hw8

December 3, 2024

1 Topic Modeling for Fun and Profit

Source

In this notebook we'll

- vectorize a streamed corpus
- run topic modeling on streamed vectors, using gensim
- explore how to choose, evaluate and tweak topic modeling parameters
- persist trained models to disk, for later re-use
- In the previous notebook 1 - Streamed Corpora we used the 20newsgroups corpus to demonstrate data preprocessing and streaming.

Now we'll switch to the English Wikipedia and do some topic modeling. Link: https://radimrehurek.com/gensim/auto_examples/core/run_corpora_and_vector_spaces.html#sphx-glr-auto-examples-core-run-corpora-and-vector-spaces-py

[1]:
```
from datetime import datetime

# datetime object containing current date and time
now = datetime.now()

print("Begun at", now)
```

```
Begun at 2024-12-03 07:05:15.629148
```

[2]:
```
!pip install six cython numpy scipy ipython[notebook]
!pip install nltk gensim pattern requests textblob
!python -m textblob.download_corpora lite
!pip install --upgrade gensim
!pip install --upgrade smart_open
```

```
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages
(1.16.0)
Requirement already satisfied: cython in /usr/local/lib/python3.10/dist-packages
(3.0.11)
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages
(1.26.4)
Requirement already satisfied: scipy in /usr/local/lib/python3.10/dist-packages
(1.13.1)
```

```
Requirement already satisfied: ipython[notebook] in
/usr/local/lib/python3.10/dist-packages (7.34.0)
Requirement already satisfied: setuptools>=18.5 in
/usr/local/lib/python3.10/dist-packages (from ipython[notebook]) (75.1.0)
Collecting jedi>=0.16 (from ipython[notebook])
  Downloading jedi-0.19.2-py2.py3-none-any.whl.metadata (22 kB)
Requirement already satisfied: decorator in /usr/local/lib/python3.10/dist-
packages (from ipython[notebook]) (4.4.2)
Requirement already satisfied: pickleshare in /usr/local/lib/python3.10/dist-
packages (from ipython[notebook]) (0.7.5)
Requirement already satisfied: traitlets>=4.2 in /usr/local/lib/python3.10/dist-
packages (from ipython[notebook]) (5.7.1)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from ipython[notebook]) (3.0.48)
Requirement already satisfied: pygments in /usr/local/lib/python3.10/dist-
packages (from ipython[notebook]) (2.18.0)
Requirement already satisfied: backcall in /usr/local/lib/python3.10/dist-
packages (from ipython[notebook]) (0.2.0)
Requirement already satisfied: matplotlib-inline in
/usr/local/lib/python3.10/dist-packages (from ipython[notebook]) (0.1.7)
Requirement already satisfied: pexpect>4.3 in /usr/local/lib/python3.10/dist-
packages (from ipython[notebook]) (4.9.0)
Requirement already satisfied: notebook in /usr/local/lib/python3.10/dist-
packages (from ipython[notebook]) (6.5.5)
Requirement already satisfied: ipywidgets in /usr/local/lib/python3.10/dist-
packages (from ipython[notebook]) (7.7.1)
Requirement already satisfied: parso<0.9.0,>=0.8.4 in
/usr/local/lib/python3.10/dist-packages (from jedi>=0.16->ipython[notebook])
(0.8.4)
Requirement already satisfied: ptyprocess>=0.5 in
/usr/local/lib/python3.10/dist-packages (from pexpect>4.3->ipython[notebook])
(0.7.0)
Requirement already satisfied: wcwidth in /usr/local/lib/python3.10/dist-
packages (from prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython[notebook])
(0.2.13)
Requirement already satisfied: ipykernel>=4.5.1 in
/usr/local/lib/python3.10/dist-packages (from ipywidgets->ipython[notebook])
(5.5.6)
Requirement already satisfied: ipython-genutils~=0.2.0 in
/usr/local/lib/python3.10/dist-packages (from ipywidgets->ipython[notebook])
(0.2.0)
Requirement already satisfied: widgetsnbextension~=3.6.0 in
/usr/local/lib/python3.10/dist-packages (from ipywidgets->ipython[notebook])
(3.6.10)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in
/usr/local/lib/python3.10/dist-packages (from ipywidgets->ipython[notebook])
(3.0.13)
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages
```

```
(from notebook->ipython[notebook]) (3.1.4)
Requirement already satisfied: tornado>=6.1 in /usr/local/lib/python3.10/dist-
packages (from notebook->ipython[notebook]) (6.3.3)
Requirement already satisfied: pyzmq<25,>=17 in /usr/local/lib/python3.10/dist-
packages (from notebook->ipython[notebook]) (24.0.1)
Requirement already satisfied: argon2-cffi in /usr/local/lib/python3.10/dist-
packages (from notebook->ipython[notebook]) (23.1.0)
Requirement already satisfied: jupyter-core>=4.6.1 in
/usr/local/lib/python3.10/dist-packages (from notebook->ipython[notebook])
(5.7.2)
Requirement already satisfied: jupyter-client<8,>=5.3.4 in
/usr/local/lib/python3.10/dist-packages (from notebook->ipython[notebook])
(6.1.12)
Requirement already satisfied: nbformat in /usr/local/lib/python3.10/dist-
packages (from notebook->ipython[notebook]) (5.10.4)
Requirement already satisfied: nbconvert>=5 in /usr/local/lib/python3.10/dist-
packages (from notebook->ipython[notebook]) (7.16.4)
Requirement already satisfied: nest-asyncio>=1.5 in
/usr/local/lib/python3.10/dist-packages (from notebook->ipython[notebook])
(1.6.0)
Requirement already satisfied: Send2Trash>=1.8.0 in
/usr/local/lib/python3.10/dist-packages (from notebook->ipython[notebook])
(1.8.3)
Requirement already satisfied: terminado>=0.8.3 in
/usr/local/lib/python3.10/dist-packages (from notebook->ipython[notebook])
(0.18.1)
Requirement already satisfied: prometheus-client in
/usr/local/lib/python3.10/dist-packages (from notebook->ipython[notebook])
(0.21.0)
Requirement already satisfied: nbclassic>=0.4.7 in
/usr/local/lib/python3.10/dist-packages (from notebook->ipython[notebook])
(1.1.0)
Requirement already satisfied: python-dateutil>=2.1 in
/usr/local/lib/python3.10/dist-packages (from jupyter-
client<8,>=5.3.4->notebook->ipython[notebook]) (2.8.2)
Requirement already satisfied: platformdirs>=2.5 in
/usr/local/lib/python3.10/dist-packages (from jupyter-
core>=4.6.1->notebook->ipython[notebook]) (4.3.6)
Requirement already satisfied: notebook-shim>=0.2.3 in
/usr/local/lib/python3.10/dist-packages (from
nbclassic>=0.4.7->notebook->ipython[notebook]) (0.2.4)
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.10/dist-
packages (from nbconvert>=5->notebook->ipython[notebook]) (4.12.3)
Requirement already satisfied: bleach!=5.0.0 in /usr/local/lib/python3.10/dist-
packages (from nbconvert>=5->notebook->ipython[notebook]) (6.2.0)
Requirement already satisfied: defusedxml in /usr/local/lib/python3.10/dist-
packages (from nbconvert>=5->notebook->ipython[notebook]) (0.7.1)
Requirement already satisfied: jupyterlab-pygments in
```

```
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook->ipython[notebook]) (0.3.0)
Requirement already satisfied: markupsafe>=2.0 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook->ipython[notebook]) (3.0.2)
Requirement already satisfied: mistune<4,>=2.0.3 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook->ipython[notebook]) (3.0.2)
Requirement already satisfied: nbclient>=0.5.0 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook->ipython[notebook]) (0.10.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-
packages (from nbconvert>=5->notebook->ipython[notebook]) (24.2)
Requirement already satisfied: pandocfilters>=1.4.1 in
/usr/local/lib/python3.10/dist-packages (from
nbconvert>=5->notebook->ipython[notebook]) (1.5.1)
Requirement already satisfied: tinycss2 in /usr/local/lib/python3.10/dist-
packages (from nbconvert>=5->notebook->ipython[notebook]) (1.4.0)
Requirement already satisfied: fastjsonschema>=2.15 in
/usr/local/lib/python3.10/dist-packages (from
nbformat->notebook->ipython[notebook]) (2.20.0)
Requirement already satisfied: jsonschema>=2.6 in
/usr/local/lib/python3.10/dist-packages (from
nbformat->notebook->ipython[notebook]) (4.23.0)
Requirement already satisfied: argon2-cffi-bindings in
/usr/local/lib/python3.10/dist-packages (from
argon2-cffi->notebook->ipython[notebook]) (21.2.0)
Requirement already satisfied: webencodings in /usr/local/lib/python3.10/dist-
packages (from bleach!=5.0.0->nbconvert>=5->notebook->ipython[notebook]) (0.5.1)
Requirement already satisfied: attrs>=22.2.0 in /usr/local/lib/python3.10/dist-
packages (from jsonschema>=2.6->nbformat->notebook->ipython[notebook]) (24.2.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
/usr/local/lib/python3.10/dist-packages (from
jsonschema>=2.6->nbformat->notebook->ipython[notebook]) (2024.10.1)
Requirement already satisfied: referencing>=0.28.4 in
/usr/local/lib/python3.10/dist-packages (from
jsonschema>=2.6->nbformat->notebook->ipython[notebook]) (0.35.1)
Requirement already satisfied: rpds-py>=0.7.1 in /usr/local/lib/python3.10/dist-
packages (from jsonschema>=2.6->nbformat->notebook->ipython[notebook]) (0.21.0)
Requirement already satisfied: jupyter-server<3,>=1.8 in
/usr/local/lib/python3.10/dist-packages (from notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook->ipython[notebook]) (1.24.0)
Requirement already satisfied: cffi>=1.0.1 in /usr/local/lib/python3.10/dist-
packages (from argon2-cffi-bindings->argon2-cffi->notebook->ipython[notebook])
(1.17.1)
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-
packages (from beautifulsoup4->nbconvert>=5->notebook->ipython[notebook]) (2.6)
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-
```

```
packages (from cffi>=1.0.1->argon2-cffi-
bindings->argon2-cffi->notebook->ipython[notebook]) (2.22)
Requirement already satisfied: anyio<4,>=3.1.0 in
/usr/local/lib/python3.10/dist-packages (from jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook->ipython[notebook]) (3.7.1)
Requirement already satisfied: websocket-client in
/usr/local/lib/python3.10/dist-packages (from jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook->ipython[notebook]) (1.8.0)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook->ipython[notebook]) (3.10)
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook->ipython[notebook]) (1.3.1)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-
packages (from anyio<4,>=3.1.0->jupyter-server<3,>=1.8->notebook-
shim>=0.2.3->nbclassic>=0.4.7->notebook->ipython[notebook]) (1.2.2)
Downloading jedi-0.19.2-py2.py3-none-any.whl (1.6 MB)
                         1.6/1.6 MB
16.8 MB/s eta 0:00:00
Installing collected packages: jedi
Successfully installed jedi-0.19.2
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages
(3.9.1)
Requirement already satisfied: gensim in /usr/local/lib/python3.10/dist-packages
(4.3.3)
ERROR: Could not find a version that satisfies the requirement pattern

(from versions: none)

ERROR: No matching distribution found for pattern

[nltk_data] Downloading package brown to /root/nltk_data…
[nltk_data]   Unzipping corpora/brown.zip.
[nltk_data] Downloading package punkt to /root/nltk_data…
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data…
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data…
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
Finished.
Requirement already satisfied: gensim in /usr/local/lib/python3.10/dist-packages
(4.3.3)
Requirement already satisfied: numpy<2.0,>=1.18.5 in
/usr/local/lib/python3.10/dist-packages (from gensim) (1.26.4)
Requirement already satisfied: scipy<1.14.0,>=1.7.0 in
/usr/local/lib/python3.10/dist-packages (from gensim) (1.13.1)
Requirement already satisfied: smart-open>=1.8.1 in
/usr/local/lib/python3.10/dist-packages (from gensim) (7.0.5)
```

```
Requirement already satisfied: wrapt in /usr/local/lib/python3.10/dist-packages
(from smart-open>=1.8.1->gensim) (1.16.0)
Requirement already satisfied: smart_open in /usr/local/lib/python3.10/dist-
packages (7.0.5)
Requirement already satisfied: wrapt in /usr/local/lib/python3.10/dist-packages
(from smart_open) (1.16.0)
```

[3]:
```
!rm -f download_data.py && wget 'https://raw.githubusercontent.com/piskvorky/
 ↪topic_modeling_tutorial/master/download_data.py'
#
# The older datasets are no longer available, use the latest one.
!sed -i 's/20140623/latest/g' download_data.py
#
# wikimedia sometimes refuses to connect due to excessive load
# use a mirror site instead. see https://dumps.wikimedia.org/mirrors.html
!sed -i 's|dumps.wikimedia.org|dumps.wikimedia.your.org|g' download_data.py
```

```
--2024-12-03 07:05:40--  https://raw.githubusercontent.com/piskvorky/topic_model
ing_tutorial/master/download_data.py
Resolving raw.githubusercontent.com (raw.githubusercontent.com)…
185.199.108.133, 185.199.109.133, 185.199.110.133, …
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.108.133|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 2101 (2.1K) [text/plain]
Saving to: ‘download_data.py’

 download_data.py      0%[                    ]       0  --.-KB/s
download_data.py    100%[===================>]   2.05K  --.-KB/s    in 0s

2024-12-03 07:05:40 (35.6 MB/s) - ‘download_data.py’ saved [2101/2101]
```

[4]:
```
!rm -rf ./data
!mkdir ./data
!python download_data.py ./data
```

```
2024-12-03 07:05:41,568 : MainThread : INFO : running download_data.py ./data
2024-12-03 07:05:41,568 : MainThread : INFO : downloading
http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz into
./data/20news-bydate.tar.gz
2024-12-03 07:05:43,582 : MainThread : INFO : downloaded 14464277 bytes
2024-12-03 07:05:43,603 : MainThread : INFO : downloading
http://dumps.wikimedia.your.org/simplewiki/latest/simplewiki-latest-pages-
articles.xml.bz2 into ./data/simplewiki-latest-pages-articles.xml.bz2
2024-12-03 07:05:48,592 : MainThread : INFO : downloaded 235367506 bytes
2024-12-03 07:05:48,592 : MainThread : INFO : finished running download_data.py
```

[5]:
```
# import and setup modules we'll be using in this notebook
import logging
import itertools

import numpy as np
import gensim

logging.basicConfig(format='%(levelname)s : %(message)s', level=logging.INFO)
logging.root.level = logging.INFO  # ipython sometimes messes up the logging␣
 ↪setup; restore

def head(stream, n=10):
    """Convenience fnc: return the first `n` elements of the stream, as plain␣
 ↪list."""
    return list(itertools.islice(stream, n))
```

[6]:
```
# import and setup modules we'll be using in this notebook
import logging
import itertools

import numpy as np
import gensim

logging.basicConfig(format='%(levelname)s : %(message)s', level=logging.INFO)
logging.root.level = logging.INFO  # ipython sometimes messes up the logging␣
 ↪setup; restore

def head(stream, n=10):
    """Convenience fnc: return the first `n` elements of the stream, as plain␣
 ↪list."""
    return list(itertools.islice(stream, n))
```

[7]:
```
from gensim.test.utils import datapath, get_tmpfile
from gensim.corpora import WikiCorpus, MmCorpus
path_to_wiki_dump = datapath("enwiki-latest-pages-articles1.
 ↪xml-p000000010p000030302-shortened.bz2")
corpus_path = get_tmpfile("wiki-corpus.mm")
wiki = WikiCorpus(path_to_wiki_dump)  # create word->word_id mapping, ~8h on␣
 ↪full wiki
MmCorpus.serialize(corpus_path, wiki)  # another 8h, creates a file in␣
 ↪MatrixMarket format and mapping

texts = [' '.join(txt) for txt in wiki.get_texts()]
print(texts[0])
print(texts[1])
```

```
INFO:gensim.corpora.dictionary:adding document #0 to Dictionary<0 unique tokens:
```

```
[]>
INFO:gensim.corpora.dictionary:built Dictionary<12 unique tokens: ['computer',
'human', 'interface', 'response', 'survey']…> from 9 documents (total 29
corpus positions)
INFO:gensim.utils:Dictionary lifecycle event {'msg': "built Dictionary<12 unique
tokens: ['computer', 'human', 'interface', 'response', 'survey']…> from 9
documents (total 29 corpus positions)", 'datetime':
'2024-12-03T07:05:51.158392', 'gensim': '4.3.3', 'python': '3.10.12 (main, Nov
6 2024, 20:22:13) [GCC 11.4.0]', 'platform': 'Linux-6.1.85+-x86_64-with-
glibc2.35', 'event': 'created'}
INFO:gensim.corpora.dictionary:adding document #0 to Dictionary<0 unique tokens:
[]>
INFO:gensim.corpora.wikicorpus:finished iterating over Wikipedia corpus of 106
documents with 452944 positions (total 206 articles, 453267 positions before
pruning articles shorter than 50 words)
INFO:gensim.corpora.dictionary:built Dictionary<34212 unique tokens:
['abandoned', 'abandoning', 'abandonment', 'ability', 'able']…> from 106
documents (total 452944 corpus positions)
INFO:gensim.utils:Dictionary lifecycle event {'msg': "built Dictionary<34212
unique tokens: ['abandoned', 'abandoning', 'abandonment', 'ability', 'able']…>
from 106 documents (total 452944 corpus positions)", 'datetime':
'2024-12-03T07:05:55.758814', 'gensim': '4.3.3', 'python': '3.10.12 (main, Nov
6 2024, 20:22:13) [GCC 11.4.0]', 'platform': 'Linux-6.1.85+-x86_64-with-
glibc2.35', 'event': 'created'}
INFO:gensim.corpora.mmcorpus:storing corpus in Matrix Market format to
/tmp/tmpphnmltpp/wiki-corpus.mm
INFO:gensim.matutils:saving sparse matrix to /tmp/tmpphnmltpp/wiki-corpus.mm
INFO:gensim.matutils:PROGRESS: saving document #0
INFO:gensim.corpora.wikicorpus:finished iterating over Wikipedia corpus of 106
documents with 452944 positions (total 206 articles, 453267 positions before
pruning articles shorter than 50 words)
INFO:gensim.matutils:saved 106x34212 matrix, density=3.537% (128253/3626472)
INFO:gensim.corpora.indexedcorpus:saving MmCorpus index to
/tmp/tmpphnmltpp/wiki-corpus.mm.index
INFO:gensim.corpora.wikicorpus:finished iterating over Wikipedia corpus of 106
documents with 452944 positions (total 206 articles, 453267 positions before
pruning articles shorter than 50 words)
```

```
anarchism is political philosophy that advocates self governed societies based
on voluntary institutions these are often described as stateless societies
although several authors have defined them more specifically as institutions
based on non hierarchical free associations anarchism considers the state to be
undesirable unnecessary and harmful while anti statism is central anarchism
entails opposing authority or hierarchical organisation in the conduct of all
human relations including but not limited to the state system anarchism draws on
many currents of thought and strategy anarchism does not offer fixed body of
doctrine from single particular world view instead fluxing and flowing as
philosophy many types and traditions of anarchism exist not all of which are
```

mutually exclusive anarchist schools of thought can differ fundamentally supporting anything from extreme individualism to complete collectivism strains of anarchism have often been divided into the categories of social and individualist anarchism or similar dual classifications anarchism is usually considered radical left wing ideology and much of anarchist economics and anarchist legal philosophy reflect anti authoritarian interpretations of communism collectivism syndicalism mutualism or participatory economics etymology and terminology the term anarchism is compound word composed from the word anarchy and the suffix ism themselves derived respectively from the greek anarchy from anarchos meaning one without rulers from the privative prefix an without and archos leader ruler cf archon or arkhē authority sovereignty realm magistracy and the suffix or ismos isma from the verbal infinitive suffix izein the first known use of this word was in various factions within the french revolution labelled opponents as anarchists as robespierre did the hébertists although few shared many views of later anarchists there would be many revolutionaries of the early nineteenth century who contributed to the anarchist doctrines of the next generation such as william godwin and wilhelm weitling but they did not use the word anarchist or anarchism in describing themselves or their beliefs the first political philosopher to call himself an anarchist was pierre joseph proudhon marking the formal birth of anarchism in the mid nineteenth century since the and beginning in france the term libertarianism has often been used as synonym for anarchism and was used almost exclusively in this sense until the in the united states its use as synonym is still common outside the united states on the other hand some use libertarianism to refer to individualistic free market philosophy only referring to free market anarchism as libertarian anarchism history origins woodcut from diggers document by william everard the earliest anarchist themes can be found in the th century bc among the works of taoist philosopher laozi and in later centuries by zhuangzi and bao jingyan zhuangzi philosophy has been described by various sources as anarchist zhuangzi wrote petty thief is put in jail great brigand becomes ruler of nation diogenes of sinope and the cynics their contemporary zeno of citium the founder of stoicism also introduced similar topics jesus is sometimes considered the first anarchist in the christian anarchist tradition georges lechartier wrote that the true founder of anarchy was jesus christ and the first anarchist society was that of the apostles in early islamic history some manifestations of anarchic thought are found during the islamic civil war over the caliphate where the kharijites insisted that the imamate is right for each individual within the islamic society later some muslim scholars such as amer al basri and abu hanifa led movements of boycotting the rulers paving the way to the waqf endowments tradition which served as an alternative to and asylum from the centralized authorities of the emirs but such interpretations reverberates subversive religious conceptions like the aforementioned seemingly anarchistic taoist teachings and that of other anti authoritarian religious traditions creating complex relationship regarding the question as to whether or not anarchism and religion are compatible this is exemplified when the glorification of the state is viewed as form of sinful idolatry the french renaissance political philosopher étienne de la boétie wrote in his most famous work the discourse on voluntary servitude what some historians consider an important

anarchist precedent the radical protestant christian gerrard winstanley and his group the diggers are cited by various authors as proposing anarchist social measures in the th century in england the term anarchist first entered the english language in during the english civil war as term of abuse used by royalists against their roundhead opponents by the time of the french revolution some such as the enragés began to use the term positively in opposition to jacobin centralisation of power seeing revolutionary government as oxymoronic by the turn of the th century the english word anarchism had lost its initial negative connotation modern anarchism sprang from the secular or religious thought of the enlightenment particularly jean jacques rousseau arguments for the moral centrality of freedom as part of the political turmoil of the in the wake of the french revolution william godwin developed the first expression of modern anarchist thought godwin was according to peter kropotkin the first to formulate the political and economical conceptions of anarchism even though he did not give that name to the ideas developed in his work while godwin attached his anarchist ideas to an early edmund burke william godwin the first to formulate the political and economical conceptions of anarchism even though he did not give that name to the ideas developed in his work godwin is generally regarded as the founder of the school of thought known as philosophical anarchism he argued in political justice that government has an inherently malevolent influence on society and that it perpetuates dependency and ignorance he thought that the spread of the use of reason to the masses would eventually cause government to wither away as an unnecessary force although he did not accord the state with moral legitimacy he was against the use of revolutionary tactics for removing the government from power rather he advocated for its replacement through process of peaceful evolution his aversion to the imposition of rules based society led him to denounce as manifestation of the people mental enslavement the foundations of law property rights and even the institution of marriage he considered the basic foundations of society as constraining the natural development of individuals to use their powers of reasoning to arrive at mutually beneficial method of social organization in each case government and its institutions are shown to constrain the development of our capacity to live wholly in accordance with the full and free exercise of private judgment the french pierre joseph proudhon is regarded as the first self proclaimed anarchist label he adopted in his groundbreaking work what is property published in it is for this reason that some claim proudhon as the founder of modern anarchist theory he developed the theory of spontaneous order in society where organisation emerges without central coordinator imposing its own idea of order against the wills of individuals acting in their own interests his famous quote on the matter is liberty is the mother not the daughter of order in what is property proudhon answers with the famous accusation property is theft in this work he opposed the institution of decreed property propriété where owners have complete rights to use and abuse their property as they wish he contrasted this with what he called possession or limited ownership of resources and goods only while in more or less continuous use later however proudhon added that property is liberty and argued that it was bulwark against state power his opposition to the state organised religion and certain capitalist practices inspired subsequent anarchists and made him one of the leading social thinkers of his

time the anarcho communist joseph déjacque was the first person to describe himself as libertarian unlike pierre joseph proudhon he argued that it is not the product of his or her labour that the worker has right to but to the satisfaction of his or her needs whatever may be their nature in in germany the post hegelian philosopher max stirner published the book the ego and its own which would later be considered an influential early text of individualist anarchism french anarchists active in the revolution included anselme bellegarrigue ernest coeurderoy joseph déjacque and pierre joseph proudhon first international and the paris commune collectivist anarchist mikhail bakunin opposed the marxist aim of dictatorship of the proletariat in favour of universal rebellion and allied himself with the federalists in the first international before his expulsion by the marxists in europe harsh reaction followed the revolutions of during which ten countries had experienced brief or long term social upheaval as groups carried out nationalist uprisings after most of these attempts at systematic change ended in failure conservative elements took advantage of the divided groups of socialists anarchists liberals and nationalists to prevent further revolt in spain ramón de la sagra established the anarchist journal el porvenir in la coruña in which was inspired by proudhon ideas the catalan politician francesc pi margall became the principal translator of proudhon works into spanish and later briefly became president of spain in while being the leader of the democratic republican federal party according to george woodcock these translations were to have profound and lasting effect on the development of spanish anarchism after but before that time proudhonian ideas as interpreted by pi already provided much of the inspiration for the federalist movement which sprang up in the early according to the encyclopedia britannica during the spanish revolution of pi margall attempted to establish decentralized or cantonalist political system on proudhonian lines in the international workingmen association sometimes called the first international united diverse revolutionary currents including french followers of proudhon blanquists philadelphes english trade unionists socialists and social democrats due to its links to active workers movements the international became significant organisation karl marx became leading figure in the international and member of its general council proudhon followers the mutualists opposed marx state socialism advocating political abstentionism and small property holdings woodcock also reports that the american individualist anarchists lysander spooner and william greene had been members of the first international in following their unsuccessful participation in the league of peace and freedom lpf russian revolutionary mikhail bakunin and his collectivist anarchist associates joined the first international which had decided not to get involved with the lpf they allied themselves with the federalist socialist sections of the international who advocated the revolutionary overthrow of the state and the of property at first the collectivists worked with the marxists to push the first international in more revolutionary socialist direction subsequently the international became polarised into two camps with marx and bakunin as their respective figureheads bakunin characterised marx ideas as centralist and predicted that if marxist party came to power its leaders would simply take the place of the ruling class they had fought against anarchist historian george woodcock reports that the annual congress of the international had not taken

place in owing to the outbreak of the paris commune and in the general council called only special conference in london one delegate was able to attend from spain and none from italy while technical excuse that they had split away from the fédération romande was used to avoid inviting bakunin swiss supporters thus only tiny minority of anarchists was present and the general council resolutions passed almost unanimously most of them were clearly directed against bakunin and his followers in the conflict climaxed with final split between the two groups at the hague congress where bakunin and james guillaume were expelled from the international and its headquarters were transferred to new york in response the federalist sections formed their own international at the st imier congress adopting revolutionary anarchist program the paris commune was government that briefly ruled paris from march more formally from march to may the commune was the result of an uprising in paris after france was defeated in the franco prussian war anarchists participated actively in the establishment of the paris commune they included george woodcock states organised labour the anti authoritarian sections of the first international were the precursors of the anarcho syndicalists seeking to replace the privilege and authority of the state with the free and spontaneous organisation of labour in the federation of organized trades and labor unions fotlu of the united states and canada unanimously set may as the date by which the eight hour work day would become standard sympathetic engraving by walter crane of the executed anarchists of chicago after the haymarket affair the haymarket affair is generally considered the most significant event for the origin of international may day observances in response unions across the united states prepared general strike in support of the event on may in chicago fight broke out when strikebreakers attempted to cross the picket line and two workers died when police opened fire upon the crowd the next day may anarchists staged rally at chicago haymarket square bomb was thrown by an unknown party near the conclusion of the rally killing an officer in the ensuing panic police opened fire on the crowd and each other seven police officers and at least four workers were killed eight anarchists directly and indirectly related to the organisers of the rally were arrested and charged with the murder of the deceased officer the men became international political celebrities among the labour movement four of the men were executed and fifth committed suicide prior to his own execution the incident became known as the haymarket affair and was setback for the labour movement and the struggle for the eight hour day in second attempt this time international in scope to organise for the eight hour day was made the event also had the secondary purpose of memorializing workers killed as result of the haymarket affair although it had initially been conceived as once off event by the following year the celebration of international workers day on may day had become firmly established as an international worker holiday in the international anarchist congress of amsterdam gathered delegates from different countries among which important figures of the anarchist movement including errico malatesta pierre monatte luigi fabbri benoît broutchoux emma goldman rudolf rocker and christiaan cornelissen various themes were treated during the congress in particular concerning the organisation of the anarchist movement popular education issues the general strike or antimilitarism central debate concerned the relation between anarchism and syndicalism or trade unionism malatesta and monatte were

in particular disagreement themselves on this issue as the latter thought that syndicalism was revolutionary and would create the conditions of social revolution while malatesta did not consider syndicalism by itself sufficient he thought that the trade union movement was reformist and even conservative citing as essentially bourgeois and anti worker the phenomenon of professional union officials malatesta warned that the syndicalists aims were in perpetuating syndicalism itself whereas anarchists must always have anarchy as their end and consequently refrain from committing to any particular method of achieving it the spanish workers federation in was the first major anarcho syndicalist movement anarchist trade union federations were of special importance in spain the most successful was the confederación nacional del trabajo national confederation of labour cnt founded in before the the cnt was the major force in spanish working class politics attracting million members at one point and playing major role in the spanish civil war the cnt was affiliated with the international workers association federation of anarcho syndicalist trade unions founded in with delegates representing two million workers from countries in europe and latin america in latin america in particular the anarchists quickly became active in organizing craft and industrial workers throughout south and central america and until the early most of the trade unions in mexico brazil peru chile and argentina were anarcho syndicalist in general outlook the prestige of the spanish as revolutionary organization was undoubtedly to great extent responsible for this situation the largest and most militant of these organizations was the federación obrera regional argentina it grew quickly to membership of nearly quarter of million which dwarfed the rival unions propaganda of the deed and illegalism italian american anarchist luigi galleani his followers known as galleanists carried out series of bombings and assassination attempts from to in what they saw as attacks on tyrants and enemies of the people some anarchists such as johann most advocated publicizing violent acts of retaliation against counter revolutionaries because we preach not only action in and for itself but also action as propaganda by the people inside and outside the anarchist movement began to use the slogan propaganda of the deed to refer to individual bombings regicides and tyrannicides from onwards the russian counterparts of these anti syndicalist anarchist communists become partisans of economic terrorism and illegal expropriations illegalism as practice emerged and within it the acts of the anarchist bombers and assassins propaganda by the deed and the anarchist burglars individual reappropriation expressed their desperation and their personal violent rejection of an intolerable society moreover they were clearly meant to be exemplary invitations to revolt france bonnot gang was the most famous group to embrace illegalism however as soon as important figures in the anarchist movement distanced themselves from such individual acts peter kropotkin thus wrote that year in le révolté that structure based on centuries of history cannot be destroyed with few kilos of dynamite variety of anarchists advocated the abandonment of these sorts of tactics in favour of collective revolutionary action for example through the trade union movement the anarcho syndicalist fernand pelloutier argued in for renewed anarchist involvement in the labour movement on the basis that anarchism could do very well without the individual dynamiter state repression including the infamous french lois scélérates of the anarchist and

labour movements following the few successful bombings and assassinations may have contributed to the abandonment of these kinds of tactics although reciprocally state repression in the first place may have played role in these isolated acts the dismemberment of the french socialist movement into many groups and following the suppression of the paris commune the execution and exile of many communards to penal colonies favoured individualist political expression and acts numerous heads of state were assassinated between and by members of the anarchist movement including tsar alexander ii of russia president sadi carnot of france empress elisabeth of austria king umberto of italy president william mckinley of the united states king carlos of portugal and king george of greece mckinley assassin leon czolgosz claimed to have been influenced by anarchist and feminist emma goldman propaganda of the deed was abandoned by the vast majority of the anarchist movement after world war and the october revolution russian revolution and other uprisings of the nestor makhno with members of the anarchist revolutionary insurrectionary army of ukraine anarchists participated alongside the bolsheviks in both february and october revolutions and were initially enthusiastic about the bolshevik revolution however following political falling out with the bolsheviks by the anarchists and other left wing opposition the conflict culminated in the kronstadt rebellion which the new government repressed anarchists in central russia were either imprisoned driven underground or joined the victorious bolsheviks the anarchists from petrograd and moscow fled to ukraine there in the free territory they fought in the civil war against the whites grouping of monarchists and other opponents of the october revolution and then the bolsheviks as part of the revolutionary insurrectionary army of ukraine led by nestor makhno who established an anarchist society in the region for number of months expelled american anarchists emma goldman and alexander berkman were amongst those agitating in response to bolshevik policy and the suppression of the kronstadt uprising before they left russia both wrote accounts of their experiences in russia criticising the amount of control the bolsheviks exercised for them bakunin predictions about the consequences of marxist rule that the rulers of the new socialist marxist state would become new elite had proved all too true the victory of the bolsheviks in the october revolution and the resulting russian civil war did serious damage to anarchist movements internationally many workers and activists saw bolshevik success as setting an example communist parties grew at the expense of anarchism and other socialist movements in france and the united states for example members of the major syndicalist movements of the cgt and iww left the organisations and joined the communist international the revolutionary wave of saw the active participation of anarchists in varying degrees of protagonism in the german uprising known as the german revolution of which established the bavarian soviet republic the anarchists gustav landauer silvio gesell and erich mühsam had important leadership positions within the revolutionary councilist structures in the italian events known as the biennio rosso the anarcho syndicalist trade union unione sindacale italiana grew to members and the influence of the italian anarchist union members plus umanita nova its daily paper grew accordingly anarchists were the first to suggest occupying workplaces in the mexican revolution the mexican liberal party was established and during the early it led series of military offensives leading to

the conquest and occupation of certain towns and districts in baja california with the leadership of anarcho communist ricardo flores magón in paris the dielo truda group of russian anarchist exiles which included nestor makhno concluded that anarchists needed to develop new forms of organisation in response to the structures of bolshevism their manifesto called the organisational platform of the general union of anarchists draft was supported platformist groups active today include the workers solidarity movement in ireland and the north eastern federation of anarchist communists of north america synthesis anarchism emerged as an organisational alternative to platformism that tries to join anarchists of different tendencies under the principles of anarchism without adjectives in the this form found as its main proponents volin and sebastien faure it is the main principle behind the anarchist federations grouped around the contemporary global international of anarchist federations conflicts with european fascist regimes in the and the rise of fascism in europe transformed anarchism conflict with the state italy saw the first struggles between anarchists and fascists italian anarchists played key role in the anti fascist organisation arditi del popolo which was strongest in areas with anarchist traditions and achieved some success in their activism such as repelling blackshirts in the anarchist stronghold of parma in august the veteran italian anarchist luigi fabbri was one of the first critical theorists of fascism describing it as the preventive counter revolution in france where the far right leagues came close to insurrection in the february riots anarchists divided over united front policy anarchists in france and italy were active in the resistance during world war ii in germany the anarchist erich mühsam was arrested on charges unknown in the early morning hours of february within few hours after the reichstag fire in berlin joseph goebbels the nazi propaganda minister labelled him as one of those jewish subversives over the next seventeen months he would be imprisoned in the concentration camps at sonnenburg brandenburg and finally oranienburg on february mühsam was transferred to the concentration camp at oranienburg when finally on the night of july mühsam was tortured and murdered by the guards his battered corpse found hanging in latrine the next morning spanish revolution in spain the national anarcho syndicalist trade union confederación nacional del trabajo initially refused to join popular front electoral alliance and abstention by cnt supporters led to right wing election victory but in the cnt changed its policy and anarchist votes helped bring the popular front back to power months later the former ruling class responded with an attempted coup causing the spanish civil war in response to the army rebellion an anarchist inspired movement of peasants and workers supported by armed militias took control of barcelona and of large areas of rural spain where they collectivised the land but even before the fascist victory in the anarchists were losing ground in bitter struggle with the stalinists who controlled much of the distribution of military aid to the republican cause from the soviet union the events known as the spanish revolution was workers social revolution that began during the outbreak of the spanish civil war in and resulted in the widespread implementation of anarchist and more broadly libertarian socialist organisational principles throughout various portions of the country for two to three years primarily catalonia aragon andalusia and parts of the levante much of spain economy was put under worker control in anarchist strongholds like

catalonia the figure was as high as but lower in areas with heavy communist party of spain influence as the soviet allied party actively resisted attempts at enactment factories were run through worker committees agrarian areas became collectivised and run as libertarian communes anarchist historian sam dolgoff estimated that about eight million people participated directly or at least indirectly in the spanish revolution which he claimed came closer to realizing the ideal of the free stateless society on vast scale than any other revolution in history spanish communist party led troops suppressed the collectives and persecuted both dissident marxists and anarchists the prominent italian anarchist camillo berneri who volunteered to fight against franco was killed instead in spain by gunmen associated with the spanish communist party the city of madrid was turned over to the francoist forces by the last non francoist mayor of the city the anarchist melchor rodríguez garcía post war years anarchism sought to reorganise itself after the war and in this context the organisational debate between synthesis anarchism and platformism took importance once again especially in the anarchist movements of italy and france the mexican anarchist federation was established in after the anarchist federation of the centre united with the anarchist federation of the federal district in the early the antifascist international solidarity and the federation of anarchist groups of cuba merged into the large national organisation asociación libertaria de cuba cuban libertarian association from to the bulgarian anarchist communist federation reemerged as part of factory and workplace committee movement but was repressed by the new communist regime in in france the fédération anarchiste and the trade union confédération nationale du travail was established in the next year while the also synthesist federazione anarchica italiana was founded in italy korean anarchists formed the league of free social constructors in september and in the japanese anarchist federation was founded an international anarchist congress with delegates from across europe was held in paris in may after world war ii an appeal in the fraye arbeter shtime detailing the plight of german anarchists and called for americans to support them by february the sending of aid parcels to anarchists in germany was large scale operation the federation of libertarian socialists was founded in germany in and rudolf rocker wrote for its organ die freie gesellschaft which survived until in the uruguayan anarchist federation was founded in the anarcho communist federation of argentina renamed itself as the argentine libertarian federation the syndicalist workers federation was syndicalist group in active in post war britain and one of solidarity federation earliest predecessors it was formed in by members of the dissolved anarchist federation of britain unlike the afb which was influenced by anarcho syndicalist ideas but ultimately not syndicalist itself the swf decided to pursue more definitely syndicalist worker centred strategy from the outset anarchism continued to influence important literary and intellectual personalities of the time such as albert camus herbert read paul goodman dwight macdonald allen ginsberg george woodcock leopold kohr julian beck john cage and the french surrealist group led by andré breton which now openly embraced anarchism and collaborated in the fédération anarchiste anarcho pacifism became influential in the anti nuclear movement and anti war movements of the time as can be seen in the activism and writings of the english anarchist member of campaign for

nuclear disarmament alex comfort or the similar activism of the american catholic anarcho pacifists ammon hennacy and dorothy day anarcho pacifism became basis for critique of militarism on both sides of the cold war the resurgence of anarchist ideas during this period is well documented in robert graham anarchism documentary history of libertarian ideas volume two the emergence of the new anarchism contemporary anarchism squat near parc güell overlooking barcelona squatting was prominent part of the emergence of renewed anarchist movement from the counterculture of the and on the roof occupy and resist surge of popular interest in anarchism occurred in western nations during the and anarchism was influential in the counterculture of the and anarchists actively participated in the late sixties students and workers revolts in in carrara italy the international of anarchist federations was founded during an international anarchist conference held there in by the three existing european federations of france the fédération anarchiste the federazione anarchica italiana of italy and the iberian anarchist federation as well as the bulgarian federation in french exile in the united kingdom in the this was associated with the punk rock movement as exemplified by bands such as crass and the sex pistols the housing and employment crisis in most of western europe led to the formation of communes and squatter movements like that of barcelona spain in denmark squatters occupied disused military base and declared the freetown christiania an autonomous haven in central copenhagen since the revival of anarchism in the mid th century number of new movements and schools of thought emerged although feminist tendencies have always been part of the anarchist movement in the form of anarcha feminism they returned with vigour during the second wave of feminism in the anarchist anthropologist david graeber and anarchist historian andrej grubacic have posited rupture between generations of anarchism with those who often still have not shaken the sectarian habits of the th century contrasted with the younger activists who are much more informed among other elements by indigenous feminist ecological and cultural critical ideas and who by the turn of the st century formed by far the majority of anarchists around the turn of the st century anarchism grew in popularity and influence as part of the anti war anti capitalist and anti globalisation movements anarchists became known for their involvement in protests against the meetings of the world trade organization wto group of eight and the world economic forum some anarchist factions at these protests engaged in rioting property destruction and violent confrontations with police these actions were precipitated by ad hoc leaderless anonymous cadres known as black blocs other organisational tactics pioneered in this time include security culture affinity groups and the use of decentralised technologies such as the internet significant event of this period was the confrontations at wto conference in seattle in according to anarchist scholar simon critchley contemporary anarchism can be seen as powerful critique of the pseudo libertarianism of contemporary neo liberalism one might say that contemporary anarchism is about responsibility whether sexual ecological or socio economic it flows from an experience of conscience about the manifold ways in which the west ravages the rest it is an ethical outrage at the yawning inequality impoverishment and that is so palpable locally and globally international anarchist federations in existence include the international of anarchist federations the international workers association and international

libertarian solidarity the largest organised anarchist movement today is in spain in the form of the confederación general del trabajo cgt and the cnt cgt membership was estimated at around for other active syndicalist movements include in sweden the central organisation of the workers of sweden and the swedish anarcho syndicalist youth federation the cnt ait in france the unione sindicale italiana in italy in the us workers solidarity alliance and the uk solidarity federation and anarchist federation the revolutionary industrial unionist industrial workers of the world claiming paying members and the international workers association an anarcho syndicalist successor to the first international also remain active anarchist schools of thought portrait of philosopher pierre joseph proudhon by gustave courbet proudhon was the primary proponent of anarchist mutualism and influenced many later individualist anarchist and social anarchist thinkers anarchist schools of thought had been generally grouped in two main historical traditions individualist anarchism and social anarchism which have some different origins values and evolution the individualist wing of anarchism emphasises negative liberty opposition to state or social control over the individual while those in the social wing emphasise positive liberty to achieve one potential and argue that humans have needs that society ought to fulfill recognizing equality of entitlement in chronological and theoretical sense there are classical those created throughout the th century and post classical anarchist schools those created since the mid th century and after beyond the specific factions of anarchist thought is philosophical anarchism which embodies the theoretical stance that the state lacks moral legitimacy without accepting the imperative of revolution to eliminate it component especially of individualist anarchism philosophical anarchism may accept the existence of minimal state as unfortunate and usually temporary necessary evil but argue that citizens do not have moral obligation to obey the state when its laws conflict with individual autonomy one reaction against sectarianism within the anarchist milieu was anarchism without adjectives call for toleration first adopted by fernando tarrida del mármol in in response to the bitter debates of anarchist theory at the time in abandoning the hyphenated anarchisms collectivist communist mutualist and individualist anarchism it sought to emphasise the anti authoritarian beliefs common to all anarchist schools of thought classical anarchist schools of thought mutualism mutualism began in th century english and french labour movements before taking an anarchist form associated with pierre joseph proudhon in france and others in the united states proudhon proposed spontaneous order whereby organisation emerges without central authority positive anarchy where order arises when everybody does what he wishes and only what he wishes and where business transactions alone produce the social order it is important to recognize that proudhon distinguished between ideal political possibilities and practical governance for this reason much in contrast to some of his theoretical statements concerning ultimate spontaneous self governance proudhon was heavily involved in french parliamentary politics and allied himself not with anarchist but socialist factions of workers movements and in addition to advocating state protected charters for worker owned cooperatives promoted certain nationalization schemes during his life of public service mutualist anarchism is concerned with reciprocity free association voluntary contract federation and

credit and currency reform according to the american mutualist william batchelder greene each worker in the mutualist system would receive just and exact pay for his work services equivalent in cost being exchangeable for services equivalent in cost without profit or discount mutualism has been retrospectively characterised as ideologically situated between individualist and collectivist forms of anarchism blackwell encyclopaedia of political thought blackwell publishing isbn proudhon first characterised his goal as third form of society the synthesis of communism and property individualist anarchism individualist anarchism refers to several traditions of thought within the anarchist movement that emphasize the individual and their will over any kinds of external determinants such as groups society traditions and ideological systems individualist anarchism is not single philosophy but refers to group of individualistic philosophies that sometimes are in conflict in william godwin who has often been cited as the first anarchist wrote political justice which some consider the first expression of anarchism godwin philosophical anarchist from rationalist and utilitarian basis opposed revolutionary action and saw minimal state as present necessary evil that would become increasingly irrelevant and powerless by the gradual spread of knowledge godwin advocated individualism proposing that all cooperation in labour be eliminated on the premise that this would be most conducive with the general good th century philosopher max stirner usually considered prominent early individualist anarchist sketch by friedrich engels an influential form of individualist anarchism called egoism or egoist anarchism was expounded by one of the earliest and best known proponents of individualist anarchism the german max stirner stirner the ego and its own published in is founding text of the philosophy according to stirner the only limitation on the rights of individuals is their power to obtain what they desire without regard for god state or morality to stirner rights were spooks in the mind and he held that society does not exist but the individuals are its reality stirner advocated self assertion and foresaw unions of egoists non systematic associations continually renewed by all parties support through an act of will which stirner proposed as form of organisation in place of the state egoist anarchists argue that egoism will foster genuine and spontaneous union between individuals egoism has inspired many interpretations of stirner philosophy it was re discovered and promoted by german philosophical anarchist and lgbt activist john henry mackay josiah warren is widely regarded as the first american anarchist and the four page weekly paper he edited during the peaceful revolutionist was the first anarchist periodical published for american anarchist historian eunice minette schuster it is apparent that proudhonian anarchism was to be found in the united states at least as early as and that it was not conscious of its affinity to the individualist anarchism of josiah warren and stephen pearl andrews william greene presented this proudhonian mutualism in its purest and most systematic form henry david thoreau was an important early influence in individualist anarchist thought in the united states and europe thoreau was an american author poet naturalist tax resister development critic surveyor historian philosopher and leading he is best known for his books walden reflection upon simple living in natural surroundings and his essay civil disobedience an argument for individual resistance to civil government in moral opposition to an unjust state later

benjamin tucker fused stirner egoism with the economics of warren and proudhon in his eclectic influential publication liberty from these early influences individualist anarchism in different countries attracted small but diverse following of bohemian artists and intellectuals free love and birth control advocates see anarchism and issues related to love and sex individualist naturists nudists see anarcho naturism freethought and anti clerical activists as well as young anarchist outlaws in what became known as illegalism and individual reclamation see european individualist anarchism and individualist anarchism in france these authors and activists included oscar wilde emile armand han ryner henri zisly renzo novatore miguel gimenez igualada adolf brand and lev chernyi among others social anarchism social anarchism calls for system with common ownership of means of production and democratic control of all organisations without any government authority or coercion it is the largest school of thought in anarchism social anarchism rejects private property seeing it as source of social inequality while retaining respect for personal property and emphasises cooperation and mutual aid collectivist anarchism collectivist anarchism also referred to as revolutionary socialism or form of such is revolutionary form of anarchism commonly associated with mikhail bakunin and johann most collectivist anarchists oppose all private ownership of the means of production instead advocating that ownership be collectivised this was to be achieved through violent revolution first starting with small cohesive group through acts of violence or propaganda by the deed which would inspire the workers as whole to revolt and forcibly collectivise the means of production however was not to be extended to the distribution of income as workers would be paid according to time worked rather than receiving goods being distributed according to need as in anarcho communism this position was criticised by anarchist communists as effectively upholding the wages system collectivist anarchism arose with marxism but opposed the marxist dictatorship of the proletariat despite the stated marxist goal of collectivist stateless society anarchist communist and collectivist ideas are not mutually exclusive although the collectivist anarchists advocated compensation for labour some held out the possibility of post revolutionary transition to communist system of distribution according to need anarcho communism anarchist communism also known as anarcho communism libertarian communism and occasionally as free communism is theory of anarchism that advocates abolition of the state markets money private property while retaining respect for personal property and capitalism in favour of common ownership of the means of production direct democracy and horizontal network of voluntary associations and workers councils with production and consumption based on the guiding principle from each according to his ability to each according to his need russian theorist peter kropotkin who was influential in the development of anarchist communism some forms of anarchist communism such as insurrectionary anarchism are strongly influenced by egoism and radical individualism believing anarcho communism is the best social system for the realization of individual freedom most anarcho communists view anarcho communism as way of reconciling the opposition between the individual and society anarcho communism developed out of radical socialist currents after the french revolution but was first formulated as such in the italian section of the first international the theoretical work of peter kropotkin took importance later as

it expanded and developed pro and insurrectionary anti sections to date the best known examples of an anarchist communist society established around the ideas as they exist today and achieving worldwide attention and knowledge in the historical canon are the anarchist territories during the spanish revolution and the free territory during the russian revolution through the efforts and influence of the spanish anarchists during the spanish revolution within the spanish civil war starting in anarchist communism existed in most of aragon parts of the levante and andalusia as well as in the stronghold of anarchist catalonia before being crushed by the combined forces of the regime that won the war hitler mussolini spanish communist party repression backed by the ussr as well as economic and armaments blockades from the capitalist countries and the spanish republic itself during the russian revolution anarchists such as nestor makhno worked to create and defend through the revolutionary insurrectionary army of ukraine anarchist communism in the free territory of the ukraine from before being conquered by the bolsheviks in anarcho syndicalism may day demonstration of spanish anarcho syndicalist trade union cnt in bilbao basque country in anarcho syndicalism is branch of anarchism that focuses on the labour movement anarcho syndicalists view labour unions as potential force for revolutionary social change replacing capitalism and the state with new society democratically self managed by workers the basic principles of anarcho syndicalism are workers solidarity direct action and workers self management anarcho syndicalists believe that only direct action that is action concentrated on directly attaining goal as opposed to indirect action such as electing representative to government position will allow workers to liberate themselves moreover anarcho syndicalists believe that workers organisations the organisations that struggle against the wage system which in anarcho syndicalist theory will eventually form the basis of new society should be self managing they should not have bosses or business agents rather the workers should be able to make all the decisions that affect them themselves rudolf rocker was one of the most popular voices in the anarcho syndicalist movement he outlined view of the origins of the movement what it sought and why it was important to the future of labour in his pamphlet anarcho syndicalism the international workers association is an international anarcho syndicalist federation of various labour unions from different countries the spanish confederación nacional del trabajo played and still plays major role in the spanish labour movement it was also an important force in the spanish civil war post classical schools of thought lawrence jarach left and john zerzan right two prominent contemporary anarchist authors zerzan is known as prominent voice within anarcho primitivism while jarach is noted advocate of post left anarchy anarchism continues to generate many philosophies and movements at times eclectic drawing upon various sources and syncretic combining disparate concepts to create new philosophical approaches green anarchism or eco anarchism is school of thought within anarchism that emphasizes environmental issues with an important precedent in anarcho naturism and whose main contemporary currents are anarcho primitivism and social ecology anarcha feminism also called anarchist feminism and anarcho feminism combines anarchism with feminism it generally views patriarchy as manifestation of involuntary coercive hierarchy that should be replaced by decentralised free association anarcha feminists believe that the struggle

against patriarchy is an essential part of class struggle and the anarchist struggle against the state in essence the philosophy sees anarchist struggle as necessary component of feminist struggle and vice versa susan brown claims that as anarchism is political philosophy that opposes all relationships of power it is inherently feminist anarcha feminism began with the late th century writings of early feminist anarchists such as emma goldman and voltairine de cleyre anarcho pacifism is tendency that rejects violence in the struggle for social change see non violence it developed mostly in the netherlands britain and the united states before and during the second world war christian anarchism is movement in political theology that combines anarchism and christianity its main proponents included leo tolstoy dorothy day ammon hennacy and jacques ellul platformism is tendency within the wider anarchist movement based on the organisational theories in the tradition of dielo truda organisational platform of the general union of anarchists draft the document was based on the experiences of russian anarchists in the october revolution which led eventually to the victory of the bolsheviks over the anarchists and other groups the platform attempted to address and explain the anarchist movement failures during the russian revolution synthesis anarchism is form of anarchism that tries to join anarchists of different tendencies under the principles of anarchism without adjectives in the this form found as its main proponents the anarcho communists voline and sébastien faure it is the main principle behind the anarchist federations grouped around the contemporary global international of anarchist federations post left anarchy is recent current in anarchist thought that promotes critique of anarchism relationship to traditional left wing politics some post leftists seek to escape the confines of ideology in general also presenting critique of organisations and morality influenced by the work of max stirner and by the marxist situationist international post left anarchy is marked by focus on social insurrection and rejection of leftist social organisation insurrectionary anarchism is revolutionary theory practice and tendency within the anarchist movement which emphasizes insurrection within anarchist practice it is critical of formal organisations such as labour unions and federations that are based on political programme and periodic congresses instead insurrectionary anarchists advocate informal organisation and small affinity group based organisation insurrectionary anarchists put value in attack permanent class conflict and refusal to negotiate or compromise with class enemies post anarchism is theoretical move towards synthesis of classical anarchist theory and thought drawing from diverse ideas including post modernism autonomist marxism post left anarchy situationist international and postcolonialism left wing market anarchism strongly affirm the classical liberal ideas of self ownership and free markets while maintaining that taken to their logical conclusions these ideas support strongly anti corporatist anti hierarchical pro labor positions and anti capitalism in economics and anti imperialism in foreign policy anarcho capitalism advocates the elimination of the state in favour of individual sovereignty in free market anarcho capitalism developed from radical anti state libertarianism and individualist anarchism drawing from austrian school economics study of law and economics and public choice theory there is strong current within anarchism which does not consider that anarcho capitalism can be considered part of the anarchist movement due to

the fact that anarchism has historically been an anti capitalist movement and for definitional reasons which see anarchism incompatible with capitalist forms internal issues and debates consistent with anarchist values is controversial subject among anarchists anarchism is philosophy that embodies many diverse attitudes tendencies and schools of thought as such disagreement over questions of values ideology and tactics is common the compatibility of capitalism nationalism and religion with anarchism is widely disputed similarly anarchism enjoys complex relationships with ideologies such as marxism communism and capitalism anarchists may be motivated by humanism divine authority enlightened self interest veganism or any number of alternative ethical doctrines phenomena such as civilization technology within anarcho primitivism and the democratic process may be sharply criticised within some anarchist tendencies and simultaneously lauded in others on tactical level while propaganda of the deed was tactic used by anarchists in the th century the nihilist movement some contemporary anarchists espouse alternative direct action methods such as nonviolence counter economics and anti state cryptography to bring about an anarchist society about the scope of an anarchist society some anarchists advocate global one while others do so by local ones the diversity in anarchism has led to widely different use of identical terms among different anarchist traditions which has led to many definitional concerns in anarchist theory topics of interest intersecting and overlapping between various schools of thought certain topics of interest and internal disputes have proven perennial within anarchist theory free love french individualist anarchist emile armand who propounded the virtues of free love in the parisian anarchist milieu of the early th century an important current within anarchism is free love free love advocates sometimes traced their roots back to josiah warren and to experimental communities viewed sexual freedom as clear direct expression of an individual sovereignty free love particularly stressed women rights since most sexual laws discriminated against women for example marriage laws and anti birth control measures the most important american free love journal was lucifer the lightbearer edited by moses harman and lois waisbrooker but also there existed ezra heywood and angela heywood the word free society as the firebrand as free society was major anarchist newspaper in the united states at the end of the th and beginning of the th centuries the publication advocated free love and women rights and critiqued comstockery censorship of sexual information also lazarus was an important american individualist anarchist who promoted free love in new york city greenwich village bohemian feminists and socialists advocated self realisation and pleasure for women and also men in the here and now they encouraged playing with sexual roles and sexuality and the openly bisexual radical edna st vincent millay and the lesbian anarchist margaret anderson were prominent among them discussion groups organised by the villagers were frequented by emma goldman among others magnus hirschfeld noted in that goldman has campaigned boldly and steadfastly for individual rights and especially for those deprived of their rights thus it came about that she was the first and only woman indeed the first and only american to take up the defense of homosexual love before the general public in fact before goldman heterosexual anarchist robert reitzel spoke positively of homosexuality from the beginning of the in his detroit based german language journal der arme teufel english the

poor devil in argentina anarcha feminist virginia bolten published the newspaper called english the woman voice which was published nine times in rosario between january and january and was revived briefly in in europe the main propagandist of free love within individualist anarchism was emile armand he proposed the concept of la camaraderie amoureuse to speak of free love as the possibility of voluntary sexual encounter between consenting adults he was also consistent proponent of polyamory in germany the stirnerists adolf brand and john henry mackay were pioneering campaigners for the acceptance of male bisexuality and homosexuality mujeres libres was an anarchist women organisation in spain that aimed to empower working class women it was founded in by lucía sánchez saornil mercedes comaposada and amparo poch gascón and had approximately members the organisation was based on the idea of double struggle for women liberation and social revolution and argued that the two objectives were equally important and should be pursued in parallel in order to gain mutual support they created networks of women anarchists lucía sánchez saornil was main founder of the spanish anarcha feminist federation mujeres libres who was open about her lesbianism she was published in variety of literary journals where working under male pen name she was able to explore lesbian themes at time when homosexuality was criminalized and subject to censorship and punishment more recently the british anarcho pacifist alex comfort gained notoriety during the sexual revolution for writing the bestseller sex manual the joy of sex the issue of free love has dedicated treatment in the work of french anarcho hedonist philosopher michel onfray in such works as théorie du corps amoureux pour une érotique solaire and invention du plaisir fragments cyréaniques libertarian education and freethought francesc ferrer guàrdia catalan anarchist pedagogue and free thinker for english anarchist william godwin education was the main means by which change would be achieved godwin saw that the main goal of education should be the promotion of happiness for godwin education had to have respect for the child autonomy which precluded any form of coercion pedagogy that respected this and sought to build on the child own motivation and initiatives and concern about the child capacity to resist an ideology transmitted through the school in his political justice he criticises state sponsored schooling on account of its obvious alliance with national government early american anarchist josiah warren advanced alternative education experiences in the libertarian communities he established max stirner wrote in long essay on education called the false principle of our education in it stirner names his educational principle personalist explaining that self understanding consists in hourly self creation education for him is to create free men sovereign characters by which he means eternal characters who are therefore eternal because they form themselves each moment in the united states freethought was basically anti christian anti clerical movement whose purpose was to make the individual politically and spiritually free to decide for himself on religious matters number of contributors to liberty anarchist publication were prominent figures in both freethought and anarchism the individualist anarchist george macdonald was co editor of freethought and for time the truth seeker walker was co editor of the excellent free thought free love journal lucifer the light bearer many of the anarchists were ardent freethinkers reprints from freethought papers such as lucifer the light bearer

freethought and the truth seeker appeared in liberty the church was viewed as common ally of the state and as repressive force in and of itself in catalan anarchist and free thinker francesc ferrer guàrdia established modern or progressive schools in barcelona in defiance of an educational system controlled by the catholic church the schools stated goal was to educate the working class in rational secular and non coercive setting fiercely anti clerical ferrer believed in freedom in education education free from the authority of church and state murray bookchin wrote this period was the heyday of libertarian schools and pedagogical projects in all areas of the country where anarchists exercised some degree of influence perhaps the best known effort in this field was francisco ferrer modern school escuela moderna project which exercised considerable influence on catalan education and on experimental techniques of teaching generally la escuela moderna and ferrer ideas generally formed the inspiration for series of modern schools in the united states cuba south america and london the first of these was started in new york city in it also inspired the italian newspaper università popolare founded in russian christian anarchist leo tolstoy established school for peasant children on his estate tolstoy educational experiments were short lived due to harassment by the tsarist secret police tolstoy established conceptual difference between education and culture he thought that education is the tendency of one man to make another just like himself education is culture under restraint culture is free education is when the teaching is forced upon the pupil and when then instruction is exclusive that is when only those subjects are taught which the educator regards as necessary for him without compulsion education was transformed into culture more recent libertarian tradition on education is that of unschooling and the free school in which child led activity replaces pedagogic approaches experiments in germany led to neill founding what became summerhill school in summerhill is often cited as an example of anarchism in practice however although summerhill and other free schools are radically libertarian they differ in principle from those of ferrer by not advocating an overtly political class struggle approach in addition to organising schools according to libertarian principles anarchists have also questioned the concept of schooling per se the term deschooling was popularised by ivan illich who argued that the school as an institution is dysfunctional for self determined learning and serves the creation of consumer society instead criticisms criticisms of anarchism include moral criticisms and pragmatic criticisms anarchism is often evaluated as unfeasible or utopian by its critics european history professor carl landauer in his book european socialism argued that social anarchism is unrealistic and that government is lesser evil than society without repressive force he also argued that ill intentions will cease if repressive force disappears is an absurdity references further reading barclay harold people without government an anthropology of anarchy nd ed left bank books isbn blumenfeld jacob bottici chiara critchley simon eds the anarchist turn pluto press march isbn carter april the political theory of anarchism harper row isbn gordon uri anarchy alive london pluto press graeber david fragments of an anarchist anthropology chicago prickly paradigm press graham robert ed anarchism documentary history of libertarian ideas volume one from anarchy to anarchism ce to black rose books montréal and london isbn volume two the anarchist current black rose books montréal isbn guerin daniel

anarchism from theory to practice monthly review press isbn harper clifford anarchy graphic guide camden press an overview updating woodcock classic and illustrated throughout by harper woodcut style artwork mckay iain ed an anarchist faq volume ak press oakland edinburgh pages isbn volume ii ak press oakland edinburgh pages isbn mclaughlin paul anarchism and authority philosophical introduction to classical anarchism ashgate isbn marshall peter demanding the impossible history of anarchism pm press isbn nettlau max anarchy through the times gordon press isbn razsa maple bastards of utopia living radical politics after socialism bloomington indiana university press scott james two cheers for anarchism six easy pieces on autonomy dignity and meaningful work and play princeton nj princeton university press isbn woodcock george anarchism history of libertarian ideas and movements penguin books isbn woodcock george ed the anarchist reader fontana collins isbn an anthology of writings from anarchist thinkers and activists including proudhon kropotkin bakunin malatesta bookchin goldman and many others david van deusen the rise and fall of the green mountain anarchist collective external links an anarchist faq webpage an anarchist faq anarchism bibliography anarchist theory faq by bryan caplan anarchy archives information relating to famous anarchists including their writings see anarchy archives daily bleed anarchist encyclopedia entries with short biographies links and dedicated pages net website of the kate sharpley library containing many historical documents pertaining to anarchism infoshop org an online collection of news and information about anarchism the anarchist library large online library with texts from anarchist authors they lie we die anarchist virtual library containing books booklets and texts autism is disorder characterized by impaired social interaction verbal and non verbal communication and restricted and repetitive behavior parents usually notice signs in the first two years of their child life these signs often develop gradually though some children with autism reach their developmental milestones at normal pace and then regress the diagnostic criteria require that symptoms become apparent in early childhood typically before age three while autism is highly heritable researchers suspect both environmental and genetic factors as causes in rare cases autism is strongly associated with agents that cause birth defects controversies surround other proposed environmental causes for example the vaccine hypotheses have been disproven autism affects information processing in the brain by altering how nerve cells and their synapses connect and organize how this occurs is not well understood it is one of three recognized disorders in the autism spectrum asds the other two being asperger syndrome which lacks delays in cognitive development and language and pervasive developmental disorder not otherwise specified commonly abbreviated as pdd nos which is diagnosed when the full set of criteria for autism or asperger syndrome are not met early speech or behavioral interventions can help children with autism gain self care social and communication skills although there is no known cure there have been reported cases of children who recovered not many children with autism live independently after reaching adulthood though some become successful an autistic culture has developed with some individuals seeking cure and others believing autism should be accepted as difference and not treated as disorder globally autism is estimated to affect million people as of as of the number of people affected is estimated at about per worldwide it

occurs four to five times more often in boys than girls about of children in the united states one in are diagnosed with asd increase from one in in the rate of autism among adults aged years and over in the united kingdom is the number of people diagnosed has been increasing dramatically since the partly due to changes in diagnostic practice and government subsidized financial incentives for named diagnoses the question of whether actual rates have increased is unresolved characteristics autism spectrum disorder video autism is highly variable disorder that first appears during infancy or childhood and generally follows steady course without remission people with autism may be severely impaired in some respects but normal or even superior in others overt symptoms gradually begin after the age of six months become established by age two or three years and tend to continue through adulthood although often in more muted form it is distinguished not by single symptom but by characteristic triad of symptoms impairments in social interaction impairments in communication and restricted interests and repetitive behavior other aspects such as atypical eating are also common but are not essential for diagnosis autism individual symptoms occur in the general population and appear not to associate highly without sharp line separating pathologically severe from common traits social development social deficits distinguish autism and the related autism spectrum disorders asd see classification from other developmental disorders people with autism have social impairments and often lack the intuition about others that many people take for granted noted autistic temple grandin described her inability to understand the social communication of neurotypicals or people with normal neural development as leaving her feeling like an anthropologist on mars unusual social development becomes apparent early in childhood autistic infants show less attention to social stimuli smile and look at others less often and respond less to their own name autistic toddlers differ more strikingly from social norms for example they have less eye contact and turn taking and do not have the ability to use simple movements to express themselves such as pointing at things three to five year old children with autism are less likely to exhibit social understanding approach others spontaneously imitate and respond to emotions communicate nonverbally and take turns with others however they do form attachments to their primary caregivers most children with autism display moderately less attachment security than neurotypical children although this difference disappears in children with higher mental development or less severe asd older children and adults with asd perform worse on tests of face and emotion recognition although this may be partly due to lower ability to define person own emotions children with high functioning autism suffer from more intense and frequent loneliness compared to non autistic peers despite the common belief that children with autism prefer to be alone making and maintaining friendships often proves to be difficult for those with autism for them the quality of friendships not the number of friends predicts how lonely they feel functional friendships such as those resulting in invitations to parties may affect the quality of life more deeply there are many anecdotal reports but few systematic studies of aggression and violence in individuals with asd the limited data suggest that in children with intellectual disability autism is associated with aggression destruction of property and tantrums communication about third to half of individuals with autism do not develop

enough natural speech to meet their daily communication needs differences in communication may be present from the first year of life and may include delayed onset of babbling unusual gestures diminished responsiveness and vocal patterns that are not synchronized with the caregiver in the second and third years children with autism have less frequent and less diverse babbling consonants words and word combinations their gestures are less often integrated with words children with autism are less likely to make requests or share experiences and are more likely to simply repeat others words echolalia or reverse pronouns joint attention seems to be necessary for functional speech and deficits in joint attention seem to distinguish infants with asd for example they may look at pointing hand instead of the pointed at object and they consistently fail to point at objects in order to comment on or share an experience children with autism may have difficulty with imaginative play and with developing symbols into language in pair of studies high functioning children with autism aged performed equally well as and adults better than individually matched controls at basic language tasks involving vocabulary and spelling both autistic groups performed worse than controls at complex language tasks such as figurative language comprehension and inference as people are often sized up initially from their basic language skills these studies suggest that people speaking to autistic individuals are more likely to overestimate what their audience comprehends repetitive behavior autistic individuals display many forms of repetitive or restricted behavior which the repetitive behavior scale revised rbs categorizes as follows young boy with autism who has arranged his toys in row stereotypy is repetitive movement such as hand flapping head rolling or body rocking compulsive behavior is intended and appears to follow rules such as arranging objects in stacks or lines sameness is resistance to change for example insisting that the furniture not be moved or refusing to be interrupted ritualistic behavior involves an unvarying pattern of daily activities such as an unchanging menu or dressing ritual this is closely associated with sameness and an independent validation has suggested combining the two factors restricted behavior is limited in focus interest or activity such as preoccupation with single television program toy or game self injury includes movements that injure or can injure the person such as eye poking skin picking hand biting and head banging no single repetitive or self injurious behavior seems to be specific to autism but autism appears to have an elevated pattern of occurrence and severity of these behaviors other symptoms autistic individuals may have symptoms that are independent of the diagnosis but that can affect the individual or the family an estimated to of individuals with asd show unusual abilities ranging from splinter skills such as the memorization of trivia to the extraordinarily rare talents of prodigious autistic savants many individuals with asd show superior skills in perception and attention relative to the general population sensory abnormalities are found in over of those with autism and are considered core features by some although there is no good evidence that sensory symptoms differentiate autism from other developmental disorders differences are greater for under responsivity for example walking into things than for over responsivity for example distress from loud noises or for sensation seeking for example rhythmic movements an estimated of autistic people have motor signs that include poor muscle tone poor motor planning and toe walking deficits in motor

coordination are pervasive across asd and are greater in autism proper unusual eating behavior occurs in about three quarters of children with asd to the extent that it was formerly diagnostic indicator selectivity is the most common problem although eating rituals and food refusal also occur this does not appear to result in malnutrition although some children with autism also have symptoms there is lack of published rigorous data to support the theory that children with autism have more or different symptoms than usual studies report conflicting results and the relationship between problems and asd is unclear parents of children with asd have higher levels of stress siblings of children with asd report greater admiration of and less conflict with the affected sibling than siblings of unaffected children and were similar to siblings of children with down syndrome in these aspects of the sibling relationship however they reported lower levels of closeness and intimacy than siblings of children with down syndrome siblings of individuals with asd have greater risk of negative well being and poorer sibling relationships as adults causes it has long been presumed that there is common cause at the genetic cognitive and neural levels for autism characteristic triad of symptoms however there is increasing suspicion that autism is instead complex disorder whose core aspects have distinct causes that often co occur deletion duplication and inversion are all chromosome abnormalities that have been implicated in autism autism has strong genetic basis although the genetics of autism are complex and it is unclear whether asd is explained more by rare mutations with major effects or by rare multigene interactions of common genetic variants complexity arises due to interactions among multiple genes the environment and epigenetic factors which do not change dna sequencing but are heritable and influence gene expression many genes have been associated with autism through sequencing the genomes of affected individuals and their parents studies of twins suggest that heritability is for autism and as high as for asd and siblings of those with autism are about times more likely to be autistic than the general population however most of the mutations that increase autism risk have not been identified typically autism cannot be traced to mendelian single gene mutation or to single chromosome abnormality and none of the genetic syndromes associated with asds have been shown to selectively cause asd numerous candidate genes have been located with only small effects attributable to any particular gene most loci individually explain less than of cases of autism the large number of autistic individuals with unaffected family members may result from spontaneous structural variation such as deletions duplications or inversions in genetic material during meiosis hence substantial fraction of autism cases may be traceable to genetic causes that are highly heritable but not inherited that is the mutation that causes the autism is not present in the parental genome several lines of evidence point to synaptic dysfunction as cause of autism some rare mutations may lead to autism by disrupting some synaptic pathways such as those involved with cell adhesion gene replacement studies in mice suggest that autistic symptoms are closely related to later developmental steps that depend on activity in synapses and on activity dependent changes all known teratogens agents that cause birth defects related to the risk of autism appear to act during the first eight weeks from conception and though this does not exclude the possibility that autism can be initiated or affected later there is strong

evidence that autism arises very early in development exposure to air pollution during pregnancy especially heavy metals and particulates may increase the risk of autism although no links have been found and some have been completely disproven environmental factors that have been claimed to contribute to or exacerbate autism include certain foods infectious diseases solvents diesel exhaust pcbs phthalates and phenols used in plastic products pesticides brominated flame retardants alcohol smoking illicit drugs vaccines and prenatal stress parents may first become aware of autistic symptoms in their child around the time of routine vaccination this has led to unsupported theories blaming vaccine overload vaccine preservative or the mmr vaccine for causing autism the latter theory was supported by litigation funded study that has since been shown to have been an elaborate fraud although these theories lack convincing scientific evidence and are biologically implausible parental concern about potential vaccine link with autism has led to lower rates of childhood immunizations outbreaks of previously controlled childhood diseases in some countries and the preventable deaths of several children mechanism autism symptoms result from maturation related changes in various systems of the brain how autism occurs is not well understood its mechanism can be divided into two areas the pathophysiology of brain structures and processes associated with autism and the linkages between brain structures and behaviors the behaviors appear to have multiple pathophysiology autism affects the amygdala cerebellum and many other parts of the brain unlike many other brain disorders such as parkinson autism does not have clear unifying mechanism at either the molecular cellular or systems level it is not known whether autism is few disorders caused by mutations converging on few common molecular pathways or is like intellectual disability large set of disorders with diverse mechanisms autism appears to result from developmental factors that affect many or all functional brain systems and to disturb the timing of brain development more than the final product neuroanatomical studies and the associations with teratogens strongly suggest that autism mechanism includes alteration of brain development soon after conception this anomaly appears to start cascade of pathological events in the brain that are significantly influenced by environmental factors just after birth the brains of children with autism tend to grow faster than usual followed by normal or relatively slower growth in childhood it is not known whether early overgrowth occurs in all children with autism it seems to be most prominent in brain areas underlying the development of higher cognitive specialization hypotheses for the cellular and molecular bases of pathological early overgrowth include the following an excess of neurons that causes local in key brain regions disturbed neuronal migration during early gestation unbalanced excitatory inhibitory networks abnormal formation of synapses and dendritic spines for example by modulation of the neurexin neuroligin cell adhesion system or by poorly regulated synthesis of synaptic proteins disrupted synaptic development may also contribute to epilepsy which may explain why the two conditions are associated the immune system is thought to play an important role in autism children with autism have been found by researchers to have inflammation of both the peripheral and central immune systems as indicated by increased levels of pro inflammatory cytokines and significant activation of microglia biomarkers of abnormal immune function have also been associated with

increased impairments in behaviors that are characteristic of the core features of autism such as deficits in social interactions and communication interactions between the immune system and the nervous system begin early during the embryonic stage of life and successful depends on balanced immune response it is thought that activation of pregnant mother immune system such as from environmental toxicants or infection can contribute to causing autism through causing disruption of brain development this is supported by recent studies that have found that infection during pregnancy is associated with an increased risk of autism the relationship of neurochemicals to autism is not well understood several have been investigated with the most evidence for the role of serotonin and of genetic differences in its transport the role of group metabotropic glutamate receptors mglur in the pathogenesis of fragile syndrome the most common identified genetic cause of autism has led to interest in the possible implications for future autism research into this pathway some data suggests neuronal overgrowth potentially related to an increase in several growth hormones or to impaired regulation of growth factor receptors also some inborn errors of metabolism are associated with autism but probably account for less than of cases the mirror neuron system mns theory of autism hypothesizes that distortion in the development of the mns interferes with imitation and leads to autism core features of social impairment and communication difficulties the mns operates when an animal performs an action or observes another animal perform the same action the mns may contribute to an individual understanding of other people by enabling the modeling of their behavior via embodied simulation of their actions intentions and emotions several studies have tested this hypothesis by demonstrating structural abnormalities in mns regions of individuals with asd delay in the activation in the core circuit for imitation in individuals with asperger syndrome and correlation between reduced mns activity and severity of the syndrome in children with asd however individuals with autism also have abnormal brain activation in many circuits outside the mns and the mns theory does not explain the normal performance of children with autism on imitation tasks that involve goal or object autistic individuals tend to use different areas of the brain yellow for movement task compared to control group blue asd related patterns of low function and aberrant activation in the brain differ depending on whether the brain is doing social or nonsocial tasks in autism there is evidence for reduced functional connectivity of the default network large scale brain network involved in social and emotional processing with intact connectivity of the task positive network used in sustained attention and goal directed thinking in people with autism the two networks are not negatively correlated in time suggesting an imbalance in toggling between the two networks possibly reflecting disturbance of self referential thought the theory of autism hypothesizes that autism is marked by high level neural connections and synchronization along with an excess of low level processes evidence for this theory has been found in functional neuroimaging studies on autistic individuals and by brainwave study that suggested that adults with asd have local in the cortex and weak functional connections between the frontal lobe and the rest of the cortex other evidence suggests the is mainly within each hemisphere of the cortex and that autism is disorder of the association cortex from studies based on event related potentials transient changes to the

brain electrical activity in response to stimuli there is considerable evidence for differences in autistic individuals with respect to attention orientation to auditory and visual stimuli novelty detection language and face processing and information storage several studies have found preference for nonsocial stimuli for example studies have found evidence in children with autism of delayed responses in the brain processing of auditory signals in the genetic area relations have been found between autism and schizophrenia based on duplications and deletions of chromosomes research showed that schizophrenia and autism are significantly more common in combination with deletion syndrome research on autism schizophrenia relations for chromosome chromosome and chromosome are inconclusive functional connectivity studies have found both hypo and hyper connectivity in brains of people with autism hypo connectivity seems to dominate especially for and cortico cortical functional connectivity neuropsychology two major categories of cognitive theories have been proposed about the links between autistic brains and behavior the first category focuses on deficits in social cognition simon baron cohen empathizing systemizing theory postulates that autistic individuals can systemize that is they can develop internal rules of operation to handle events inside the brain but are less effective at empathizing by handling events generated by other agents an extension the extreme male brain theory hypothesizes that autism is an extreme case of the male brain defined as individuals in whom systemizing is better than empathizing these theories are somewhat related to baron cohen earlier theory of mind approach which hypothesizes that autistic behavior arises from an inability to ascribe mental states to oneself and others the theory of mind hypothesis is supported by the atypical responses of children with autism to the sally anne test for reasoning about others motivations and the mirror neuron system theory of autism described in pathophysiology maps well to the hypothesis however most studies have found no evidence of impairment in autistic individuals ability to understand other people basic intentions or goals instead data suggests that impairments are found in understanding more complex social emotions or in considering others viewpoints the second category focuses on nonsocial or general processing the executive functions such as working memory planning inhibition in his review kenworthy states that the claim of executive dysfunction as causal factor in autism is controversial however it is clear that executive dysfunction plays role in the social and cognitive deficits observed in individuals with autism tests of core executive processes such as eye movement tasks indicate improvement from late childhood to adolescence but performance never reaches typical adult levels strength of the theory is predicting stereotyped behavior and narrow interests two weaknesses are that executive function is hard to measure and that executive function deficits have not been found in young children with autism weak central coherence theory hypothesizes that limited ability to see the big picture underlies the central disturbance in autism one strength of this theory is predicting special talents and peaks in performance in autistic people related theory enhanced perceptual functioning focuses more on the superiority of locally oriented and perceptual operations in autistic individuals these theories map well from the theory of autism neither category is satisfactory on its own social cognition theories poorly address autism rigid and repetitive behaviors while the nonsocial

theories have difficulty explaining social impairment and communication difficulties combined theory based on multiple deficits may prove to be more useful diagnosis diagnosis is based on behavior not cause or mechanism under the dsm autism is characterized by persistent deficits in social communication and interaction across multiple contexts as well as restricted repetitive patterns of behavior interests or activities these deficits are present in early childhood typically before age three and lead to clinically significant functional impairment sample symptoms include lack of social or emotional reciprocity stereotyped and repetitive use of language or idiosyncratic language and persistent preoccupation with unusual objects the disturbance must not be better accounted for by rett syndrome intellectual disability or global developmental delay icd uses essentially the same definition several diagnostic instruments are available two are commonly used in autism research the autism diagnostic interview revised adi is semistructured parent interview and the autism diagnostic observation schedule ados uses observation and interaction with the child the childhood autism rating scale cars is used widely in clinical environments to assess severity of autism based on observation of children pediatrician commonly performs preliminary investigation by taking developmental history and physically examining the child if warranted diagnosis and evaluations are conducted with help from asd specialists observing and assessing cognitive communication family and other factors using standardized tools and taking into account any associated medical conditions pediatric is often asked to assess behavior and cognitive skills both to aid diagnosis and to help recommend educational interventions differential diagnosis for asd at this stage might also consider intellectual disability hearing impairment and specific language impairment such as landau kleffner syndrome the presence of autism can make it harder to diagnose coexisting psychiatric disorders such as depression clinical genetics evaluations are often done once asd is diagnosed particularly when other symptoms already suggest genetic cause although genetic technology allows clinical geneticists to link an estimated of cases to genetic causes consensus guidelines in the us and uk are limited to high resolution chromosome and fragile testing genotype first model of diagnosis has been proposed which would routinely assess the genome copy number variations as new genetic tests are developed several ethical legal and social issues will emerge commercial availability of tests may precede adequate understanding of how to use test results given the complexity of autism genetics metabolic and neuroimaging tests are sometimes helpful but are not routine asd can sometimes be diagnosed by age months although diagnosis becomes increasingly stable over the first three years of life for example one year old who meets diagnostic criteria for asd is less likely than three year old to continue to do so few years later in the uk the national autism plan for children recommends at most weeks from first concern to completed diagnosis and assessment though few cases are handled that quickly in practice although the symptoms of autism and asd begin early in childhood they are sometimes missed years later adults may seek diagnoses to help them or their friends and family understand themselves to help their employers make adjustments or in some locations to claim disability living allowances or other benefits underdiagnosis and overdiagnosis are problems in marginal cases and much of the recent increase in the number of reported asd cases is likely due to

changes in diagnostic practices the increasing popularity of drug treatment options and the expansion of benefits has given providers incentives to diagnose asd resulting in some overdiagnosis of children with uncertain symptoms conversely the cost of screening and diagnosis and the challenge of obtaining payment can inhibit or delay diagnosis it is particularly hard to diagnose autism among the visually impaired partly because some of its diagnostic criteria depend on vision and partly because autistic symptoms overlap with those of common blindness syndromes or blindisms classification autism is one of the five pervasive developmental disorders pdd which are characterized by widespread abnormalities of social interactions and communication and severely restricted interests and highly repetitive behavior these symptoms do not imply sickness fragility or emotional disturbance of the five pdd forms asperger syndrome is closest to autism in signs and likely causes rett syndrome and childhood disintegrative disorder share several signs with autism but may have unrelated causes pdd not otherwise specified pdd nos also called atypical autism is diagnosed when the criteria are not met for more specific disorder unlike with autism people with asperger syndrome have no substantial delay in language development the terminology of autism can be bewildering with autism asperger syndrome and pdd nos often called the autism spectrum disorders asd or sometimes the autistic disorders whereas autism itself is often called autistic disorder childhood autism or infantile autism in this article autism refers to the classic autistic disorder in clinical practice though autism asd and pdd are often used interchangeably asd in turn is subset of the broader autism phenotype which describes individuals who may not have asd but do have autistic like traits such as avoiding eye contact the manifestations of autism cover wide spectrum ranging from individuals with severe impairments who may be silent developmentally disabled and locked into hand flapping and rocking to high functioning individuals who may have active but distinctly odd social approaches narrowly focused interests and verbose pedantic communication because the behavior spectrum is continuous boundaries between diagnostic categories are necessarily somewhat arbitrary sometimes the syndrome is divided into low medium or high functioning autism lfa mfa and hfa based on iq thresholds or on how much support the individual requires in daily life these subdivisions are not standardized and are controversial autism can also be divided into syndromal and non syndromal autism the syndromal autism is associated with severe or profound intellectual disability or congenital syndrome with physical symptoms such as tuberous sclerosis although individuals with asperger syndrome tend to perform better cognitively than those with autism the extent of the overlap between asperger syndrome hfa and non syndromal autism is unclear some studies have reported diagnoses of autism in children due to loss of language or social skills as opposed to failure to make progress typically from to months of age the validity of this distinction remains controversial it is possible that regressive autism is specific subtype or that there is continuum of behaviors between autism with and without regression research into causes has been hampered by the inability to identify biologically meaningful subgroups within the autistic population and by the traditional boundaries between the disciplines of psychiatry psychology neurology and pediatrics newer technologies such as fmri and diffusion tensor imaging can help identify biologically

relevant phenotypes observable traits that can be viewed on brain scans to help further neurogenetic studies of autism one example is lowered activity in the fusiform face area of the brain which is associated with impaired perception of people versus objects it has been proposed to classify autism using genetics as well as behavior screening about half of parents of children with asd notice their child unusual behaviors by age months and about four fifths notice by age months according to an article failure to meet any of the following milestones is an absolute indication to proceed with further evaluations delay in referral for such testing may delay early diagnosis and treatment and affect the long term outcome no babbling by months no gesturing pointing waving etc by months no single words by months no two word spontaneous not just echolalic phrases by months any loss of any language or social skills at any age the united states preventative services task force in found it was unclear if screening was beneficial or harmful among children in whom there is no concerns the japanese practice is to screen all children for asd at and months using autism specific formal screening tests in contrast in the uk children whose families or doctors recognize possible signs of autism are screened it is not known which approach is more effective screening tools include the modified checklist for autism in toddlers chat the early screening of autistic traits questionnaire and the first year inventory initial data on chat and its predecessor the checklist for autism in toddlers chat on children aged months suggests that it is best used in clinical setting and that it has low sensitivity many false negatives but good specificity few false positives it may be more accurate to precede these tests with broadband screener that does not distinguish asd from other developmental disorders screening tools designed for one culture norms for behaviors like eye contact may be inappropriate for different culture although genetic screening for autism is generally still impractical it can be considered in some cases such as children with neurological symptoms and dysmorphic features prevention infection with rubella during pregnancy causes fewer than of cases of autism vaccination against rubella can prevent many of those cases management three year old with autism points to fish in an aquarium as part of an experiment on the effect of intensive shared attention training on language development the main goals when treating children with autism are to lessen associated deficits and family distress and to increase quality of life and functional independence in general higher iqs are correlated with greater responsiveness to treatment and improved treatment outcomes no single treatment is best and treatment is typically tailored to the child needs families and the educational system are the main resources for treatment studies of interventions have methodological problems that prevent definitive conclusions about efficacy however the development of evidence based interventions has advanced in recent years although many psychosocial interventions have some positive evidence suggesting that some form of treatment is preferable to no treatment the methodological quality of systematic reviews of these studies has generally been poor their clinical results are mostly tentative and there is little evidence for the relative effectiveness of treatment options intensive sustained special education programs and behavior therapy early in life can help children acquire self care social and job skills and often improve functioning and decrease symptom severity and maladaptive behaviors claims that intervention by around

age three years is crucial are not substantiated available approaches include applied behavior analysis aba developmental models structured teaching speech and language therapy social skills therapy and occupational therapy among these approaches interventions either treat autistic features comprehensively or focalize treatment on specific area of deficit there is some evidence that early intensive behavioral intervention eibi an early intervention model based on aba for to hours week for multiple years is an effective treatment for some children with asd two theoretical frameworks outlined for early childhood intervention include applied behavioral analysis aba and developmental social pragmatic models dsp one interventional strategy utilizes parent training model which teaches parents how to implement various aba and dsp techniques allowing for parents to disseminate interventions themselves various dsp programs have been developed to explicitly deliver intervention systems through at home parent implementation despite the recent development of parent training models these interventions have demonstrated effectiveness in numerous studies being evaluated as probable efficacious mode of treatment education educational interventions can be effective to varying degrees in most children intensive aba treatment has demonstrated effectiveness in enhancing global functioning in preschool children and is well established for improving intellectual performance of young children similarly teacher implemented intervention that utilizes an aba combined with developmental social pragmatic approach has been found to be well established treatment in improving social communication skills in young children although there is less evidence in its treatment of global symptoms reports are often poorly communicated to educators resulting in gap between what report recommends and what education is provided it is not known whether treatment programs for children lead to significant improvements after the children grow up and the limited research on the effectiveness of adult residential programs shows mixed results the appropriateness of including children with varying severity of autism spectrum disorders in the general education population is subject of current debate among educators and researchers medication many medications are used to treat asd symptoms that interfere with integrating child into home or school when behavioral treatment fails more than half of us children diagnosed with asd are prescribed psychoactive drugs or anticonvulsants with the most common drug classes being antidepressants stimulants and antipsychotics antipsychotics such as risperidone and aripiprazole have been found to be useful for treating irritability repetitive behavior and sleeplessness that often occurs with autism however their side effects must be weighed against their potential benefits and people with autism may respond atypically there is scant reliable research about the effectiveness or safety of drug treatments for adolescents and adults with asd no known medication relieves autism core symptoms of social and communication impairments experiments in mice have reversed or reduced some symptoms related to autism by replacing or modulating gene function suggesting the possibility of targeting therapies to specific rare mutations known to cause autism alternative medicine although many alternative therapies and interventions are available few are supported by scientific studies treatment approaches have little empirical support in quality of life contexts and many programs focus on success measures that lack predictive validity and real world relevance scientific evidence

appears to matter less to service providers than program marketing training availability and parent requests some alternative treatments may place the child at risk study found that compared to their peers autistic boys have significantly thinner bones if on casein free diets in botched chelation therapy killed five year old child with autism there has been early research looking at hyperbaric treatments in children with autism although popularly used as an alternative treatment for people with autism there is no good evidence that gluten free diet is of benefit in the subset of people who have gluten sensitivity there is limited evidence that suggests that gluten free diet may improve some autistic behaviours cost treatment is expensive indirect costs are more so for someone born in us study estimated an average lifetime cost of net present value in dollars inflation adjusted from estimate with about medical care extra education and other care and lost economic productivity publicly supported programs are often inadequate or inappropriate for given child and unreimbursed out of pocket medical or therapy expenses are associated with likelihood of family financial problems one us study found average loss of annual income in families of children with asd and related study found that asd is associated with higher probability that child care problems will greatly affect parental employment us states increasingly require private health insurance to cover autism services shifting costs from publicly funded education programs to privately funded health insurance after childhood key treatment issues include residential care job training and placement sexuality social skills and estate planning society and culture the rainbow colored infinity is often used as symbol for the diversity of the autism spectrum as well as neurodiversity in general the emergence of the autism rights movement has served as an attempt to encourage people to be more tolerant of those with autism through this movement people hope to cause others to think of autism as difference instead of disease proponents of this movement wish to seek acceptance not cures there have also been many worldwide events promoting autism awareness such as world autism awareness day light it up blue autism sunday autistic pride day autreat and others there have also been many organizations dedicated to increasing the awareness of autism and the effects that autism has on someone life these organizations include autism speaks autism national committee autism society of america and many others social science scholars have had an increased focused on studying those with autism in hopes to learn more about autism as culture transcultural comparisons and research on social movements media has had an influence on how the public perceives those with autism rain man film that won oscars depicts character with autism who has incredible talents and abilities while many autistics don have these special abilities there are some autistic individuals who have been successful in their fields prognosis there is no known cure children recover occasionally so that they lose their diagnosis of asd this occurs sometimes after intensive treatment and sometimes not it is not known how often recovery happens reported rates in unselected samples of children with asd have ranged from to most children with autism acquire language by age five or younger though few have developed communication skills in later years most children with autism lack social support meaningful relationships future employment opportunities or self determination although core difficulties tend to persist symptoms often become

less severe with age few high quality studies address long term prognosis some adults show modest improvement in communication skills but few decline no study has focused on autism after midlife acquiring language before age six having an iq above and having marketable skill all predict better outcomes independent living is unlikely with severe autism most people with autism face significant obstacles in transitioning to adulthood epidemiology reports of autism cases per children grew dramatically in the us from to it is unknown how much if any growth came from changes in rates of autism most recent reviews tend to estimate prevalence of per for autism and close to per for asd and per children in the united states for asd as of because of inadequate data these numbers may underestimate asd true rate globally autism affects an estimated million people as of while asperger syndrome affects further million in the nhs estimated that the overall prevalence of autism among adults aged years and over in the uk was rates of pdd nos has been estimated at per asperger syndrome at roughly per and childhood disintegrative disorder at per cdc most recent estimate is that out of every children or per has an asd as of the number of reported cases of autism increased dramatically in the and early this increase is largely attributable to changes in diagnostic practices referral patterns availability of services age at diagnosis and public awareness though unidentified environmental risk factors cannot be ruled out the available evidence does not rule out the possibility that autism true prevalence has increased real increase would suggest directing more attention and funding toward changing environmental factors instead of continuing to focus on genetics boys are at higher risk for asd than girls the sex ratio averages and is greatly modified by cognitive impairment it may be close to with intellectual disability and more than without several theories about the higher prevalence in males have been investigated but the cause of the difference is unconfirmed one theory is that females are underdiagnosed although the evidence does not implicate any single pregnancy related risk factor as cause of autism the risk of autism is associated with advanced age in either parent and with diabetes bleeding and use of psychiatric drugs in the mother during pregnancy the risk is greater with older fathers than with older mothers two potential explanations are the known increase in mutation burden in older sperm and the hypothesis that men marry later if they carry genetic liability and show some signs of autism most professionals believe that race ethnicity and socioeconomic background do not affect the occurrence of autism several other conditions are common in children with autism they include genetic disorders about of autism cases have an identifiable mendelian single gene condition chromosome abnormality or other genetic syndrome and asd is associated with several genetic disorders intellectual disability the percentage of autistic individuals who also meet criteria for intellectual disability has been reported as anywhere from to wide variation illustrating the difficulty of assessing autistic intelligence in comparison for pdd nos the association with intellectual disability is much weaker and by definition the diagnosis of asperger excludes intellectual disability anxiety disorders are common among children with asd there are no firm data but studies have reported prevalences ranging from to many anxiety disorders have symptoms that are better explained by asd itself or are hard to distinguish from asd symptoms epilepsy with variations in risk of epilepsy due to age cognitive level and type of language

disorder several metabolic defects such as phenylketonuria are associated with autistic symptoms minor physical anomalies are significantly increased in the autistic population preempted diagnoses although the dsm iv rules out concurrent diagnosis of many other conditions along with autism the full criteria for attention deficit hyperactivity disorder adhd tourette syndrome and other of these conditions are often present and these comorbid diagnoses are increasingly accepted sleep problems affect about two thirds of individuals with asd at some point in childhood these most commonly include symptoms of insomnia such as difficulty in falling asleep frequent nocturnal awakenings and early morning awakenings sleep problems are associated with difficult behaviors and family stress and are often focus of clinical attention over and above the primary asd diagnosis history leo kanner introduced the label early infantile autism in few examples of autistic symptoms and treatments were described long before autism was named the table talk of martin luther compiled by his notetaker mathesius contains the story of year old boy who may have been severely autistic luther reportedly thought the boy was soulless mass of flesh possessed by the devil and suggested that he be suffocated although later critic has cast doubt on the veracity of this report the earliest well documented case of autism is that of hugh blair of borgue as detailed in court case in which his brother successfully petitioned to annul blair marriage to gain blair inheritance the wild boy of aveyron feral child caught in showed several signs of autism the medical student jean itard treated him with behavioral program designed to help him form social attachments and to induce speech via imitation the new latin word autismus english translation autism was coined by the swiss psychiatrist eugen bleuler in as he was defining symptoms of schizophrenia he derived it from the greek word autós meaning self and used it to mean morbid self admiration referring to autistic withdrawal of the patient to his fantasies against which any influence from outside becomes an intolerable disturbance the word autism first took its modern sense in when hans asperger of the vienna university hospital adopted bleuler terminology autistic psychopaths in lecture in german about child psychology asperger was investigating an asd now known as asperger syndrome though for various reasons it was not widely recognized as separate diagnosis until leo kanner of the johns hopkins hospital first used autism in its modern sense in english when he introduced the label early infantile autism in report of children with striking behavioral similarities almost all the characteristics described in kanner first paper on the subject notably autistic aloneness and insistence on sameness are still regarded as typical of the autistic spectrum of disorders it is not known whether kanner derived the term independently of asperger kanner reuse of autism led to decades of confused terminology like infantile schizophrenia and child psychiatry focus on maternal deprivation led to misconceptions of autism as an infant response to refrigerator mothers starting in the late autism was established as separate syndrome by demonstrating that it is lifelong distinguishing it from intellectual disability and schizophrenia and from other developmental disorders and demonstrating the benefits of involving parents in active programs of therapy as late as the mid there was little evidence of genetic role in autism now it is thought to be one of the most heritable of all psychiatric conditions although the rise of parent organizations and the of childhood asd have deeply affected how we view asd

parents continue to feel social stigma in situations where their child autistic behavior is perceived negatively by others and many primary care physicians and medical specialists still express some beliefs consistent with outdated autism research the internet has helped autistic individuals bypass nonverbal cues and emotional sharing that they find so hard to deal with and has given them way to form online communities and work remotely sociological and cultural aspects of autism have developed some in the community seek cure while others believe that autism is simply another way of being references further reading external links

[8]:
```
# import gensim.utils as utils
from smart_open import smart_open
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import STOPWORDS
from gensim.corpora.wikicorpus import _extract_pages, filter_wiki

def tokenize(text):
    return [token for token in simple_preprocess(text) if token not in␣
 ↪STOPWORDS]

def iter_wiki(dump_file):
    """Yield each article from the Wikipedia dump, as a `(title, tokens)`␣
 ↪2-tuple."""
    ignore_namespaces = 'Wikipedia Category File Portal Template MediaWiki User␣
 ↪Help Book Draft'.split()
    for title, text, pageid in _extract_pages(smart_open(dump_file)):
        text = filter_wiki(text)
        tokens = tokenize(text)
        if len(tokens) < 50 or any(title.startswith(ns + ':') for ns in␣
 ↪ignore_namespaces):
            continue  # ignore short articles and various meta-articles
        yield title, tokens
```

[9]:
```
# only use simplewiki in this tutorial (fewer documents)
# the full wiki dump is exactly the same format, but larger
wiki_file = './data/simplewiki-latest-pages-articles.xml.bz2'
stream = iter_wiki(wiki_file)
for title, tokens in itertools.islice(iter_wiki(wiki_file), 8):
    print (title, tokens[:10])  # print the article title and its first ten␣
 ↪tokens
```

```
April ['april', 'fourth', 'month', 'year', 'julian', 'gregorian', 'calendars',
'comes', 'march', 'months']
August ['august', 'aug', 'eighth', 'month', 'year', 'gregorian', 'calendar',
'coming', 'july', 'september']
Art ['painting', 'renoir', 'work', 'art', 'art', 'creative', 'activity',
'expresses', 'imaginative', 'technical']
A ['writing', 'cursive', 'font', 'letter', 'english', 'alphabet', 'small',
```

```
'letter', 'lower', 'case']
Air ['fan', 'air', 'air', 'refers', 'earth', 'atmosphere', 'air', 'mixture',
'gases', 'tiny']
Autonomous communities of Spain ['spain', 'divided', 'parts', 'called',
'autonomous', 'communities', 'autonomous', 'means', 'autonomous', 'communities']
Alan Turing ['statue', 'alan', 'turing', 'turing', 'idea', 'bombe',
'mechanical', 'details', 'added', 'built']
Alanis Morissette ['alanis', 'nadine', 'morissette', 'born', 'june', 'grammy',
'award', 'winning', 'canadian', 'american']
```

[10]:
```
id2word = {0: u'word', 2: u'profit', 300: u'another_word'}
```

[11]:
```
doc_stream = (tokens for _, tokens in iter_wiki(wiki_file))
```

[12]:
```
%time id2word_wiki = gensim.corpora.Dictionary(doc_stream)
print(id2word_wiki)
```

```
INFO:gensim.corpora.dictionary:adding document #0 to Dictionary<0 unique tokens:
[]>
INFO:gensim.corpora.dictionary:adding document #10000 to Dictionary<168992
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:adding document #20000 to Dictionary<246465
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:adding document #30000 to Dictionary<307692
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:adding document #40000 to Dictionary<366887
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:adding document #50000 to Dictionary<433236
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:adding document #60000 to Dictionary<469090
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:adding document #70000 to Dictionary<525543
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:adding document #80000 to Dictionary<581339
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:adding document #90000 to Dictionary<643679
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…>
INFO:gensim.corpora.dictionary:built Dictionary<650295 unique tokens:
['abdicated', 'abdicates', 'abraham', 'additionally', 'adolf']…> from 91800
```

```
documents (total 19901910 corpus positions)
INFO:gensim.utils:Dictionary lifecycle event {'msg': "built Dictionary<650295
unique tokens: ['abdicated', 'abdicates', 'abraham', 'additionally',
'adolf']…> from 91800 documents (total 19901910 corpus positions)",
'datetime': '2024-12-03T07:17:38.654762', 'gensim': '4.3.3', 'python': '3.10.12
(main, Nov  6 2024, 20:22:13) [GCC 11.4.0]', 'platform':
'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'created'}

CPU times: user 11min 26s, sys: 1.48 s, total: 11min 27s
Wall time: 11min 32s
Dictionary<650295 unique tokens: ['abdicated', 'abdicates', 'abraham',
'additionally', 'adolf']…>
```

[13]:
```
# ignore words that appear in less than 20 documents or more than 10% documents
id2word_wiki.filter_extremes(no_below=20, no_above=0.1)
print(id2word_wiki)
```

```
INFO:gensim.corpora.dictionary:discarding 610151 tokens: [('alvares', 4),
('american', 20610), ('aperire', 1), ('april', 10648), ('arbroath', 17),
('born', 24070), ('chakri', 16), ('city', 15421), ('cosmonauts', 18),
('davidians', 7)]…
INFO:gensim.corpora.dictionary:keeping 40144 tokens which were in no less than
20 and no more than 9180 (=10.0%) documents
INFO:gensim.corpora.dictionary:resulting dictionary: Dictionary<40144 unique
tokens: ['abdicated', 'abdicates', 'abraham', 'additionally', 'adolf']…>

Dictionary<40144 unique tokens: ['abdicated', 'abdicates', 'abraham',
'additionally', 'adolf']…>
```

[14]:
```
now = datetime.now()

print("Done with SimpleWiki at", now)
```

```
Done with SimpleWiki at 2024-12-03 07:17:39.986532
```

Question 1: Print all words and their ids from id2word_wiki where the word starts with "human".

Note for advanced users: In fully online scenarios, where the documents can only be streamed once (no repeating the stream), we can't exhaust the document stream just to build a dictionary. In this case we can map strings directly into their integer hash, using a hashing function such as MurmurHash or MD5. This is called the "hashing trick". A dictionary built this way is more difficult to debug, because there may be hash collisions: multiple words represented by a single id. See the documentation of HashDictionary for more details.

[45]:
```
# Iterate through the items in the id2word_wiki dictionary
for id, word in id2word_wiki.items():
    # Check if the word starts with "human"
    if word.startswith("human"):
        # Print the word and its corresponding ID in a formatted string
        print(f"Word: {word} (ID: {id})")
```

```
Word: human (ID: 296)
Word: humanitarian (ID: 735)
Word: humans (ID: 953)
Word: humanity (ID: 2910)
Word: humanism (ID: 7356)
Word: humankind (ID: 9270)
Word: humanities (ID: 16429)
Word: humanistic (ID: 24754)
Word: humanist (ID: 26705)
Word: humanoid (ID: 30593)
Word: humane (ID: 32096)
```

1.1 Vectorization

A streamed corpus and a dictionary is all we need to create bag-of-words vectors:

[15]:
```
doc = "A blood cell, also called a hematocyte, is a cell produced by␣
 ↪hematopoiesis and normally found in blood."
bow = id2word_wiki.doc2bow(tokenize(doc))
print(bow)
```

```
[(989, 1), (1176, 2), (1262, 1), (3368, 2)]
```

[16]:
```
print(id2word_wiki[10882])
```

```
naruhito
```

[17]:
```
class WikiCorpus(object):
    def __init__(self, dump_file, dictionary, clip_docs=None):
        """
        Parse the first `clip_docs` Wikipedia documents from file `dump_file`.
        Yield each document in turn, as a list of tokens (unicode strings).

        """
        self.dump_file = dump_file
        self.dictionary = dictionary
        self.clip_docs = clip_docs

    def __iter__(self):
        self.titles = []
        for title, tokens in itertools.islice(iter_wiki(self.dump_file), self.
 ↪clip_docs):
            self.titles.append(title)
            yield self.dictionary.doc2bow(tokens)

    def __len__(self):
        return self.clip_docs
```

```
# create a stream of bag-of-words vectors
wiki_corpus = WikiCorpus(wiki_file, id2word_wiki)
vector = next(iter(wiki_corpus))
print(vector)  # print the first vector in the stream
```

```
[(0, 1), (1, 2), (2, 1), (3, 1), (4, 2), (5, 1), (6, 2), (7, 1), (8, 1), (9, 2),
(10, 2), (11, 3), (12, 1), (13, 1), (14, 1), (15, 1), (16, 2), (17, 1), (18, 5),
(19, 1), (20, 1), (21, 1), (22, 1), (23, 1), (24, 1), (25, 1), (26, 2), (27, 4),
(28, 1), (29, 1), (30, 1), (31, 2), (32, 1), (33, 1), (34, 1), (35, 3), (36, 3),
(37, 1), (38, 1), (39, 2), (40, 1), (41, 1), (42, 1), (43, 1), (44, 1), (45, 1),
(46, 1), (47, 2), (48, 1), (49, 1), (50, 5), (51, 1), (52, 1), (53, 1), (54, 1),
(55, 1), (56, 1), (57, 1), (58, 1), (59, 1), (60, 10), (61, 2), (62, 1), (63,
1), (64, 1), (65, 1), (66, 1), (67, 2), (68, 1), (69, 1), (70, 2), (71, 2), (72,
1), (73, 2), (74, 1), (75, 1), (76, 1), (77, 1), (78, 1), (79, 1), (80, 1), (81,
1), (82, 1), (83, 1), (84, 2), (85, 1), (86, 2), (87, 1), (88, 2), (89, 1), (90,
2), (91, 1), (92, 2), (93, 1), (94, 1), (95, 1), (96, 1), (97, 2), (98, 1), (99,
2), (100, 2), (101, 2), (102, 4), (103, 2), (104, 1), (105, 1), (106, 2), (107,
1), (108, 1), (109, 2), (110, 1), (111, 1), (112, 1), (113, 6), (114, 2), (115,
2), (116, 3), (117, 2), (118, 1), (119, 2), (120, 1), (121, 2), (122, 4), (123,
1), (124, 1), (125, 8), (126, 1), (127, 2), (128, 1), (129, 1), (130, 1), (131,
1), (132, 1), (133, 1), (134, 1), (135, 3), (136, 1), (137, 2), (138, 1), (139,
1), (140, 1), (141, 2), (142, 2), (143, 2), (144, 1), (145, 2), (146, 1), (147,
1), (148, 4), (149, 1), (150, 3), (151, 5), (152, 1), (153, 1), (154, 1), (155,
1), (156, 1), (157, 1), (158, 1), (159, 3), (160, 1), (161, 1), (162, 1), (163,
3), (164, 1), (165, 1), (166, 5), (167, 1), (168, 2), (169, 1), (170, 1), (171,
1), (172, 1), (173, 2), (174, 1), (175, 3), (176, 4), (177, 12), (178, 1), (179,
1), (180, 1), (181, 1), (182, 1), (183, 1), (184, 2), (185, 1), (186, 2), (187,
4), (188, 4), (189, 1), (190, 1), (191, 2), (192, 1), (193, 1), (194, 1), (195,
3), (196, 7), (197, 6), (198, 2), (199, 1), (200, 1), (201, 1), (202, 1), (203,
1), (204, 1), (205, 1), (206, 1), (207, 2), (208, 1), (209, 6), (210, 1), (211,
1), (212, 1), (213, 1), (214, 2), (215, 3), (216, 1), (217, 2), (218, 1), (219,
5), (220, 1), (221, 1), (222, 2), (223, 5), (224, 1), (225, 5), (226, 2), (227,
1), (228, 6), (229, 1), (230, 1), (231, 1), (232, 1), (233, 4), (234, 1), (235,
2), (236, 4), (237, 3), (238, 1), (239, 1), (240, 1), (241, 1), (242, 2), (243,
1), (244, 1), (245, 1), (246, 1), (247, 2), (248, 3), (249, 1), (250, 1), (251,
3), (252, 3), (253, 2), (254, 1), (255, 1), (256, 1), (257, 1), (258, 1), (259,
2), (260, 1), (261, 1), (262, 2), (263, 1), (264, 1), (265, 1), (266, 1), (267,
1), (268, 2), (269, 1), (270, 1), (271, 1), (272, 1), (273, 1), (274, 1), (275,
2), (276, 5), (277, 1), (278, 3), (279, 1), (280, 1), (281, 1), (282, 1), (283,
2), (284, 2), (285, 1), (286, 1), (287, 1), (288, 1), (289, 1), (290, 1), (291,
2), (292, 1), (293, 2), (294, 1), (295, 1), (296, 1), (297, 1), (298, 1), (299,
6), (300, 1), (301, 1), (302, 1), (303, 6), (304, 1), (305, 1), (306, 1), (307,
8), (308, 3), (309, 1), (310, 2), (311, 1), (312, 5), (313, 1), (314, 1), (315,
1), (316, 2), (317, 1), (318, 2), (319, 3), (320, 3), (321, 1), (322, 1), (323,
1), (324, 1), (325, 2), (326, 2), (327, 1), (328, 1), (329, 1), (330, 1), (331,
1), (332, 1), (333, 3), (334, 1), (335, 1), (336, 1), (337, 3), (338, 4), (339,
1), (340, 11), (341, 5), (342, 1), (343, 2), (344, 4), (345, 2), (346, 5), (347,
```

```
1), (348, 1), (349, 1), (350, 1), (351, 1), (352, 1), (353, 1), (354, 1), (355,
8), (356, 1), (357, 1), (358, 1), (359, 3), (360, 1), (361, 1), (362, 1), (363,
2), (364, 1), (365, 2), (366, 1), (367, 1), (368, 1), (369, 1), (370, 1), (371,
1), (372, 1), (373, 1), (374, 1), (375, 1), (376, 2), (377, 2), (378, 1), (379,
2), (380, 1), (381, 1), (382, 2), (383, 1), (384, 2), (385, 2), (386, 1), (387,
1), (388, 1), (389, 2), (390, 1), (391, 3), (392, 1), (393, 1), (394, 2), (395,
1), (396, 1), (397, 2), (398, 1), (399, 9), (400, 5), (401, 1), (402, 1), (403,
1), (404, 1), (405, 1), (406, 1), (407, 1), (408, 1), (409, 1), (410, 1), (411,
2), (412, 1), (413, 1), (414, 2), (415, 1), (416, 14), (417, 1), (418, 1), (419,
3), (420, 6), (421, 1), (422, 1), (423, 2), (424, 1), (425, 2), (426, 2), (427,
1), (428, 2), (429, 1), (430, 1), (431, 1), (432, 1), (433, 1), (434, 2), (435,
1), (436, 1), (437, 2), (438, 2), (439, 1), (440, 1), (441, 1), (442, 1), (443,
1), (444, 1), (445, 1), (446, 1), (447, 1), (448, 2), (449, 1), (450, 1), (451,
1), (452, 1), (453, 1), (454, 1), (455, 1), (456, 1), (457, 1), (458, 1), (459,
1), (460, 1), (461, 1), (462, 1), (463, 1), (464, 1), (465, 2), (466, 1), (467,
3), (468, 2), (469, 2), (470, 6), (471, 6), (472, 1), (473, 1), (474, 1), (475,
1), (476, 1), (477, 1), (478, 1), (479, 1), (480, 1), (481, 8), (482, 1), (483,
1), (484, 1), (485, 1), (486, 1), (487, 1), (488, 1), (489, 1), (490, 1), (491,
1), (492, 1), (493, 1), (494, 1), (495, 1), (496, 1), (497, 1), (498, 1), (499,
1), (500, 4), (501, 2), (502, 1), (503, 1), (504, 1), (505, 2), (506, 3), (507,
1), (508, 1), (509, 1), (510, 1), (511, 1), (512, 1), (513, 1), (514, 2), (515,
1), (516, 1), (517, 1), (518, 1), (519, 3), (520, 1), (521, 4), (522, 1), (523,
3), (524, 1), (525, 1), (526, 1), (527, 1), (528, 1), (529, 2), (530, 1), (531,
1), (532, 1), (533, 2), (534, 1), (535, 1), (536, 1), (537, 1), (538, 2), (539,
1), (540, 1), (541, 1), (542, 3), (543, 1), (544, 1), (545, 1), (546, 1), (547,
1), (548, 1), (549, 1), (550, 1), (551, 3), (552, 1), (553, 1), (554, 2), (555,
4), (556, 1), (557, 1), (558, 3), (559, 1), (560, 3), (561, 1), (562, 1), (563,
5), (564, 4), (565, 1), (566, 1), (567, 1), (568, 1), (569, 4), (570, 1), (571,
1), (572, 2), (573, 4), (574, 2), (575, 1), (576, 2), (577, 1), (578, 1), (579,
2), (580, 1), (581, 1), (582, 2), (583, 1), (584, 1), (585, 1), (586, 1), (587,
2), (588, 4), (589, 1), (590, 2), (591, 1), (592, 3), (593, 1), (594, 1), (595,
1), (596, 2), (597, 1), (598, 1), (599, 1), (600, 2), (601, 2), (602, 1), (603,
1), (604, 1), (605, 1), (606, 1), (607, 1), (608, 1), (609, 1), (610, 1), (611,
1), (612, 1), (613, 2), (614, 2), (615, 1), (616, 1), (617, 2), (618, 1), (619,
1), (620, 1), (621, 1), (622, 2), (623, 1), (624, 1), (625, 1), (626, 2), (627,
1), (628, 15), (629, 1), (630, 5), (631, 1), (632, 3), (633, 2), (634, 2), (635,
1), (636, 2), (637, 1), (638, 2), (639, 1), (640, 1), (641, 3), (642, 1), (643,
1), (644, 2), (645, 1), (646, 5), (647, 2), (648, 1)]
```

[18]:
```
len(vector)
max([pair[1] for pair in vector])

index = [pair[1] for pair in vector].index(15)
index
```

[18]: 628

[19]:
```
# what is the most common word in that first article?

(most_index, most_count) = max(vector, key=lambda pair: pair[1])
print(id2word_wiki[most_index], most_count)
```

```
week 15
```

[20]:
```
%time gensim.corpora.MmCorpus.serialize('./data/wiki_bow.mm', wiki_corpus)
```

```
INFO:gensim.corpora.mmcorpus:storing corpus in Matrix Market format to
./data/wiki_bow.mm
INFO:gensim.matutils:saving sparse matrix to ./data/wiki_bow.mm
INFO:gensim.matutils:PROGRESS: saving document #0
INFO:gensim.matutils:PROGRESS: saving document #1000
INFO:gensim.matutils:PROGRESS: saving document #2000
INFO:gensim.matutils:PROGRESS: saving document #3000
INFO:gensim.matutils:PROGRESS: saving document #4000
INFO:gensim.matutils:PROGRESS: saving document #5000
INFO:gensim.matutils:PROGRESS: saving document #6000
INFO:gensim.matutils:PROGRESS: saving document #7000
INFO:gensim.matutils:PROGRESS: saving document #8000
INFO:gensim.matutils:PROGRESS: saving document #9000
INFO:gensim.matutils:PROGRESS: saving document #10000
INFO:gensim.matutils:PROGRESS: saving document #11000
INFO:gensim.matutils:PROGRESS: saving document #12000
INFO:gensim.matutils:PROGRESS: saving document #13000
INFO:gensim.matutils:PROGRESS: saving document #14000
INFO:gensim.matutils:PROGRESS: saving document #15000
INFO:gensim.matutils:PROGRESS: saving document #16000
INFO:gensim.matutils:PROGRESS: saving document #17000
INFO:gensim.matutils:PROGRESS: saving document #18000
INFO:gensim.matutils:PROGRESS: saving document #19000
INFO:gensim.matutils:PROGRESS: saving document #20000
INFO:gensim.matutils:PROGRESS: saving document #21000
INFO:gensim.matutils:PROGRESS: saving document #22000
INFO:gensim.matutils:PROGRESS: saving document #23000
INFO:gensim.matutils:PROGRESS: saving document #24000
INFO:gensim.matutils:PROGRESS: saving document #25000
INFO:gensim.matutils:PROGRESS: saving document #26000
INFO:gensim.matutils:PROGRESS: saving document #27000
INFO:gensim.matutils:PROGRESS: saving document #28000
INFO:gensim.matutils:PROGRESS: saving document #29000
INFO:gensim.matutils:PROGRESS: saving document #30000
INFO:gensim.matutils:PROGRESS: saving document #31000
INFO:gensim.matutils:PROGRESS: saving document #32000
INFO:gensim.matutils:PROGRESS: saving document #33000
INFO:gensim.matutils:PROGRESS: saving document #34000
INFO:gensim.matutils:PROGRESS: saving document #35000
```

```
INFO:gensim.matutils:PROGRESS: saving document #36000
INFO:gensim.matutils:PROGRESS: saving document #37000
INFO:gensim.matutils:PROGRESS: saving document #38000
INFO:gensim.matutils:PROGRESS: saving document #39000
INFO:gensim.matutils:PROGRESS: saving document #40000
INFO:gensim.matutils:PROGRESS: saving document #41000
INFO:gensim.matutils:PROGRESS: saving document #42000
INFO:gensim.matutils:PROGRESS: saving document #43000
INFO:gensim.matutils:PROGRESS: saving document #44000
INFO:gensim.matutils:PROGRESS: saving document #45000
INFO:gensim.matutils:PROGRESS: saving document #46000
INFO:gensim.matutils:PROGRESS: saving document #47000
INFO:gensim.matutils:PROGRESS: saving document #48000
INFO:gensim.matutils:PROGRESS: saving document #49000
INFO:gensim.matutils:PROGRESS: saving document #50000
INFO:gensim.matutils:PROGRESS: saving document #51000
INFO:gensim.matutils:PROGRESS: saving document #52000
INFO:gensim.matutils:PROGRESS: saving document #53000
INFO:gensim.matutils:PROGRESS: saving document #54000
INFO:gensim.matutils:PROGRESS: saving document #55000
INFO:gensim.matutils:PROGRESS: saving document #56000
INFO:gensim.matutils:PROGRESS: saving document #57000
INFO:gensim.matutils:PROGRESS: saving document #58000
INFO:gensim.matutils:PROGRESS: saving document #59000
INFO:gensim.matutils:PROGRESS: saving document #60000
INFO:gensim.matutils:PROGRESS: saving document #61000
INFO:gensim.matutils:PROGRESS: saving document #62000
INFO:gensim.matutils:PROGRESS: saving document #63000
INFO:gensim.matutils:PROGRESS: saving document #64000
INFO:gensim.matutils:PROGRESS: saving document #65000
INFO:gensim.matutils:PROGRESS: saving document #66000
INFO:gensim.matutils:PROGRESS: saving document #67000
INFO:gensim.matutils:PROGRESS: saving document #68000
INFO:gensim.matutils:PROGRESS: saving document #69000
INFO:gensim.matutils:PROGRESS: saving document #70000
INFO:gensim.matutils:PROGRESS: saving document #71000
INFO:gensim.matutils:PROGRESS: saving document #72000
INFO:gensim.matutils:PROGRESS: saving document #73000
INFO:gensim.matutils:PROGRESS: saving document #74000
INFO:gensim.matutils:PROGRESS: saving document #75000
INFO:gensim.matutils:PROGRESS: saving document #76000
INFO:gensim.matutils:PROGRESS: saving document #77000
INFO:gensim.matutils:PROGRESS: saving document #78000
INFO:gensim.matutils:PROGRESS: saving document #79000
INFO:gensim.matutils:PROGRESS: saving document #80000
INFO:gensim.matutils:PROGRESS: saving document #81000
INFO:gensim.matutils:PROGRESS: saving document #82000
INFO:gensim.matutils:PROGRESS: saving document #83000
```

```
INFO:gensim.matutils:PROGRESS: saving document #84000
INFO:gensim.matutils:PROGRESS: saving document #85000
INFO:gensim.matutils:PROGRESS: saving document #86000
INFO:gensim.matutils:PROGRESS: saving document #87000
INFO:gensim.matutils:PROGRESS: saving document #88000
INFO:gensim.matutils:PROGRESS: saving document #89000
INFO:gensim.matutils:PROGRESS: saving document #90000
INFO:gensim.matutils:PROGRESS: saving document #91000
INFO:gensim.matutils:saved 91800x40144 matrix, density=0.238%
(8783660/3685219200)
INFO:gensim.corpora.indexedcorpus:saving MmCorpus index to
./data/wiki_bow.mm.index

CPU times: user 11min 31s, sys: 2.8 s, total: 11min 34s
Wall time: 11min 39s
```

[21]:
```
mm_corpus = gensim.corpora.MmCorpus('./data/wiki_bow.mm')
print(mm_corpus)
```

```
INFO:gensim.corpora.indexedcorpus:loaded corpus index from
./data/wiki_bow.mm.index
INFO:gensim.corpora._mmreader:initializing cython corpus reader from
./data/wiki_bow.mm
INFO:gensim.corpora._mmreader:accepted corpus with 91800 documents, 40144
features, 8783660 non-zero entries

MmCorpus(91800 documents, 40144 features, 8783660 non-zero entries)
```

[22]:
```
print(next(iter(mm_corpus)))
```

```
[(0, 1.0), (1, 2.0), (2, 1.0), (3, 1.0), (4, 2.0), (5, 1.0), (6, 2.0), (7, 1.0),
(8, 1.0), (9, 2.0), (10, 2.0), (11, 3.0), (12, 1.0), (13, 1.0), (14, 1.0), (15,
1.0), (16, 2.0), (17, 1.0), (18, 5.0), (19, 1.0), (20, 1.0), (21, 1.0), (22,
1.0), (23, 1.0), (24, 1.0), (25, 1.0), (26, 2.0), (27, 4.0), (28, 1.0), (29,
1.0), (30, 1.0), (31, 2.0), (32, 1.0), (33, 1.0), (34, 1.0), (35, 3.0), (36,
3.0), (37, 1.0), (38, 1.0), (39, 2.0), (40, 1.0), (41, 1.0), (42, 1.0), (43,
1.0), (44, 1.0), (45, 1.0), (46, 1.0), (47, 2.0), (48, 1.0), (49, 1.0), (50,
5.0), (51, 1.0), (52, 1.0), (53, 1.0), (54, 1.0), (55, 1.0), (56, 1.0), (57,
1.0), (58, 1.0), (59, 1.0), (60, 10.0), (61, 2.0), (62, 1.0), (63, 1.0), (64,
1.0), (65, 1.0), (66, 1.0), (67, 2.0), (68, 1.0), (69, 1.0), (70, 2.0), (71,
2.0), (72, 1.0), (73, 2.0), (74, 1.0), (75, 1.0), (76, 1.0), (77, 1.0), (78,
1.0), (79, 1.0), (80, 1.0), (81, 1.0), (82, 1.0), (83, 1.0), (84, 2.0), (85,
1.0), (86, 2.0), (87, 1.0), (88, 2.0), (89, 1.0), (90, 2.0), (91, 1.0), (92,
2.0), (93, 1.0), (94, 1.0), (95, 1.0), (96, 1.0), (97, 2.0), (98, 1.0), (99,
2.0), (100, 2.0), (101, 2.0), (102, 4.0), (103, 2.0), (104, 1.0), (105, 1.0),
(106, 2.0), (107, 1.0), (108, 1.0), (109, 2.0), (110, 1.0), (111, 1.0), (112,
1.0), (113, 6.0), (114, 2.0), (115, 2.0), (116, 3.0), (117, 2.0), (118, 1.0),
(119, 2.0), (120, 1.0), (121, 2.0), (122, 4.0), (123, 1.0), (124, 1.0), (125,
8.0), (126, 1.0), (127, 2.0), (128, 1.0), (129, 1.0), (130, 1.0), (131, 1.0),
```

(132, 1.0), (133, 1.0), (134, 1.0), (135, 3.0), (136, 1.0), (137, 2.0), (138,
1.0), (139, 1.0), (140, 1.0), (141, 2.0), (142, 2.0), (143, 2.0), (144, 1.0),
(145, 2.0), (146, 1.0), (147, 1.0), (148, 4.0), (149, 1.0), (150, 3.0), (151,
5.0), (152, 1.0), (153, 1.0), (154, 1.0), (155, 1.0), (156, 1.0), (157, 1.0),
(158, 1.0), (159, 3.0), (160, 1.0), (161, 1.0), (162, 1.0), (163, 3.0), (164,
1.0), (165, 1.0), (166, 5.0), (167, 1.0), (168, 2.0), (169, 1.0), (170, 1.0),
(171, 1.0), (172, 1.0), (173, 2.0), (174, 1.0), (175, 3.0), (176, 4.0), (177,
12.0), (178, 1.0), (179, 1.0), (180, 1.0), (181, 1.0), (182, 1.0), (183, 1.0),
(184, 2.0), (185, 1.0), (186, 2.0), (187, 4.0), (188, 4.0), (189, 1.0), (190,
1.0), (191, 2.0), (192, 1.0), (193, 1.0), (194, 1.0), (195, 3.0), (196, 7.0),
(197, 6.0), (198, 2.0), (199, 1.0), (200, 1.0), (201, 1.0), (202, 1.0), (203,
1.0), (204, 1.0), (205, 1.0), (206, 1.0), (207, 2.0), (208, 1.0), (209, 6.0),
(210, 1.0), (211, 1.0), (212, 1.0), (213, 1.0), (214, 2.0), (215, 3.0), (216,
1.0), (217, 2.0), (218, 1.0), (219, 5.0), (220, 1.0), (221, 1.0), (222, 2.0),
(223, 5.0), (224, 1.0), (225, 5.0), (226, 2.0), (227, 1.0), (228, 6.0), (229,
1.0), (230, 1.0), (231, 1.0), (232, 1.0), (233, 4.0), (234, 1.0), (235, 2.0),
(236, 4.0), (237, 3.0), (238, 1.0), (239, 1.0), (240, 1.0), (241, 1.0), (242,
2.0), (243, 1.0), (244, 1.0), (245, 1.0), (246, 1.0), (247, 2.0), (248, 3.0),
(249, 1.0), (250, 1.0), (251, 3.0), (252, 3.0), (253, 2.0), (254, 1.0), (255,
1.0), (256, 1.0), (257, 1.0), (258, 1.0), (259, 2.0), (260, 1.0), (261, 1.0),
(262, 2.0), (263, 1.0), (264, 1.0), (265, 1.0), (266, 1.0), (267, 1.0), (268,
2.0), (269, 1.0), (270, 1.0), (271, 1.0), (272, 1.0), (273, 1.0), (274, 1.0),
(275, 2.0), (276, 5.0), (277, 1.0), (278, 3.0), (279, 1.0), (280, 1.0), (281,
1.0), (282, 1.0), (283, 2.0), (284, 2.0), (285, 1.0), (286, 1.0), (287, 1.0),
(288, 1.0), (289, 1.0), (290, 1.0), (291, 2.0), (292, 1.0), (293, 2.0), (294,
1.0), (295, 1.0), (296, 1.0), (297, 1.0), (298, 1.0), (299, 6.0), (300, 1.0),
(301, 1.0), (302, 1.0), (303, 6.0), (304, 1.0), (305, 1.0), (306, 1.0), (307,
8.0), (308, 3.0), (309, 1.0), (310, 2.0), (311, 1.0), (312, 5.0), (313, 1.0),
(314, 1.0), (315, 1.0), (316, 2.0), (317, 1.0), (318, 2.0), (319, 3.0), (320,
3.0), (321, 1.0), (322, 1.0), (323, 1.0), (324, 1.0), (325, 2.0), (326, 2.0),
(327, 1.0), (328, 1.0), (329, 1.0), (330, 1.0), (331, 1.0), (332, 1.0), (333,
3.0), (334, 1.0), (335, 1.0), (336, 1.0), (337, 3.0), (338, 4.0), (339, 1.0),
(340, 11.0), (341, 5.0), (342, 1.0), (343, 2.0), (344, 4.0), (345, 2.0), (346,
5.0), (347, 1.0), (348, 1.0), (349, 1.0), (350, 1.0), (351, 1.0), (352, 1.0),
(353, 1.0), (354, 1.0), (355, 8.0), (356, 1.0), (357, 1.0), (358, 1.0), (359,
3.0), (360, 1.0), (361, 1.0), (362, 1.0), (363, 2.0), (364, 1.0), (365, 2.0),
(366, 1.0), (367, 1.0), (368, 1.0), (369, 1.0), (370, 1.0), (371, 1.0), (372,
1.0), (373, 1.0), (374, 1.0), (375, 1.0), (376, 2.0), (377, 2.0), (378, 1.0),
(379, 2.0), (380, 1.0), (381, 1.0), (382, 2.0), (383, 1.0), (384, 2.0), (385,
2.0), (386, 1.0), (387, 1.0), (388, 1.0), (389, 2.0), (390, 1.0), (391, 3.0),
(392, 1.0), (393, 1.0), (394, 2.0), (395, 1.0), (396, 1.0), (397, 2.0), (398,
1.0), (399, 9.0), (400, 5.0), (401, 1.0), (402, 1.0), (403, 1.0), (404, 1.0),
(405, 1.0), (406, 1.0), (407, 1.0), (408, 1.0), (409, 1.0), (410, 1.0), (411,
2.0), (412, 1.0), (413, 1.0), (414, 2.0), (415, 1.0), (416, 14.0), (417, 1.0),
(418, 1.0), (419, 3.0), (420, 6.0), (421, 1.0), (422, 1.0), (423, 2.0), (424,
1.0), (425, 2.0), (426, 2.0), (427, 1.0), (428, 2.0), (429, 1.0), (430, 1.0),
(431, 1.0), (432, 1.0), (433, 1.0), (434, 2.0), (435, 1.0), (436, 1.0), (437,
2.0), (438, 2.0), (439, 1.0), (440, 1.0), (441, 1.0), (442, 1.0), (443, 1.0),

```
(444, 1.0), (445, 1.0), (446, 1.0), (447, 1.0), (448, 2.0), (449, 1.0), (450,
1.0), (451, 1.0), (452, 1.0), (453, 1.0), (454, 1.0), (455, 1.0), (456, 1.0),
(457, 1.0), (458, 1.0), (459, 1.0), (460, 1.0), (461, 1.0), (462, 1.0), (463,
1.0), (464, 1.0), (465, 2.0), (466, 1.0), (467, 3.0), (468, 2.0), (469, 2.0),
(470, 6.0), (471, 6.0), (472, 1.0), (473, 1.0), (474, 1.0), (475, 1.0), (476,
1.0), (477, 1.0), (478, 1.0), (479, 1.0), (480, 1.0), (481, 8.0), (482, 1.0),
(483, 1.0), (484, 1.0), (485, 1.0), (486, 1.0), (487, 1.0), (488, 1.0), (489,
1.0), (490, 1.0), (491, 1.0), (492, 1.0), (493, 1.0), (494, 1.0), (495, 1.0),
(496, 1.0), (497, 1.0), (498, 1.0), (499, 1.0), (500, 4.0), (501, 2.0), (502,
1.0), (503, 1.0), (504, 1.0), (505, 2.0), (506, 3.0), (507, 1.0), (508, 1.0),
(509, 1.0), (510, 1.0), (511, 1.0), (512, 1.0), (513, 1.0), (514, 2.0), (515,
1.0), (516, 1.0), (517, 1.0), (518, 1.0), (519, 3.0), (520, 1.0), (521, 4.0),
(522, 1.0), (523, 3.0), (524, 1.0), (525, 1.0), (526, 1.0), (527, 1.0), (528,
1.0), (529, 2.0), (530, 1.0), (531, 1.0), (532, 1.0), (533, 2.0), (534, 1.0),
(535, 1.0), (536, 1.0), (537, 1.0), (538, 2.0), (539, 1.0), (540, 1.0), (541,
1.0), (542, 3.0), (543, 1.0), (544, 1.0), (545, 1.0), (546, 1.0), (547, 1.0),
(548, 1.0), (549, 1.0), (550, 1.0), (551, 3.0), (552, 1.0), (553, 1.0), (554,
2.0), (555, 4.0), (556, 1.0), (557, 1.0), (558, 3.0), (559, 1.0), (560, 3.0),
(561, 1.0), (562, 1.0), (563, 5.0), (564, 4.0), (565, 1.0), (566, 1.0), (567,
1.0), (568, 1.0), (569, 4.0), (570, 1.0), (571, 1.0), (572, 2.0), (573, 4.0),
(574, 2.0), (575, 1.0), (576, 2.0), (577, 1.0), (578, 1.0), (579, 2.0), (580,
1.0), (581, 1.0), (582, 2.0), (583, 1.0), (584, 1.0), (585, 1.0), (586, 1.0),
(587, 2.0), (588, 4.0), (589, 1.0), (590, 2.0), (591, 1.0), (592, 3.0), (593,
1.0), (594, 1.0), (595, 1.0), (596, 2.0), (597, 1.0), (598, 1.0), (599, 1.0),
(600, 2.0), (601, 2.0), (602, 1.0), (603, 1.0), (604, 1.0), (605, 1.0), (606,
1.0), (607, 1.0), (608, 1.0), (609, 1.0), (610, 1.0), (611, 1.0), (612, 1.0),
(613, 2.0), (614, 2.0), (615, 1.0), (616, 1.0), (617, 2.0), (618, 1.0), (619,
1.0), (620, 1.0), (621, 1.0), (622, 2.0), (623, 1.0), (624, 1.0), (625, 1.0),
(626, 2.0), (627, 1.0), (628, 15.0), (629, 1.0), (630, 5.0), (631, 1.0), (632,
3.0), (633, 2.0), (634, 2.0), (635, 1.0), (636, 2.0), (637, 1.0), (638, 2.0),
(639, 1.0), (640, 1.0), (641, 3.0), (642, 1.0), (643, 1.0), (644, 2.0), (645,
1.0), (646, 5.0), (647, 2.0), (648, 1.0)]
```

1.2 Semantic transformations

Topic modeling in gensim is realized via transformations. A transformation is something that takes a corpus and spits out another corpus on output, using `corpus_out = transformation_object[corpus_in]` syntax. What exactly happens in between is determined by what kind of transformation we're using – options are Latent Semantic Indexing (LSI), Latent Dirichlet Allocation (LDA), Random Projections (RP) etc.

Some transformations need to be initialized (=trained) before they can be used. For example, let's train an LDA transformation model, using our bag-of-words WikiCorpus as training data:

[23]:
```
from gensim.utils import SaveLoad
class ClippedCorpus(SaveLoad):
    def __init__(self, corpus, max_docs=None):
        """
```

```
        Return a corpus that is the "head" of input iterable `corpus`.

        Any documents after `max_docs` are ignored. This effectively limits the
        length of the returned corpus to <= `max_docs`. Set `max_docs=None` for
        "no limit", effectively wrapping the entire input corpus.

        """
        self.corpus = corpus
        self.max_docs = max_docs

    def __iter__(self):
        return itertools.islice(self.corpus, self.max_docs)

    def __len__(self):
        return min(self.max_docs, len(self.corpus))

clipped_corpus = gensim.utils.ClippedCorpus(mm_corpus, 4000)  # use fewer␣
 ↪documents during training, LDA is slow
# ClippedCorpus new in gensim 0.10.1
# copy&paste it from https://github.com/piskvorky/gensim/blob/0.10.1/gensim/
 ↪utils.py#L467 if necessary (or upgrade your gensim)
%time lda_model = gensim.models.LdaModel(clipped_corpus, num_topics=10,␣
 ↪id2word=id2word_wiki, passes=4)
```

```
INFO:gensim.models.ldamodel:using symmetric alpha at 0.1
INFO:gensim.models.ldamodel:using symmetric eta at 0.1
INFO:gensim.models.ldamodel:using serial LDA version on this node
INFO:gensim.models.ldamodel:running online (multi-pass) LDA training, 10 topics,
4 passes over the supplied corpus of 4000 documents, updating model once every
2000 documents, evaluating perplexity every 4000 documents, iterating 50x with a
convergence threshold of 0.001000
WARNING:gensim.models.ldamodel:too few updates, training might not converge;
consider increasing the number of passes or iterations to improve accuracy
INFO:gensim.models.ldamodel:PROGRESS: pass 0, at document #2000/4000
INFO:gensim.models.ldamodel:merging changes from 2000 documents into a model of
4000 documents
INFO:gensim.models.ldamodel:topic #0 (0.100): 0.003*"president" + 0.003*"words"
+ 0.002*"league" + 0.002*"word" + 0.002*"person" + 0.002*"king" +
0.002*"countries" + 0.002*"number" + 0.002*"police" + 0.002*"example"
INFO:gensim.models.ldamodel:topic #5 (0.100): 0.003*"light" + 0.003*"language" +
0.002*"lake" + 0.002*"water" + 0.002*"countries" + 0.002*"usually" +
0.002*"mario" + 0.002*"rgb" + 0.002*"large" + 0.002*"hex"
INFO:gensim.models.ldamodel:topic #4 (0.100): 0.003*"actor" + 0.003*"french" +
0.003*"german" + 0.003*"politician" + 0.003*"actress" + 0.002*"person" +
0.002*"singer" + 0.002*"country" + 0.002*"president" + 0.002*"language"
INFO:gensim.models.ldamodel:topic #9 (0.100): 0.003*"great" + 0.003*"things" +
0.002*"actress" + 0.002*"president" + 0.002*"example" + 0.002*"party" +
```

```
0.002*"british" + 0.002*"usually" + 0.002*"countries" + 0.002*"country"
INFO:gensim.models.ldamodel:topic #3 (0.100): 0.004*"hex" + 0.004*"rgb" +
0.003*"country" + 0.003*"league" + 0.003*"water" + 0.003*"government" +
0.003*"usually" + 0.002*"example" + 0.002*"important" + 0.002*"light"
INFO:gensim.models.ldamodel:topic diff=4.475298, rho=1.000000
INFO:gensim.models.ldamodel:-9.625 per-word bound, 789.4 perplexity estimate
based on a held-out corpus of 2000 documents with 873767 words
INFO:gensim.models.ldamodel:PROGRESS: pass 0, at document #4000/4000
INFO:gensim.models.ldamodel:merging changes from 2000 documents into a model of
4000 documents
INFO:gensim.models.ldamodel:topic #4 (0.100): 0.012*"actor" + 0.011*"politician"
+ 0.010*"singer" + 0.009*"actress" + 0.009*"german" + 0.009*"footballer" +
0.008*"french" + 0.008*"player" + 0.007*"writer" + 0.007*"british"
INFO:gensim.models.ldamodel:topic #2 (0.100): 0.004*"number" + 0.003*"tower" +
0.003*"country" + 0.003*"jpg" + 0.002*"mast" + 0.002*"example" + 0.002*"numbers"
+ 0.002*"player" + 0.002*"god" + 0.002*"kansas"
INFO:gensim.models.ldamodel:topic #9 (0.100): 0.003*"things" + 0.003*"island" +
0.002*"person" + 0.002*"usually" + 0.002*"great" + 0.002*"party" +
0.002*"example" + 0.002*"country" + 0.002*"countries" + 0.002*"british"
INFO:gensim.models.ldamodel:topic #6 (0.100): 0.005*"rural" + 0.004*"jpg" +
0.004*"germany" + 0.004*"river" + 0.003*"file" + 0.003*"german" + 0.002*"king" +
0.002*"capital" + 0.002*"urban" + 0.002*"island"
INFO:gensim.models.ldamodel:topic #8 (0.100): 0.004*"windows" + 0.003*"country"
+ 0.003*"park" + 0.002*"countries" + 0.002*"british" + 0.002*"bridge" +
0.002*"century" + 0.002*"french" + 0.002*"microsoft" + 0.002*"population"
INFO:gensim.models.ldamodel:topic diff=1.185392, rho=0.707107
INFO:gensim.models.ldamodel:PROGRESS: pass 1, at document #2000/4000
INFO:gensim.models.ldamodel:merging changes from 2000 documents into a model of
4000 documents
INFO:gensim.models.ldamodel:topic #2 (0.100): 0.006*"number" + 0.006*"tower" +
0.006*"mast" + 0.005*"transmission" + 0.005*"uhf" + 0.004*"numbers" +
0.003*"example" + 0.003*"country" + 0.003*"england" + 0.003*"game"
INFO:gensim.models.ldamodel:topic #3 (0.100): 0.010*"hex" + 0.010*"rgb" +
0.005*"color" + 0.004*"water" + 0.003*"light" + 0.003*"usually" +
0.003*"country" + 0.003*"blue" + 0.003*"countries" + 0.003*"league"
INFO:gensim.models.ldamodel:topic #5 (0.100): 0.004*"mario" + 0.004*"light" +
0.003*"music" + 0.003*"usually" + 0.003*"birds" + 0.003*"metal" + 0.002*"large"
+ 0.002*"water" + 0.002*"chemical" + 0.002*"game"
INFO:gensim.models.ldamodel:topic #1 (0.100): 0.005*"water" + 0.003*"earth" +
0.003*"person" + 0.003*"example" + 0.003*"government" + 0.003*"things" +
0.003*"country" + 0.002*"countries" + 0.002*"study" + 0.002*"usually"
INFO:gensim.models.ldamodel:topic #0 (0.100): 0.005*"god" + 0.003*"person" +
0.003*"words" + 0.003*"word" + 0.003*"president" + 0.003*"books" +
0.003*"believe" + 0.003*"church" + 0.002*"father" + 0.002*"said"
INFO:gensim.models.ldamodel:topic diff=0.812787, rho=0.500000
INFO:gensim.models.ldamodel:-8.921 per-word bound, 484.6 perplexity estimate
based on a held-out corpus of 2000 documents with 873767 words
INFO:gensim.models.ldamodel:PROGRESS: pass 1, at document #4000/4000
```

```
INFO:gensim.models.ldamodel:merging changes from 2000 documents into a model of
4000 documents
INFO:gensim.models.ldamodel:topic #3 (0.100): 0.012*"rgb" + 0.012*"hex" +
0.006*"color" + 0.004*"usually" + 0.003*"water" + 0.003*"light" + 0.003*"red" +
0.003*"blue" + 0.003*"country" + 0.003*"countries"
INFO:gensim.models.ldamodel:topic #8 (0.100): 0.005*"windows" + 0.005*"country"
+ 0.004*"language" + 0.003*"park" + 0.003*"microsoft" + 0.003*"countries" +
0.003*"population" + 0.003*"century" + 0.003*"bc" + 0.003*"capital"
INFO:gensim.models.ldamodel:topic #1 (0.100): 0.004*"person" + 0.004*"water" +
0.003*"example" + 0.003*"things" + 0.003*"earth" + 0.003*"women" +
0.002*"countries" + 0.002*"government" + 0.002*"human" + 0.002*"body"
INFO:gensim.models.ldamodel:topic #7 (0.100): 0.013*"president" + 0.005*"king" +
0.004*"england" + 0.004*"henry" + 0.004*"france" + 0.003*"queen" +
0.003*"british" + 0.003*"league" + 0.003*"kansas" + 0.003*"reagan"
INFO:gensim.models.ldamodel:topic #9 (0.100): 0.003*"island" + 0.003*"things" +
0.003*"great" + 0.003*"usually" + 0.003*"penis" + 0.003*"person" +
0.003*"islands" + 0.003*"body" + 0.002*"example" + 0.002*"countries"
INFO:gensim.models.ldamodel:topic diff=0.689439, rho=0.500000
INFO:gensim.models.ldamodel:PROGRESS: pass 2, at document #2000/4000
INFO:gensim.models.ldamodel:merging changes from 2000 documents into a model of
4000 documents
INFO:gensim.models.ldamodel:topic #2 (0.100): 0.008*"tower" + 0.007*"number" +
0.007*"mast" + 0.006*"transmission" + 0.006*"uhf" + 0.004*"numbers" +
0.004*"game" + 0.003*"games" + 0.003*"example" + 0.003*"player"
INFO:gensim.models.ldamodel:topic #7 (0.100): 0.012*"president" + 0.006*"king" +
0.006*"league" + 0.005*"england" + 0.005*"reagan" + 0.004*"queen" +
0.004*"france" + 0.004*"henry" + 0.004*"premier" + 0.003*"kingdom"
INFO:gensim.models.ldamodel:topic #4 (0.100): 0.012*"actor" + 0.011*"politician"
+ 0.010*"actress" + 0.010*"singer" + 0.009*"german" + 0.009*"footballer" +
0.008*"french" + 0.008*"player" + 0.008*"british" + 0.007*"writer"
INFO:gensim.models.ldamodel:topic #3 (0.100): 0.013*"rgb" + 0.012*"hex" +
0.006*"color" + 0.005*"water" + 0.004*"usually" + 0.004*"blue" + 0.004*"light" +
0.003*"green" + 0.003*"korea" + 0.003*"red"
INFO:gensim.models.ldamodel:topic #0 (0.100): 0.006*"god" + 0.004*"words" +
0.004*"person" + 0.003*"books" + 0.003*"church" + 0.003*"said" + 0.003*"father"
+ 0.003*"word" + 0.003*"death" + 0.003*"believe"
INFO:gensim.models.ldamodel:topic diff=0.621670, rho=0.447214
INFO:gensim.models.ldamodel:-8.808 per-word bound, 448.2 perplexity estimate
based on a held-out corpus of 2000 documents with 873767 words
INFO:gensim.models.ldamodel:PROGRESS: pass 2, at document #4000/4000
INFO:gensim.models.ldamodel:merging changes from 2000 documents into a model of
4000 documents
INFO:gensim.models.ldamodel:topic #2 (0.100): 0.007*"number" + 0.006*"tower" +
0.005*"mast" + 0.005*"game" + 0.005*"transmission" + 0.004*"player" +
0.004*"uhf" + 0.004*"numbers" + 0.004*"games" + 0.003*"players"
INFO:gensim.models.ldamodel:topic #1 (0.100): 0.005*"person" + 0.004*"water" +
0.004*"things" + 0.004*"example" + 0.003*"earth" + 0.003*"women" + 0.003*"human"
+ 0.003*"body" + 0.003*"study" + 0.003*"way"
```

```
INFO:gensim.models.ldamodel:topic #5 (0.100): 0.006*"music" + 0.004*"band" +
0.004*"album" + 0.004*"rock" + 0.003*"released" + 0.003*"light" + 0.003*"live" +
0.003*"game" + 0.003*"metal" + 0.003*"mario"
INFO:gensim.models.ldamodel:topic #6 (0.100): 0.007*"river" + 0.006*"jpg" +
0.006*"germany" + 0.005*"capital" + 0.005*"government" + 0.005*"country" +
0.004*"rural" + 0.004*"file" + 0.004*"cities" + 0.004*"empire"
INFO:gensim.models.ldamodel:topic #0 (0.100): 0.006*"god" + 0.004*"said" +
0.003*"death" + 0.003*"person" + 0.003*"books" + 0.003*"father" + 0.003*"man" +
0.003*"movie" + 0.003*"children" + 0.003*"book"
INFO:gensim.models.ldamodel:topic diff=0.482295, rho=0.447214
INFO:gensim.models.ldamodel:PROGRESS: pass 3, at document #2000/4000
INFO:gensim.models.ldamodel:merging changes from 2000 documents into a model of
4000 documents
INFO:gensim.models.ldamodel:topic #6 (0.100): 0.007*"river" + 0.006*"country" +
0.006*"government" + 0.006*"jpg" + 0.006*"germany" + 0.005*"capital" +
0.005*"union" + 0.004*"empire" + 0.004*"cities" + 0.004*"party"
INFO:gensim.models.ldamodel:topic #0 (0.100): 0.006*"god" + 0.004*"books" +
0.004*"words" + 0.004*"person" + 0.003*"said" + 0.003*"church" + 0.003*"death" +
0.003*"father" + 0.003*"book" + 0.003*"movie"
INFO:gensim.models.ldamodel:topic #3 (0.100): 0.015*"rgb" + 0.015*"hex" +
0.007*"color" + 0.006*"water" + 0.004*"blue" + 0.004*"food" + 0.004*"green" +
0.004*"usually" + 0.004*"red" + 0.004*"korea"
INFO:gensim.models.ldamodel:topic #5 (0.100): 0.005*"music" + 0.004*"mario" +
0.004*"light" + 0.004*"rock" + 0.003*"metal" + 0.003*"animals" +
0.003*"chemical" + 0.003*"live" + 0.003*"usually" + 0.003*"water"
INFO:gensim.models.ldamodel:topic #1 (0.100): 0.005*"person" + 0.005*"earth" +
0.005*"water" + 0.004*"things" + 0.004*"example" + 0.003*"energy" +
0.003*"study" + 0.003*"theory" + 0.003*"human" + 0.003*"way"
INFO:gensim.models.ldamodel:topic diff=0.446956, rho=0.408248
INFO:gensim.models.ldamodel:-8.752 per-word bound, 431.2 perplexity estimate
based on a held-out corpus of 2000 documents with 873767 words
INFO:gensim.models.ldamodel:PROGRESS: pass 3, at document #4000/4000
INFO:gensim.models.ldamodel:merging changes from 2000 documents into a model of
4000 documents
INFO:gensim.models.ldamodel:topic #7 (0.100): 0.014*"president" + 0.006*"league"
+ 0.006*"king" + 0.006*"england" + 0.004*"henry" + 0.004*"kansas" + 0.004*"bush"
+ 0.004*"reagan" + 0.004*"queen" + 0.004*"france"
INFO:gensim.models.ldamodel:topic #0 (0.100): 0.006*"god" + 0.004*"movie" +
0.004*"said" + 0.003*"love" + 0.003*"book" + 0.003*"death" + 0.003*"music" +
0.003*"books" + 0.003*"man" + 0.003*"album"
INFO:gensim.models.ldamodel:topic #9 (0.100): 0.004*"island" + 0.004*"penis" +
0.004*"body" + 0.004*"blood" + 0.003*"things" + 0.003*"usually" + 0.003*"great"
+ 0.003*"islands" + 0.003*"man" + 0.003*"person"
INFO:gensim.models.ldamodel:topic #8 (0.100): 0.008*"language" + 0.006*"country"
+ 0.005*"windows" + 0.004*"languages" + 0.004*"lake" + 0.004*"africa" +
0.003*"countries" + 0.003*"india" + 0.003*"microsoft" + 0.003*"internet"
INFO:gensim.models.ldamodel:topic #5 (0.100): 0.006*"music" + 0.005*"band" +
0.005*"rock" + 0.004*"album" + 0.004*"released" + 0.004*"live" + 0.003*"light" +
```

```
0.003*"game" + 0.003*"metal" + 0.003*"species"
INFO:gensim.models.ldamodel:topic diff=0.337317, rho=0.408248
INFO:gensim.utils:LdaModel lifecycle event {'msg': 'trained
LdaModel<num_terms=40144, num_topics=10, decay=0.5, chunksize=2000> in 47.06s',
'datetime': '2024-12-03T07:30:06.479900', 'gensim': '4.3.3', 'python': '3.10.12
(main, Nov  6 2024, 20:22:13) [GCC 11.4.0]', 'platform':
'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'created'}

CPU times: user 45 s, sys: 22.7 s, total: 1min 7s
Wall time: 47.1 s
```

[24]:
```
_ = lda_model.print_topics(-1)  # print a few most important words for each LDA
↪topic
```

```
INFO:gensim.models.ldamodel:topic #0 (0.100): 0.006*"god" + 0.004*"movie" +
0.004*"said" + 0.003*"love" + 0.003*"book" + 0.003*"death" + 0.003*"music" +
0.003*"books" + 0.003*"man" + 0.003*"album"
INFO:gensim.models.ldamodel:topic #1 (0.100): 0.006*"person" + 0.004*"water" +
0.004*"things" + 0.004*"example" + 0.004*"earth" + 0.003*"human" + 0.003*"study"
+ 0.003*"body" + 0.003*"way" + 0.003*"women"
INFO:gensim.models.ldamodel:topic #2 (0.100): 0.007*"tower" + 0.007*"number" +
0.005*"game" + 0.005*"mast" + 0.005*"transmission" + 0.005*"player" +
0.005*"uhf" + 0.004*"numbers" + 0.004*"games" + 0.004*"players"
INFO:gensim.models.ldamodel:topic #3 (0.100): 0.016*"rgb" + 0.016*"hex" +
0.008*"color" + 0.005*"water" + 0.005*"food" + 0.004*"red" + 0.004*"blue" +
0.004*"usually" + 0.004*"green" + 0.003*"light"
INFO:gensim.models.ldamodel:topic #4 (0.100): 0.013*"actor" + 0.012*"politician"
+ 0.011*"singer" + 0.011*"actress" + 0.010*"german" + 0.010*"footballer" +
0.009*"french" + 0.009*"player" + 0.008*"writer" + 0.008*"british"
INFO:gensim.models.ldamodel:topic #5 (0.100): 0.006*"music" + 0.005*"band" +
0.005*"rock" + 0.004*"album" + 0.004*"released" + 0.004*"live" + 0.003*"light" +
0.003*"game" + 0.003*"metal" + 0.003*"species"
INFO:gensim.models.ldamodel:topic #6 (0.100): 0.007*"river" + 0.006*"jpg" +
0.006*"country" + 0.006*"capital" + 0.006*"germany" + 0.005*"government" +
0.004*"empire" + 0.004*"cities" + 0.004*"file" + 0.004*"east"
INFO:gensim.models.ldamodel:topic #7 (0.100): 0.014*"president" + 0.006*"league"
+ 0.006*"king" + 0.006*"england" + 0.004*"henry" + 0.004*"kansas" + 0.004*"bush"
+ 0.004*"reagan" + 0.004*"queen" + 0.004*"france"
INFO:gensim.models.ldamodel:topic #8 (0.100): 0.008*"language" + 0.006*"country"
+ 0.005*"windows" + 0.004*"languages" + 0.004*"lake" + 0.004*"africa" +
0.003*"countries" + 0.003*"india" + 0.003*"microsoft" + 0.003*"internet"
INFO:gensim.models.ldamodel:topic #9 (0.100): 0.004*"island" + 0.004*"penis" +
0.004*"body" + 0.004*"blood" + 0.003*"things" + 0.003*"usually" + 0.003*"great"
+ 0.003*"islands" + 0.003*"man" + 0.003*"person"
```

[25]:
```
now = datetime.now()

print("LDA Topic Models computed at", now)
```

```
LDA Topic Models computed at 2024-12-03 07:30:06.558728
```

More info on model parameters in gensim docs.

Transformation can be stacked. For example, here we'll train a TFIDF model, and then train Latent Semantic Analysis on top of TFIDF:

[26]:
```
%time tfidf_model = gensim.models.TfidfModel(mm_corpus, id2word=id2word_wiki)
```

```
INFO:gensim.models.tfidfmodel:collecting document frequencies
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #0
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #10000
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #20000
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #30000
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #40000
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #50000
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #60000
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #70000
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #80000
INFO:gensim.models.tfidfmodel:PROGRESS: processing document #90000
INFO:gensim.utils:TfidfModel lifecycle event {'msg': 'calculated IDF weights for
91800 documents and 40144 features (8783660 matrix non-zeros)', 'datetime':
'2024-12-03T07:30:17.265825', 'gensim': '4.3.3', 'python': '3.10.12 (main, Nov
6 2024, 20:22:13) [GCC 11.4.0]', 'platform': 'Linux-6.1.85+-x86_64-with-
glibc2.35', 'event': 'initialize'}

CPU times: user 9.86 s, sys: 223 ms, total: 10.1 s
Wall time: 10.7 s
```

[27]:
```
%time lsi_model = gensim.models.LsiModel(tfidf_model[mm_corpus],␣
↪id2word=id2word_wiki, num_topics=200)
```

```
INFO:gensim.models.lsimodel:using serial LSI version on this node
INFO:gensim.models.lsimodel:updating model with new documents
INFO:gensim.models.lsimodel:preparing a new chunk of documents
INFO:gensim.models.lsimodel:using 100 extra samples and 2 power iterations
INFO:gensim.models.lsimodel:1st phase: constructing (40144, 300) action matrix
INFO:gensim.models.lsimodel:orthonormalizing (40144, 300) action matrix
INFO:gensim.models.lsimodel:2nd phase: running dense svd on (300, 20000) matrix
INFO:gensim.models.lsimodel:computing the final decomposition
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 15.008% of energy
spectrum)
INFO:gensim.models.lsimodel:processed documents up to #20000
INFO:gensim.models.lsimodel:topic #0(15.817): 0.225*"footballer" + 0.225*"actor"
+ 0.218*"politician" + 0.206*"actress" + 0.188*"german" + 0.185*"singer" +
0.165*"french" + 0.160*"writer" + 0.144*"player" + 0.139*"british"
INFO:gensim.models.lsimodel:topic #1(10.686): -0.183*"footballer" +
-0.170*"politician" + -0.160*"actor" + -0.151*"actress" + 0.148*"music" +
-0.118*"singer" + -0.105*"writer" + -0.083*"player" + 0.083*"jpg" + 0.082*"band"
INFO:gensim.models.lsimodel:topic #2(8.313): 0.304*"music" + -0.239*"district" +
```

```
0.228*"band" + -0.206*"coat" + 0.203*"album" + -0.198*"arms" +
-0.146*"municipalities" + -0.138*"county" + -0.118*"river" + -0.117*"towns"
INFO:gensim.models.lsimodel:topic #3(7.695): 0.340*"district" + 0.316*"coat" +
0.302*"arms" + -0.231*"king" + 0.212*"municipalities" + 0.154*"towns" +
0.139*"band" + 0.131*"county" + -0.129*"emperor" + 0.124*"districts"
INFO:gensim.models.lsimodel:topic #4(7.479): 0.271*"music" + 0.255*"king" +
0.159*"band" + 0.135*"album" + 0.132*"england" + 0.129*"emperor" + 0.119*"coat"
+ 0.116*"arms" + 0.109*"district" + -0.109*"water"
INFO:gensim.models.lsimodel:preparing a new chunk of documents
INFO:gensim.models.lsimodel:using 100 extra samples and 2 power iterations
INFO:gensim.models.lsimodel:1st phase: constructing (40144, 300) action matrix
INFO:gensim.models.lsimodel:orthonormalizing (40144, 300) action matrix
INFO:gensim.models.lsimodel:2nd phase: running dense svd on (300, 20000) matrix
INFO:gensim.models.lsimodel:computing the final decomposition
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 13.662% of energy
spectrum)
INFO:gensim.models.lsimodel:merging projections: (40144, 200) + (40144, 200)
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 13.373% of energy
spectrum)
INFO:gensim.models.lsimodel:processed documents up to #40000
INFO:gensim.models.lsimodel:topic #0(18.799): 0.212*"league" + 0.148*"japan" +
0.134*"player" + 0.132*"actor" + 0.125*"footballer" + 0.122*"german" +
0.120*"actress" + 0.119*"politician" + 0.118*"played" + 0.115*"team"
INFO:gensim.models.lsimodel:topic #1(15.407): -0.515*"league" + -0.257*"japan" +
-0.249*"club" + -0.213*"cup" + -0.198*"played" + -0.195*"team" +
-0.164*"football" + -0.120*"division" + -0.105*"goals" + -0.104*"statistics"
INFO:gensim.models.lsimodel:topic #2(13.072): 0.223*"footballer" +
0.213*"politician" + 0.205*"actor" + 0.188*"actress" + -0.173*"album" +
-0.142*"band" + 0.141*"german" + 0.140*"writer" + 0.138*"singer" +
-0.132*"music"
INFO:gensim.models.lsimodel:topic #3(11.788): -0.362*"album" + -0.289*"band" +
-0.203*"music" + -0.182*"song" + -0.160*"released" + -0.146*"albums" +
-0.146*"guitar" + -0.124*"chart" + -0.120*"songs" + 0.119*"county"
INFO:gensim.models.lsimodel:topic #4(10.381): -0.369*"nhl" + 0.325*"japan" +
-0.276*"hurricane" + -0.221*"hockey" + -0.210*"tropical" + -0.203*"storm" +
0.149*"emperor" + -0.149*"season" + 0.125*"province" + -0.112*"montreal"
INFO:gensim.models.lsimodel:preparing a new chunk of documents
INFO:gensim.models.lsimodel:using 100 extra samples and 2 power iterations
INFO:gensim.models.lsimodel:1st phase: constructing (40144, 300) action matrix
INFO:gensim.models.lsimodel:orthonormalizing (40144, 300) action matrix
INFO:gensim.models.lsimodel:2nd phase: running dense svd on (300, 20000) matrix
INFO:gensim.models.lsimodel:computing the final decomposition
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 14.802% of energy
spectrum)
INFO:gensim.models.lsimodel:merging projections: (40144, 200) + (40144, 200)
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 11.757% of energy
spectrum)
INFO:gensim.models.lsimodel:processed documents up to #60000
```

```
INFO:gensim.models.lsimodel:topic #0(21.839): 0.158*"league" + 0.123*"japan" +
0.120*"actor" + 0.116*"played" + 0.112*"player" + 0.109*"team" + 0.105*"album" +
0.104*"actress" + 0.100*"movie" + 0.098*"singer"
INFO:gensim.models.lsimodel:topic #1(17.038): -0.473*"league" + -0.259*"japan" +
-0.230*"team" + -0.229*"club" + -0.222*"played" + -0.221*"cup" +
-0.179*"football" + -0.114*"goals" + -0.113*"games" + -0.112*"nhl"
INFO:gensim.models.lsimodel:topic #2(15.008): -0.421*"album" + -0.281*"band" +
-0.212*"released" + -0.206*"song" + -0.183*"albums" + -0.183*"chart" +
-0.171*"music" + -0.148*"guitar" + -0.137*"vocals" + 0.120*"politician"
INFO:gensim.models.lsimodel:topic #3(14.132): -0.225*"actor" + 0.216*"river" +
-0.200*"actress" + -0.180*"politician" + -0.176*"footballer" + -0.167*"singer" +
0.148*"jpg" + 0.142*"county" + -0.126*"writer" + -0.116*"player"
INFO:gensim.models.lsimodel:topic #4(12.715): -0.351*"wrestling" +
-0.331*"championship" + -0.301*"wwe" + -0.289*"match" + -0.225*"defeated" +
-0.191*"tag" + 0.171*"japan" + 0.158*"league" + -0.129*"heavyweight" +
0.129*"album"
INFO:gensim.models.lsimodel:preparing a new chunk of documents
INFO:gensim.models.lsimodel:using 100 extra samples and 2 power iterations
INFO:gensim.models.lsimodel:1st phase: constructing (40144, 300) action matrix
INFO:gensim.models.lsimodel:orthonormalizing (40144, 300) action matrix
INFO:gensim.models.lsimodel:2nd phase: running dense svd on (300, 20000) matrix
INFO:gensim.models.lsimodel:computing the final decomposition
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 14.380% of energy
spectrum)
INFO:gensim.models.lsimodel:merging projections: (40144, 200) + (40144, 200)
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 11.077% of energy
spectrum)
INFO:gensim.models.lsimodel:processed documents up to #80000
INFO:gensim.models.lsimodel:topic #0(24.430): 0.133*"league" + 0.114*"actor" +
0.111*"movie" + 0.103*"player" + 0.103*"played" + 0.099*"politician" +
0.099*"album" + 0.097*"team" + 0.097*"actress" + 0.094*"japan"
INFO:gensim.models.lsimodel:topic #1(18.297): -0.475*"league" + -0.243*"cup" +
-0.238*"team" + -0.233*"club" + -0.225*"played" + -0.219*"japan" +
-0.195*"football" + -0.142*"goals" + -0.113*"championship" + -0.113*"season"
INFO:gensim.models.lsimodel:topic #2(16.736): -0.362*"album" + 0.250*"county" +
-0.212*"song" + -0.212*"band" + -0.200*"released" + 0.176*"px" + -0.168*"chart"
+ -0.165*"music" + -0.155*"albums" + 0.128*"river"
INFO:gensim.models.lsimodel:topic #3(15.796): -0.330*"county" + 0.207*"actor" +
0.200*"politician" + -0.182*"river" + 0.178*"actress" + -0.176*"album" +
0.148*"footballer" + -0.125*"jpg" + 0.121*"minister" + -0.121*"district"
INFO:gensim.models.lsimodel:topic #4(14.449): -0.519*"county" + -0.427*"px" +
-0.159*"album" + -0.155*"party" + 0.142*"jpg" + 0.128*"bar" + 0.110*"file" +
-0.106*"election" + -0.106*"democratic" + -0.094*"republican"
INFO:gensim.models.lsimodel:preparing a new chunk of documents
INFO:gensim.models.lsimodel:using 100 extra samples and 2 power iterations
INFO:gensim.models.lsimodel:1st phase: constructing (40144, 300) action matrix
INFO:gensim.models.lsimodel:orthonormalizing (40144, 300) action matrix
INFO:gensim.models.lsimodel:2nd phase: running dense svd on (300, 11800) matrix
```

```
INFO:gensim.models.lsimodel:computing the final decomposition
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 12.984% of energy
spectrum)
INFO:gensim.models.lsimodel:merging projections: (40144, 200) + (40144, 200)
INFO:gensim.models.lsimodel:keeping 200 factors (discarding 8.407% of energy
spectrum)
INFO:gensim.models.lsimodel:processed documents up to #91800
INFO:gensim.models.lsimodel:topic #0(26.039): 0.155*"league" + 0.125*"team" +
0.124*"japan" + 0.122*"played" + 0.108*"movie" + 0.107*"player" + 0.104*"actor"
+ 0.094*"club" + 0.094*"politician" + 0.094*"cup"
INFO:gensim.models.lsimodel:topic #1(20.409): -0.419*"league" + -0.266*"japan" +
-0.255*"team" + -0.240*"cup" + -0.231*"club" + -0.227*"played" +
-0.185*"football" + -0.167*"goals" + -0.110*"apps" + -0.108*"championship"
INFO:gensim.models.lsimodel:topic #2(17.646): -0.340*"album" + 0.251*"px" +
-0.217*"song" + -0.199*"released" + 0.189*"county" + -0.189*"band" +
-0.168*"music" + -0.160*"chart" + -0.142*"albums" + 0.134*"party"
INFO:gensim.models.lsimodel:topic #3(16.406): -0.260*"county" +
0.201*"politician" + 0.197*"actor" + -0.184*"river" + -0.170*"jpg" +
0.168*"actress" + 0.133*"footballer" + -0.133*"file" + 0.130*"minister" +
-0.126*"album"
INFO:gensim.models.lsimodel:topic #4(15.586): -0.613*"px" + -0.253*"album" +
-0.218*"party" + -0.146*"democratic" + -0.137*"band" + -0.137*"election" +
-0.137*"song" + -0.130*"chart" + 0.126*"movie" + -0.124*"republican"
INFO:gensim.utils:LsiModel lifecycle event {'msg': 'trained
LsiModel<num_terms=40144, num_topics=200, decay=1.0, chunksize=20000> in
108.32s', 'datetime': '2024-12-03T07:32:05.606132', 'gensim': '4.3.3', 'python':
'3.10.12 (main, Nov  6 2024, 20:22:13) [GCC 11.4.0]', 'platform':
'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'created'}

CPU times: user 2min 11s, sys: 8.97 s, total: 2min 20s
Wall time: 1min 48s
```

The LSI transformation goes from a space of high dimensionality (~TFIDF, tens of thousands) into a space of low dimensionality (a few hundreds; here 200). For this reason it can also seen as **dimensionality reduction**.

As always, the transformations are applied "lazily", so the resulting output corpus is streamed as well:

[28]:
```
print(next(iter(lsi_model[tfidf_model[mm_corpus]])))
```

```
[(0, 0.22393276989999275), (1, 0.07501314293211767), (2, 0.07275384672982214),
(3, -0.001961009056115981), (4, 0.04583826729681909), (5, 0.049213755133140975),
(6, 0.07508945225959064), (7, 0.05778063755480169), (8, 0.019264736674517585),
(9, 0.020731427194454692), (10, 0.06754124807391265), (11,
-0.005845872717793255), (12, 0.06350898211848328), (13, 0.03445918305580428),
(14, -0.016220735885074996), (15, -0.011463269089408892), (16,
0.045255754301228475), (17, -0.002746651617059838), (18, 0.049425340683087314),
(19, 0.007842511625561623), (20, 0.05545586003595362), (21,
```

0.07409170720137478), (22, 0.006452992076814771), (23, 0.04075030802361773),
(24, -0.04620519626374119), (25, -0.03800413976917934), (26,
0.038658858099167456), (27, 0.025385442571694675), (28, -0.02827900805216565),
(29, -0.04496748293415219), (30, 0.014515694550262498), (31,
-0.012112955007708292), (32, -0.03008396533293991), (33, 0.04132342439135069),
(34, -0.0014373067311077908), (35, 0.001002569602345121), (36,
-0.02115817672414308), (37, 0.004168869925649381), (38, -0.002781369561528),
(39, 0.01588339360126847), (40, -0.034496120395062314), (41,
0.019524002234163458), (42, -0.02146900780323419), (43, 0.0115289085919467),
(44, -0.028845916150439844), (45, 0.02919316420978608), (46,
0.015316724961770002), (47, 0.02549757235424949), (48, 0.015268887895742026),
(49, -0.0006024746670320056), (50, -0.059912143840196844), (51,
-0.004552083005811392), (52, 0.01852412171929603), (53, -0.002401769386574912),
(54, -0.002445039953896686), (55, -0.02191893280300557), (56,
0.024249090163370797), (57, 0.00289611106018478), (58, 0.00912582128036506),
(59, 0.05389278010795387), (60, -0.0007843780993712056), (61,
0.019467092352574475), (62, 0.02747305851538176), (63, 0.026198493965950996),
(64, -0.018882686048552495), (65, 0.032735118562836965), (66,
0.0066557202062515215), (67, -0.01390973475211677), (68, -0.008281527789140285),
(69, 0.009978718691513257), (70, 0.0014846050102702174), (71,
0.0016856623043096366), (72, 0.015705151997557162), (73, 0.02086283295158448),
(74, 0.006912288402862519), (75, -0.0434373433499784), (76,
0.005140163021175716), (77, -0.0390832313832008), (78, -0.044692502957108415),
(79, 0.014397847481329109), (80, -0.05371943959345275), (81,
-0.002689638329404038), (82, 0.0039199735542547985), (83,
-0.019991652295075678), (84, 0.033120425295335224), (85, -0.05471829049537553),
(86, -0.05371219873519711), (87, -0.007064970462560247), (88,
-0.013792349391980025), (89, 0.007505040701827598), (90, 0.008137641300177625),
(91, 0.013035034408912407), (92, -0.021056225681208983), (93,
-0.0040275611088436405), (94, -0.0566559910629513), (95, 0.017002164516669055),
(96, -0.04227303976297809), (97, 0.015067850865716262), (98,
-0.008251538870278227), (99, 0.0010931037366001236), (100,
-0.017297799669164796), (101, 0.042999107947893994), (102,
-0.01228274007348246), (103, -0.026220893826243232), (104,
0.015721322997120105), (105, -0.046502166519594754), (106,
-0.02281765154365831), (107, -0.01979861495544322), (108, -0.03654092864024723),
(109, 0.01720797614072933), (110, -0.01989862343861351), (111,
0.004160703216709971), (112, -0.022230126563972996), (113,
-0.008620714262559021), (114, -0.05708416786449753), (115,
-0.039674677641061244), (116, 0.025357901490931277), (117,
-0.004842289669099021), (118, 0.008727593934270873), (119,
-0.06629841452846334), (120, -0.019156881373177614), (121,
-0.014926727583575517), (122, -0.023330525778154275), (123,
0.003622801640923261), (124, -0.01582757493163932), (125,
-0.009368664284156706), (126, -0.05341910886719008), (127,
-0.03048953636193779), (128, 0.01675018114358188), (129, 0.009874022432889267),
(130, 0.008306383364887433), (131, -0.049176924026671544), (132,
0.02914431281110004), (133, 0.005902618847000011), (134, -0.01646367515647596),

```
(135, 0.01636897319330848), (136, -0.07389545938133936), (137,
-0.0636115925070111), (138, -0.03789465239113594), (139, 0.017473462799466818),
(140, -0.02168547878156946), (141, -0.020416030132667778), (142,
-0.02896344261299046), (143, -0.010722248771624669), (144,
0.004664901266674675), (145, 0.022080092840441157), (146,
0.0017504374177559027), (147, 0.040071273580944346), (148,
0.007011119516756241), (149, 0.03355627772723473), (150, 0.027615288969716694),
(151, -0.0067989058811471), (152, 0.015019092948152958), (153,
0.011220026317505677), (154, -2.1162159988055774e-06), (155,
0.035787493514356575), (156, 0.017243302036371106), (157, -0.0216828257880564),
(158, 0.011299288826883486), (159, -0.01901716413196372), (160,
0.04836292596000211), (161, 0.057813948847975055), (162, 0.038762252191298224),
(163, 0.016473679981506156), (164, 0.006002568702656044), (165,
0.014618293735914733), (166, -0.009973207438297082), (167, 0.0704403120639505),
(168, 0.01761132755750609), (169, 0.12602325768697817), (170,
0.03260523430861253), (171, -0.038177778279078), (172, 0.019369567753962272),
(173, -0.0058351336001990954), (174, -0.003090105355751445), (175,
-0.009672893981252179), (176, -0.05096410080338765), (177, 0.01242766211514128),
(178, 0.013341299250303322), (179, -0.02804353804762845), (180,
0.037626199249081956), (181, -0.0046036191907624716), (182,
0.03399553824299241), (183, -0.013428231523204307), (184,
-0.0051892557193116366), (185, -0.02611604858323847), (186,
-0.006538484729465975), (187, 0.023171767832333396), (188,
0.010437778342618977), (189, 0.003084608647941442), (190,
-0.012962822921219136), (191, 0.0223314053957382), (192,
-0.0035047277895373257), (193, -0.024977400033845243), (194,
0.00016931639499144236), (195, 0.03463321867093954), (196,
0.020133699398957015), (197, 0.01317734726126596), (198, 0.006713893377591693),
(199, 0.006843365937687729)]
```

[29]:
```
# cache the transformed corpora to disk, for use in later notebooks
%time gensim.corpora.MmCorpus.serialize('./data/wiki_tfidf.mm',␣
 ↪tfidf_model[mm_corpus])
%time gensim.corpora.MmCorpus.serialize('./data/wiki_lsa.mm',␣
 ↪lsi_model[tfidf_model[mm_corpus]])
# gensim.corpora.MmCorpus.serialize('./data/wiki_lda.mm', lda_model[mm_corpus])
```

```
INFO:gensim.corpora.mmcorpus:storing corpus in Matrix Market format to
./data/wiki_tfidf.mm
INFO:gensim.matutils:saving sparse matrix to ./data/wiki_tfidf.mm
INFO:gensim.matutils:PROGRESS: saving document #0
INFO:gensim.matutils:PROGRESS: saving document #1000
INFO:gensim.matutils:PROGRESS: saving document #2000
INFO:gensim.matutils:PROGRESS: saving document #3000
INFO:gensim.matutils:PROGRESS: saving document #4000
INFO:gensim.matutils:PROGRESS: saving document #5000
INFO:gensim.matutils:PROGRESS: saving document #6000
INFO:gensim.matutils:PROGRESS: saving document #7000
```

```
INFO:gensim.matutils:PROGRESS: saving document #8000
INFO:gensim.matutils:PROGRESS: saving document #9000
INFO:gensim.matutils:PROGRESS: saving document #10000
INFO:gensim.matutils:PROGRESS: saving document #11000
INFO:gensim.matutils:PROGRESS: saving document #12000
INFO:gensim.matutils:PROGRESS: saving document #13000
INFO:gensim.matutils:PROGRESS: saving document #14000
INFO:gensim.matutils:PROGRESS: saving document #15000
INFO:gensim.matutils:PROGRESS: saving document #16000
INFO:gensim.matutils:PROGRESS: saving document #17000
INFO:gensim.matutils:PROGRESS: saving document #18000
INFO:gensim.matutils:PROGRESS: saving document #19000
INFO:gensim.matutils:PROGRESS: saving document #20000
INFO:gensim.matutils:PROGRESS: saving document #21000
INFO:gensim.matutils:PROGRESS: saving document #22000
INFO:gensim.matutils:PROGRESS: saving document #23000
INFO:gensim.matutils:PROGRESS: saving document #24000
INFO:gensim.matutils:PROGRESS: saving document #25000
INFO:gensim.matutils:PROGRESS: saving document #26000
INFO:gensim.matutils:PROGRESS: saving document #27000
INFO:gensim.matutils:PROGRESS: saving document #28000
INFO:gensim.matutils:PROGRESS: saving document #29000
INFO:gensim.matutils:PROGRESS: saving document #30000
INFO:gensim.matutils:PROGRESS: saving document #31000
INFO:gensim.matutils:PROGRESS: saving document #32000
INFO:gensim.matutils:PROGRESS: saving document #33000
INFO:gensim.matutils:PROGRESS: saving document #34000
INFO:gensim.matutils:PROGRESS: saving document #35000
INFO:gensim.matutils:PROGRESS: saving document #36000
INFO:gensim.matutils:PROGRESS: saving document #37000
INFO:gensim.matutils:PROGRESS: saving document #38000
INFO:gensim.matutils:PROGRESS: saving document #39000
INFO:gensim.matutils:PROGRESS: saving document #40000
INFO:gensim.matutils:PROGRESS: saving document #41000
INFO:gensim.matutils:PROGRESS: saving document #42000
INFO:gensim.matutils:PROGRESS: saving document #43000
INFO:gensim.matutils:PROGRESS: saving document #44000
INFO:gensim.matutils:PROGRESS: saving document #45000
INFO:gensim.matutils:PROGRESS: saving document #46000
INFO:gensim.matutils:PROGRESS: saving document #47000
INFO:gensim.matutils:PROGRESS: saving document #48000
INFO:gensim.matutils:PROGRESS: saving document #49000
INFO:gensim.matutils:PROGRESS: saving document #50000
INFO:gensim.matutils:PROGRESS: saving document #51000
INFO:gensim.matutils:PROGRESS: saving document #52000
INFO:gensim.matutils:PROGRESS: saving document #53000
INFO:gensim.matutils:PROGRESS: saving document #54000
INFO:gensim.matutils:PROGRESS: saving document #55000
```

```
INFO:gensim.matutils:PROGRESS: saving document #56000
INFO:gensim.matutils:PROGRESS: saving document #57000
INFO:gensim.matutils:PROGRESS: saving document #58000
INFO:gensim.matutils:PROGRESS: saving document #59000
INFO:gensim.matutils:PROGRESS: saving document #60000
INFO:gensim.matutils:PROGRESS: saving document #61000
INFO:gensim.matutils:PROGRESS: saving document #62000
INFO:gensim.matutils:PROGRESS: saving document #63000
INFO:gensim.matutils:PROGRESS: saving document #64000
INFO:gensim.matutils:PROGRESS: saving document #65000
INFO:gensim.matutils:PROGRESS: saving document #66000
INFO:gensim.matutils:PROGRESS: saving document #67000
INFO:gensim.matutils:PROGRESS: saving document #68000
INFO:gensim.matutils:PROGRESS: saving document #69000
INFO:gensim.matutils:PROGRESS: saving document #70000
INFO:gensim.matutils:PROGRESS: saving document #71000
INFO:gensim.matutils:PROGRESS: saving document #72000
INFO:gensim.matutils:PROGRESS: saving document #73000
INFO:gensim.matutils:PROGRESS: saving document #74000
INFO:gensim.matutils:PROGRESS: saving document #75000
INFO:gensim.matutils:PROGRESS: saving document #76000
INFO:gensim.matutils:PROGRESS: saving document #77000
INFO:gensim.matutils:PROGRESS: saving document #78000
INFO:gensim.matutils:PROGRESS: saving document #79000
INFO:gensim.matutils:PROGRESS: saving document #80000
INFO:gensim.matutils:PROGRESS: saving document #81000
INFO:gensim.matutils:PROGRESS: saving document #82000
INFO:gensim.matutils:PROGRESS: saving document #83000
INFO:gensim.matutils:PROGRESS: saving document #84000
INFO:gensim.matutils:PROGRESS: saving document #85000
INFO:gensim.matutils:PROGRESS: saving document #86000
INFO:gensim.matutils:PROGRESS: saving document #87000
INFO:gensim.matutils:PROGRESS: saving document #88000
INFO:gensim.matutils:PROGRESS: saving document #89000
INFO:gensim.matutils:PROGRESS: saving document #90000
INFO:gensim.matutils:PROGRESS: saving document #91000
INFO:gensim.matutils:saved 91800x40144 matrix, density=0.238%
(8783660/3685219200)
INFO:gensim.corpora.indexedcorpus:saving MmCorpus index to
./data/wiki_tfidf.mm.index
INFO:gensim.corpora.mmcorpus:storing corpus in Matrix Market format to
./data/wiki_lsa.mm
INFO:gensim.matutils:saving sparse matrix to ./data/wiki_lsa.mm

CPU times: user 43.9 s, sys: 1.52 s, total: 45.5 s
Wall time: 46.1 s

INFO:gensim.matutils:PROGRESS: saving document #0
INFO:gensim.matutils:PROGRESS: saving document #1000
```

```
INFO:gensim.matutils:PROGRESS: saving document #2000
INFO:gensim.matutils:PROGRESS: saving document #3000
INFO:gensim.matutils:PROGRESS: saving document #4000
INFO:gensim.matutils:PROGRESS: saving document #5000
INFO:gensim.matutils:PROGRESS: saving document #6000
INFO:gensim.matutils:PROGRESS: saving document #7000
INFO:gensim.matutils:PROGRESS: saving document #8000
INFO:gensim.matutils:PROGRESS: saving document #9000
INFO:gensim.matutils:PROGRESS: saving document #10000
INFO:gensim.matutils:PROGRESS: saving document #11000
INFO:gensim.matutils:PROGRESS: saving document #12000
INFO:gensim.matutils:PROGRESS: saving document #13000
INFO:gensim.matutils:PROGRESS: saving document #14000
INFO:gensim.matutils:PROGRESS: saving document #15000
INFO:gensim.matutils:PROGRESS: saving document #16000
INFO:gensim.matutils:PROGRESS: saving document #17000
INFO:gensim.matutils:PROGRESS: saving document #18000
INFO:gensim.matutils:PROGRESS: saving document #19000
INFO:gensim.matutils:PROGRESS: saving document #20000
INFO:gensim.matutils:PROGRESS: saving document #21000
INFO:gensim.matutils:PROGRESS: saving document #22000
INFO:gensim.matutils:PROGRESS: saving document #23000
INFO:gensim.matutils:PROGRESS: saving document #24000
INFO:gensim.matutils:PROGRESS: saving document #25000
INFO:gensim.matutils:PROGRESS: saving document #26000
INFO:gensim.matutils:PROGRESS: saving document #27000
INFO:gensim.matutils:PROGRESS: saving document #28000
INFO:gensim.matutils:PROGRESS: saving document #29000
INFO:gensim.matutils:PROGRESS: saving document #30000
INFO:gensim.matutils:PROGRESS: saving document #31000
INFO:gensim.matutils:PROGRESS: saving document #32000
INFO:gensim.matutils:PROGRESS: saving document #33000
INFO:gensim.matutils:PROGRESS: saving document #34000
INFO:gensim.matutils:PROGRESS: saving document #35000
INFO:gensim.matutils:PROGRESS: saving document #36000
INFO:gensim.matutils:PROGRESS: saving document #37000
INFO:gensim.matutils:PROGRESS: saving document #38000
INFO:gensim.matutils:PROGRESS: saving document #39000
INFO:gensim.matutils:PROGRESS: saving document #40000
INFO:gensim.matutils:PROGRESS: saving document #41000
INFO:gensim.matutils:PROGRESS: saving document #42000
INFO:gensim.matutils:PROGRESS: saving document #43000
INFO:gensim.matutils:PROGRESS: saving document #44000
INFO:gensim.matutils:PROGRESS: saving document #45000
INFO:gensim.matutils:PROGRESS: saving document #46000
INFO:gensim.matutils:PROGRESS: saving document #47000
INFO:gensim.matutils:PROGRESS: saving document #48000
INFO:gensim.matutils:PROGRESS: saving document #49000
```

```
INFO:gensim.matutils:PROGRESS: saving document #50000
INFO:gensim.matutils:PROGRESS: saving document #51000
INFO:gensim.matutils:PROGRESS: saving document #52000
INFO:gensim.matutils:PROGRESS: saving document #53000
INFO:gensim.matutils:PROGRESS: saving document #54000
INFO:gensim.matutils:PROGRESS: saving document #55000
INFO:gensim.matutils:PROGRESS: saving document #56000
INFO:gensim.matutils:PROGRESS: saving document #57000
INFO:gensim.matutils:PROGRESS: saving document #58000
INFO:gensim.matutils:PROGRESS: saving document #59000
INFO:gensim.matutils:PROGRESS: saving document #60000
INFO:gensim.matutils:PROGRESS: saving document #61000
INFO:gensim.matutils:PROGRESS: saving document #62000
INFO:gensim.matutils:PROGRESS: saving document #63000
INFO:gensim.matutils:PROGRESS: saving document #64000
INFO:gensim.matutils:PROGRESS: saving document #65000
INFO:gensim.matutils:PROGRESS: saving document #66000
INFO:gensim.matutils:PROGRESS: saving document #67000
INFO:gensim.matutils:PROGRESS: saving document #68000
INFO:gensim.matutils:PROGRESS: saving document #69000
INFO:gensim.matutils:PROGRESS: saving document #70000
INFO:gensim.matutils:PROGRESS: saving document #71000
INFO:gensim.matutils:PROGRESS: saving document #72000
INFO:gensim.matutils:PROGRESS: saving document #73000
INFO:gensim.matutils:PROGRESS: saving document #74000
INFO:gensim.matutils:PROGRESS: saving document #75000
INFO:gensim.matutils:PROGRESS: saving document #76000
INFO:gensim.matutils:PROGRESS: saving document #77000
INFO:gensim.matutils:PROGRESS: saving document #78000
INFO:gensim.matutils:PROGRESS: saving document #79000
INFO:gensim.matutils:PROGRESS: saving document #80000
INFO:gensim.matutils:PROGRESS: saving document #81000
INFO:gensim.matutils:PROGRESS: saving document #82000
INFO:gensim.matutils:PROGRESS: saving document #83000
INFO:gensim.matutils:PROGRESS: saving document #84000
INFO:gensim.matutils:PROGRESS: saving document #85000
INFO:gensim.matutils:PROGRESS: saving document #86000
INFO:gensim.matutils:PROGRESS: saving document #87000
INFO:gensim.matutils:PROGRESS: saving document #88000
INFO:gensim.matutils:PROGRESS: saving document #89000
INFO:gensim.matutils:PROGRESS: saving document #90000
INFO:gensim.matutils:PROGRESS: saving document #91000
INFO:gensim.matutils:saved 91800x200 matrix, density=100.000%
(18360000/18360000)
INFO:gensim.corpora.indexedcorpus:saving MmCorpus index to
./data/wiki_lsa.mm.index

CPU times: user 1min 23s, sys: 3.13 s, total: 1min 26s
```

```
Wall time: 1min 27s
```

[30]:
```
tfidf_corpus = gensim.corpora.MmCorpus('./data/wiki_tfidf.mm')
# `tfidf_corpus` is now exactly the same as `tfidf_model[wiki_corpus]`
print(tfidf_corpus)

lsi_corpus = gensim.corpora.MmCorpus('./data/wiki_lsa.mm')
# and `lsi_corpus` now equals `lsi_model[tfidf_model[wiki_corpus]]` =
 ↪`lsi_model[tfidf_corpus]`
print(lsi_corpus)
```

```
INFO:gensim.corpora.indexedcorpus:loaded corpus index from
./data/wiki_tfidf.mm.index
INFO:gensim.corpora._mmreader:initializing cython corpus reader from
./data/wiki_tfidf.mm
INFO:gensim.corpora._mmreader:accepted corpus with 91800 documents, 40144
features, 8783660 non-zero entries
INFO:gensim.corpora.indexedcorpus:loaded corpus index from
./data/wiki_lsa.mm.index
INFO:gensim.corpora._mmreader:initializing cython corpus reader from
./data/wiki_lsa.mm
INFO:gensim.corpora._mmreader:accepted corpus with 91800 documents, 200
features, 18360000 non-zero entries

MmCorpus(91800 documents, 40144 features, 8783660 non-zero entries)
MmCorpus(91800 documents, 200 features, 18360000 non-zero entries)
```

[31]:
```
now = datetime.now()

print("LSI Topic Models computed at", now)
```

```
LSI Topic Models computed at 2024-12-03 07:34:19.981992
```

1.3 Transforming unseen documents

We can use the trained models to transform new, unseen documents into the semantic space:

[32]:
```
text = "A blood cell, also called a hematocyte, is a cell produced by
 ↪hematopoiesis and normally found in blood."

# transform text into the bag-of-words space
bow_vector = id2word_wiki.doc2bow(tokenize(text))
print([(id2word_wiki[id], count) for id, count in bow_vector])
```

```
[('normally', 1), ('blood', 2), ('produced', 1), ('cell', 2)]
```

[33]:
```
# transform into LDA space
lda_vector = lda_model[bow_vector]
print(lda_vector)
```

```
# print the document's single most prominent LDA topic
print(lda_model.print_topic(max(lda_vector, key=lambda item: item[1])[0]))
```

```
[(0, 0.014289127), (1, 0.014290458), (2, 0.014288421), (3, 0.014289067), (4,
0.014287789), (5, 0.45924333), (6, 0.014287876), (7, 0.0142878), (8,
0.014288354), (9, 0.4264478)]
0.006*"music" + 0.005*"band" + 0.005*"rock" + 0.004*"album" + 0.004*"released" +
0.004*"live" + 0.003*"light" + 0.003*"game" + 0.003*"metal" + 0.003*"species"
```

Question 2: print text transformed into TFIDF space.

For stacked transformations, apply the same stack during transformation as was applied during training:

[46]:
```
# Step 1: Transform the bag-of-words vector into the TF-IDF space
tfidf_vector = tfidf_model[bow_vector]

# Step 2: Print the words and their corresponding TF-IDF values
word_tfidf_pairs = [(id2word_wiki[word_id], value) for word_id, value in␣
 ↪tfidf_vector]
print("Words and their TF-IDF values:")
for word, tfidf_value in word_tfidf_pairs:
    print(f"Word: {word}, TF-IDF Value: {tfidf_value}")

# Example output
# [('normally', 0.3314205262599425), ('blood', 0.5989905558961313),
# ('produced', 0.23752276286313315), ('cell', 0.6891688369642741)]

# Step 3: Transform the TF-IDF vector into the LSI space
lsi_vector = lsi_model[tfidf_vector]

# Step 4: Print the LSI vector
print("\nLSI Vector:")
for topic_id, topic_value in lsi_vector:
    print(f"Topic ID: {topic_id}, Topic Value: {topic_value}")

# Step 5: Determine the document's single most prominent LSI topic
# (Note: Topics are not interpretable like LDA topics)
most_prominent_topic = max(lsi_vector, key=lambda item: abs(item[1]))
topic_id = most_prominent_topic[0]
print("\nMost prominent LSI topic:")
print(lsi_model.print_topic(topic_id))
```

```
Words and their TF-IDF values:
Word: normally, TF-IDF Value: 0.3314205262599425
Word: blood, TF-IDF Value: 0.5989905558961313
Word: produced, TF-IDF Value: 0.23752276286313315
Word: cell, TF-IDF Value: 0.6891688369642741
```

```
LSI Vector:
Topic ID: 0, Topic Value: 0.020771135766093504
Topic ID: 1, Topic Value: 0.013537780757873662
Topic ID: 2, Topic Value: -0.009321309152130077
Topic ID: 3, Topic Value: -0.01559526890515083
Topic ID: 4, Topic Value: 0.01232793414820253
Topic ID: 5, Topic Value: 0.0229804553583646
Topic ID: 6, Topic Value: -0.02197305832214112
Topic ID: 7, Topic Value: 0.01926553166661128
Topic ID: 8, Topic Value: -0.0008783947593611081
Topic ID: 9, Topic Value: 0.0006519541961327018
Topic ID: 10, Topic Value: 0.002825176543564574
Topic ID: 11, Topic Value: -0.004016475074507546
Topic ID: 12, Topic Value: -0.001623739529448502
Topic ID: 13, Topic Value: 0.017035121555942813
Topic ID: 14, Topic Value: -0.015067992900782843
Topic ID: 15, Topic Value: -0.0038279656684423336
Topic ID: 16, Topic Value: -0.006499647879520855
Topic ID: 17, Topic Value: -0.0056433778766763285
Topic ID: 18, Topic Value: -0.010605510261206395
Topic ID: 19, Topic Value: -0.017781923846529733
Topic ID: 20, Topic Value: -0.015359566967121055
Topic ID: 21, Topic Value: -0.007941889976168594
Topic ID: 22, Topic Value: -0.0412862801607349
Topic ID: 23, Topic Value: 0.002961021271260929
Topic ID: 24, Topic Value: 0.04891783124716678
Topic ID: 25, Topic Value: 0.010471173850135384
Topic ID: 26, Topic Value: 0.027725098651374683
Topic ID: 27, Topic Value: -0.030635720263554357
Topic ID: 28, Topic Value: -0.009313559752918085
Topic ID: 29, Topic Value: 0.014431813075590622
Topic ID: 30, Topic Value: -0.018182395493397684
Topic ID: 31, Topic Value: 0.009204823414894037
Topic ID: 32, Topic Value: 0.0014415474402023377
Topic ID: 33, Topic Value: -0.027480912706255807
Topic ID: 34, Topic Value: -0.025913440177139473
Topic ID: 35, Topic Value: 0.051969780443060554
Topic ID: 36, Topic Value: 0.07114096017048382
Topic ID: 37, Topic Value: 0.002024293910243614
Topic ID: 38, Topic Value: 0.0042478035551312
Topic ID: 39, Topic Value: 0.0067684189729090155
Topic ID: 40, Topic Value: 0.0014318319090879262
Topic ID: 41, Topic Value: 0.01824468544029175
Topic ID: 42, Topic Value: 0.009862879053279594
Topic ID: 43, Topic Value: 0.011628409157771488
Topic ID: 44, Topic Value: 0.0037227999431635158
Topic ID: 45, Topic Value: 0.007582436681952647
Topic ID: 46, Topic Value: -0.028091833971861054
```

Topic ID: 47, Topic Value: -0.001114549805431919
Topic ID: 48, Topic Value: -0.021302491433901928
Topic ID: 49, Topic Value: 0.0004687851010451624
Topic ID: 50, Topic Value: 0.022504882682886154
Topic ID: 51, Topic Value: -0.0114725684484739
Topic ID: 52, Topic Value: 0.041368790499288585
Topic ID: 53, Topic Value: 0.01957938064467706
Topic ID: 54, Topic Value: 0.11368602371286521
Topic ID: 55, Topic Value: -0.01929395496028261
Topic ID: 56, Topic Value: -0.032899032121236904
Topic ID: 57, Topic Value: 0.06040833022392278
Topic ID: 58, Topic Value: 0.0454235036341607
Topic ID: 59, Topic Value: -0.04097455786937229
Topic ID: 60, Topic Value: 0.03873563046364316
Topic ID: 61, Topic Value: -0.05565752626367871
Topic ID: 62, Topic Value: 0.003206100339403892
Topic ID: 63, Topic Value: 0.017657608907276846
Topic ID: 64, Topic Value: 0.007442134985876923
Topic ID: 65, Topic Value: 0.044810215547976956
Topic ID: 66, Topic Value: 0.013147496916564805
Topic ID: 67, Topic Value: 0.0438917750831865
Topic ID: 68, Topic Value: -0.06514343438850345
Topic ID: 69, Topic Value: 0.02724882883732946
Topic ID: 70, Topic Value: -0.011852114106271086
Topic ID: 71, Topic Value: -0.020024404171267884
Topic ID: 72, Topic Value: 0.0988459728926261
Topic ID: 73, Topic Value: 0.024778984419879635
Topic ID: 74, Topic Value: 0.013516060176463628
Topic ID: 75, Topic Value: -0.09122778105773166
Topic ID: 76, Topic Value: 0.0766229495331158
Topic ID: 77, Topic Value: 0.002833964897346196
Topic ID: 78, Topic Value: -0.06515825115457521
Topic ID: 79, Topic Value: 0.04021425692411049
Topic ID: 80, Topic Value: -0.04708469581412111
Topic ID: 81, Topic Value: -0.037956579698985624
Topic ID: 82, Topic Value: -0.00809944882496409
Topic ID: 83, Topic Value: 0.03772721514903877
Topic ID: 84, Topic Value: -0.06528185124553672
Topic ID: 85, Topic Value: -0.01465359795662559
Topic ID: 86, Topic Value: 0.012082296504098793
Topic ID: 87, Topic Value: -0.09192978477406412
Topic ID: 88, Topic Value: -0.0754463622419324
Topic ID: 89, Topic Value: 0.05279384299130911
Topic ID: 90, Topic Value: 0.1516931388759638
Topic ID: 91, Topic Value: -0.02526101401139489
Topic ID: 92, Topic Value: -0.06265338788966587
Topic ID: 93, Topic Value: 0.0015743775143401417
Topic ID: 94, Topic Value: 0.091349857360144

Topic ID: 95, Topic Value: 0.019305648150600326
Topic ID: 96, Topic Value: 0.023686916538191383
Topic ID: 97, Topic Value: 0.030546656498686892
Topic ID: 98, Topic Value: 0.027857282667025723
Topic ID: 99, Topic Value: -0.03636354602988921
Topic ID: 100, Topic Value: -0.05128863935559473
Topic ID: 101, Topic Value: -0.0641971805590143
Topic ID: 102, Topic Value: 0.05215088571648143
Topic ID: 103, Topic Value: 0.017184922349606373
Topic ID: 104, Topic Value: -0.030717889243114242
Topic ID: 105, Topic Value: -0.008251792783359344
Topic ID: 106, Topic Value: 0.028253123256559216
Topic ID: 107, Topic Value: -0.02648811660470758
Topic ID: 108, Topic Value: 0.017202530772992122
Topic ID: 109, Topic Value: 0.018534431182961074
Topic ID: 110, Topic Value: 0.024124607150756976
Topic ID: 111, Topic Value: 0.03023498675482124
Topic ID: 112, Topic Value: -0.026156398461172264
Topic ID: 113, Topic Value: 0.006512246833304543
Topic ID: 114, Topic Value: 0.006538687453533033
Topic ID: 115, Topic Value: 0.03450764065526578
Topic ID: 116, Topic Value: 0.020129446719686778
Topic ID: 117, Topic Value: 0.00653230518993729
Topic ID: 118, Topic Value: -0.0375227187967279
Topic ID: 119, Topic Value: -0.06501056537586847
Topic ID: 120, Topic Value: 0.003857053037608655
Topic ID: 121, Topic Value: -0.024296280065495033
Topic ID: 122, Topic Value: -0.03602622962071163
Topic ID: 123, Topic Value: -0.02163781098584823
Topic ID: 124, Topic Value: 0.021935912754052187
Topic ID: 125, Topic Value: 0.041913524059483974
Topic ID: 126, Topic Value: -0.04197976103510592
Topic ID: 127, Topic Value: -0.0296736864820507
Topic ID: 128, Topic Value: -0.013689854808938431
Topic ID: 129, Topic Value: -0.0016432382305463754
Topic ID: 130, Topic Value: -0.02080480521822372
Topic ID: 131, Topic Value: 0.021014873978046
Topic ID: 132, Topic Value: -0.013173254519361969
Topic ID: 133, Topic Value: -0.018583374590170447
Topic ID: 134, Topic Value: 0.0030254238087297214
Topic ID: 135, Topic Value: 0.023867383152881567
Topic ID: 136, Topic Value: 0.042586295346710164
Topic ID: 137, Topic Value: 0.023297318759272014
Topic ID: 138, Topic Value: -0.0020046633249193486
Topic ID: 139, Topic Value: 0.03334592540039151
Topic ID: 140, Topic Value: 0.023869611466142724
Topic ID: 141, Topic Value: 0.0033902857140355094
Topic ID: 142, Topic Value: 0.001993044650939045

Topic ID: 143, Topic Value: 0.030789568633698937
Topic ID: 144, Topic Value: 0.03511830335543919
Topic ID: 145, Topic Value: 0.01724379322005411
Topic ID: 146, Topic Value: -0.006710630323549807
Topic ID: 147, Topic Value: 0.05225463416091126
Topic ID: 148, Topic Value: -0.026144073059781548
Topic ID: 149, Topic Value: 0.020615009386882516
Topic ID: 150, Topic Value: -0.014480328656628742
Topic ID: 151, Topic Value: -0.02074174144165653
Topic ID: 152, Topic Value: -0.01501158217690229
Topic ID: 153, Topic Value: -0.013124811745949618
Topic ID: 154, Topic Value: -0.0021629355721367316
Topic ID: 155, Topic Value: 0.018174194752765674
Topic ID: 156, Topic Value: 0.029196316119539028
Topic ID: 157, Topic Value: -0.057663489751152236
Topic ID: 158, Topic Value: 0.029228795721640582
Topic ID: 159, Topic Value: -0.01897307465344189
Topic ID: 160, Topic Value: -0.04917333763280769
Topic ID: 161, Topic Value: -0.0004969048623828416
Topic ID: 162, Topic Value: -0.04618610352133509
Topic ID: 163, Topic Value: -0.023193286668604478
Topic ID: 164, Topic Value: -0.014813425546944502
Topic ID: 165, Topic Value: -0.016282847300171606
Topic ID: 166, Topic Value: -0.04971071656041473
Topic ID: 167, Topic Value: -0.06563662812438133
Topic ID: 168, Topic Value: 0.0031596055624197247
Topic ID: 169, Topic Value: -0.007116412936324257
Topic ID: 170, Topic Value: -0.03342750581369191
Topic ID: 171, Topic Value: 0.023532656546973348
Topic ID: 172, Topic Value: 0.1038220680992846
Topic ID: 173, Topic Value: 0.023262432815109348
Topic ID: 174, Topic Value: -0.024867013242782423
Topic ID: 175, Topic Value: 0.05373323389500819
Topic ID: 176, Topic Value: -0.06527355285223925
Topic ID: 177, Topic Value: -0.002464541122409205
Topic ID: 178, Topic Value: -0.0242932328043677
Topic ID: 179, Topic Value: -0.03894789994942411
Topic ID: 180, Topic Value: -0.04132934455937976
Topic ID: 181, Topic Value: -0.025717049957492515
Topic ID: 182, Topic Value: -0.008899278759845555
Topic ID: 183, Topic Value: 0.012091924024502996
Topic ID: 184, Topic Value: -0.037332690448036755
Topic ID: 185, Topic Value: -0.05627647013435717
Topic ID: 186, Topic Value: -0.005676822720036512
Topic ID: 187, Topic Value: 0.005267770467613528
Topic ID: 188, Topic Value: -0.028280347972816568
Topic ID: 189, Topic Value: 0.02681593900806952
Topic ID: 190, Topic Value: 0.010248187017871165

```
Topic ID: 191, Topic Value: -0.0007512142045434968
Topic ID: 192, Topic Value: 0.01392645674086292
Topic ID: 193, Topic Value: 0.00040362360791757444
Topic ID: 194, Topic Value: -0.02443114401263285
Topic ID: 195, Topic Value: 0.04390925278330525
Topic ID: 196, Topic Value: 0.0005840213558438116
Topic ID: 197, Topic Value: 0.036419901881359136
Topic ID: 198, Topic Value: -0.006094682162891634
Topic ID: 199, Topic Value: 0.012740629715525961

Most prominent LSI topic:
0.222*"cells" + 0.199*"cell" + -0.179*"kw" + 0.167*"australia" + -0.167*"hp" +
-0.150*"election" + 0.142*"windows" + 0.136*"bridge" + -0.118*"covid" +
-0.112*"episode"
```

[35]:
```
# store all trained models to disk
lda_model.save('./data/lda_wiki.model')
lsi_model.save('./data/lsi_wiki.model')
tfidf_model.save('./data/tfidf_wiki.model')
id2word_wiki.save('./data/wiki.dictionary')
```

```
INFO:gensim.utils:LdaState lifecycle event {'fname_or_handle':
'./data/lda_wiki.model.state', 'separately': 'None', 'sep_limit': 10485760,
'ignore': frozenset(), 'datetime': '2024-12-03T07:34:20.044237', 'gensim':
'4.3.3', 'python': '3.10.12 (main, Nov  6 2024, 20:22:13) [GCC 11.4.0]',
'platform': 'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'saving'}
INFO:gensim.utils:saved ./data/lda_wiki.model.state
INFO:gensim.utils:LdaModel lifecycle event {'fname_or_handle':
'./data/lda_wiki.model', 'separately': "['expElogbeta', 'sstats']", 'sep_limit':
10485760, 'ignore': ['id2word', 'dispatcher', 'state'], 'datetime':
'2024-12-03T07:34:20.101246', 'gensim': '4.3.3', 'python': '3.10.12 (main, Nov
6 2024, 20:22:13) [GCC 11.4.0]', 'platform': 'Linux-6.1.85+-x86_64-with-
glibc2.35', 'event': 'saving'}
INFO:gensim.utils:storing np array 'expElogbeta' to
./data/lda_wiki.model.expElogbeta.npy
INFO:gensim.utils:not storing attribute id2word
INFO:gensim.utils:not storing attribute dispatcher
INFO:gensim.utils:not storing attribute state
INFO:gensim.utils:saved ./data/lda_wiki.model
INFO:gensim.utils:Projection lifecycle event {'fname_or_handle':
'./data/lsi_wiki.model.projection', 'separately': 'None', 'sep_limit': 10485760,
'ignore': frozenset(), 'datetime': '2024-12-03T07:34:20.131977', 'gensim':
'4.3.3', 'python': '3.10.12 (main, Nov  6 2024, 20:22:13) [GCC 11.4.0]',
'platform': 'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'saving'}
INFO:gensim.utils:saved ./data/lsi_wiki.model.projection
INFO:gensim.utils:LsiModel lifecycle event {'fname_or_handle':
'./data/lsi_wiki.model', 'separately': 'None', 'sep_limit': 10485760, 'ignore':
['projection', 'dispatcher'], 'datetime': '2024-12-03T07:34:20.304952',
```

```
'gensim': '4.3.3', 'python': '3.10.12 (main, Nov  6 2024, 20:22:13) [GCC
11.4.0]', 'platform': 'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'saving'}
INFO:gensim.utils:not storing attribute projection
INFO:gensim.utils:not storing attribute dispatcher
INFO:gensim.utils:saved ./data/lsi_wiki.model
INFO:gensim.utils:TfidfModel lifecycle event {'fname_or_handle':
'./data/tfidf_wiki.model', 'separately': 'None', 'sep_limit': 10485760,
'ignore': frozenset(), 'datetime': '2024-12-03T07:34:20.356613', 'gensim':
'4.3.3', 'python': '3.10.12 (main, Nov  6 2024, 20:22:13) [GCC 11.4.0]',
'platform': 'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'saving'}
INFO:gensim.utils:saved ./data/tfidf_wiki.model
INFO:gensim.utils:Dictionary lifecycle event {'fname_or_handle':
'./data/wiki.dictionary', 'separately': 'None', 'sep_limit': 10485760, 'ignore':
frozenset(), 'datetime': '2024-12-03T07:34:20.644261', 'gensim': '4.3.3',
'python': '3.10.12 (main, Nov  6 2024, 20:22:13) [GCC 11.4.0]', 'platform':
'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'saving'}
INFO:gensim.utils:saved ./data/wiki.dictionary
```

[36]:
```
# load the same model back; the result is equal to `lda_model`
same_lda_model = gensim.models.LdaModel.load('./data/lda_wiki.model')
```

```
INFO:gensim.utils:loading LdaModel object from ./data/lda_wiki.model
INFO:gensim.utils:loading expElogbeta from ./data/lda_wiki.model.expElogbeta.npy
with mmap=None
INFO:gensim.utils:setting ignored attribute id2word to None
INFO:gensim.utils:setting ignored attribute dispatcher to None
INFO:gensim.utils:setting ignored attribute state to None
INFO:gensim.utils:LdaModel lifecycle event {'fname': './data/lda_wiki.model',
'datetime': '2024-12-03T07:34:20.720253', 'gensim': '4.3.3', 'python': '3.10.12
(main, Nov  6 2024, 20:22:13) [GCC 11.4.0]', 'platform':
'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'loaded'}
INFO:gensim.utils:loading LdaState object from ./data/lda_wiki.model.state
INFO:gensim.utils:LdaState lifecycle event {'fname':
'./data/lda_wiki.model.state', 'datetime': '2024-12-03T07:34:20.725293',
'gensim': '4.3.3', 'python': '3.10.12 (main, Nov  6 2024, 20:22:13) [GCC
11.4.0]', 'platform': 'Linux-6.1.85+-x86_64-with-glibc2.35', 'event': 'loaded'}
```

1.4 Evaluation

Topic modeling is an **unsupervised task**; we do not know in advance what the topics ought to look like. This makes evaluation tricky: whereas in supervised learning (classification, regression) we simply compare predicted labels to expected labels, there are no "expected labels" in topic modeling.

Each topic modeling method (LSI, LDA…) its own way of measuring internal quality (perplexity, reconstruction error…). But these are an artifact of the particular approach taken (bayesian training, matrix factorization…), and mostly of academic interest. There's no way to compare such scores across different types of topic models, either. The best way to really evaluate quality of unsupervised tasks is to **evaluate how they improve the superordinate task, the one we're actually**

training them for.

For example, when the ultimate goal is to retrieve semantically similar documents, we manually tag a set of similar documents and then see how well a given semantic model maps those similar documents together.

Such manual tagging can be resource intensive, so people hae been looking for clever ways to automate it. In Reading tea leaves: How humans interpret topic models, Wallach *et al* suggest a "word intrusion" method that works well for models where the topics are meant to be "human interpretable", such as LDA. For each trained topic, they take its first ten words, then substitute one of them with another, randomly chosen word (intruder!) and see whether a human can reliably tell which one it was. If so, the trained topic is **topically coherent** (good); if not, the topic has no discernible theme (bad):

1.5 Misplaced Words

[37]:
```
# select top 50 words for each of the 20 LDA topics
top_words = [[word for _, word in lda_model.show_topic(topicno, topn=50)] for␣
 ↪topicno in range(lda_model.num_topics)]
print(top_words)
```

```
[[0.005583402, 0.0036500155, 0.003641657, 0.0034257483, 0.003396844,
0.0033896745, 0.0032755192, 0.003272386, 0.0032516439, 0.0031186368,
0.0030405226, 0.0030360084, 0.0030282533, 0.0029203445, 0.0029130506,
0.0028747066, 0.0026654536, 0.0025563538, 0.0025524597, 0.0024486298,
0.0023646306, 0.0023521848, 0.0023513022, 0.0023480568, 0.002314297,
0.0022802134, 0.0022517277, 0.0022330114, 0.0021148075, 0.0020592168,
0.0020121406, 0.0019839504, 0.0019552512, 0.0019535483, 0.0019272179,
0.0018685394, 0.0018241741, 0.0018097537, 0.0017639016, 0.0016940563,
0.0016832809, 0.0016699083, 0.0016423594, 0.0016397523, 0.001639515,
0.0016358288, 0.001630218, 0.0016256218, 0.0016223363, 0.001615119],
[0.0060351687, 0.0042303563, 0.0040747183, 0.0040338393, 0.0039944365,
0.0031268704, 0.0030895914, 0.003055896, 0.002857495, 0.002841373, 0.00283547,
0.0026984068, 0.0024614409, 0.0023666848, 0.0023084946, 0.0022425712,
0.0022327115, 0.00220799, 0.0021012125, 0.0020701843, 0.0020603838,
0.0020259062, 0.0020122656, 0.001993242, 0.001991297, 0.0019716762,
0.0019658569, 0.001963204, 0.0019630466, 0.0019608254, 0.0018849726,
0.0018770742, 0.0018507167, 0.0018097309, 0.0017699221, 0.0017483768,
0.0017213775, 0.0016867985, 0.0016863163, 0.0016725394, 0.0016703457,
0.001668379, 0.0016434379, 0.001617666, 0.0015993428, 0.0015633757,
0.0015601036, 0.0015499752, 0.0015372896, 0.0014529909], [0.0072909147,
0.007053169, 0.005435358, 0.0053279744, 0.0052708704, 0.0047052093,
0.0046567004, 0.0043572816, 0.004239192, 0.0037538428, 0.0035929724,
0.0034971545, 0.0028771209, 0.0024616006, 0.0024548809, 0.002424563,
0.0024135048, 0.0023821283, 0.0023782812, 0.002339949, 0.0022992315,
0.002233107, 0.0022128602, 0.0021503733, 0.0020745771, 0.0020529402,
0.0020256408, 0.0018742288, 0.0018542241, 0.0017799401, 0.001729779,
0.0017182768, 0.0016902528, 0.0016676013, 0.0016669723, 0.0016633703,
0.0016520689, 0.0016485121, 0.0016405553, 0.0016305026, 0.0015787682,
```

0.001577818, 0.0015558823, 0.0015384032, 0.0015317192, 0.0015316984,
0.0015301245, 0.001522073, 0.0014992312, 0.0014767523], [0.015758948,
0.015580533, 0.008245165, 0.0050915256, 0.0047229417, 0.0043692375, 0.004190899,
0.0041905115, 0.003914692, 0.003493336, 0.003483429, 0.0034614615, 0.0032117406,
0.0030454942, 0.0030163492, 0.0029822737, 0.0028350682, 0.0027882373,
0.0027497825, 0.002703099, 0.0025518995, 0.0025388957, 0.0025103043,
0.0024197663, 0.002397081, 0.0023176286, 0.00224179, 0.0022283327, 0.0022044617,
0.0021705024, 0.002083709, 0.0020717927, 0.002071539, 0.0020136554,
0.0019966308, 0.0019606308, 0.0019601013, 0.0019515422, 0.0019337856,
0.0018888896, 0.00185424, 0.0018130493, 0.001812936, 0.0018095695, 0.0017778128,
0.001762935, 0.0017552534, 0.0017364871, 0.0017269535, 0.0017153593],
[0.013267425, 0.01188509, 0.010871651, 0.010819224, 0.009928567, 0.009634195,
0.009024038, 0.008875162, 0.00825516, 0.008239236, 0.006525847, 0.0062657446,
0.0047279536, 0.004614731, 0.0044843596, 0.004372473, 0.004199333, 0.004112858,
0.0040586935, 0.0037622019, 0.0034877332, 0.0034711107, 0.003149663,
0.0030928664, 0.0030728, 0.0028603133, 0.002742654, 0.002723987, 0.0026815152,
0.0026071577, 0.002463217, 0.0024554757, 0.0024503956, 0.0023501287,
0.0022793622, 0.0022660017, 0.0022424909, 0.0022266505, 0.0022211082,
0.0021976067, 0.0021745292, 0.0021540124, 0.0021446948, 0.002123015,
0.0020661245, 0.0020501374, 0.0019582608, 0.0018778238, 0.0018713292,
0.0018644257], [0.0057136063, 0.0048436685, 0.004536515, 0.004062322,
0.0036818245, 0.0035725979, 0.0033278032, 0.0031719734, 0.0031162282,
0.0030199117, 0.0029915886, 0.0029118026, 0.0028850958, 0.0028301806,
0.0028249351, 0.0026861038, 0.002501062, 0.0024917668, 0.0024061853,
0.0023920718, 0.0023654813, 0.0023653752, 0.0021410224, 0.0021068917,
0.0020821667, 0.0020775457, 0.0020767776, 0.0020466733, 0.0020033917,
0.0019551218, 0.0019268772, 0.0018983352, 0.001862713, 0.0018520859,
0.0018371393, 0.0018201683, 0.0018102031, 0.0017677003, 0.0017507678,
0.001749999, 0.0017149454, 0.0016795133, 0.0016750118, 0.001657525,
0.0016202627, 0.0016042959, 0.0015898682, 0.0015594758, 0.0015242774,
0.0015177797], [0.006710382, 0.0061116354, 0.005756131, 0.0055593126,
0.0055181496, 0.0054955212, 0.004330181, 0.004317533, 0.004132553, 0.004055705,
0.0040360447, 0.00393665, 0.0038331556, 0.00376762, 0.0037090501, 0.00364654,
0.0035958146, 0.0035814277, 0.0034456465, 0.003389268, 0.003329664,
0.0032630798, 0.0031757844, 0.0029681416, 0.0029461926, 0.0028842136,
0.0028762242, 0.0028190606, 0.0027324874, 0.0026535816, 0.0025139381,
0.0024803136, 0.0023336897, 0.0023304205, 0.002308282, 0.0022780134,
0.0022413884, 0.0022082618, 0.002207782, 0.0021859463, 0.002172062,
0.0021616318, 0.0020976574, 0.002095628, 0.0020450891, 0.0020140752,
0.0019974625, 0.0019567562, 0.0019040674, 0.0017928123], [0.014256958,
0.0064785443, 0.006055501, 0.0056825364, 0.0043962942, 0.004092535,
0.0039251936, 0.0037132583, 0.0036912158, 0.0036164536, 0.0032582304,
0.003172365, 0.0030470467, 0.0029479794, 0.002916546, 0.002770842, 0.0027330273,
0.0026730534, 0.0026098285, 0.002539092, 0.0025344305, 0.0024364304,
0.0023697382, 0.0023547483, 0.0022637267, 0.0021810378, 0.002118538,
0.0020927356, 0.0020804764, 0.0020713476, 0.0020473488, 0.002016728,
0.0020076428, 0.0019904887, 0.0019817168, 0.0019612082, 0.0019603642,
0.0019557436, 0.0019243937, 0.0018845221, 0.0018754442, 0.0018475282,

```
0.0018335017, 0.0017855283, 0.0017641936, 0.0017597748, 0.00175004,
0.0017461084, 0.0017322907, 0.0017298689], [0.007673, 0.0062189293,
0.0054936344, 0.0038335184, 0.003786035, 0.003657901, 0.003443309, 0.0033892766,
0.0033043595, 0.0032798592, 0.0030010906, 0.002972878, 0.0029236346,
0.0029190653, 0.0028713893, 0.0027455979, 0.0026168386, 0.002545007,
0.002513952, 0.0025069464, 0.0024050274, 0.0023397692, 0.0022406837,
0.0022137857, 0.0021751397, 0.0021515386, 0.0021074796, 0.0020808661,
0.0020315885, 0.0019967817, 0.001958853, 0.0019431071, 0.0019169563,
0.0018941564, 0.0018487597, 0.0018394743, 0.0018145947, 0.0017935147,
0.001748185, 0.0017357809, 0.0017143644, 0.0017117291, 0.0017096458,
0.0017052997, 0.0016591153, 0.0016516673, 0.0016467272, 0.0016389407,
0.0016373033, 0.0016181006], [0.003957242, 0.0037778786, 0.003641941,
0.0036362761, 0.0034926147, 0.0034594543, 0.0034001283, 0.0033580917,
0.0033225273, 0.0031117196, 0.002712865, 0.002699031, 0.0025966694,
0.0025417388, 0.0025000533, 0.0021209884, 0.0020992262, 0.0020680162,
0.0020154866, 0.0020133012, 0.001988912, 0.0019736027, 0.0019335988,
0.0019133647, 0.0018489723, 0.0018390205, 0.0018255186, 0.0018244947,
0.001819559, 0.0017960789, 0.0017725051, 0.0017549279, 0.0017545096,
0.0016813328, 0.0016777716, 0.001659559, 0.001652672, 0.0016282989,
0.0016118569, 0.0015994163, 0.001585802, 0.0015658232, 0.0015594539,
0.0015466312, 0.0015231832, 0.0015111156, 0.0014541209, 0.0014533082,
0.0014410603, 0.0014373705]]
```

1.6 Question 03. [12 points] Identify the source of difference and [16 points] change it so they are equivalent.

[47]:
```
# Radim's output for comparison
Radim_output = [
    ['album', 'band', 'released', 'movie', 'music', 'island', 'york', 'award',␣
 ↪'series', 'song',
     'won', 'albums', 'president', 'game', 'rock', 'british', 'england',␣
 ↪'king', 'popular',
     'video', 'sold', 'million', 'songs', 'awards', 'married', 'tour',␣
 ↪'jackson', 'live', 'mother',
     'father', 'career', 'movies', 'australia', 'games', 'said', 'came',␣
 ↪'left', 'white', 'home',
     'death', 'went', 'ford', 'got', 'single', 'bush', 'children', 'record',␣
 ↪'played', 'george',
     'love'],
    ['rgb', 'hex', 'color', 'blood', 'body', 'disease', 'person', 'blue',␣
 ↪'red', 'green', 'cells',
     'light', 'pink', 'heart', 'bc', 'woman', 'web', 'women', 'purple',␣
 ↪'cause', 'colors',
     'diseases', 'abortion', 'sex', 'cancer', 'man', 'crayola', 'ff',␣
 ↪'doctors', 'yellow', 'penis',
     'malaria', 'men', 'means', 'pain', 'male', 'violet', 'com', 'orange',␣
 ↪'immune', 'medical',
```

```
    'sexual', 'types', 'causes', 'semen', 'common', 'magenta', 'bacteria',␣
↪'brain', 'dark'],
   ['god', 'tower', 'mast', 'transmission', 'left', 'book', 'books',␣
↪'believe', 'school', 'mount',
    'church', 'jesus', 'said', 'party', 'bible', 'earth', 'religion', 'built',␣
↪'al', 'east',
    'align', 'country', 'muslims', 'things', 'christian', 'building',␣
↪'middle', 'largest',
    'children', 'written', 'roman', 'ancient', 'radio', 'kansas', 'empire',␣
↪'cities', 'live',
    'began', 'father', 'july', 'religious', 'moon', 'death', 'man',␣
↪'estimate', 'holy',
    'religions', 'government', 'today', 'king'],
   ['light', 'game', 'league', 'earth', 'energy', 'example', 'player', 'team',␣
↪'games', 'football',
    'point', 'space', 'numbers', 'mass', 'players', 'universe', 'speed',␣
↪'things', 'theory', 'sun',
    'object', 'park', 'line', 'play', 'means', 'distance', 'africa', 'ball',␣
↪'right', 'field',
    'physics', 'matter', 'club', 'force', 'black', 'stars', 'star', 'premier',␣
↪'moving', 'teams',
    'change', 'units', 'position', 'particles', 'special', 'atoms',␣
↪'electrons', 'iron',
    'scientists', 'big'],
   ['actor', 'german', 'british', 'singer', 'french', 'footballer', 'actress',␣
↪'writer',
    'politician', 'player', 'italian', 'president', 'musician', 'composer',␣
↪'king', 'ii',
    'minister', 'russian', 'prime', 'canadian', 'japanese', 'director',␣
↪'poet', 'battle',
    'governor', 'france', 'william', 'spanish', 'general', 'emperor',␣
↪'charles', 'killing',
    'painter', 'songwriter', 'george', 'movie', 'henry', 'england',␣
↪'scottish', 'james',
    'physicist', 'robert', 'queen', 'dutch', 'mathematician', 'leader',␣
↪'austrian', 'swedish',
    'ice', 'producer'],
   ['water', 'jpg', 'bridge', 'species', 'image', 'animals', 'live', 'food',␣
↪'plants', 'air',
    'birds', 'mario', 'sea', 'eat', 'file', 'living', 'plant', 'land', 'body',␣
↪'chemical', 'tree',
    'cell', 'grow', 'trees', 'common', 'inside', 'cells', 'white', 'makes',␣
↪'america', 'largest',
    'island', 'animal', 'built', 'things', 'forest', 'types', 'parts', 'form',␣
↪'places', 'fruit',
```

```
    'example', 'fish', 'big', 'ground', 'compounds', 'leaves', 'evolution',␣
↪'eggs', 'london'],
   ['president', 'government', 'country', 'union', 'july', 'party', 'korea',␣
↪'april', 'army',
    'countries', 'germany', 'british', 'december', 'international', 'al',␣
↪'january', 'usa',
    'soviet', 'february', 'independence', 'russia', 'baltimore', 'election',␣
↪'kingdom', 'france',
    'military', 'civil', 'republic', 'elected', 'french', 'usb', 'washington',␣
↪'nations',
    'capital', 'killed', 'ii', 'japan', 'britain', 'democratic', 'general',␣
↪'november', 'september',
    'vice', 'virginia', 'rights', 'house', 'october', 'political', 'minister',␣
↪'august'],
   ['language', 'word', 'languages', 'river', 'windows', 'words', 'means',␣
↪'country', 'internet',
    'example', 'lake', 'church', 'countries', 'information', 'software',␣
↪'microsoft', 'latin',
    'version', 'computers', 'person', 'things', 'population', 'free', 'web',␣
↪'program', 'pope',
    'million', 'written', 'operating', 'spoken', 'speak', 'uses', 'parts',␣
↪'file', 'europe',
    'america', 'programs', 'largest', 'catholic', 'data', 'today', 'came',␣
↪'spanish', 'change',
    'say', 'republic', 'rivers', 'user', 'released', 'greek'],
   ['music', 'person', 'countries', 'things', 'country', 'government',␣
↪'money', 'china', 'good',
    'example', 'think', 'wrote', 'say', 'word', 'said', 'means', 'popular',␣
↪'chinese', 'human',
    'want', 'common', 'fish', 'include', 'thought', 'right', 'ideas',␣
↪'modern', 'power', 'women',
    'today', 'food', 'man', 'play', 'society', 'political', 'lot', 'capital',␣
↪'social',
    'instruments', 'ancient', 'age', 'help', 'groups', 'written', 'bass',␣
↪'period', 'making',
    'guitar', 'types', 'law'],
   ['january', 'november', 'december', 'february', 'october', 'august',␣
↪'april', 'september',
    'actor', 'movie', 'july', 'german', 'germany', 'rural', 'actress',␣
↪'president', 'king',
    'singer', 'love', 'television', 'movies', 'writer', 'british', 'calendar',␣
↪'award', 'chicago',
    'disney', 'french', 'film', 'france', 'minister', 'band', 'george', 'ii',␣
↪'paul', 'rock',
```

```
    'kingdom', 'prime', 'urban', 'roman', 'man', 'james', 'music', 'director',␣
 ↪'william', 'events',
    'bavaria', 'musician', 'japan', 'india']
]

# Print Radim's output for each topic
print("Top words for each topic (Radim's Output):")
for topic_index, words in enumerate(Radim_output):
    print(f"Topic {topic_index + 1}: {', '.join(words)}")

# Check dimensions of Radim's output
num_topics_radim = len(Radim_output)
num_words_radim = len(Radim_output[0]) if num_topics_radim > 0 else 0

print("\nComparison of Topic Dimensions:")
print(f"- Radim's output contains {num_topics_radim} topics, each with␣
 ↪{num_words_radim} words.")
```

```
Top words for each topic (Radim's Output):
Topic 1: album, band, released, movie, music, island, york, award, series, song,
won, albums, president, game, rock, british, england, king, popular, video,
sold, million, songs, awards, married, tour, jackson, live, mother, father,
career, movies, australia, games, said, came, left, white, home, death, went,
ford, got, single, bush, children, record, played, george, love
Topic 2: rgb, hex, color, blood, body, disease, person, blue, red, green, cells,
light, pink, heart, bc, woman, web, women, purple, cause, colors, diseases,
abortion, sex, cancer, man, crayola, ff, doctors, yellow, penis, malaria, men,
means, pain, male, violet, com, orange, immune, medical, sexual, types, causes,
semen, common, magenta, bacteria, brain, dark
Topic 3: god, tower, mast, transmission, left, book, books, believe, school,
mount, church, jesus, said, party, bible, earth, religion, built, al, east,
align, country, muslims, things, christian, building, middle, largest, children,
written, roman, ancient, radio, kansas, empire, cities, live, began, father,
july, religious, moon, death, man, estimate, holy, religions, government, today,
king
Topic 4: light, game, league, earth, energy, example, player, team, games,
football, point, space, numbers, mass, players, universe, speed, things, theory,
sun, object, park, line, play, means, distance, africa, ball, right, field,
physics, matter, club, force, black, stars, star, premier, moving, teams,
change, units, position, particles, special, atoms, electrons, iron, scientists,
big
Topic 5: actor, german, british, singer, french, footballer, actress, writer,
politician, player, italian, president, musician, composer, king, ii, minister,
russian, prime, canadian, japanese, director, poet, battle, governor, france,
william, spanish, general, emperor, charles, killing, painter, songwriter,
george, movie, henry, england, scottish, james, physicist, robert, queen, dutch,
mathematician, leader, austrian, swedish, ice, producer
```

```
Topic 6: water, jpg, bridge, species, image, animals, live, food, plants, air,
birds, mario, sea, eat, file, living, plant, land, body, chemical, tree, cell,
grow, trees, common, inside, cells, white, makes, america, largest, island,
animal, built, things, forest, types, parts, form, places, fruit, example, fish,
big, ground, compounds, leaves, evolution, eggs, london
Topic 7: president, government, country, union, july, party, korea, april, army,
countries, germany, british, december, international, al, january, usa, soviet,
february, independence, russia, baltimore, election, kingdom, france, military,
civil, republic, elected, french, usb, washington, nations, capital, killed, ii,
japan, britain, democratic, general, november, september, vice, virginia,
rights, house, october, political, minister, august
Topic 8: language, word, languages, river, windows, words, means, country,
internet, example, lake, church, countries, information, software, microsoft,
latin, version, computers, person, things, population, free, web, program, pope,
million, written, operating, spoken, speak, uses, parts, file, europe, america,
programs, largest, catholic, data, today, came, spanish, change, say, republic,
rivers, user, released, greek
Topic 9: music, person, countries, things, country, government, money, china,
good, example, think, wrote, say, word, said, means, popular, chinese, human,
want, common, fish, include, thought, right, ideas, modern, power, women, today,
food, man, play, society, political, lot, capital, social, instruments, ancient,
age, help, groups, written, bass, period, making, guitar, types, law
Topic 10: january, november, december, february, october, august, april,
september, actor, movie, july, german, germany, rural, actress, president, king,
singer, love, television, movies, writer, british, calendar, award, chicago,
disney, french, film, france, minister, band, george, ii, paul, rock, kingdom,
prime, urban, roman, man, james, music, director, william, events, bavaria,
musician, japan, india

Comparison of Topic Dimensions:
- Radim's output contains 10 topics, each with 50 words.
```

1.6.1 Question 03. (Part 3) [12 points] Identify the source of difference.

Both outputs are printing the same information dimensions, but different information. Radim's notebook is returning the top 50 words of each topic, but this notebook is returning numbers. These numbers are actually the weights acossiated with each word, which we were trying to ignore, but somehow the words and weight got switched.

After the code runs, on part 03, we will observe words instead of numbers. While the words and topics do not perfectly match Radim's output, this is expected due to the inherent stochasticity in these models. Additionally, the Simple Wikipedia dataset has likely changed since Radim's notebook was created. This is evident from the earlier comparison of the `lda_vector` outputs, where discrepancies were already noted. These differences are sufficiently equivalent for practical purposes. Other potential reasons for the variations include differences in software versions between the notebooks. Radim's notebook likely uses older versions, as its code is no longer compatible with current environments. Furthermore, the outputs in Radim's notebook include the u prefix for non-ASCII strings, a characteristic of Python 2.x, whereas this notebook utilizes Python 3.x.

1.6.2 Question 03. (Part 3) [16 points] change it so they are equivalent.

[48]:
```
# Adjust the order of the wildcard placement to correctly extract words
top_words = []

# Iterate through all topics in the LDA model
for topic_number in range(lda_model.num_topics):
    # Extract the top 50 words for the current topic
    words = [word for word, _ in lda_model.show_topic(topic_number, topn=50)]
    top_words.append(words)

# Print the top words for each topic
print("Top words for each topic:")
for topic_index, words in enumerate(top_words):
    print(f"Topic {topic_index + 1}: {', '.join(words)}")
```

```
Top words for each topic:
Topic 1: god, movie, said, love, book, death, music, books, man, album, father,
children, person, church, wrote, words, movies, series, school, believe, word,
woman, mother, written, story, famous, heart, doctor, married, way, award, pope,
things, good, song, went, men, played, jesus, son, bible, friends, women, means,
child, jackson, popular, role, young, awards
Topic 2: person, water, things, example, earth, human, study, body, way, women,
energy, theory, usually, change, science, law, light, means, important,
countries, cause, gender, right, mass, disease, common, thought, object, sexual,
space, sun, problems, think, rights, scientists, idea, universe, makes, force,
brain, help, blood, small, speed, social, include, laws, good, symptoms, medical
Topic 3: tower, number, game, mast, transmission, player, uhf, numbers, games,
players, team, example, ball, england, jpg, football, play, usually, point,
county, text, line, radio, century, way, written, town, file, bar, value, teams,
square, stone, ring, rings, zero, ireland, built, played, times, ancient, rules,
roman, olympic, cards, means, tv, mythology, points, board
Topic 4: rgb, hex, color, water, food, red, blue, usually, green, light, fruit,
pink, purple, korea, tree, web, countries, ice, plants, white, ff, coffee,
plant, common, small, trees, crayola, air, temperature, yellow, annual,
chocolate, types, violet, word, china, grow, milk, example, dark, cold, chinese,
flowers, japanese, leaves, type, eat, com, magenta, country
Topic 5: actor, politician, singer, actress, german, footballer, french, player,
writer, british, italian, president, musician, composer, minister, canadian,
russian, director, prime, japanese, ii, king, poet, songwriter, scottish,
producer, spanish, australian, governor, general, indian, movie, william,
battle, painter, george, charles, james, france, dutch, killing, robert,
journalist, emperor, swedish, author, physicist, ice, football, baseball
Topic 6: music, band, rock, album, released, live, light, game, metal, species,
mario, animals, black, usually, small, water, birds, large, cells, video,
number, chemical, white, games, cell, lead, body, jpg, film, file, song, things,
songs, form, makes, common, iron, nintendo, types, energy, compounds, example,
red, albums, usb, popular, atoms, way, animation, guitar
```

```
Topic 7: river, jpg, country, capital, germany, government, empire, cities,
file, east, union, rural, party, island, population, largest, west, century,
countries, german, soviet, republic, important, sea, china, europe, great,
famous, built, army, million, land, power, ii, live, kingdom, region, king,
roman, france, large, town, russia, york, political, western, took, urban,
center, communist
Topic 8: president, league, king, england, henry, kansas, bush, reagan, queen,
france, house, government, election, premier, kingdom, george, party, british,
london, presidential, elected, california, york, william, carter, usa,
republican, britain, ii, vocals, home, club, washington, st, married, white,
vice, cup, governor, texas, minister, scotland, football, elizabeth, chicago,
dole, prime, office, union, great
Topic 9: language, country, windows, languages, lake, africa, countries, india,
microsoft, internet, park, software, population, century, largest, bc, version,
capital, republic, america, spanish, million, computers, sea, region, east,
free, independence, morocco, spoken, operating, european, common, west, land,
web, central, europe, word, music, popular, mountains, french, ocean, linux,
explorer, released, northern, san, data
Topic 10: island, penis, body, blood, things, usually, great, islands, man,
person, money, example, mount, word, countries, means, semen, sex, weapons,
good, sea, cells, country, cancer, cousin, language, food, animals, men, sperm,
nudity, woman, living, disease, land, paper, testicles, live, car, natural,
common, makes, male, removed, parts, half, vagina, small, denmark, sexual
```

[38]:
```python
# get all top 50 words in all 20 topics, as one large set
all_words = set(itertools.chain.from_iterable(top_words))

print("Can you spot the misplaced word in each topic?")

# for each topic, replace a word at a different index, to make it more␣
 ↪interesting
replace_index = np.random.randint(0, 10, lda_model.num_topics)

replacements = []
for topicno, words in enumerate(top_words):
    other_words = all_words.difference(words)
    replacement = np.random.choice(list(other_words))
    replacements.append((words[replace_index[topicno]], replacement))
    words[replace_index[topicno]] = replacement
    print (topicno, ' '.join([str(w) for w in words[:10]]))
    # print("%i: %s" % (topicno, ' '.join(words[:10])))
```

```
Can you spot the misplaced word in each topic?
0 0.005583402 0.0036500155 0.003641657 0.0019606308 0.003396844 0.0033896745
0.0032755192 0.003272386 0.0032516439 0.0031186368
1 0.0060351687 0.0024135048 0.0040747183 0.0040338393 0.0039944365 0.0031268704
0.0030895914 0.003055896 0.002857495 0.002841373
2 0.0072909147 0.007053169 0.005435358 0.0053279744 0.0052708704 0.0047052093
```

```
0.0046567004 0.0043572816 0.004239192 0.003657901
3 0.015758948 0.015580533 0.008245165 0.001659559 0.0047229417 0.0043692375
0.004190899 0.0041905115 0.003914692 0.003493336
4 0.013267425 0.01188509 0.010871651 0.010819224 0.009928567 0.009634195
0.009024038 0.008875162 0.0017799401 0.008239236
5 0.0057136063 0.0048436685 0.004536515 0.004062322 0.0036818245 0.0035725979
0.0033278032 0.0029130506 0.0031162282 0.0030199117
6 0.006710382 0.0061116354 0.0034926147 0.0055593126 0.0055181496 0.0054955212
0.004330181 0.004317533 0.004132553 0.004055705
7 0.014256958 0.0064785443 0.006055501 0.0016423594 0.0043962942 0.004092535
0.0039251936 0.0037132583 0.0036912158 0.0036164536
8 0.007673 0.0062189293 0.003483429 0.0038335184 0.003786035 0.003657901
0.003443309 0.0033892766 0.0033043595 0.0032798592
9 0.003957242 0.003493336 0.003641941 0.0036362761 0.0034926147 0.0034594543
0.0034001283 0.0033580917 0.0033225273 0.0031117196
```

[39]:
```
print("Actual replacements were:")
print(list(enumerate(replacements)))
```

```
Actual replacements were:
[(0, (0.0034257483, 0.0019606308)), (1, (0.0042303563, 0.0024135048)), (2,
(0.0037538428, 0.003657901)), (3, (0.0050915256, 0.001659559)), (4, (0.00825516,
0.0017799401)), (5, (0.0031719734, 0.0029130506)), (6, (0.005756131,
0.0034926147)), (7, (0.0056825364, 0.0016423594)), (8, (0.0054936344,
0.003483429)), (9, (0.0037778786, 0.003493336))]
```

[40]:
```
# evaluate on 1k documents **not** used in LDA training
doc_stream = (tokens for _, tokens in iter_wiki(wiki_file))  # generator
test_docs = list(itertools.islice(doc_stream, 8000, 9000))
```

[41]:
```
def intra_inter(model, test_docs, num_pairs=10000):
    # split each test document into two halves and compute topics for each half
    half = int(len(test_docs)/2)
    part1 = [model[id2word_wiki.doc2bow(tokens[: half])] for tokens in␣
 ↪test_docs]
    part2 = [model[id2word_wiki.doc2bow(tokens[half :])] for tokens in␣
 ↪test_docs]

    # print computed similarities (uses cossim)
    print("average cosine similarity between corresponding parts (higher is␣
 ↪better):")
    print(np.mean([gensim.matutils.cossim(p1, p2) for p1, p2 in zip(part1,␣
 ↪part2)]))

    random_pairs = np.random.randint(0, len(test_docs), size=(num_pairs, 2))
    print("average cosine similarity between 10,000 random parts (lower is␣
 ↪better):")
```

```
    print(np.mean([gensim.matutils.cossim(part1[i[0]], part2[i[1]]) for i in␣
 ↪random_pairs]))
```

[42]:
```
print("LDA results:")
intra_inter(lda_model, test_docs)
```

```
LDA results:
average cosine similarity between corresponding parts (higher is better):
0.517788666250796
average cosine similarity between 10,000 random parts (lower is better):
0.4677868095303634
```

[43]:
```
print("LSI results:")
intra_inter(lsi_model, test_docs)
```

```
LSI results:
average cosine similarity between corresponding parts (higher is better):
0.06449505593668768
average cosine similarity between 10,000 random parts (lower is better):
0.009181873947877051
```

[44]:
```
now = datetime.now()

print("Ended at", now)
```

```
Ended at 2024-12-03 07:35:31.793933
```

---

2 Topic Tagging

2.1 Question 01. [16 points] Misplaced word technique

[61]:
```
# Misplaced Word Technique: Identify the misplaced word in each topic
# Topics and words from the dataset
topics = {
    0: ["god", "political", "person", "words", "things", "word", "book",␣
 ↪"books", "said", "languages"],
    1: ["henry", "government", "countries", "capital", "river", "union",␣
 ↪"party", "republic", "east", "island"],
    2: ["music", "game", "movie", "series", "award", "player", "movies",␣
 ↪"film", "sexual", "released"],
    3: ["caffeine", "jpg", "park", "bc", "century", "file", "language",␣
 ↪"great", "built", "london"],
    4: ["actor", "politician", "actress", "women", "german", "footballer",␣
 ↪"french", "player", "british", "writer"],
```

```
    5: ["body", "person", "blood", "cells", "usually", "water", "disease",␣
 ↪"evolution", "common", "sexual"],
    6: ["water", "earth", "light", "number", "species", "example", "energy",␣
 ↪"small", "writer", "numbers"],
    7: ["president", "actor", "operating", "actress", "album", "band", "king",␣
 ↪"henry", "politician", "german"],
    8: ["speed", "hex", "color", "web", "green", "blue", "red", "pink",␣
 ↪"purple", "fruit"],
    9: ["tower", "windows", "rural", "mast", "transmission", "uhf", "kansas",␣
 ↪"microsoft", "internet", "school"]
}

# Our guesses for misplaced words in each topic
our_guesses = {
    0: "political",  #  relate to language
    1: "river",      #  relate to politics
    2: "sexual",     #  rekate to television/movies
    3: "park",       #  relate to digital humanities
    4: "footballer",      #  relate to descriptive/adjectives
    5: "usually",    #  relate to biology
    6: "writer",     #  relate to environment
    7: "operating",  #  relate to occupations/roles
    8: "fruit",      #  relate to colors
    9: "internet"      #  relate to radio transmission
}

# Compare our guesses to the provided replacements and calculate the accuracy
correct_answers = [
    (0, "political"),
    (1, "river"),
    (2, "sexual"),
    (3, "flower"),  # Incorrect in this version
    (4, "footballer"),
    (5, "earth"),  # Incorrect in this version
    (6, "writer"),
    (7, "operating"),
    (8, "fruit"),      # Incorrect in this version
    (9, "internet")
]

# Calculate the score
correct_count = sum(1 for idx, word in correct_answers if our_guesses[idx] ==␣
 ↪word)
total_count = len(correct_answers)
accuracy = correct_count / total_count
```

```
# Output results
print(f"Our misplaced guessing accuracy is {accuracy * 100}%.")
```

```
Our misplaced guessing accuracy is 80.0%.
```

I am genuinely satisfied with achieving a 80% accuracy in this exercise. Identifying misplaced words within topics was no easy task, especially given the subtle nuances in word associations. This process required analyzing the broader context of each topic while carefully considering which word might not belong—a challenge that grows significantly more complex as the number of words in each topic increases.

What reassures me is that, aside from #3, the replacement words provided confirmed my overall understanding of the topics. This indicates that my interpretation of the themes was mostly on target, even if I occasionally misjudged which specific words were out of place. It's fascinating how some replacement words feel like they belong due to their contextual relevance, making the distinction between misplaced and fitting words particularly challenging.

This exercise has highlighted the intricacy of topic modeling and the subjective nature of interpreting themes, especially when faced with edge cases where words can seem equally plausible. Overall, I'm proud of this result and see it as an opportunity to deepen my understanding of the nuances in natural language processing.

2.1.1 Question 02. [16 points] Half & half technique: split each document into two parts, and check that topics of the first half are similar to topics of the second halves of different documents are mostly dissimilar.

[62]:
```
intra_inter(lda_model, test_docs)
```

```
average cosine similarity between corresponding parts (higher is better):
0.5178777573748987
average cosine similarity between 10,000 random parts (lower is better):
0.4695026392042921
```

The half & half technique evaluates the quality of a topic modeling algorithm by splitting each document into two halves and comparing the topic distributions of these halves. The method assumes that the two halves of the same document (intra-document similarity) should exhibit higher similarity than halves from different documents (inter-document similarity). The goal is to verify the model's ability to capture meaningful and consistent topics within documents while ensuring distinctiveness between unrelated documents.

In this case, the results are as follows:

- Average intra-document cosine similarity: 0.5179 This indicates that the two halves of the same document have a relatively strong similarity in their topic distributions, as expected. The higher the intra-document similarity, the better the model's coherence and ability to capture consistent topics within a document.
- Average inter-document cosine similarity: 0.4695 This value reflects the similarity between halves of different documents. A lower inter-document similarity is desired, as it shows that the model effectively differentiates between unrelated documents.

The observed difference between intra-document and inter-document similarity (0.5179 vs. 0.4695) demonstrates that the LDA model is capturing coherent and distinct topics. Although the difference is not vast, it is sufficient to conclude that the model performs reasonably well in distinguishing topics while maintaining internal consistency. This performance may vary depending on the dataset, preprocessing steps, or model parameters, but the results suggest that the model is effective for this application.

I believe the lack of extreme values in the similarity scores is partly due to the nature of the corpus used—Simple Wikipedia. This version of Wikipedia is designed to be easy to understand, featuring shorter sentences and a higher prevalence of common words. As a result, many of the more distinctive and hyperspecific terms typically found in the standard Wikipedia corpus are absent in this dataset. This limitation reduces the potential for dramatic differences in topic distributions, effectively creating a dataset with more stopword-like content than usual, which dampens the influence on similarity scores.

2.1.2 Question 03. [14 points] Which algorithm, LSI or LDA, performs better for this dataset. Please justify your answer.

[63]:
```
# Results for LDA
lda_results = {
    "cos_corr": 0.5046553942494272,  # Average cosine similarity between␣
 ↪corresponding parts
    "cos_rand": 0.4509096882892372   # Average cosine similarity between 10,000␣
 ↪random parts
}

# Results for LSI
lsi_results = {
    "cos_corr": 0.06456262774242728,  # Average cosine similarity between␣
 ↪corresponding parts
    "cos_rand": 0.007273692121527065  # Average cosine similarity between␣
 ↪10,000 random parts
}

# Calculate differences
cos_corr_diff = abs(lda_results["cos_corr"] - lsi_results["cos_corr"])
cos_rand_diff = abs(lda_results["cos_rand"] - lsi_results["cos_rand"])

# Determine which model performed better for each metric
cos_corr_better = "LDA" if lda_results["cos_corr"] > lsi_results["cos_corr"]␣
 ↪else "LSI"
cos_rand_better = "LDA" if lda_results["cos_rand"] < lsi_results["cos_rand"]␣
 ↪else "LSI"

# Print results with explanations
print(f"The cosine similarity between corresponding parts was higher for the␣
 ↪{cos_corr_better} model, "
```

```
      f"with a difference of {cos_corr_diff:.4f}.")
print(f"The cosine similarity between random parts was lower for the␣
 ↪{cos_rand_better} model, "
      f"with a difference of {cos_rand_diff:.4f}.")
```

```
The cosine similarity between corresponding parts was higher for the LDA model,
with a difference of 0.4401.
The cosine similarity between random parts was lower for the LSI model, with a
difference of 0.4436.
```

Based on the cosine similarity results, the LDA model performs better for this dataset in terms of capturing coherent topics within documents:

- Cosine Similarity Between Corresponding Parts: The LDA model achieves a significantly higher average cosine similarity between corresponding parts (0.5047) compared to the LSI model (0.0646), with a difference of 0.4401. This indicates that LDA is much better at preserving the thematic coherence within individual documents.
- Cosine Similarity Between Random Parts: The LSI model produces a lower average cosine similarity between random parts (0.0073) compared to LDA (0.4509), with a difference of 0.4436. While this suggests that LSI is more distinct in separating unrelated parts, the primary goal of topic modeling is to identify meaningful and consistent topics within documents, which is better reflected by the corresponding part similarity.

The LDA model is better suited for this dataset as it strikes a good balance between maintaining thematic coherence and differentiating between unrelated topics. LSI's extremely low corresponding part similarity indicates that it struggles to capture the structure and context of topics effectively. The results demonstrate that LDA is more robust and reliable for this type of text data.

[ ]:

[70]:
```
# ! apt-get install -y pandoc

#! apt-get install -y texlive-xetex texlive-fonts-recommended␣
 ↪texlive-plain-generic


! jupyter nbconvert --to pdf "/content/drive/MyDrive/Colab Notebooks/hw8.ipynb"
```

```
[NbConvertApp] WARNING | pattern '/content/drive/MyDrive/Colab
Notebooks/hw8.ipynb' matched no files
This application is used to convert notebook files (*.ipynb)
        to various other formats.

        WARNING: THE COMMANDLINE INTERFACE MAY CHANGE IN FUTURE RELEASES.

Options
=======
The options below are convenience aliases to configurable class-options,
```

```
as listed in the "Equivalent to" description-line of the aliases.
To see all configurable class-options for some <cmd>, use:
    <cmd> --help-all

--debug
    set log level to logging.DEBUG (maximize logging output)
    Equivalent to: [--Application.log_level=10]
--show-config
    Show the application's configuration (human-readable format)
    Equivalent to: [--Application.show_config=True]
--show-config-json
    Show the application's configuration (json format)
    Equivalent to: [--Application.show_config_json=True]
--generate-config
    generate default config file
    Equivalent to: [--JupyterApp.generate_config=True]
-y
    Answer yes to any questions instead of prompting.
    Equivalent to: [--JupyterApp.answer_yes=True]
--execute
    Execute the notebook prior to export.
    Equivalent to: [--ExecutePreprocessor.enabled=True]
--allow-errors
    Continue notebook execution even if one of the cells throws an error and
include the error message in the cell output (the default behaviour is to abort
conversion). This flag is only relevant if '--execute' was specified, too.
    Equivalent to: [--ExecutePreprocessor.allow_errors=True]
--stdin
    read a single notebook file from stdin. Write the resulting notebook with
default basename 'notebook.*'
    Equivalent to: [--NbConvertApp.from_stdin=True]
--stdout
    Write notebook output to stdout instead of files.
    Equivalent to: [--NbConvertApp.writer_class=StdoutWriter]
--inplace
    Run nbconvert in place, overwriting the existing notebook (only
            relevant when converting to notebook format)
    Equivalent to: [--NbConvertApp.use_output_suffix=False
--NbConvertApp.export_format=notebook --FilesWriter.build_directory=]
--clear-output
    Clear output of current file and save in place,
            overwriting the existing notebook.
    Equivalent to: [--NbConvertApp.use_output_suffix=False
--NbConvertApp.export_format=notebook --FilesWriter.build_directory=
--ClearOutputPreprocessor.enabled=True]
--coalesce-streams
    Coalesce consecutive stdout and stderr outputs into one stream (within each
cell).
```

```
    Equivalent to: [--NbConvertApp.use_output_suffix=False
--NbConvertApp.export_format=notebook --FilesWriter.build_directory=
--CoalesceStreamsPreprocessor.enabled=True]
--no-prompt
    Exclude input and output prompts from converted document.
    Equivalent to: [--TemplateExporter.exclude_input_prompt=True
--TemplateExporter.exclude_output_prompt=True]
--no-input
    Exclude input cells and output prompts from converted document.
            This mode is ideal for generating code-free reports.
    Equivalent to: [--TemplateExporter.exclude_output_prompt=True
--TemplateExporter.exclude_input=True
--TemplateExporter.exclude_input_prompt=True]
--allow-chromium-download
    Whether to allow downloading chromium if no suitable version is found on the
system.
    Equivalent to: [--WebPDFExporter.allow_chromium_download=True]
--disable-chromium-sandbox
    Disable chromium security sandbox when converting to PDF..
    Equivalent to: [--WebPDFExporter.disable_sandbox=True]
--show-input
    Shows code input. This flag is only useful for dejavu users.
    Equivalent to: [--TemplateExporter.exclude_input=False]
--embed-images
    Embed the images as base64 dataurls in the output. This flag is only useful
for the HTML/WebPDF/Slides exports.
    Equivalent to: [--HTMLExporter.embed_images=True]
--sanitize-html
    Whether the HTML in Markdown cells and cell outputs should be sanitized..
    Equivalent to: [--HTMLExporter.sanitize_html=True]
--log-level=<Enum>
    Set the log level by value or name.
    Choices: any of [0, 10, 20, 30, 40, 50, 'DEBUG', 'INFO', 'WARN', 'ERROR',
'CRITICAL']
    Default: 30
    Equivalent to: [--Application.log_level]
--config=<Unicode>
    Full path of a config file.
    Default: ''
    Equivalent to: [--JupyterApp.config_file]
--to=<Unicode>
    The export format to be used, either one of the built-in formats
            ['asciidoc', 'custom', 'html', 'latex', 'markdown', 'notebook',
'pdf', 'python', 'qtpdf', 'qtpng', 'rst', 'script', 'slides', 'webpdf']
            or a dotted object name that represents the import path for an
            ``Exporter`` class
    Default: ''
    Equivalent to: [--NbConvertApp.export_format]
```

```
--template=<Unicode>
    Name of the template to use
    Default: ''
    Equivalent to: [--TemplateExporter.template_name]
--template-file=<Unicode>
    Name of the template file to use
    Default: None
    Equivalent to: [--TemplateExporter.template_file]
--theme=<Unicode>
    Template specific theme(e.g. the name of a JupyterLab CSS theme distributed
    as prebuilt extension for the lab template)
    Default: 'light'
    Equivalent to: [--HTMLExporter.theme]
--sanitize_html=<Bool>
    Whether the HTML in Markdown cells and cell outputs should be sanitized.This
    should be set to True by nbviewer or similar tools.
    Default: False
    Equivalent to: [--HTMLExporter.sanitize_html]
--writer=<DottedObjectName>
    Writer class used to write the
                                        results of the conversion
    Default: 'FilesWriter'
    Equivalent to: [--NbConvertApp.writer_class]
--post=<DottedOrNone>
    PostProcessor class used to write the
                                        results of the conversion
    Default: ''
    Equivalent to: [--NbConvertApp.postprocessor_class]
--output=<Unicode>
    Overwrite base name use for output files.
                Supports pattern replacements '{notebook_name}'.
    Default: '{notebook_name}'
    Equivalent to: [--NbConvertApp.output_base]
--output-dir=<Unicode>
    Directory to write output(s) to. Defaults
                                    to output to the directory of each notebook.
To recover
                                    previous default behaviour (outputting to the
current
                                    working directory) use . as the flag value.
    Default: ''
    Equivalent to: [--FilesWriter.build_directory]
--reveal-prefix=<Unicode>
    The URL prefix for reveal.js (version 3.x).
            This defaults to the reveal CDN, but can be any url pointing to a
copy
            of reveal.js.
            For speaker notes to work, this must be a relative path to a local
```

```
            copy of reveal.js: e.g., "reveal.js".
            If a relative path is given, it must be a subdirectory of the
            current directory (from which the server is run).
            See the usage documentation
            (https://nbconvert.readthedocs.io/en/latest/usage.html#reveal-js-
html-slideshow)
            for more details.
    Default: ''
    Equivalent to: [--SlidesExporter.reveal_url_prefix]
--nbformat=<Enum>
    The nbformat version to write.
            Use this to downgrade notebooks.
    Choices: any of [1, 2, 3, 4]
    Default: 4
    Equivalent to: [--NotebookExporter.nbformat_version]

Examples
--------

    The simplest way to use nbconvert is

            > jupyter nbconvert mynotebook.ipynb --to html

            Options include ['asciidoc', 'custom', 'html', 'latex', 'markdown',
'notebook', 'pdf', 'python', 'qtpdf', 'qtpng', 'rst', 'script', 'slides',
'webpdf'].

            > jupyter nbconvert --to latex mynotebook.ipynb

            Both HTML and LaTeX support multiple output templates. LaTeX
includes
            'base', 'article' and 'report'.  HTML includes 'basic', 'lab' and
            'classic'. You can specify the flavor of the format used.

            > jupyter nbconvert --to html --template lab mynotebook.ipynb

            You can also pipe the output to stdout, rather than a file

            > jupyter nbconvert mynotebook.ipynb --stdout

            PDF is generated via latex

            > jupyter nbconvert mynotebook.ipynb --to pdf

            You can get (and serve) a Reveal.js-powered slideshow

            > jupyter nbconvert myslides.ipynb --to slides --post serve
```

```
            Multiple notebooks can be given at the command line in a couple of
            different ways:

            > jupyter nbconvert notebook*.ipynb
            > jupyter nbconvert notebook1.ipynb notebook2.ipynb

            or you can specify the notebooks list in a config file, containing::

                c.NbConvertApp.notebooks = ["my_notebook.ipynb"]

            > jupyter nbconvert --config mycfg.py

To see all available configurables, use `--help-all`.
```