 1
Dogs -- Fiction 1
Physicians -- Fiction 1
Comedies 1
Supernatural -- Fiction 1
Islands -- Fiction 1
Friendship -- Fiction 1
Metamorphosis -- Fiction 1
State, The -- Early works to 1800 1
Detective and mystery stories, English 1
Murder -- Fiction 1
African Americans 1
Short stories 1
Dublin (Ireland) -- Fiction 1
Philippines 1
Fathers and sons -- Fiction 1
Brothers -- Fiction 1
Incest -- Fiction 1
Erotic stories 1
Married people -- Fiction 1
Male friendship -- Fiction 1
Japan -- Fiction 1
Political satire, English 1
Religious satire, English 1
Ireland -- Politics and government -- 18th century -- Humor 1
Imperialism -- Fiction 1
Fantasy fiction 1
Imaginary places -- Juvenile fiction 1
Children's stories 1
Women -- Japan -- Social conditions 1
```

```
Gothic fiction 1
Vampires -- Fiction 1
Aristocracy (Social class) -- Russia -- Fiction 1
Ethics 1
Horror Tales 1
Gothic Fiction 1
Scientists -- Fiction 1
Love 1
Sex 1
Solitude 1
Natural history -- Massachusetts -- Walden Woods 1
Married women -- Fiction 1
First loves -- Fiction 1
Paranormal fiction 1
Japan -- Social life and customs -- Fiction 1
Prose poems, American 1
Mysticism -- Poetry 1
Married women -- Psychology -- Fiction 1
Spiritual life -- Fiction 1
Men -- Psychology -- Fiction 1
Spain -- Social life and customs -- 16th century -- Fiction 1
```

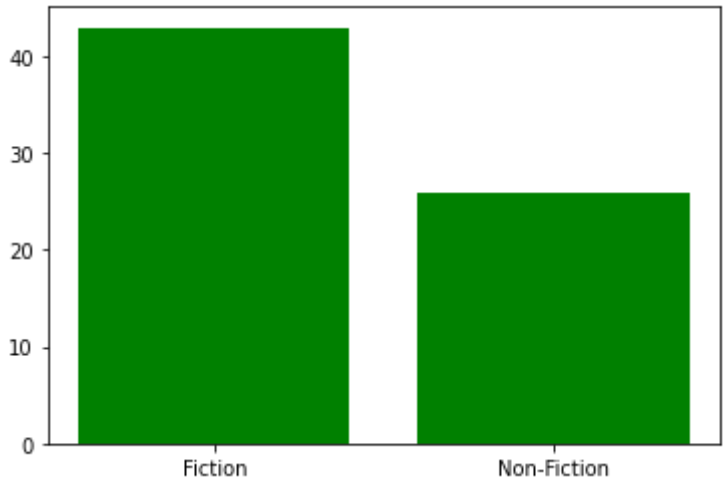
In [5]:

```
#Let's see how many books are Fiction related vs non Fiction
fiction_dicc = {}
fiction_dicc["Fiction"] = 0
fiction_dicc["Non-Fiction"] = 0
for k,v in books_genres.items():
    isFiction = False
    for genres in v:
        if "fiction" in genres.lower():
            isFiction = True

    if isFiction:
        fiction_dicc["Fiction"] += 1
    else:
        fiction_dicc["Non-Fiction"] += 1


print("#Fiction Books: ",fiction_dicc["Fiction"])
print("#Non-Fiction Books: ",fiction_dicc["Non-Fiction"])

limit = 20

plt.bar(fiction_dicc.keys(), fiction_dicc.values(), color='g')
```

```
#Fiction Books:  43
#Non-Fiction Books:  26
```

Out[5]:

```
<BarContainer object of 2 artists>
```



In [6]:
```
#Method for reading a file into gensim model, with a tag
def read_corpus(fname, tag):
    with smart_open.open(fname, encoding="utf-8") as f:
        for i, line in enumerate(f):
            tokens = gensim.utils.simple_preprocess(line)
            yield gensim.models.doc2vec.TaggedDocument(tokens, tag)
```

In [7]:
```
#We now create a corpus for each fiction and non-fiction genres.
#We use the first 26 Fiction books from Fiction to address for class imbalance.

train_corpus = []
test_corpus = []

counter_nf = 0
counter_f = 0

for k,v in books_genres.items():
    isFiction = False
    for genres in v:
        if "fiction" in genres.lower():
            isFiction = True

    if isFiction and counter_f <= limit:
        train_corpus.extend(list(read_corpus(books_path+k+".txt", [1])))
        print("Processed into Fiction: ",k," Genres: ",v)
        counter_f += 1

    elif not isFiction and counter_f <= limit:
        train_corpus.extend(list(read_corpus(books_path+k+".txt", [0])))
        print("Processed into Non-Fiction: ",k," Genres: ",v)
        counter_nf += 1

    else:
        train_corpus.extend(list(read_corpus(books_path+k+".txt", ["{}".format(k
        print("Processed TEST: ",k," Genres: ",v)
```

```
Processed into Non-Fiction:  The Republic by Plato (241)  Genres:  ['Classical l
iterature', 'Political science -- Early works to 1800']
Processed into Fiction:  Moby Dick; Or, The Whale by Herman Melville (769)  Genr
```

```
es:  ['Psychological fiction', 'Adventure stories', 'Sea stories', 'Whaling -- F
iction', 'Ship captains -- Fiction']
Processed into Non-Fiction:  Grimms' Fairy Tales by Jacob Grimm and Wilhelm Grim
m (451)  Genres:  ['Fairy tales -- Germany']
Processed into Fiction:  Peter Pan by J. M.  Barrie (284)  Genres:  ['Fantasy li
terature', 'Fiction', 'Fairies -- Fiction', 'Pirates -- Fiction']
Processed into Fiction:  The Hound of the Baskervilles by Arthur Conan Doyle (24
3)  Genres:  ['Detective and mystery stories', 'Fiction', 'Dogs -- Fiction']
Processed into Fiction:  The Scarlet Letter by Nathaniel Hawthorne (891)  Genre
s:  ['Adultery -- Fiction', 'Historical fiction', 'Psychological fiction']
Processed into Fiction:  The Awakening, and Selected Short Stories by Kate Chopi
n (306)  Genres:  ['Adultery -- Fiction']
Processed into Fiction:  The Strange Case of Dr. Jekyll and Mr. Hyde by Robert L
ouis Stevenson (444)  Genres:  ['Science fiction', 'Horror tales', 'Physicians -
- Fiction', 'Psychological fiction']
Processed into Fiction:  The Wonderful Wizard of Oz by L. Frank  Baum (246)  Gen
res:  ['Fantasy literature', 'Juvenile fiction']
Processed into Fiction:  The Youngest Girl in the School by Evelyn Sharp (433)
Genres:  ['Juvenile fiction']
Processed into Non-Fiction:  Second Treatise of Government by John Locke (216)
Genres:  ['Political science -- Early works to 1800']
Processed into Fiction:  Pride and Prejudice by Jane Austen (1715)  Genres:  ['Y
oung women -- Fiction', 'Love stories', 'Domestic fiction']
Processed into Fiction:  Treasure Island by Robert Louis Stevenson (219)  Genre
s:  ['Sea stories', 'Pirates -- Fiction']
Processed into Non-Fiction:  The Iliad by Homer (389)  Genres:  ['Epic poetry']
Processed into Fiction:  A Christmas Carol in Prose; Being a Ghost Story of Chri
stmas by Charles Dickens (1921)  Genres:  ['Christmas stories', 'England -- Fict
ion', 'Ghost stories']
Processed into Non-Fiction:  The Translations of Beowulf: A Critical Bibliograph
y by Chauncey Brewster Tinker (315)  Genres:  ['Epic poetry']
Processed into Non-Fiction:  A Gamekeeper's Note-book by Owen Jones and Marcus W
oodward (744)  Genres:  []
Processed into Non-Fiction:  The Importance of Being Earnest: A Trivial Comedy f
or Serious People by Oscar Wilde (555)  Genres:  ['Comedies', 'Drama']
Processed into Fiction:  The Picture of Dorian Gray by Oscar Wilde (601)  Genre
s:  ['Didactic fiction', 'Supernatural -- Fiction']
Processed into Fiction:  Little Women by Louisa May Alcott (361)  Genres:  ['You
ng women -- Fiction', 'Domestic fiction', 'Biography', 'Bildungsromans']
Processed into Fiction:  Anne of Green Gables by L. M.  Montgomery (278)  Genre
s:  ['Orphans -- Fiction', 'Islands -- Fiction', 'Friendship -- Fiction', 'Bildu
ngsromans']
Processed into Fiction:  Metamorphosis by Franz Kafka (423)  Genres:  ['Psycholo
gical fiction', 'Metamorphosis -- Fiction']
Processed into Non-Fiction:  The Prince by Niccolò Machiavelli (493)  Genres:
['State, The -- Early works to 1800', 'Philosophy']
Processed into Fiction:  The Adventures of Sherlock Holmes by Arthur Conan Doyle
(669)  Genres:  ['England -- Fiction', 'Detective and mystery stories, English']
Processed into Fiction:  Crime and Punishment by Fyodor Dostoyevsky (464)  Genre
s:  ['Detective and mystery stories', 'Psychological fiction', 'Murder -- Fictio
n']
Processed into Non-Fiction:  The Souls of Black Folk by W. E. B.  Du Bois (297)
Genres:  ['African Americans']
Processed into Non-Fiction:  Joseph and his Brethren by W. K.  Tweedie (480)  Ge
nres:  []
Processed into Fiction:  Adventures of Huckleberry Finn by Mark Twain (355)  Gen
res:  ['Humorous stories', 'Adventure stories', 'Boys -- Fiction', 'Bildungsroma
ns']
Processed into Fiction:  Dubliners by James Joyce (341)  Genres:  ['Short storie
s', 'Dublin (Ireland) -- Fiction']
```

```
Processed into Fiction:  A Christmas Carol by Charles Dickens (263)  Genres:
['Christmas stories', 'England -- Fiction', 'Ghost stories']
Processed into Non-Fiction:  The Philippines a Century Hence by José Rizal (306)
Genres:  ['Philippines']
Processed into Fiction:  The Brothers Karamazov by Fyodor Dostoyevsky (235)  Gen
res:  ['Didactic fiction', 'Fathers and sons -- Fiction', 'Brothers -- Fiction']
Processed TEST:  The Romance of Lust: A classic Victorian erotic novel by Anonym
ous (220)  Genres:  ['Incest -- Fiction', 'Erotic stories']
Processed TEST:  A Pickle for the Knowing Ones by Timothy Dexter (486)  Genres:
['Biography']
Processed TEST:  Märchen der Gebrüder Grimm 2 by Jacob Grimm and Wilhelm Grimm
(220)  Genres:  ['Fairy tales -- Germany']
Processed TEST:  Anna Karenina by graf Leo Tolstoy (238)  Genres:  ['Adultery --
Fiction', 'Didactic fiction', 'Love stories']
Processed TEST:  Wuthering Heights by Emily Brontë (268)  Genres:  ['Psychologic
al fiction', 'Love stories', 'Domestic fiction']
Processed TEST:  Ulysses by James Joyce (390)  Genres:  ['Psychological fictio
n', 'Domestic fiction', 'Epic literature', 'Married people -- Fiction', 'Male fr
iendship -- Fiction']
Processed TEST:  Tractatus Logico-Philosophicus by Ludwig Wittgenstein (320)  Ge
nres:  ['Philosophy']
Processed TEST:  The American Diary of a Japanese Girl by Yoné Noguchi (623)  Ge
nres:  ['Japan -- Fiction']
Processed TEST:  A Modest Proposal by Jonathan Swift (607)  Genres:  ['Political
satire, English', 'Religious satire, English', 'Ireland -- Politics and governme
nt -- 18th century -- Humor']
Processed TEST:  Heart of Darkness by Joseph Conrad (508)  Genres:  ['Psychologi
cal fiction', 'Imperialism -- Fiction']
Processed TEST:  A Warning to the Curious and Other Ghost Stories by M. R.  Jame
s (425)  Genres:  []
Processed TEST:  A Tale of Two Cities by Charles Dickens (556)  Genres:  ['War s
tories', 'Historical fiction', 'England -- Fiction']
Processed TEST:  The Odyssey by Homer (248)  Genres:  ['Epic poetry']
Processed TEST:  Alice's Adventures in Wonderland by Lewis Carroll (748)  Genre
s:  ['Fantasy fiction', 'Imaginary places -- Juvenile fiction', "Children's stor
ies"]
Processed TEST:  Japanese Girls and Women by Alice Mabel Bacon (404)  Genres:
['Women -- Japan -- Social conditions']
Processed TEST:  Dracula by Bram Stoker (620)  Genres:  ['Horror tales', 'Gothic
fiction', 'Vampires -- Fiction']
Processed TEST:  War and Peace by graf Leo Tolstoy (412)  Genres:  ['Historical
fiction', 'War stories', 'Aristocracy (Social class) -- Russia -- Fiction']
Processed TEST:  Beyond Good and Evil by Friedrich Wilhelm Nietzsche (275)  Genr
es:  ['Ethics', 'Philosophy']
Processed TEST:  Les Misérables by Victor Hugo (217)  Genres:  ['Historical fict
ion', 'Epic literature', 'Orphans -- Fiction']
Processed TEST:  Essays of Michel de Montaigne — Complete by Michel de Montaigne
(254)  Genres:  []
Processed TEST:  Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft
Shelley (1963)  Genres:  ['Science fiction', 'Horror Tales', 'Gothic Fiction',
'Scientists -- Fiction']
Processed TEST:  The Kama Sutra of Vatsyayana by Vatsyayana (235)  Genres:  ['Lo
ve', 'Sex']
Processed TEST:  Walden, and On The Duty Of Civil Disobedience by Henry David Th
oreau (341)  Genres:  ['Biography', 'Solitude', 'Natural history -- Massachusett
s -- Walden Woods']
Processed TEST:  The Great Gatsby by F. Scott  Fitzgerald (510)  Genres:  ['Psyc
hological fiction', 'Married women -- Fiction', 'First loves -- Fiction']
Processed TEST:  Kwaidan: Stories and Studies of Strange Things by Lafcadio Hear
n (259)  Genres:  ['Ghost stories', 'Paranormal fiction', 'Japan -- Social life
```

```
and customs -- Fiction']
Processed TEST:  The Prophet by Kahlil Gibran (282)  Genres:  ['Prose poems, Ame
rican', 'Mysticism -- Poetry']
Processed TEST:  A Doll's House : a play by Henrik Ibsen (599)  Genres:  ['Dram
a']
Processed TEST:  Great Expectations by Charles Dickens (446)  Genres:  ['Orphans
-- Fiction', 'England -- Fiction', 'Revenge -- Fiction', 'Bildungsromans']
Processed TEST:  Narrative of the Life of Frederick Douglass, an American Slave
by Frederick Douglass (289)  Genres:  ['Biography']
Processed TEST:  The Yellow Wallpaper by Charlotte Perkins Gilman (503)  Genres:
['Psychological fiction', 'Married women -- Psychology -- Fiction']
Processed TEST:  Siddhartha by Hermann Hesse (299)  Genres:  ['Spiritual life --
Fiction']
Processed TEST:  The Count of Monte Cristo, Illustrated by Alexandre Dumas (333)
Genres:  ['Historical fiction', 'Revenge -- Fiction', 'Adventure stories', 'Pira
tes -- Fiction']
Processed TEST:  The Adventures of Tom Sawyer, Complete by Mark Twain (238)  Gen
res:  ['Humorous stories', 'Boys -- Fiction', 'Adventure stories', 'Bildungsroma
ns']
Processed TEST:  Anthem by Ayn Rand (218)  Genres:  ['Science fiction', 'Psychol
ogical fiction', 'Love stories', 'Men -- Psychology -- Fiction']
Processed TEST:  Don Quixote by Miguel de Cervantes Saavedra (249)  Genres:  ['S
pain -- Social life and customs -- 16th century -- Fiction']
Processed TEST:  Summer by Romain Rolland (361)  Genres:  []
Processed TEST:  Jane Eyre: An Autobiography by Charlotte Brontë (500)  Genres:
['Orphans -- Fiction', 'England -- Fiction', 'Young women -- Fiction', 'Love sto
ries', 'Bildungsromans']
```

In [8]:

```
#We create and train our model
model = gensim.models.doc2vec.Doc2Vec(vector_size=100, min_count=2, epochs=15)
model.build_vocab(train_corpus)
model.train(train_corpus, total_examples=model.corpus_count, epochs=model.epochs
```

In [12]:

```
#This is the list of books we didn't tag into our fiction and non-fiction vector
#calculate similarity against.
model.docvecs.index_to_key[2:]
```

Out[12]:

```
['The Romance of Lust: A classic Victorian erotic novel by Anonymous (220)',
 'A Pickle for the Knowing Ones by Timothy Dexter (486)',
 'Märchen der Gebrüder Grimm 2 by Jacob Grimm and Wilhelm Grimm (220)',
 'Anna Karenina by graf Leo Tolstoy (238)',
 'Wuthering Heights by Emily Brontë (268)',
 'Ulysses by James Joyce (390)',
 'Tractatus Logico-Philosophicus by Ludwig Wittgenstein (320)',
 'The American Diary of a Japanese Girl by Yoné Noguchi (623)',
 'A Modest Proposal by Jonathan Swift (607)',
 'Heart of Darkness by Joseph Conrad (508)',
 'A Warning to the Curious and Other Ghost Stories by M. R.  James (425)',
 'A Tale of Two Cities by Charles Dickens (556)',
 'The Odyssey by Homer (248)',
 "Alice's Adventures in Wonderland by Lewis Carroll (748)",
 'Japanese Girls and Women by Alice Mabel Bacon (404)',
 'Dracula by Bram Stoker (620)',
 'War and Peace by graf Leo Tolstoy (412)',
 'Beyond Good and Evil by Friedrich Wilhelm Nietzsche (275)',
 'Les Misérables by Victor Hugo (217)',
 'Essays of Michel de Montaigne — Complete by Michel de Montaigne (254)',
 'Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley (196
```

```
3)',
 'The Kama Sutra of Vatsyayana by Vatsyayana (235)',
 'Walden, and On The Duty Of Civil Disobedience by Henry David Thoreau (341)',
 'The Great Gatsby by F. Scott  Fitzgerald (510)',
 'Kwaidan: Stories and Studies of Strange Things by Lafcadio Hearn (259)',
 'The Prophet by Kahlil Gibran (282)',
 "A Doll's House : a play by Henrik Ibsen (599)",
 'Great Expectations by Charles Dickens (446)',
 'Narrative of the Life of Frederick Douglass, an American Slave by Frederick Do
uglass (289)',
 'The Yellow Wallpaper by Charlotte Perkins Gilman (503)',
 'Siddhartha by Hermann Hesse (299)',
 'The Count of Monte Cristo, Illustrated by Alexandre Dumas (333)',
 'The Adventures of Tom Sawyer, Complete by Mark Twain (238)',
 'Anthem by Ayn Rand (218)',
 'Don Quixote by Miguel de Cervantes Saavedra (249)',
 'Summer by Romain Rolland (361)',
 'Jane Eyre: An Autobiography by Charlotte Brontë (500)']
```

In [13]:

```
#We extract our fiction and non-fiction vectors, and we compute similarity for e
fiction_vec = model.docvecs[1]
non_fiction_vec = model.docvecs[0]

print("Cosine Similarities with Fiction vs Non-Fiction Embeddings:")
print()

for i in range(2,len(model.docvecs)):
    print()
    print(model.docvecs.index_to_key[i],"  ",books_genres[model.docvecs.index_to

    print("Fiction similarity: ",cosine_similarity([model.docvecs[i]], [fiction_
    print("Non-Fiction similarity: ",cosine_similarity([model.docvecs[i]], [non_
```

```
Cosine Similarities with Fiction vs Non-Fiction Embeddings:


The Romance of Lust: A classic Victorian erotic novel by Anonymous (220)    ['In
cest -- Fiction', 'Erotic stories']
Fiction similarity:  [[-0.05316773]]
Non-Fiction similarity:  [[0.02405052]]

A Pickle for the Knowing Ones by Timothy Dexter (486)    ['Biography']
Fiction similarity:  [[0.03112186]]
Non-Fiction similarity:  [[0.27475834]]

Märchen der Gebrüder Grimm 2 by Jacob Grimm and Wilhelm Grimm (220)    ['Fairy t
ales -- Germany']
Fiction similarity:  [[0.08288112]]
Non-Fiction similarity:  [[0.36124396]]

Anna Karenina by graf Leo Tolstoy (238)    ['Adultery -- Fiction', 'Didactic fic
tion', 'Love stories']
Fiction similarity:  [[0.39001006]]
Non-Fiction similarity:  [[0.22117396]]

Wuthering Heights by Emily Brontë (268)    ['Psychological fiction', 'Love stori
es', 'Domestic fiction']
```

```
Fiction similarity:  [[0.0976584]]
Non-Fiction similarity:  [[0.19259836]]

Ulysses by James Joyce (390)    ['Psychological fiction', 'Domestic fiction', 'E
pic literature', 'Married people -- Fiction', 'Male friendship -- Fiction']
Fiction similarity:  [[0.03295119]]
Non-Fiction similarity:  [[0.08621626]]

Tractatus Logico-Philosophicus by Ludwig Wittgenstein (320)    ['Philosophy']
Fiction similarity:  [[0.07726576]]
Non-Fiction similarity:  [[0.35948765]]

The American Diary of a Japanese Girl by Yoné Noguchi (623)    ['Japan -- Fictio
n']
Fiction similarity:  [[0.18793815]]
Non-Fiction similarity:  [[0.24376105]]

A Modest Proposal by Jonathan Swift (607)    ['Political satire, English', 'Reli
gious satire, English', 'Ireland -- Politics and government -- 18th century -- H
umor']
Fiction similarity:  [[0.02830998]]
Non-Fiction similarity:  [[0.48431265]]

Heart of Darkness by Joseph Conrad (508)    ['Psychological fiction', 'Imperiali
sm -- Fiction']
Fiction similarity:  [[0.36236417]]
Non-Fiction similarity:  [[0.09613796]]

A Warning to the Curious and Other Ghost Stories by M. R.  James (425)    []
Fiction similarity:  [[0.11689119]]
Non-Fiction similarity:  [[0.19405797]]

A Tale of Two Cities by Charles Dickens (556)    ['War stories', 'Historical fic
tion', 'England -- Fiction']
Fiction similarity:  [[0.0930782]]
Non-Fiction similarity:  [[0.12418421]]

The Odyssey by Homer (248)    ['Epic poetry']
Fiction similarity:  [[0.04639429]]
Non-Fiction similarity:  [[0.3002144]]

Alice's Adventures in Wonderland by Lewis Carroll (748)    ['Fantasy fiction',
'Imaginary places -- Juvenile fiction', "Children's stories"]
Fiction similarity:  [[0.3130059]]
Non-Fiction similarity:  [[0.18933827]]

Japanese Girls and Women by Alice Mabel Bacon (404)    ['Women -- Japan -- Socia
l conditions']
Fiction similarity:  [[0.13491642]]
Non-Fiction similarity:  [[0.34866852]]

Dracula by Bram Stoker (620)    ['Horror tales', 'Gothic fiction', 'Vampires --
Fiction']
Fiction similarity:  [[0.09649402]]
Non-Fiction similarity:  [[0.19946308]]

War and Peace by graf Leo Tolstoy (412)    ['Historical fiction', 'War stories',
'Aristocracy (Social class) -- Russia -- Fiction']
Fiction similarity:  [[0.17590179]]
Non-Fiction similarity:  [[0.27332434]]
```

Beyond Good and Evil by Friedrich Wilhelm Nietzsche (275) ['Ethics', 'Philoso
phy']
Fiction similarity: [[0.15141302]]
Non-Fiction similarity: [[0.23496476]]

Les Misérables by Victor Hugo (217) ['Historical fiction', 'Epic literature',
'Orphans -- Fiction']
Fiction similarity: [[0.06488585]]
Non-Fiction similarity: [[0.14127941]]

Essays of Michel de Montaigne — Complete by Michel de Montaigne (254) []
Fiction similarity: [[0.02146232]]
Non-Fiction similarity: [[0.12770727]]

Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley (1963)
['Science fiction', 'Horror Tales', 'Gothic Fiction', 'Scientists -- Fiction']
Fiction similarity: [[0.10958128]]
Non-Fiction similarity: [[0.17222267]]

The Kama Sutra of Vatsyayana by Vatsyayana (235) ['Love', 'Sex']
Fiction similarity: [[0.08646447]]
Non-Fiction similarity: [[0.20152023]]

Walden, and On The Duty Of Civil Disobedience by Henry David Thoreau (341)
['Biography', 'Solitude', 'Natural history -- Massachusetts -- Walden Woods']
Fiction similarity: [[0.1117556]]
Non-Fiction similarity: [[0.32351252]]

The Great Gatsby by F. Scott Fitzgerald (510) ['Psychological fiction', 'Mar
ried women -- Fiction', 'First loves -- Fiction']
Fiction similarity: [[0.32229045]]
Non-Fiction similarity: [[0.05365625]]

Kwaidan: Stories and Studies of Strange Things by Lafcadio Hearn (259) ['Ghos
t stories', 'Paranormal fiction', 'Japan -- Social life and customs -- Fiction']
Fiction similarity: [[0.1615097]]
Non-Fiction similarity: [[0.21151352]]

The Prophet by Kahlil Gibran (282) ['Prose poems, American', 'Mysticism -- Po
etry']
Fiction similarity: [[0.09554582]]
Non-Fiction similarity: [[0.2847405]]

A Doll's House : a play by Henrik Ibsen (599) ['Drama']
Fiction similarity: [[0.36909798]]
Non-Fiction similarity: [[0.12462879]]

Great Expectations by Charles Dickens (446) ['Orphans -- Fiction', 'England -
- Fiction', 'Revenge -- Fiction', 'Bildungsromans']
Fiction similarity: [[0.23567179]]
Non-Fiction similarity: [[0.14792575]]

Narrative of the Life of Frederick Douglass, an American Slave by Frederick Doug
lass (289) ['Biography']
Fiction similarity: [[0.14226636]]
Non-Fiction similarity: [[0.3139574]]

The Yellow Wallpaper by Charlotte Perkins Gilman (503) ['Psychological fictio
n', 'Married women -- Psychology -- Fiction']

```
Fiction similarity:  [[0.35056472]]
Non-Fiction similarity:  [[0.19185829]]

Siddhartha by Hermann Hesse (299)    ['Spiritual life -- Fiction']
Fiction similarity:  [[0.19157973]]
Non-Fiction similarity:  [[0.12734298]]

The Count of Monte Cristo, Illustrated by Alexandre Dumas (333)    ['Historical
fiction', 'Revenge -- Fiction', 'Adventure stories', 'Pirates -- Fiction']
Fiction similarity:  [[0.05591396]]
Non-Fiction similarity:  [[0.11180107]]

The Adventures of Tom Sawyer, Complete by Mark Twain (238)    ['Humorous storie
s', 'Boys -- Fiction', 'Adventure stories', 'Bildungsromans']
Fiction similarity:  [[0.35497108]]
Non-Fiction similarity:  [[0.11962791]]

Anthem by Ayn Rand (218)    ['Science fiction', 'Psychological fiction', 'Love s
tories', 'Men -- Psychology -- Fiction']
Fiction similarity:  [[0.21538979]]
Non-Fiction similarity:  [[0.205314]]

Don Quixote by Miguel de Cervantes Saavedra (249)    ['Spain -- Social life and
customs -- 16th century -- Fiction']
Fiction similarity:  [[0.01779339]]
Non-Fiction similarity:  [[0.03409424]]

Summer by Romain Rolland (361)    []
Fiction similarity:  [[0.2446959]]
Non-Fiction similarity:  [[0.1727797]]

Jane Eyre: An Autobiography by Charlotte Brontë (500)    ['Orphans -- Fiction',
'England -- Fiction', 'Young women -- Fiction', 'Love stories', 'Bildungsroman
s']
Fiction similarity:  [[0.04781731]]
Non-Fiction similarity:  [[0.20992048]]
```

RESULTS

We can see most of fiction categories were assigned properly.

Only 13 of a total of 38 books were not classified correctly, yielding a 65.8% success rate.

The expections were the following:

- Jane Eyre: An Autobiography by Charlotte Brontë
- Don Quixote by Miguel de Cervantes Saavedra
- A Doll's House : a play by Henrik Ibsen
- Kwaidan: Stories and Studies of Strange Things by Lafcadio Hearn
- Frankenstein; Or, The Modern Prometheus by Mary Wollstonecraft Shelley
- Les Misérables by Victor Hugo
- War and Peace by graf Leo Tolstoy
- Dracula by Bram Stoker
- A Tale of Two Cities by Charles Dickens
- The American Diary of a Japanese Girl by Yoné Noguchi

- Ulysses by James Joyce
- Wuthering Heights by Emily Brontë
- The Romance of Lust: A classic Victorian erotic novel by Anonymous

This is a simple example of categories evaluation by using Ginsem doc2vec embeddings. Categories with more specificity can be studied with this same model but a much larger data set must be extracted to obtain representative genre vectors that can be used for evaluating books genres.