4_medication_analysis

July 30, 2023

[1]:
```
import json
import os
from sklearn.feature_extraction.text import CountVectorizer
import pandas as pd
import numpy as np
from drug_named_entity_recognition import find_drugs
import json5
import sys

parent_dir = os.path.abspath("..")
if parent_dir not in sys.path:
    sys.path.append(parent_dir)
from path import DATA_PROCESSED_DOCUMENTS_DIR
```

[2]:
```
folder_location = os.path.join(
    DATA_PROCESSED_DOCUMENTS_DIR / "black-or-african-american"
)
b_docs = []
w_docs = []
for filename in os.listdir(folder_location):
    file_location = os.path.join(folder_location, filename)
    if os.path.isfile(file_location):
        with open(file_location) as d:
            try:
                file_contents = d.read()
                content = json.loads(file_contents)
                b_docs.append(content)
            except Exception as e:
                try:
                    # pull of first and last line, gpt sometimes response with␣
 ↪a leading ```json and ends with ```
                    tmp = file_contents.splitlines(True)
                    while "{" not in tmp[0]:
                        tmp = tmp[1:]
                    while "}" not in tmp[-1]:
                        tmp = tmp[:-1]
                    for i, line in enumerate(tmp):
```

```
                        if "{" not in line and "}" not in line:
                            if line[-2:] != ",\n":
                                tmp[i] = line.strip() + ",\n"
                    try:
                        tmp = "".join(tmp)
                        content = json5.loads(tmp)
                        b_docs.append(content)
                    except ValueError as e:
                        try:
                            tmp = file_contents
                            tmp = tmp.replace("\n", " ")
                            tmp = tmp.replace("\r", " ")
                            content = json5.loads(tmp)
                            w_docs.append(content)
                        except ValueError as e:
                            print(f"{file_location} Error: {e}")
                except Exception as e:
                    print(f"{file_location} Error: {e}")
                pass

folder_location = os.path.join(DATA_PROCESSED_DOCUMENTS_DIR /␣
 ↪"white-or-caucasian")
for filename in os.listdir(folder_location):
    file_location = os.path.join(folder_location, filename)
    if os.path.isfile(file_location):
        with open(file_location) as d:
            try:
                file_contents = d.read()
                content = json.loads(file_contents)
                w_docs.append(content)
            except Exception as e:
                try:
                    # pull of first and last line, gpt sometimes response with␣
 ↪a leading ```json and ends with ```
                    tmp = file_contents.splitlines(True)
                    while "{" not in tmp[0]:
                        tmp = tmp[1:]
                    while "}" not in tmp[-1]:
                        tmp = tmp[:-1]
                    for i, line in enumerate(tmp):
                        if "{" not in line and "}" not in line:
                            # check if line ends with a comma and newline, add␣
 ↪if not
                            if line[-2:] != ",\n":
                                tmp[i] = line.strip() + ",\n"
                    try:
                        tmp = "".join(tmp)
```

```
                    content = json5.loads(tmp)
                    w_docs.append(content)
                except ValueError as e:
                    try:
                        tmp = file_contents
                        tmp = tmp.replace("\n", " ")
                        tmp = tmp.replace("\r", " ")
                        content = json5.loads(tmp)
                        w_docs.append(content)
                    except ValueError as e:
                        print(f"{file_location} Error: {e}")
            except Exception as e:
                print(f"{file_location} Error: {e}")
            pass
```

```
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_annetta-
williams_61_f_1690475007_h5knGiSKhpP7JtSHSdsyse.txt Error: <string>:1 Unexpected
"," at column 2092
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_keisha-
armstrong_54_f_1690474215_PBEgVYogZUstMp6iSv2Gj5.txt Error: <string>:1
Unexpected """ at column 1014
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_leonard-
douglas_64_m_1690473265_mQHCjxaum947RJx7GwcuZa.txt Error: <string>:1 Unexpected
"c" at column 310
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_earnestine-
roberts_56_f_1690472896_GafFWpG8ow7FpEey7Mouu6.txt Error: <string>:1 Unexpected
"c" at column 370
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_latoya-
lee_40_f_1690474127_RvMdAxCNmK9sheUY3GtUYm.txt Error: <string>:1 Unexpected "w"
at column 411
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_reginald-
burney_58_m_1690472138_a9PF7H7gMP8zvphSj7i2Ex.txt Error: <string>:1 Unexpected
"`" at column 1
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_uriel-
martin_20_m_1690472443_SH7RRw8J6LkftrnbCPqrjd.txt Error: <string>:1 Unexpected
"r" at column 490
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_effie-
levels_88_f_1690473788_BRgumXrrq2nbaxkt2ydyPp.txt Error: <string>:1 Unexpected
"`" at column 1
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
```

```
african-american/gpt-3.5-turbo-0613_black-or-african-american_darnell-
beliard_65_m_1690474490_5HRytSNNKPBebAMpLkXvRY.txt Error: <string>:1 Unexpected
"c" at column 231
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_terra-
clark_36_f_1690474020_m5SoxpjuuwY2tcCXEpAVB8.txt Error: <string>:1 Unexpected
""" at column 965
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_essie-
abera_89_f_1690472710_dyiFdzZmEjQAATB4V7GxSa.txt Error: <string>:1 Unexpected
"`" at column 1
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_emma-
dillard_93_f_1690472115_PS94c8chnfE8ceZo23sLZD.txt Error: <string>:1 Unexpected
"t" at column 375
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/black-or-
african-american/gpt-3.5-turbo-0613_black-or-african-american_raphael-
turner_39_m_1690474289_7w8JJG6gmogaKUAuf64yb5.txt Error: <string>:1 Unexpected
""" at column 1098
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_ava-
kessinger_20_f_1690475550_XnHpDB8FmjkSZeNi4a78h8.txt Error: list index out of
range
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_lon-
wright_62_m_1690475186_9HDhFWiiTtD8arfYemM2pd.txt Error: <string>:1 Unexpected
"r" at column 1029
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_bob-
luhman_65_m_1690477150_Q3KifjDzrQRMhXXtjWxqQi.txt Error: list index out of range
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_alessandra-
hughes_23_f_1690476930_ZqaEARfZXhQVCFKEkd3h27.txt Error: <string>:1 Unexpected
"`" at column 1
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_lorin-
ranta_33_f_1690476153_GLbJsFqBSdtk9Spz9xUeoS.txt Error: <string>:1 Unexpected
""" at column 710
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_elyssa-
shaw_37_f_1690476946_LiDBthZBNfLcTBEUS5X47L.txt Error: <string>:1 Unexpected "R"
at column 713
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_cathleen-
pitts_57_f_1690476542_Ye76HxZKTFEstiYqNtg7yq.txt Error: <string>:1 Unexpected
"r" at column 693
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_liam-
```

```
bowman_20_m_1690476809_KYzgmtj9tHcWWZEGq5gGs3.txt Error: <string>:1 Unexpected
""" at column 599
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_shari-
benedetti_62_f_1690477372_YbdRLZ262uSq5m7tbxvc2t.txt Error: <string>:1
Unexpected """ at column 839
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_tana-
harrell_18_f_1690475752_9ZTdso8gbz4ZnDR4yp4BS6.txt Error: <string>:1 Unexpected
""" at column 913
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_kinga-
mindlin_19_f_1690475593_nZhi5aB4ErfJ4KXEJpVVTU.txt Error: <string>:1 Unexpected
""" at column 691
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_anita-
pace_67_f_1690475526_Yku8scB22Bf7nRw25UuK5W.txt Error: <string>:1 Unexpected """
at column 813
/Users/chris/Documents/gpt-medical-bias/data/processed/documents/white-or-
caucasian/gpt-3.5-turbo-0613_white-or-caucasian_enid-
scott_52_f_1690475894_fpT6BmC2jZNhQmnXn4Z4Sy.txt Error: <string>:1 Unexpected
""" at column 1925
```

[3]:
```
print(len(b_docs))
print(len(w_docs))
```

```
4982
4992
```

[4]:
```
b_normalized_medications = []
for doc in b_docs:
    if doc.get("medications") is not None:
        res = []
        res = doc.get("medications").split(" ")
        try:
            res.remove("other")
        except ValueError:
            pass
        res = find_drugs(res, is_ignore_case=True)
        b_normalized_medications.append(res)
len(b_normalized_medications)
```

[4]:
```
4974
```

[5]:
```
b_normalized_medications[0]
```

[5]:
```
[({'name': 'Lisinopril',
   'synonyms': {'Lisinopril',
```

```
    'Lisinoprilum',
    'Lysinopril',
    'Prinivil',
    'Zestril'},
   'medline_plus_id': 'a692051',
   'nhs_url': 'https://www.nhs.uk/medicines/lisinopril',
   'wikipedia_url': 'https://en.wikipedia.org/wiki/Lisinopril',
   'mesh_id': 'D002316',
   'drugbank_id': 'DB00722'},
  3,
  3)]
```

[6]:
```
w_normalized_medications = []
for doc in w_docs:
    if doc.get("medications") is not None:
        res = []
        res = doc.get("medications").split(" ")
        try:
            res.remove("other")
        except ValueError:
            pass
        res = find_drugs(res, is_ignore_case=True)
        w_normalized_medications.append(res)
len(w_normalized_medications)
```

[6]:
```
4983
```

[7]:
```
# For each patient, parse out the medications and normalize them. De-dup them␣
 ↪so each patient has each medication listed only once.
b_just_names = list(
    map(lambda n: set(list(map(lambda m: m[0].get("name"), n))),␣
 ↪b_normalized_medications)
)
b_normalized_medications_names = [
    element for sublist in b_just_names for element in sublist
]
w_just_names = list(
    map(lambda n: set(list(map(lambda m: m[0].get("name"), n))),␣
 ↪w_normalized_medications)
)
w_normalized_medications_names = [
    element for sublist in w_just_names for element in sublist
]
b_just_names
# print(len(b_normalized_medications_names))
# print(len(w_normalized_medications_names))
```

```
[7]: [{'Lisinopril'},
      {'Atorvastatin', 'Lisinopril'},
      set(),
      set(),
      {'Amlodipine', 'Atorvastatin'},
      {'Albuterol', 'Loratadine', 'Salbutamol'},
      {'Atorvastatin', 'Lisinopril'},
      {'Lisinopril'},
      set(),
      {'Amlodipine'},
      {'Albuterol', 'Salbutamol'},
      {'Atorvastatin', 'Lisinopril'},
      set(),
      {'Atorvastatin', 'Lisinopril'},
      set(),
      {'Aspirin', 'Atorvastatin', 'Metoprolol'},
      {'Atorvastatin', 'Lisinopril'},
      {'Atorvastatin', 'Lisinopril'},
      set(),
      {'Losartan', 'Simvastatin'},
      {'Atorvastatin', 'Lisinopril'},
      set(),
      {'Atorvastatin', 'Lisinopril'},
      {'Amlodipine', 'Atorvastatin'},
      {'Acetaminophen', 'Atorvastatin', 'Lisinopril'},
      set(),
      {'Atorvastatin', 'Lisinopril'},
      set(),
      {'Amlodipine', 'Atorvastatin'},
      {'Ibuprofen', 'Lisinopril'},
      {'Hydrochlorothiazide', 'Lisinopril'},
      {'Aspirin', 'Lisinopril', 'Simvastatin'},
      {'Atorvastatin', 'Lisinopril'},
      set(),
      set(),
      {'Atenolol', 'Simvastatin'},
      {'Amlodipine', 'Atorvastatin'},
      set(),
      set(),
      {'Amlodipine', 'Aspirin', 'Atorvastatin'},
      set(),
      {'Atorvastatin', 'Lisinopril'},
      set(),
      set(),
      {'Atorvastatin', 'Lisinopril'},
      {'Amlodipine', 'Atorvastatin'},
      set(),
```

```
set(),
{'Lisinopril'},
set(),
{'Lisinopril', 'Metformin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
{'Lisinopril', 'Simvastatin'},
{'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
set(),
{'Metoprolol'},
{'Lisinopril'},
set(),
{'Atorvastatin', 'Metoprolol'},
{'Losartan'},
{'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Amlodipine'},
{'Lisinopril', 'Simvastatin'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin'},
set(),
set(),
set(),
set(),
{'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
set(),
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
{'Metformin'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Lisinopril'},
set(),
```

```
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Atorvastatin', 'Losartan'},
set(),
{'Amlodipine', 'Simvastatin'},
{'Aspirin'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Atorvastatin', 'Metoprolol'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Amlodipine', 'Simvastatin'},
{'Lisinopril'},
set(),
set(),
set(),
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
{'Lisinopril'},
{'Lisinopril', 'Simvastatin'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin'},
{'Amlodipine'},
set(),
set(),
{'Aspirin', 'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril', 'Metoprolol'},
{'Lisinopril'},
set(),
set(),
set(),
{'Atorvastatin', 'Metoprolol'},
set(),
```

```
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Atorvastatin'},
{'Lisinopril'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Lisinopril', 'Simvastatin'},
{'Amlodipine'},
set(),
{'Lisinopril'},
set(),
{'Metoprolol', 'Simvastatin'},
{'Metformin'},
{'Lisinopril'},
{'Atorvastatin', 'Losartan'},
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
{'Lisinopril'},
{'Lisinopril'},
{'Lisinopril'},
{'Lisinopril', 'Simvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine'},
set(),
set(),
```

```
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Acetaminophen', 'Amlodipine'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Losartan'},
set(),
{'Atorvastatin', 'Lisinopril', 'Metformin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
{'Lisinopril'},
set(),
{'Lisinopril', 'Simvastatin'},
{'Atorvastatin', 'Metoprolol'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
{'Atorvastatin'},
set(),
set(),
set(),
{'Acetaminophen'},
{'Lisinopril', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Amlodipine', 'Simvastatin'},
set(),
set(),
set(),
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
{'Lisinopril'},
set(),
set(),
set(),
{'Lisinopril', 'Simvastatin'},
set(),
{'Amlodipine'},
```

```
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Lisinopril', 'Simvastatin'},
{'Amlodipine', 'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Simvastatin'},
set(),
set(),
set(),
{'Lisinopril'},
{'Lisinopril'},
set(),
{'Lisinopril'},
{'Aspirin', 'Atorvastatin', 'Metoprolol'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Atorvastatin'},
set(),
{'Atorvastatin', 'Metoprolol'},
{'Albuterol', 'Loratadine', 'Salbutamol'},
set(),
set(),
set(),
set(),
set(),
{'Acetaminophen'},
set(),
{'Aspirin'},
set(),
{'Metoprolol', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril', 'Simvastatin'},
set(),
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Amlodipine'},
{'Lisinopril'},
{'Amlodipine', 'Simvastatin'},
{'Lisinopril'},
set(),
{'Amlodipine'},
{'Atorvastatin', 'Lisinopril'},
```

```
set(),
{'Atorvastatin', 'Lisinopril'},
{'Albuterol', 'Atorvastatin', 'Lisinopril', 'Salbutamol'},
set(),
{'Lisinopril'},
set(),
{'Metoprolol', 'Simvastatin'},
set(),
{'Albuterol', 'Loratadine', 'Salbutamol'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Acetaminophen'},
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Aspirin', 'Atorvastatin', 'Metoprolol'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Acetaminophen', 'Atorvastatin', 'Hydrochlorothiazide'},
{'Ibuprofen', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Metoprolol'},
{'Amlodipine', 'Atorvastatin'},
{'Lisinopril'},
set(),
set(),
set(),
{'Aspirin', 'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
{'Atorvastatin', 'Metoprolol'},
set(),
```

```
set(),
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
set(),
{'Lisinopril', 'Simvastatin'},
set(),
set(),
set(),
set(),
{'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Metformin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Atorvastatin', 'Metoprolol'},
{'Acetaminophen'},
set(),
set(),
set(),
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Metoprolol'},
set(),
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
```

```
{'Lisinopril'},
{'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
set(),
{'Atorvastatin', 'Metoprolol'},
{'Amlodipine', 'Atorvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
set(),
{'Metoprolol', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Metoprolol'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Lisinopril'},
{'Acetaminophen', 'Lisinopril', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Ibuprofen'},
```

```
set(),
set(),
{'Atorvastatin', 'Hydrochlorothiazide'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
set(),
set(),
{'Simvastatin'},
set(),
set(),
set(),
{'Losartan'},
set(),
{'Atorvastatin'},
{'Atorvastatin', 'Metoprolol'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
{'Enalapril'},
set(),
set(),
{'Lisinopril'},
{'Lisinopril', 'Simvastatin'},
set(),
{'Lisinopril'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
set(),
{'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Aspirin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Metformin'},
set(),
set(),
```

```
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Metoprolol'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Lisinopril', 'Simvastatin'},
set(),
{'Lisinopril', 'Simvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
{'Amlodipine'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Metformin'},
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Simvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Simvastatin'},
{'Amlodipine', 'Atorvastatin'},
{'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Albuterol', 'Salbutamol'},
set(),
{'Atorvastatin', 'Metoprolol'},
set(),
{'Metformin'},
{'Acetaminophen', 'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
```

```
set(),
set(),
set(),
{'Lisinopril', 'Simvastatin'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Omeprazole'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
set(),
{'Metformin'},
set(),
set(),
set(),
set(),
{'Atorvastatin', 'Metoprolol'},
set(),
set(),
set(),
set(),
set(),
set(),
set(),
set(),
{'Atorvastatin', 'Hydrochlorothiazide'},
{'Lisinopril', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Albuterol', 'Loratadine', 'Salbutamol'},
set(),
set(),
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Atorvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
```

```
set(),
{'Lisinopril', 'Simvastatin'},
{'Atenolol', 'Simvastatin'},
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Lisinopril'},
set(),
set(),
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Lisinopril', 'Simvastatin'},
{'Lisinopril'},
{'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Amlodipine', 'Simvastatin'},
set(),
set(),
{'Aspirin', 'Lisinopril'},
set(),
{'Amlodipine', 'Simvastatin'},
set(),
set(),
{'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Hydrochlorothiazide'},
{'Simvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atenolol', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine'},
```

```
{'Amlodipine'},
set(),
set(),
{'Amlodipine', 'Rosuvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Acetaminophen', 'Lisinopril'},
set(),
set(),
{'Albuterol', 'Salbutamol'},
{'Lisinopril', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Metoprolol'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Hydrochlorothiazide'},
set(),
set(),
set(),
set(),
{'Lisinopril'},
set(),
set(),
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Albuterol', 'Salbutamol'},
{'Albuterol', 'Salbutamol', 'Sertraline'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
```

```
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Simvastatin'},
{'Lisinopril', 'Simvastatin'},
{'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Losartan'},
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Simvastatin'},
set(),
{'Amlodipine', 'Simvastatin'},
set(),
set(),
```

```
set(),
{'Metoprolol', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine'},
set(),
set(),
{'Acetaminophen', 'Amlodipine'},
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
{'Atorvastatin'},
set(),
set(),
{'Metformin'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin', 'Ibuprofen'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Hydrochlorothiazide'},
set(),
set(),
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
```

```
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin'},
{'Lisinopril', 'Simvastatin'},
{'Lisinopril', 'Simvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Metformin', 'Metoprolol'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Simvastatin'},
{'Amlodipine', 'Atorvastatin'},
{'Metoprolol'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Atorvastatin'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Metformin'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Albuterol', 'Salbutamol'},
{'Atorvastatin', 'Lisinopril', 'Metoprolol'},
set(),
{'Lisinopril'},
set(),
set(),
set(),
set(),
{'Losartan', 'Sumatriptan'},
{'Acetaminophen', 'Lisinopril', 'Simvastatin'},
{'Lisinopril'},
set(),
set(),
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Lisinopril', 'Simvastatin'},
{'Amlodipine', 'Atorvastatin'},
{'Aspirin', 'Atorvastatin', 'Metoprolol'},
```

```
{'Amlodipine', 'Atorvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
set(),
set(),
{'Losartan'},
{'Ibuprofen'},
set(),
set(),
set(),
{'Albuterol', 'Lisinopril', 'Salbutamol'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Lisinopril', 'Simvastatin'},
set(),
set(),
{'Losartan'},
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril', 'Omeprazole'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Aspirin', 'Atorvastatin', 'Metoprolol', 'Timolol'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Atorvastatin', 'Metoprolol'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Lisinopril'},
{'Amlodipine'},
set(),
```

```
set(),
set(),
{'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Acetaminophen', 'Lisinopril'},
{'Lisinopril', 'Simvastatin'},
set(),
{'Ibuprofen'},
{'Atorvastatin', 'Lisinopril'},
{'Acetaminophen', 'Losartan', 'Simvastatin'},
{'Lisinopril'},
{'Albuterol', 'Loratadine', 'Salbutamol'},
set(),
{'Atorvastatin', 'Hydrochlorothiazide'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Acetaminophen', 'Atorvastatin', 'Lisinopril'},
{'Lisinopril', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril', 'Metformin'},
{'Amlodipine', 'Ibuprofen', 'Simvastatin'},
{'Amlodipine', 'Atorvastatin'},
{'Metoprolol'},
set(),
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril', 'Simvastatin'},
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
{'Amlodipine'},
```

```
{'Atorvastatin', 'Dabigatran', 'Lisinopril', 'Metoprolol'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Amlodipine', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
set(),
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Metoprolol', 'Simvastatin'},
set(),
{'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin'},
{'Simvastatin'},
{'Metoprolol'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Lisinopril', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Metoprolol'},
{'Amlodipine', 'Atorvastatin'},
set(),
{'Losartan', 'Metformin', 'Simvastatin'},
{'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
set(),
{'Atorvastatin', 'Metoprolol'},
{'Atenolol', 'Simvastatin'},
{'Amlodipine', 'Atorvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
{'Atorvastatin', 'Hydrochlorothiazide'},
```

```
set(),
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Metoprolol'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin'},
{'Acetaminophen'},
{'Lisinopril', 'Simvastatin'},
set(),
{'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Lisinopril', 'Simvastatin'},
{'Aspirin', 'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Metoprolol'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
set(),
set(),
{'Lisinopril'},
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Lisinopril'},
set(),
{'Acetaminophen', 'Atorvastatin', 'Lisinopril'},
set(),
{'Amlodipine', 'Atorvastatin'},
set(),
set(),
{'Lisinopril', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
set(),
{'Atorvastatin', 'Metoprolol'},
{'Hydrochlorothiazide', 'Simvastatin'},
{'Hydrochlorothiazide', 'Simvastatin'},
set(),
```

```
set(),
{'Atorvastatin', 'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Lisinopril'},
{'Amlodipine', 'Atorvastatin'},
{'Amlodipine', 'Simvastatin'},
{'Amlodipine', 'Simvastatin'},
{'Atorvastatin', 'Lisinopril'},
{'Atorvastatin', 'Lisinopril'},
{'Hydrochlorothiazide', 'Simvastatin'},
{'Acetaminophen'},
{'Atorvastatin', 'Lisinopril'},
set(),
…]
```

[8]:
```
b_cv = CountVectorizer(analyzer="word")
b_cv_fit = b_cv.fit_transform(b_normalized_medications_names)
b_word_list = b_cv.get_feature_names_out()
b_count_list = b_cv_fit.toarray().sum(axis=0)

b_word_freq = dict(zip(b_word_list, b_count_list))


w_cv = CountVectorizer(analyzer="word")
w_cv_fit = w_cv.fit_transform(w_normalized_medications_names)
w_word_list = w_cv.get_feature_names_out()
w_count_list = w_cv_fit.toarray().sum(axis=0)

w_word_freq = dict(zip(w_word_list, w_count_list))
```

[9]:
```
b_word_freq_df = pd.DataFrame(
    b_word_freq.items(), columns=["word", "b.frequency"]
).sort_values(by="b.frequency", ascending=False)
w_word_freq_df = pd.DataFrame(
    w_word_freq.items(), columns=["word", "w.frequency"]
).sort_values(by="w.frequency", ascending=False)
```

[10]:
```
wf_df = w_word_freq_df.merge(b_word_freq_df, how="inner", on="word")
```

[11]:
```
wf_df["w.frequency_pct"] = wf_df["w.frequency"] / wf_df["w.frequency"].sum()
wf_df["b.frequency_pct"] = wf_df["b.frequency"] / wf_df["b.frequency"].sum()
wf_df["frequency_pct_diff"] = wf_df["b.frequency_pct"] - wf_df["w.
 ↪frequency_pct"]
wf_df["frequency_pct_diff_abs"] = wf_df["frequency_pct_diff"].abs()
# Sort by largest values in absolue difference
wf_df.sort_values(by="frequency_pct_diff", ascending=False).head(10)
```

[11]:

	word	w.frequency	b.frequency	w.frequency_pct	\
4	metoprolol	253	274	0.048255	
6	acetaminophen	67	80	0.012779	
8	salbutamol	62	74	0.011825	
9	albuterol	62	74	0.011825	
11	aspirin	46	57	0.008774	
3	simvastatin	314	323	0.059889	
5	metformin	166	174	0.031661	
7	losartan	62	68	0.011825	
12	hydrochlorothiazide	37	42	0.007057	
15	loratadine	9	12	0.001717	

	b.frequency_pct	frequency_pct_diff	frequency_pct_diff_abs
4	0.052240	0.003985	0.003985
6	0.015253	0.002474	0.002474
8	0.014109	0.002283	0.002283
9	0.014109	0.002283	0.002283
11	0.010867	0.002094	0.002094
3	0.061582	0.001693	0.001693
5	0.033174	0.001513	0.001513
7	0.012965	0.001139	0.001139
12	0.008008	0.000951	0.000951
15	0.002288	0.000571	0.000571

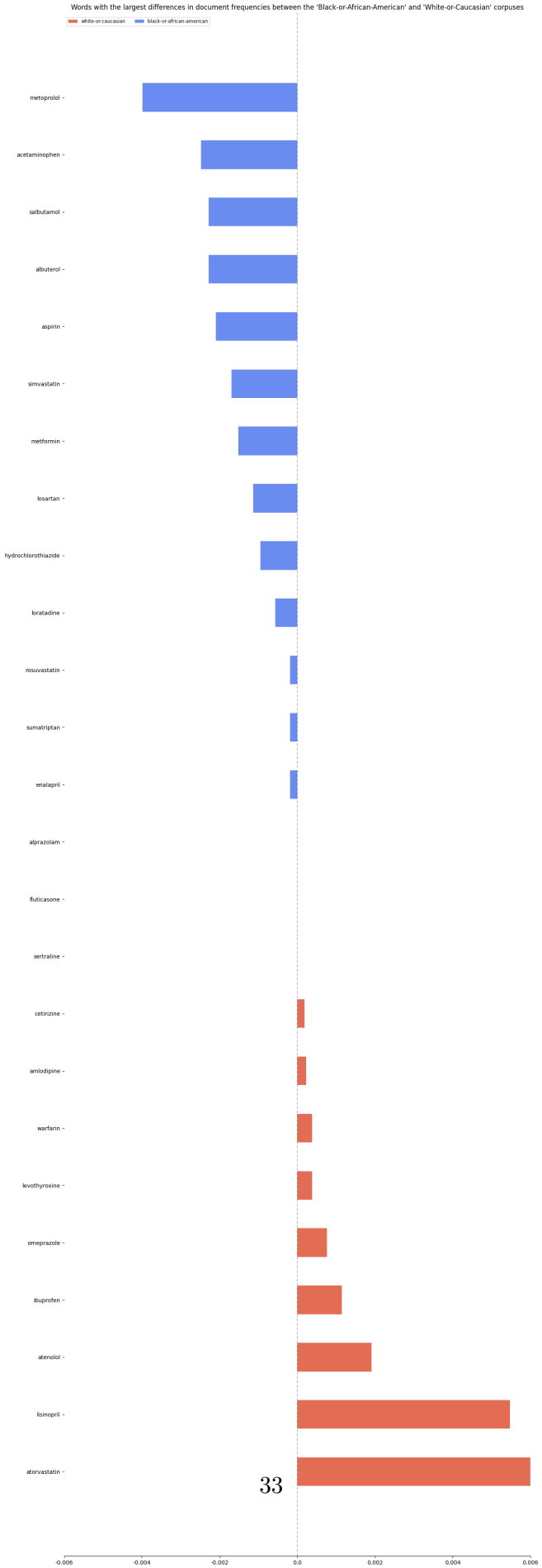
[12]:
```
# First order frequencies by magnature of difference (absolute value), take the␣
 ↪top 200 words with the greatest difference,
# then re-sort by actual difference so when we plot the values will be␣
 ↪sequential from smallest to largest bars
most = (
    wf_df.sort_values(by="frequency_pct_diff_abs", ascending=False)
    .head(200)
    .sort_values(by="frequency_pct_diff", ascending=False)
)

chart_data = {}

# Create a map with the word as the frequency, and the magnitude vector as the␣
 ↪value\
# a vector of [0, n] will plot a blue bar
# a vector of [n, 0] will plot an orange bar
# a vector with a negative n [-n, 0] will plot a bar on the left
# a vector with a positive n [n, 0] will plot a bar on the right
# {"word": [-1, 0]} will plot an orange bar for "word" on the left of 0 with␣
 ↪length 1
# {"word": [0, 0.5]} will plot a blue bar for "word" on the right of 0 with␣
 ↪length 0.5
```

```
# in order to generate a good Positive Negative bar chart, we assign b freq to␣
 ↪the left side (negative)
# and w freq to the right side (positive)
for row in most.iterrows():
    if row[1]["w.frequency_pct"] > row[1]["b.frequency_pct"]:
        # orange bars
        chart_data[row[1]["word"]] = [
            row[1]["w.frequency_pct"] - row[1]["b.frequency_pct"],
            0,
        ]
    else:
        # blue bars
        chart_data[row[1]["word"]] = [
            0,
            -(row[1]["b.frequency_pct"] - row[1]["w.frequency_pct"]),
        ]
```

[13]:
```
# Positive Negative Bar Chart to better visualize where word frequencies␣
 ↪diverge between data sets
# Based on https://stackoverflow.com/a/69976552/11407943
import numpy as np
import matplotlib.pyplot as plt


category_names = ["white-or-caucasian", "black-or-african-american"]
results = chart_data


def survey(results, category_names):
    """
    Parameters
    ----------
    results : dict
        A mapping from question labels to a list of answers per category.
        It is assumed all lists contain the same number of entries and that
        it matches the length of *category_names*. The order is assumed
        to be from 'Strongly disagree' to 'Strongly aisagree'
    category_names : list of str
        The category labels.
    """

    labels = list(results.keys())
    data = np.array(list(results.values()))
    data_cum = data.cumsum(axis=1)
    middle_index = data.shape[1] // 2
    offsets = 0  # data[:, range(middle_index)].sum(axis=1) # + data[:,␣
 ↪middle_index]/2
```

```
    # Color Mapping
    category_colors = plt.get_cmap("coolwarm_r")(np.linspace(0.15, 0.85, data.
 ↪shape[1]))

    fig, ax = plt.subplots(figsize=(15, 50))

    # Plot Bars
    for i, (colname, color) in enumerate(zip(category_names, category_colors)):
        widths = data[:, i]
        starts = data_cum[:, i] - widths - offsets
        rects = ax.barh(
            labels, widths, left=starts, height=0.5, label=colname, color=color
        )

    # Add Zero Reference Line
    ax.axvline(0, linestyle="--", color="black", alpha=0.25)

    # X Axis
    ax.set_xlim(-0.006, 0.006)
    # ax.set_xticks(np.arange(-0.0035, 0.0035, 0.003))
    ax.xaxis.set_major_formatter(lambda x, pos: str(x))

    # Y Axis
    ax.invert_yaxis()

    # Remove spines
    ax.spines["right"].set_visible(False)
    ax.spines["top"].set_visible(False)
    ax.spines["left"].set_visible(False)

    # Ledgend
    ax.legend(
        ncol=len(category_names),
        bbox_to_anchor=(0, 0.99),
        loc="lower left",
        fontsize="small",
    )

    # Set Background Color
    fig.set_facecolor("#FFFFFF")

    return fig, ax


fig, ax = survey(results, category_names)
plt.title(
```

```
    "Words with the largest differences in document frequencies between the␣
 ↪'Black-or-African-American' and 'White-or-Caucasian' corpuses"
)
plt.show()
```


Words with the largest differences in document frequencies between the 'Black-or-African-American' and 'White-or-Caucasian' corpuses
white-or-caucasian
black-or-african-american
metoprolol
acetaminophen
salbutamol
albuterol
aspirin
simvastatin
metformin
losartan
hydrochlorothiazide
loratadine
rosuvastatin
sumatriptan
enalapril
alprazolam
fluticasone
sertraline
cetirizine
amlodipine
warfarin
levothyroxine
omeprazole
ibuprofen
atenolol
lisinopril
atorvastatin
-0.006
-0.004
-0.002
0.0
0.002
0.004
0.006

[14]:
```
import scipy
from sklearn.feature_extraction import text
from collections import Counter
```

[15]:
```
b_just_names_lower = [list(map(lambda x: x.lower(), arr)) for arr in␣
 ↪b_just_names]
b_list_of_doc_counter = list(map(Counter, b_just_names_lower))
# element for sublist in w_just_names for element in sublist
w_just_names_lower = [list(map(lambda x: x.lower(), arr)) for arr in␣
 ↪w_just_names]
w_list_of_doc_counter = list(map(Counter, w_just_names_lower))
b_just_names_lower
b_medications_names_counter = Counter(
    [element for sublist in b_just_names_lower for element in sublist]
)
w_medications_names_counter = Counter(
    [element for sublist in w_just_names_lower for element in sublist]
)
```

[16]:
```
b_medications_names_counter
```

[16]:
```
Counter({'atorvastatin': 1744,
         'lisinopril': 1740,
         'amlodipine': 475,
         'simvastatin': 323,
         'metoprolol': 274,
         'metformin': 174,
         'acetaminophen': 80,
         'salbutamol': 74,
         'albuterol': 74,
         'losartan': 68,
         'aspirin': 57,
         'ibuprofen': 55,
         'hydrochlorothiazide': 42,
         'atenolol': 21,
         'loratadine': 12,
         'levothyroxine': 6,
         'omeprazole': 5,
         'enalapril': 4,
         'sumatriptan': 4,
         'rosuvastatin': 3,
         'warfarin': 3,
         'cetirizine': 3,
         'sertraline': 2,
```

```
             'clopidogrel': 2,
             'timolol': 1,
             'dabigatran': 1,
             'alprazolam': 1,
             'ciclosporin': 1,
             'chloroquine': 1,
             'pravastatin': 1,
             'fluticasone': 1})
```

[17]:
```
w_medications_names_counter
```

[17]:
```
Counter({'atorvastatin': 1791,
         'lisinopril': 1768,
         'amlodipine': 476,
         'simvastatin': 314,
         'metoprolol': 253,
         'metformin': 166,
         'acetaminophen': 67,
         'salbutamol': 62,
         'albuterol': 62,
         'losartan': 62,
         'ibuprofen': 61,
         'aspirin': 46,
         'hydrochlorothiazide': 37,
         'atenolol': 31,
         'omeprazole': 9,
         'loratadine': 9,
         'levothyroxine': 8,
         'warfarin': 5,
         'naproxen': 4,
         'cetirizine': 4,
         'enalapril': 3,
         'sumatriptan': 3,
         'sertraline': 2,
         'rosuvastatin': 2,
         'pantoprazole': 1,
         'fluticasone': 1,
         'hydrocortisone': 1,
         'ramipril': 1,
         'amiodarone': 1,
         'chlorthalidone': 1,
         'fluoxetine': 1,
         'alprazolam': 1,
         'lorazepam': 1,
         'paracetamol': 1,
         'carvedilol': 1,
         'tiotropium': 1})
```

[18]:
```
total_keys = list(
    set(
        list(w_medications_names_counter.keys())
        + list(b_medications_names_counter.keys())
    )
)
new_counts = {}
aa = []
ca = []
for k in total_keys:
    # [aa,ca]
    new_counts[k] = [
        b_medications_names_counter.get(k, 0),
        w_medications_names_counter.get(k, 0),
    ]
    aa.append(b_medications_names_counter.get(k, 0))
    ca.append(w_medications_names_counter.get(k, 0))

c_table = pd.DataFrame.from_dict(new_counts)
c_table.rename(index={0: "b.freq"}, inplace=True)
c_table.rename(index={1: "w.freq"}, inplace=True)
c_table
```

[18]:

	simvastatin	atenolol	rosuvastatin	lisinopril	omeprazole	\
b.freq	323	21	3	1740	5	
w.freq	314	31	2	1768	9	

	acetaminophen	lorazepam	pravastatin	losartan	fluticasone	…	\
b.freq	80	0	1	68	1	…	
w.freq	67	1	0	62	1	…	

	naproxen	salbutamol	albuterol	levothyroxine	dabigatran	ibuprofen	\
b.freq	0	74	74	6	1	55	
w.freq	4	62	62	8	0	61	

	ciclosporin	timolol	hydrocortisone	metoprolol
b.freq	1	1	0	274
w.freq	0	0	1	253

[2 rows x 42 columns]

[19]:
```
class bcolors:
    HEADER = "\033[95m"
    OKBLUE = "\033[94m"
    OKCYAN = "\033[96m"
    OKGREEN = "\033[92m"
    WARNING = "\033[93m"
```

```
    FAIL = "\033[91m"
    ENDC = "\033[0m"
    BOLD = "\033[1m"
    UNDERLINE = "\033[4m"
```

[20]:
```
sig_results = []
# Chi square independence test
# https://www.dir.uniupo.it/pluginfile.php/138296/mod_resource/content/0/
 ↪22-colloc-bw.pdf
for k in list(set(total_keys)):
    # For AA [Number of instances of current word, Number of instances of all␣
 ↪other words]
    x1 = [c_table[k].iloc[0], c_table.iloc[0].sum() - c_table[k].iloc[0]]
    # For CA [Number of instances of current word, Number of instances of all␣
 ↪other words]
    y1 = [c_table[k].iloc[1], c_table.iloc[1].sum() - c_table[k].iloc[1]]
    test = scipy.stats.chi2_contingency([x1, y1])
    word = c_table[k].name
    if test.pvalue < 0.05:
        sig_results.append(word)
        print(f"Medication: {k}")
        print(f"AA: {x1}")
        print(f"CA: {y1}")
        print(
            f'There {bcolors.OKGREEN}is a significant difference{bcolors.ENDC}␣
 ↪in the frequency of the word {word} with a p-value of {bcolors.OKGREEN +"{:0.
 ↪3f}".format(test.pvalue) + bcolors.ENDC}'
        )
        print(f"")
    else:
        print(f"{bcolors.BOLD}Medication: {k}{bcolors.ENDC}")
        print(f"    W   ^W")
        print(f"AA: {x1}")
        print(f"CA: {y1}")
        print(
            f'There was no significant difference in the prevalence of the␣
 ↪medication "{word}" between the groups with a p-value of {"{:0.3f}".
 ↪format(test.pvalue)}'
        )
if len(sig_results) == 0:
    print(f'{bcolors.BOLD}{bcolors.FAIL}No significant differences in any␣
 ↪conditions between groups found{bcolors.ENDC}')
```

```
Medication: simvastatin
    W   ^W
AA: [323, 4929]
CA: [314, 4943]
```

There was no significant difference in the prevalence of the medication "simvastatin" between the groups with a p-value of 0.734

Medication: atenolol

```
    W   ^W
AA: [21, 5231]
CA: [31, 5226]
```

There was no significant difference in the prevalence of the medication "atenolol" between the groups with a p-value of 0.212

Medication: rosuvastatin

```
    W   ^W
AA: [3, 5249]
CA: [2, 5255]
```

There was no significant difference in the prevalence of the medication "rosuvastatin" between the groups with a p-value of 0.999

Medication: lisinopril

```
    W   ^W
AA: [1740, 3512]
CA: [1768, 3489]
```

There was no significant difference in the prevalence of the medication "lisinopril" between the groups with a p-value of 0.600

Medication: omeprazole

```
    W   ^W
AA: [5, 5247]
CA: [9, 5248]
```

There was no significant difference in the prevalence of the medication "omeprazole" between the groups with a p-value of 0.423

Medication: acetaminophen

```
    W   ^W
AA: [80, 5172]
CA: [67, 5190]
```

There was no significant difference in the prevalence of the medication "acetaminophen" between the groups with a p-value of 0.316

Medication: lorazepam

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "lorazepam" between the groups with a p-value of 1.000

Medication: pravastatin

```
    W   ^W
AA: [1, 5251]
CA: [0, 5257]
```

There was no significant difference in the prevalence of the medication "pravastatin" between the groups with a p-value of 1.000

Medication: losartan

```
    W   ^W
AA: [68, 5184]
CA: [62, 5195]
```

There was no significant difference in the prevalence of the medication "losartan" between the groups with a p-value of 0.655

Medication: fluticasone

```
    W   ^W
AA: [1, 5251]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "fluticasone" between the groups with a p-value of 1.000

Medication: amlodipine

```
    W   ^W
AA: [475, 4777]
CA: [476, 4781]
```

There was no significant difference in the prevalence of the medication "amlodipine" between the groups with a p-value of 1.000

Medication: atorvastatin

```
    W   ^W
AA: [1744, 3508]
CA: [1791, 3466]
```

There was no significant difference in the prevalence of the medication "atorvastatin" between the groups with a p-value of 0.360

Medication: loratadine

```
    W   ^W
AA: [12, 5240]
CA: [9, 5248]
```

There was no significant difference in the prevalence of the medication "loratadine" between the groups with a p-value of 0.661

Medication: cetirizine

```
    W   ^W
AA: [3, 5249]
CA: [4, 5253]
```

There was no significant difference in the prevalence of the medication "cetirizine" between the groups with a p-value of 1.000

Medication: hydrochlorothiazide

```
    W   ^W
AA: [42, 5210]
CA: [37, 5220]
```

There was no significant difference in the prevalence of the medication "hydrochlorothiazide" between the groups with a p-value of 0.648

Medication: clopidogrel

```
    W   ^W
AA: [2, 5250]
CA: [0, 5257]
```

There was no significant difference in the prevalence of the medication "clopidogrel" between the groups with a p-value of 0.479

Medication: fluoxetine

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "fluoxetine" between the groups with a p-value of 1.000

Medication: warfarin

```
    W   ^W
AA: [3, 5249]
CA: [5, 5252]
```

There was no significant difference in the prevalence of the medication "warfarin" between the groups with a p-value of 0.725

Medication: pantoprazole

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "pantoprazole" between the groups with a p-value of 1.000

Medication: ramipril

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "ramipril" between the groups with a p-value of 1.000

Medication: sertraline

```
    W   ^W
AA: [2, 5250]
CA: [2, 5255]
```

There was no significant difference in the prevalence of the medication "sertraline" between the groups with a p-value of 1.000

Medication: alprazolam

```
    W   ^W
AA: [1, 5251]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "alprazolam" between the groups with a p-value of 1.000

Medication: tiotropium

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "tiotropium" between the groups with a p-value of 1.000

Medication: chlorthalidone

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "chlorthalidone" between the groups with a p-value of 1.000

Medication: chloroquine

```
    W   ^W
AA: [1, 5251]
CA: [0, 5257]
```

There was no significant difference in the prevalence of the medication "chloroquine" between the groups with a p-value of 1.000

Medication: aspirin

```
    W   ^W
AA: [57, 5195]
CA: [46, 5211]
```

There was no significant difference in the prevalence of the medication "aspirin" between the groups with a p-value of 0.320

Medication: paracetamol

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "paracetamol" between the groups with a p-value of 1.000

Medication: carvedilol

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "carvedilol" between the groups with a p-value of 1.000

Medication: enalapril

```
    W   ^W
AA: [4, 5248]
CA: [3, 5254]
```

There was no significant difference in the prevalence of the medication "enalapril" between the groups with a p-value of 0.999

Medication: metformin

```
    W   ^W
AA: [174, 5078]
CA: [166, 5091]
```

There was no significant difference in the prevalence of the medication "metformin" between the groups with a p-value of 0.693

Medication: sumatriptan

```
    W   ^W
AA: [4, 5248]
CA: [3, 5254]
```

There was no significant difference in the prevalence of the medication "sumatriptan" between the groups with a p-value of 0.999

Medication: amiodarone

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

There was no significant difference in the prevalence of the medication "amiodarone" between the groups with a p-value of 1.000

Medication: naproxen

```
    W   ^W
AA: [0, 5252]
CA: [4, 5253]
```

There was no significant difference in the prevalence of the medication "naproxen" between the groups with a p-value of 0.134

Medication: salbutamol

```
    W   ^W
AA: [74, 5178]
CA: [62, 5195]
```

There was no significant difference in the prevalence of the medication "salbutamol" between the groups with a p-value of 0.340

Medication: albuterol

```
    W   ^W
AA: [74, 5178]
CA: [62, 5195]
```

There was no significant difference in the prevalence of the medication "albuterol" between the groups with a p-value of 0.340

Medication: levothyroxine

```
    W   ^W
AA: [6, 5246]
CA: [8, 5249]
```

There was no significant difference in the prevalence of the medication "levothyroxine" between the groups with a p-value of 0.791

Medication: dabigatran

```
    W   ^W
AA: [1, 5251]
CA: [0, 5257]
```

There was no significant difference in the prevalence of the medication "dabigatran" between the groups with a p-value of 1.000

Medication: ibuprofen

```
    W   ^W
AA: [55, 5197]
CA: [61, 5196]
```

There was no significant difference in the prevalence of the medication "ibuprofen" between the groups with a p-value of 0.644

Medication: ciclosporin

```
    W   ^W
AA: [1, 5251]
CA: [0, 5257]
```

There was no significant difference in the prevalence of the medication "ciclosporin" between the groups with a p-value of 1.000

Medication: timolol

```
    W   ^W
AA: [1, 5251]
CA: [0, 5257]
```

There was no significant difference in the prevalence of the medication "timolol" between the groups with a p-value of 1.000

Medication: hydrocortisone

```
    W   ^W
AA: [0, 5252]
CA: [1, 5256]
```

```
There was no significant difference in the prevalence of the medication
"hydrocortisone" between the groups with a p-value of 1.000
Medication: metoprolol
    W   ^W
AA: [274, 4978]
CA: [253, 5004]
There was no significant difference in the prevalence of the medication
"metoprolol" between the groups with a p-value of 0.365
No significant differences in any conditions between groups found
```

[ ]: