In [1]:
```
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.metrics import f1_score
from itertools import cycle, islice
from operator import itemgetter
import sif_embedding_wrapper
import pandas as pd
import numpy as np
import itertools
import codecs
import utils
import os
```

In [2]:
```
# from gensim.models.keyedvectors import KeyedVectors

# model = KeyedVectors.load_word2vec_format('/home/stirunag/pre-trained_word_embeddings/PubMed-and-PMC-w2v.bin', binary=True)
# model.save_word2vec_format('/home/stirunag/pre-trained_word_embeddings/PubMed-and-PMC-w2v.txt', binary=False)

words, embs, weight4ind = sif_embedding_wrapper.load_embeddings("/home/stirunag/pre-trained_word_embeddings/PubMed-and-PMC-FS.txt",
                                                                '/home/stirunag/pre-trained_word_embeddings/wiki/enwiki_vocab_min200.txt')

# words, embs, weight4ind = sif_embedding_wrapper.load_embeddings("/home/stirunag/pre-trained_word_embeddings/glove/glove.6B.300d.txt",
#                                                                 '/home/stirunag/pre-trained_word_embeddings/wiki/enwiki_vocab_min200.txt')
```

In [3]:
```
# get the current working directory
data_path = os.path.abspath(os.path.join(os.path.dirname( '__file__' ), '..', 'Datasets'))+'/'

# Although the dataset says csv, it is tab delimited. In addition to this, they have severe codels problems.
# So best to parse throught codes first.
# UnicodeDecodeError: 'utf-8' codec can't decode byte 0xfc in position 2: invalid start byte

#open for reading with "universal" type set

doc_d_t = codecs.open(data_path+'EUADR_Corpus_IBIgroup/'+'EUADR_drug_target'+'.csv','rU','UTF-8')
EUADR_drug_target = pd.read_csv(doc_d_t, sep='\t', na_filter = False)
EUADR_drug_target['CLASS'] = 'drug_gene'

doc_t_d = codecs.open(data_path+'EUADR_Corpus_IBIgroup/'+'EUADR_target_disease'+'.csv','rU','UTF-8',errors='ignore')
EUADR_target_disease = pd.read_csv(doc_t_d, sep='\t', na_filter = False)
EUADR_target_disease['CLASS'] = 'gene_disease'

doc_d_d = codecs.open(data_path+'EUADR_Corpus_IBIgroup/'+'EUADR_drug_disease'+'.csv','rU','UTF-8')
EUADR_drug_disease = pd.read_csv(doc_d_d, sep='\t', na_filter = False)
EUADR_drug_disease['CLASS'] = 'drug_disease'
```

In [4]:
```
df = EUADR_drug_target.append(EUADR_target_disease).append(EUADR_drug
_disease)
df.reset_index(inplace=True)
```

In [5]:
```
doc_embeddings = sif_embedding_wrapper.sentences2vecs(df["SENTENCE"],
embs, words, weight4ind)
df["vector"] = pd.Series(list(doc_embeddings))
```

In [6]:
```
ground_truth = {}

for idx, row in df.iterrows():
    ground_truth[idx] = row['CLASS']

# ground_truth

# inv_map = {}
# for k, v in ground_truth.items():
#     inv_map[v] = inv_map.get(v, [])
#     inv_map[v].append(k)

# inv_map
```

In [7]:
```
categories = list(df["CLASS"].unique())
categories
```

Out[7]:
```
['drug_gene', 'gene_disease', 'drug_disease']
```

In [8]:
```
# Use LDA to get the topics and assign to class to find top performan
ing sentences

min_text_length=80
max_iter=150
batch_size=128
learning_offset=300.
n_topics = len(categories)


docs = df

unclassifiable = list(docs[docs["SENTENCE"].map(len) < min_text_lengt
h].index)
filtered = docs[~docs.index.isin(unclassifiable)]
ids = [d for d in list(filtered.index)[0:10]]

n_features = 5000
tf_vectorizer = TfidfVectorizer(
    stop_words='english',
    max_df=0.95,
    min_df=0.1,
    max_features=n_features)
tf = tf_vectorizer.fit_transform(list(filtered.loc[:, 'SENTENCE']))

lda = LatentDirichletAllocation(
    n_components=n_topics,
    max_iter=max_iter,
    batch_size=batch_size,
    learning_method='online',
    learning_offset=learning_offset,
    random_state=0)

lda.fit(tf)
doc_topics = lda.transform(tf)
```

In [9]:
```
topic_leaders = {"topic_{}".format(i): [] for i in iter(range(n_topic
s))}
doc_id = filtered.index

for idx, probs in enumerate(doc_topics):
    score = max(probs)
    topic = np.argmax(probs)
    doc_id = filtered.index[idx]
    topic_leaders["topic_{}".format(topic)].append({"doc_id": doc_id,
"score": score})

for i in iter(range(n_topics)):
    topic_leaders["topic_{}".format(i)] = sorted(
        topic_leaders["topic_{}".format(i)], key=itemgetter('score'),
reverse=True)



topic_leaders
```

```
Out[9]: {'topic_0': [{'doc_id': 260, 'score': 0.7238185053008797},
                     {'doc_id': 261, 'score': 0.7238185053008797},
                     {'doc_id': 273, 'score': 0.7238185053008797},
                     {'doc_id': 277, 'score': 0.7238185053008797},
                     {'doc_id': 278, 'score': 0.7238185053008797},
                     {'doc_id': 322, 'score': 0.7238185053008797},
                     {'doc_id': 325, 'score': 0.7238185053008797},
                     {'doc_id': 326, 'score': 0.7238185053008797},
                     {'doc_id': 327, 'score': 0.7238185053008797},
                     {'doc_id': 328, 'score': 0.7238185053008797},
                     {'doc_id': 346, 'score': 0.7238185053008797},
                     {'doc_id': 347, 'score': 0.7238185053008797},
                     {'doc_id': 359, 'score': 0.7238185053008797},
                     {'doc_id': 362, 'score': 0.7238185053008797},
                     {'doc_id': 367, 'score': 0.7238185053008797},
                     {'doc_id': 471, 'score': 0.7238185053008797},
                     {'doc_id': 472, 'score': 0.7238185053008797},
                     {'doc_id': 503, 'score': 0.7238185053008797},
                     {'doc_id': 558, 'score': 0.7238185053008797},
                     {'doc_id': 570, 'score': 0.7238185053008797},
                     {'doc_id': 226, 'score': 0.6666231079772607},
                     {'doc_id': 227, 'score': 0.6666231079772607},
                     {'doc_id': 228, 'score': 0.6666231079772607},
                     {'doc_id': 235, 'score': 0.6666231079772607},
                     {'doc_id': 248, 'score': 0.6666231079772607},
                     {'doc_id': 249, 'score': 0.6666231079772607},
                     {'doc_id': 283, 'score': 0.6666231079772607},
                     {'doc_id': 285, 'score': 0.6666231079772607},
                     {'doc_id': 293, 'score': 0.6666231079772607},
                     {'doc_id': 294, 'score': 0.6666231079772607},
                     {'doc_id': 318, 'score': 0.6666231079772607},
                     {'doc_id': 319, 'score': 0.6666231079772607},
                     {'doc_id': 320, 'score': 0.6666231079772607},
                     {'doc_id': 338, 'score': 0.6666231079772607},
                     {'doc_id': 357, 'score': 0.6666231079772607},
                     {'doc_id': 360, 'score': 0.6666231079772607},
                     {'doc_id': 363, 'score': 0.6666231079772607},
                     {'doc_id': 364, 'score': 0.6666231079772607},
                     {'doc_id': 366, 'score': 0.6666231079772607},
                     {'doc_id': 371, 'score': 0.6666231079772607},
                     {'doc_id': 372, 'score': 0.6666231079772607},
                     {'doc_id': 375, 'score': 0.6666231079772607},
                     {'doc_id': 381, 'score': 0.6666231079772607},
                     {'doc_id': 383, 'score': 0.6666231079772607},
                     {'doc_id': 384, 'score': 0.6666231079772607},
                     {'doc_id': 386, 'score': 0.6666231079772607},
                     {'doc_id': 387, 'score': 0.6666231079772607},
                     {'doc_id': 389, 'score': 0.6666231079772607},
                     {'doc_id': 415, 'score': 0.6666231079772607},
                     {'doc_id': 464, 'score': 0.6666231079772607},
                     {'doc_id': 483, 'score': 0.6666231079772607},
                     {'doc_id': 484, 'score': 0.6666231079772607},
                     {'doc_id': 487, 'score': 0.6666231079772607},
                     {'doc_id': 488, 'score': 0.6666231079772607},
                     {'doc_id': 489, 'score': 0.6666231079772607},
                     {'doc_id': 491, 'score': 0.6666231079772607},
                     {'doc_id': 500, 'score': 0.6666231079772607},
                     {'doc_id': 501, 'score': 0.6666231079772607},
                     {'doc_id': 502, 'score': 0.6666231079772607},
                     {'doc_id': 505, 'score': 0.6666231079772607},
                     {'doc_id': 507, 'score': 0.6666231079772607},
                     {'doc_id': 508, 'score': 0.6666231079772607},
                     {'doc_id': 519, 'score': 0.6666231079772607},
```

```
{'doc_id': 520, 'score': 0.6666231079772607},
{'doc_id': 524, 'score': 0.6666231079772607},
{'doc_id': 525, 'score': 0.6666231079772607},
{'doc_id': 589, 'score': 0.6666231079772607},
{'doc_id': 595, 'score': 0.6666231079772607},
{'doc_id': 603, 'score': 0.6666231079772607},
{'doc_id': 604, 'score': 0.6666231079772607},
{'doc_id': 644, 'score': 0.6666231079772607},
{'doc_id': 645, 'score': 0.6666231079772607},
{'doc_id': 653, 'score': 0.6666231079772607},
{'doc_id': 685, 'score': 0.6666231079772607},
{'doc_id': 699, 'score': 0.6666231079772607},
{'doc_id': 703, 'score': 0.6666231079772607},
{'doc_id': 704, 'score': 0.6666231079772607},
{'doc_id': 707, 'score': 0.6666231079772607},
{'doc_id': 717, 'score': 0.6666231079772607},
{'doc_id': 791, 'score': 0.6666231079772607},
{'doc_id': 792, 'score': 0.6666231079772607},
{'doc_id': 32, 'score': 0.6666173633870782},
{'doc_id': 33, 'score': 0.6666173633870782},
{'doc_id': 34, 'score': 0.6666173633870782},
{'doc_id': 38, 'score': 0.6666173633870782},
{'doc_id': 39, 'score': 0.6666173633870782},
{'doc_id': 205, 'score': 0.6666173633870782},
{'doc_id': 206, 'score': 0.6666173633870782},
{'doc_id': 210, 'score': 0.6666173633870782},
{'doc_id': 216, 'score': 0.6666173633870782},
{'doc_id': 219, 'score': 0.6666173633870782},
{'doc_id': 288, 'score': 0.6666173633870782},
{'doc_id': 316, 'score': 0.6666173633870782},
{'doc_id': 317, 'score': 0.6666173633870782},
{'doc_id': 333, 'score': 0.6666173633870782},
{'doc_id': 337, 'score': 0.6666173633870782},
{'doc_id': 349, 'score': 0.6666173633870782},
{'doc_id': 350, 'score': 0.6666173633870782},
{'doc_id': 351, 'score': 0.6666173633870782},
{'doc_id': 352, 'score': 0.6666173633870782},
{'doc_id': 353, 'score': 0.6666173633870782},
{'doc_id': 376, 'score': 0.6666173633870782},
{'doc_id': 418, 'score': 0.6666173633870782},
{'doc_id': 419, 'score': 0.6666173633870782},
{'doc_id': 420, 'score': 0.6666173633870782},
{'doc_id': 421, 'score': 0.6666173633870782},
{'doc_id': 423, 'score': 0.6666173633870782},
{'doc_id': 446, 'score': 0.6666173633870782},
{'doc_id': 452, 'score': 0.6666173633870782},
{'doc_id': 454, 'score': 0.6666173633870782},
{'doc_id': 467, 'score': 0.6666173633870782},
{'doc_id': 470, 'score': 0.6666173633870782},
{'doc_id': 509, 'score': 0.6666173633870782},
{'doc_id': 510, 'score': 0.6666173633870782},
{'doc_id': 532, 'score': 0.6666173633870782},
{'doc_id': 534, 'score': 0.6666173633870782},
{'doc_id': 536, 'score': 0.6666173633870782},
{'doc_id': 538, 'score': 0.6666173633870782},
{'doc_id': 539, 'score': 0.6666173633870782},
{'doc_id': 540, 'score': 0.6666173633870782},
{'doc_id': 567, 'score': 0.6666173633870782},
{'doc_id': 573, 'score': 0.6666173633870782},
{'doc_id': 599, 'score': 0.6666173633870782},
{'doc_id': 480, 'score': 0.5860739080994982},
{'doc_id': 481, 'score': 0.5860739080994982},
{'doc_id': 485, 'score': 0.5860739080994982},
```

```
{'doc_id': 486, 'score': 0.5860739080994982},
{'doc_id': 490, 'score': 0.5860739080994982},
{'doc_id': 492, 'score': 0.5860739080994982},
{'doc_id': 477, 'score': 0.5746127905833837},
{'doc_id': 138, 'score': 0.5370254512192199},
{'doc_id': 141, 'score': 0.5370254512192199},
{'doc_id': 392, 'score': 0.520664103202985},
{'doc_id': 393, 'score': 0.520664103202985},
{'doc_id': 395, 'score': 0.520664103202985},
{'doc_id': 396, 'score': 0.520664103202985},
{'doc_id': 397, 'score': 0.520664103202985},
{'doc_id': 398, 'score': 0.520664103202985},
{'doc_id': 400, 'score': 0.520664103202985},
{'doc_id': 401, 'score': 0.520664103202985},
{'doc_id': 402, 'score': 0.520664103202985},
{'doc_id': 403, 'score': 0.520664103202985},
{'doc_id': 404, 'score': 0.520664103202985},
{'doc_id': 405, 'score': 0.520664103202985},
{'doc_id': 407, 'score': 0.520664103202985},
{'doc_id': 408, 'score': 0.520664103202985},
{'doc_id': 409, 'score': 0.520664103202985},
{'doc_id': 410, 'score': 0.520664103202985},
{'doc_id': 411, 'score': 0.520664103202985},
{'doc_id': 600, 'score': 0.520664103202985},
{'doc_id': 256, 'score': 0.4452637462229781},
{'doc_id': 259, 'score': 0.4452637462229781},
{'doc_id': 272, 'score': 0.4452637462229781},
{'doc_id': 275, 'score': 0.4452637462229781},
{'doc_id': 331, 'score': 0.4452637462229781},
{'doc_id': 332, 'score': 0.4452637462229781},
{'doc_id': 598, 'score': 0.4438969675154334},
{'doc_id': 673, 'score': 0.4438969675154334},
{'doc_id': 674, 'score': 0.4438969675154334},
{'doc_id': 675, 'score': 0.4438969675154334},
{'doc_id': 1, 'score': 0.3333333333333333},
{'doc_id': 6, 'score': 0.3333333333333333},
{'doc_id': 16, 'score': 0.3333333333333333},
{'doc_id': 17, 'score': 0.3333333333333333},
{'doc_id': 18, 'score': 0.3333333333333333},
{'doc_id': 19, 'score': 0.3333333333333333},
{'doc_id': 20, 'score': 0.3333333333333333},
{'doc_id': 21, 'score': 0.3333333333333333},
{'doc_id': 35, 'score': 0.3333333333333333},
{'doc_id': 36, 'score': 0.3333333333333333},
{'doc_id': 37, 'score': 0.3333333333333333},
{'doc_id': 40, 'score': 0.3333333333333333},
{'doc_id': 41, 'score': 0.3333333333333333},
{'doc_id': 42, 'score': 0.3333333333333333},
{'doc_id': 43, 'score': 0.3333333333333333},
{'doc_id': 44, 'score': 0.3333333333333333},
{'doc_id': 45, 'score': 0.3333333333333333},
{'doc_id': 46, 'score': 0.3333333333333333},
{'doc_id': 47, 'score': 0.3333333333333333},
{'doc_id': 48, 'score': 0.3333333333333333},
{'doc_id': 49, 'score': 0.3333333333333333},
{'doc_id': 50, 'score': 0.3333333333333333},
{'doc_id': 53, 'score': 0.3333333333333333},
{'doc_id': 54, 'score': 0.3333333333333333},
{'doc_id': 55, 'score': 0.3333333333333333},
{'doc_id': 56, 'score': 0.3333333333333333},
{'doc_id': 57, 'score': 0.3333333333333333},
{'doc_id': 58, 'score': 0.3333333333333333},
{'doc_id': 59, 'score': 0.3333333333333333},
```

```
{'doc_id': 60, 'score': 0.3333333333333333},
{'doc_id': 61, 'score': 0.3333333333333333},
{'doc_id': 62, 'score': 0.3333333333333333},
{'doc_id': 63, 'score': 0.3333333333333333},
{'doc_id': 64, 'score': 0.3333333333333333},
{'doc_id': 65, 'score': 0.3333333333333333},
{'doc_id': 66, 'score': 0.3333333333333333},
{'doc_id': 68, 'score': 0.3333333333333333},
{'doc_id': 69, 'score': 0.3333333333333333},
{'doc_id': 70, 'score': 0.3333333333333333},
{'doc_id': 73, 'score': 0.3333333333333333},
{'doc_id': 76, 'score': 0.3333333333333333},
{'doc_id': 77, 'score': 0.3333333333333333},
{'doc_id': 78, 'score': 0.3333333333333333},
{'doc_id': 79, 'score': 0.3333333333333333},
{'doc_id': 80, 'score': 0.3333333333333333},
{'doc_id': 81, 'score': 0.3333333333333333},
{'doc_id': 82, 'score': 0.3333333333333333},
{'doc_id': 83, 'score': 0.3333333333333333},
{'doc_id': 84, 'score': 0.3333333333333333},
{'doc_id': 86, 'score': 0.3333333333333333},
{'doc_id': 87, 'score': 0.3333333333333333},
{'doc_id': 88, 'score': 0.3333333333333333},
{'doc_id': 89, 'score': 0.3333333333333333},
{'doc_id': 90, 'score': 0.3333333333333333},
{'doc_id': 91, 'score': 0.3333333333333333},
{'doc_id': 92, 'score': 0.3333333333333333},
{'doc_id': 93, 'score': 0.3333333333333333},
{'doc_id': 94, 'score': 0.3333333333333333},
{'doc_id': 95, 'score': 0.3333333333333333},
{'doc_id': 96, 'score': 0.3333333333333333},
{'doc_id': 97, 'score': 0.3333333333333333},
{'doc_id': 98, 'score': 0.3333333333333333},
{'doc_id': 99, 'score': 0.3333333333333333},
{'doc_id': 100, 'score': 0.3333333333333333},
{'doc_id': 102, 'score': 0.3333333333333333},
{'doc_id': 113, 'score': 0.3333333333333333},
{'doc_id': 114, 'score': 0.3333333333333333},
{'doc_id': 115, 'score': 0.3333333333333333},
{'doc_id': 116, 'score': 0.3333333333333333},
{'doc_id': 117, 'score': 0.3333333333333333},
{'doc_id': 118, 'score': 0.3333333333333333},
{'doc_id': 121, 'score': 0.3333333333333333},
{'doc_id': 122, 'score': 0.3333333333333333},
{'doc_id': 125, 'score': 0.3333333333333333},
{'doc_id': 126, 'score': 0.3333333333333333},
{'doc_id': 128, 'score': 0.3333333333333333},
{'doc_id': 129, 'score': 0.3333333333333333},
{'doc_id': 130, 'score': 0.3333333333333333},
{'doc_id': 131, 'score': 0.3333333333333333},
{'doc_id': 132, 'score': 0.3333333333333333},
{'doc_id': 133, 'score': 0.3333333333333333},
{'doc_id': 137, 'score': 0.3333333333333333},
{'doc_id': 139, 'score': 0.3333333333333333},
{'doc_id': 140, 'score': 0.3333333333333333},
{'doc_id': 142, 'score': 0.3333333333333333},
{'doc_id': 143, 'score': 0.3333333333333333},
{'doc_id': 144, 'score': 0.3333333333333333},
{'doc_id': 146, 'score': 0.3333333333333333},
{'doc_id': 148, 'score': 0.3333333333333333},
{'doc_id': 149, 'score': 0.3333333333333333},
{'doc_id': 151, 'score': 0.3333333333333333},
{'doc_id': 152, 'score': 0.3333333333333333},
```

```
{'doc_id': 153, 'score': 0.3333333333333333},
{'doc_id': 154, 'score': 0.3333333333333333},
{'doc_id': 155, 'score': 0.3333333333333333},
{'doc_id': 156, 'score': 0.3333333333333333},
{'doc_id': 157, 'score': 0.3333333333333333},
{'doc_id': 158, 'score': 0.3333333333333333},
{'doc_id': 159, 'score': 0.3333333333333333},
{'doc_id': 160, 'score': 0.3333333333333333},
{'doc_id': 162, 'score': 0.3333333333333333},
{'doc_id': 163, 'score': 0.3333333333333333},
{'doc_id': 164, 'score': 0.3333333333333333},
{'doc_id': 166, 'score': 0.3333333333333333},
{'doc_id': 173, 'score': 0.3333333333333333},
{'doc_id': 174, 'score': 0.3333333333333333},
{'doc_id': 175, 'score': 0.3333333333333333},
{'doc_id': 176, 'score': 0.3333333333333333},
{'doc_id': 177, 'score': 0.3333333333333333},
{'doc_id': 178, 'score': 0.3333333333333333},
{'doc_id': 179, 'score': 0.3333333333333333},
{'doc_id': 180, 'score': 0.3333333333333333},
{'doc_id': 182, 'score': 0.3333333333333333},
{'doc_id': 183, 'score': 0.3333333333333333},
{'doc_id': 185, 'score': 0.3333333333333333},
{'doc_id': 186, 'score': 0.3333333333333333},
{'doc_id': 187, 'score': 0.3333333333333333},
{'doc_id': 188, 'score': 0.3333333333333333},
{'doc_id': 190, 'score': 0.3333333333333333},
{'doc_id': 191, 'score': 0.3333333333333333},
{'doc_id': 193, 'score': 0.3333333333333333},
{'doc_id': 195, 'score': 0.3333333333333333},
{'doc_id': 196, 'score': 0.3333333333333333},
{'doc_id': 197, 'score': 0.3333333333333333},
{'doc_id': 198, 'score': 0.3333333333333333},
{'doc_id': 199, 'score': 0.3333333333333333},
{'doc_id': 200, 'score': 0.3333333333333333},
{'doc_id': 201, 'score': 0.3333333333333333},
{'doc_id': 202, 'score': 0.3333333333333333},
{'doc_id': 203, 'score': 0.3333333333333333},
{'doc_id': 204, 'score': 0.3333333333333333},
{'doc_id': 207, 'score': 0.3333333333333333},
{'doc_id': 208, 'score': 0.3333333333333333},
{'doc_id': 209, 'score': 0.3333333333333333},
{'doc_id': 211, 'score': 0.3333333333333333},
{'doc_id': 212, 'score': 0.3333333333333333},
{'doc_id': 214, 'score': 0.3333333333333333},
{'doc_id': 215, 'score': 0.3333333333333333},
{'doc_id': 217, 'score': 0.3333333333333333},
{'doc_id': 220, 'score': 0.3333333333333333},
{'doc_id': 223, 'score': 0.3333333333333333},
{'doc_id': 224, 'score': 0.3333333333333333},
{'doc_id': 225, 'score': 0.3333333333333333},
{'doc_id': 229, 'score': 0.3333333333333333},
{'doc_id': 230, 'score': 0.3333333333333333},
{'doc_id': 231, 'score': 0.3333333333333333},
{'doc_id': 232, 'score': 0.3333333333333333},
{'doc_id': 233, 'score': 0.3333333333333333},
{'doc_id': 234, 'score': 0.3333333333333333},
{'doc_id': 236, 'score': 0.3333333333333333},
{'doc_id': 237, 'score': 0.3333333333333333},
{'doc_id': 238, 'score': 0.3333333333333333},
{'doc_id': 239, 'score': 0.3333333333333333},
{'doc_id': 240, 'score': 0.3333333333333333},
{'doc_id': 241, 'score': 0.3333333333333333},
```

```
{'doc_id': 247, 'score': 0.3333333333333333},
{'doc_id': 250, 'score': 0.3333333333333333},
{'doc_id': 253, 'score': 0.3333333333333333},
{'doc_id': 254, 'score': 0.3333333333333333},
{'doc_id': 255, 'score': 0.3333333333333333},
{'doc_id': 268, 'score': 0.3333333333333333},
{'doc_id': 279, 'score': 0.3333333333333333},
{'doc_id': 282, 'score': 0.3333333333333333},
{'doc_id': 284, 'score': 0.3333333333333333},
{'doc_id': 287, 'score': 0.3333333333333333},
{'doc_id': 289, 'score': 0.3333333333333333},
{'doc_id': 290, 'score': 0.3333333333333333},
{'doc_id': 291, 'score': 0.3333333333333333},
{'doc_id': 292, 'score': 0.3333333333333333},
{'doc_id': 295, 'score': 0.3333333333333333},
{'doc_id': 297, 'score': 0.3333333333333333},
{'doc_id': 300, 'score': 0.3333333333333333},
{'doc_id': 302, 'score': 0.3333333333333333},
{'doc_id': 303, 'score': 0.3333333333333333},
{'doc_id': 304, 'score': 0.3333333333333333},
{'doc_id': 305, 'score': 0.3333333333333333},
{'doc_id': 306, 'score': 0.3333333333333333},
{'doc_id': 307, 'score': 0.3333333333333333},
{'doc_id': 308, 'score': 0.3333333333333333},
{'doc_id': 309, 'score': 0.3333333333333333},
{'doc_id': 310, 'score': 0.3333333333333333},
{'doc_id': 311, 'score': 0.3333333333333333},
{'doc_id': 312, 'score': 0.3333333333333333},
{'doc_id': 313, 'score': 0.3333333333333333},
{'doc_id': 314, 'score': 0.3333333333333333},
{'doc_id': 315, 'score': 0.3333333333333333},
{'doc_id': 321, 'score': 0.3333333333333333},
{'doc_id': 323, 'score': 0.3333333333333333},
{'doc_id': 324, 'score': 0.3333333333333333},
{'doc_id': 330, 'score': 0.3333333333333333},
{'doc_id': 334, 'score': 0.3333333333333333},
{'doc_id': 335, 'score': 0.3333333333333333},
{'doc_id': 336, 'score': 0.3333333333333333},
{'doc_id': 339, 'score': 0.3333333333333333},
{'doc_id': 340, 'score': 0.3333333333333333},
{'doc_id': 341, 'score': 0.3333333333333333},
{'doc_id': 342, 'score': 0.3333333333333333},
{'doc_id': 343, 'score': 0.3333333333333333},
{'doc_id': 344, 'score': 0.3333333333333333},
{'doc_id': 345, 'score': 0.3333333333333333},
{'doc_id': 348, 'score': 0.3333333333333333},
{'doc_id': 358, 'score': 0.3333333333333333},
{'doc_id': 361, 'score': 0.3333333333333333},
{'doc_id': 365, 'score': 0.3333333333333333},
{'doc_id': 368, 'score': 0.3333333333333333},
{'doc_id': 369, 'score': 0.3333333333333333},
{'doc_id': 370, 'score': 0.3333333333333333},
{'doc_id': 373, 'score': 0.3333333333333333},
{'doc_id': 374, 'score': 0.3333333333333333},
{'doc_id': 378, 'score': 0.3333333333333333},
{'doc_id': 379, 'score': 0.3333333333333333},
{'doc_id': 380, 'score': 0.3333333333333333},
{'doc_id': 382, 'score': 0.3333333333333333},
{'doc_id': 385, 'score': 0.3333333333333333},
{'doc_id': 388, 'score': 0.3333333333333333},
{'doc_id': 414, 'score': 0.3333333333333333},
{'doc_id': 424, 'score': 0.3333333333333333},
{'doc_id': 425, 'score': 0.3333333333333333},
```

```
{'doc_id': 426, 'score': 0.3333333333333333},
{'doc_id': 427, 'score': 0.3333333333333333},
{'doc_id': 428, 'score': 0.3333333333333333},
{'doc_id': 429, 'score': 0.3333333333333333},
{'doc_id': 430, 'score': 0.3333333333333333},
{'doc_id': 431, 'score': 0.3333333333333333},
{'doc_id': 432, 'score': 0.3333333333333333},
{'doc_id': 433, 'score': 0.3333333333333333},
{'doc_id': 434, 'score': 0.3333333333333333},
{'doc_id': 435, 'score': 0.3333333333333333},
{'doc_id': 436, 'score': 0.3333333333333333},
{'doc_id': 437, 'score': 0.3333333333333333},
{'doc_id': 438, 'score': 0.3333333333333333},
{'doc_id': 439, 'score': 0.3333333333333333},
{'doc_id': 440, 'score': 0.3333333333333333},
{'doc_id': 441, 'score': 0.3333333333333333},
{'doc_id': 442, 'score': 0.3333333333333333},
{'doc_id': 443, 'score': 0.3333333333333333},
{'doc_id': 444, 'score': 0.3333333333333333},
{'doc_id': 445, 'score': 0.3333333333333333},
{'doc_id': 447, 'score': 0.3333333333333333},
{'doc_id': 448, 'score': 0.3333333333333333},
{'doc_id': 449, 'score': 0.3333333333333333},
{'doc_id': 450, 'score': 0.3333333333333333},
{'doc_id': 453, 'score': 0.3333333333333333},
{'doc_id': 456, 'score': 0.3333333333333333},
{'doc_id': 457, 'score': 0.3333333333333333},
{'doc_id': 468, 'score': 0.3333333333333333},
{'doc_id': 469, 'score': 0.3333333333333333},
{'doc_id': 473, 'score': 0.3333333333333333},
{'doc_id': 474, 'score': 0.3333333333333333},
{'doc_id': 482, 'score': 0.3333333333333333},
{'doc_id': 493, 'score': 0.3333333333333333},
{'doc_id': 495, 'score': 0.3333333333333333},
{'doc_id': 497, 'score': 0.3333333333333333},
{'doc_id': 498, 'score': 0.3333333333333333},
{'doc_id': 499, 'score': 0.3333333333333333},
{'doc_id': 504, 'score': 0.3333333333333333},
{'doc_id': 506, 'score': 0.3333333333333333},
{'doc_id': 512, 'score': 0.3333333333333333},
{'doc_id': 522, 'score': 0.3333333333333333},
{'doc_id': 530, 'score': 0.3333333333333333},
{'doc_id': 531, 'score': 0.3333333333333333},
{'doc_id': 533, 'score': 0.3333333333333333},
{'doc_id': 535, 'score': 0.3333333333333333},
{'doc_id': 537, 'score': 0.3333333333333333},
{'doc_id': 542, 'score': 0.3333333333333333},
{'doc_id': 543, 'score': 0.3333333333333333},
{'doc_id': 544, 'score': 0.3333333333333333},
{'doc_id': 545, 'score': 0.3333333333333333},
{'doc_id': 546, 'score': 0.3333333333333333},
{'doc_id': 560, 'score': 0.3333333333333333},
{'doc_id': 563, 'score': 0.3333333333333333},
{'doc_id': 564, 'score': 0.3333333333333333},
{'doc_id': 566, 'score': 0.3333333333333333},
{'doc_id': 568, 'score': 0.3333333333333333},
{'doc_id': 569, 'score': 0.3333333333333333},
{'doc_id': 571, 'score': 0.3333333333333333},
{'doc_id': 572, 'score': 0.3333333333333333},
{'doc_id': 574, 'score': 0.3333333333333333},
{'doc_id': 575, 'score': 0.3333333333333333},
{'doc_id': 576, 'score': 0.3333333333333333},
{'doc_id': 577, 'score': 0.3333333333333333},
```

```
{'doc_id': 578, 'score': 0.3333333333333333},
{'doc_id': 579, 'score': 0.3333333333333333},
{'doc_id': 580, 'score': 0.3333333333333333},
{'doc_id': 582, 'score': 0.3333333333333333},
{'doc_id': 583, 'score': 0.3333333333333333},
{'doc_id': 584, 'score': 0.3333333333333333},
{'doc_id': 585, 'score': 0.3333333333333333},
{'doc_id': 586, 'score': 0.3333333333333333},
{'doc_id': 587, 'score': 0.3333333333333333},
{'doc_id': 588, 'score': 0.3333333333333333},
{'doc_id': 592, 'score': 0.3333333333333333},
{'doc_id': 593, 'score': 0.3333333333333333},
{'doc_id': 594, 'score': 0.3333333333333333},
{'doc_id': 597, 'score': 0.3333333333333333},
{'doc_id': 602, 'score': 0.3333333333333333},
{'doc_id': 605, 'score': 0.3333333333333333},
{'doc_id': 606, 'score': 0.3333333333333333},
{'doc_id': 608, 'score': 0.3333333333333333},
{'doc_id': 609, 'score': 0.3333333333333333},
{'doc_id': 610, 'score': 0.3333333333333333},
{'doc_id': 617, 'score': 0.3333333333333333},
{'doc_id': 618, 'score': 0.3333333333333333},
{'doc_id': 619, 'score': 0.3333333333333333},
{'doc_id': 621, 'score': 0.3333333333333333},
{'doc_id': 622, 'score': 0.3333333333333333},
{'doc_id': 623, 'score': 0.3333333333333333},
{'doc_id': 624, 'score': 0.3333333333333333},
{'doc_id': 625, 'score': 0.3333333333333333},
{'doc_id': 626, 'score': 0.3333333333333333},
{'doc_id': 627, 'score': 0.3333333333333333},
{'doc_id': 628, 'score': 0.3333333333333333},
{'doc_id': 629, 'score': 0.3333333333333333},
{'doc_id': 630, 'score': 0.3333333333333333},
{'doc_id': 631, 'score': 0.3333333333333333},
{'doc_id': 632, 'score': 0.3333333333333333},
{'doc_id': 633, 'score': 0.3333333333333333},
{'doc_id': 635, 'score': 0.3333333333333333},
{'doc_id': 636, 'score': 0.3333333333333333},
{'doc_id': 637, 'score': 0.3333333333333333},
{'doc_id': 638, 'score': 0.3333333333333333},
{'doc_id': 639, 'score': 0.3333333333333333},
{'doc_id': 640, 'score': 0.3333333333333333},
{'doc_id': 642, 'score': 0.3333333333333333},
{'doc_id': 643, 'score': 0.3333333333333333},
{'doc_id': 646, 'score': 0.3333333333333333},
{'doc_id': 647, 'score': 0.3333333333333333},
{'doc_id': 648, 'score': 0.3333333333333333},
{'doc_id': 649, 'score': 0.3333333333333333},
{'doc_id': 650, 'score': 0.3333333333333333},
{'doc_id': 651, 'score': 0.3333333333333333},
{'doc_id': 652, 'score': 0.3333333333333333},
{'doc_id': 654, 'score': 0.3333333333333333},
{'doc_id': 658, 'score': 0.3333333333333333},
{'doc_id': 661, 'score': 0.3333333333333333},
{'doc_id': 664, 'score': 0.3333333333333333},
{'doc_id': 665, 'score': 0.3333333333333333},
{'doc_id': 666, 'score': 0.3333333333333333},
{'doc_id': 667, 'score': 0.3333333333333333},
{'doc_id': 668, 'score': 0.3333333333333333},
{'doc_id': 669, 'score': 0.3333333333333333},
{'doc_id': 678, 'score': 0.3333333333333333},
{'doc_id': 679, 'score': 0.3333333333333333},
{'doc_id': 680, 'score': 0.3333333333333333},
```

```
{'doc_id': 681, 'score': 0.3333333333333333},
{'doc_id': 683, 'score': 0.3333333333333333},
{'doc_id': 691, 'score': 0.3333333333333333},
{'doc_id': 693, 'score': 0.3333333333333333},
{'doc_id': 694, 'score': 0.3333333333333333},
{'doc_id': 695, 'score': 0.3333333333333333},
{'doc_id': 696, 'score': 0.3333333333333333},
{'doc_id': 697, 'score': 0.3333333333333333},
{'doc_id': 698, 'score': 0.3333333333333333},
{'doc_id': 700, 'score': 0.3333333333333333},
{'doc_id': 714, 'score': 0.3333333333333333},
{'doc_id': 718, 'score': 0.3333333333333333},
{'doc_id': 719, 'score': 0.3333333333333333},
{'doc_id': 720, 'score': 0.3333333333333333},
{'doc_id': 721, 'score': 0.3333333333333333},
{'doc_id': 722, 'score': 0.3333333333333333},
{'doc_id': 723, 'score': 0.3333333333333333},
{'doc_id': 724, 'score': 0.3333333333333333},
{'doc_id': 725, 'score': 0.3333333333333333},
{'doc_id': 726, 'score': 0.3333333333333333},
{'doc_id': 727, 'score': 0.3333333333333333},
{'doc_id': 728, 'score': 0.3333333333333333},
{'doc_id': 729, 'score': 0.3333333333333333},
{'doc_id': 730, 'score': 0.3333333333333333},
{'doc_id': 731, 'score': 0.3333333333333333},
{'doc_id': 732, 'score': 0.3333333333333333},
{'doc_id': 733, 'score': 0.3333333333333333},
{'doc_id': 734, 'score': 0.3333333333333333},
{'doc_id': 735, 'score': 0.3333333333333333},
{'doc_id': 736, 'score': 0.3333333333333333},
{'doc_id': 738, 'score': 0.3333333333333333},
{'doc_id': 740, 'score': 0.3333333333333333},
{'doc_id': 741, 'score': 0.3333333333333333},
{'doc_id': 742, 'score': 0.3333333333333333},
{'doc_id': 744, 'score': 0.3333333333333333},
{'doc_id': 745, 'score': 0.3333333333333333},
{'doc_id': 753, 'score': 0.3333333333333333},
{'doc_id': 754, 'score': 0.3333333333333333},
{'doc_id': 775, 'score': 0.3333333333333333},
{'doc_id': 776, 'score': 0.3333333333333333},
{'doc_id': 777, 'score': 0.3333333333333333},
{'doc_id': 779, 'score': 0.3333333333333333},
{'doc_id': 786, 'score': 0.3333333333333333},
{'doc_id': 787, 'score': 0.3333333333333333},
{'doc_id': 788, 'score': 0.3333333333333333},
{'doc_id': 789, 'score': 0.3333333333333333},
{'doc_id': 790, 'score': 0.3333333333333333},
{'doc_id': 794, 'score': 0.3333333333333333},
{'doc_id': 795, 'score': 0.3333333333333333},
{'doc_id': 796, 'score': 0.3333333333333333},
{'doc_id': 797, 'score': 0.3333333333333333},
{'doc_id': 798, 'score': 0.3333333333333333},
{'doc_id': 802, 'score': 0.3333333333333333},
{'doc_id': 803, 'score': 0.3333333333333333},
{'doc_id': 807, 'score': 0.3333333333333333},
{'doc_id': 810, 'score': 0.3333333333333333},
{'doc_id': 811, 'score': 0.3333333333333333},
{'doc_id': 812, 'score': 0.3333333333333333},
{'doc_id': 813, 'score': 0.3333333333333333},
{'doc_id': 814, 'score': 0.3333333333333333},
{'doc_id': 815, 'score': 0.3333333333333333},
{'doc_id': 816, 'score': 0.3333333333333333},
{'doc_id': 817, 'score': 0.3333333333333333},
```

```
{'doc_id': 818, 'score': 0.3333333333333333},
{'doc_id': 819, 'score': 0.3333333333333333},
{'doc_id': 820, 'score': 0.3333333333333333},
{'doc_id': 821, 'score': 0.3333333333333333},
{'doc_id': 822, 'score': 0.3333333333333333},
{'doc_id': 823, 'score': 0.3333333333333333},
{'doc_id': 824, 'score': 0.3333333333333333},
{'doc_id': 825, 'score': 0.3333333333333333},
{'doc_id': 826, 'score': 0.3333333333333333},
{'doc_id': 829, 'score': 0.3333333333333333},
{'doc_id': 830, 'score': 0.3333333333333333},
{'doc_id': 831, 'score': 0.3333333333333333},
{'doc_id': 833, 'score': 0.3333333333333333},
{'doc_id': 834, 'score': 0.3333333333333333},
{'doc_id': 835, 'score': 0.3333333333333333},
{'doc_id': 836, 'score': 0.3333333333333333},
{'doc_id': 837, 'score': 0.3333333333333333},
{'doc_id': 838, 'score': 0.3333333333333333},
{'doc_id': 839, 'score': 0.3333333333333333}],
'topic_1': [{'doc_id': 0, 'score': 0.6666534244750105},
{'doc_id': 2, 'score': 0.6666534244750105},
{'doc_id': 3, 'score': 0.6666534244750105},
{'doc_id': 4, 'score': 0.6666534244750105},
{'doc_id': 5, 'score': 0.6666534244750105},
{'doc_id': 8, 'score': 0.6666534244750105},
{'doc_id': 9, 'score': 0.6666534244750105},
{'doc_id': 10, 'score': 0.6666534244750105},
{'doc_id': 11, 'score': 0.6666534244750105},
{'doc_id': 12, 'score': 0.6666534244750105},
{'doc_id': 134, 'score': 0.6666534244750105},
{'doc_id': 135, 'score': 0.6666534244750105},
{'doc_id': 147, 'score': 0.6666534244750105},
{'doc_id': 167, 'score': 0.6666534244750105},
{'doc_id': 265, 'score': 0.6666534244750105},
{'doc_id': 266, 'score': 0.6666534244750105},
{'doc_id': 267, 'score': 0.6666534244750105},
{'doc_id': 276, 'score': 0.6666534244750105},
{'doc_id': 354, 'score': 0.6666534244750105},
{'doc_id': 356, 'score': 0.6666534244750105},
{'doc_id': 394, 'score': 0.6666534244750105},
{'doc_id': 406, 'score': 0.6666534244750105},
{'doc_id': 412, 'score': 0.6666534244750105},
{'doc_id': 413, 'score': 0.6666534244750105},
{'doc_id': 416, 'score': 0.6666534244750105},
{'doc_id': 417, 'score': 0.6666534244750105},
{'doc_id': 461, 'score': 0.6666534244750105},
{'doc_id': 462, 'score': 0.6666534244750105},
{'doc_id': 463, 'score': 0.6666534244750105},
{'doc_id': 465, 'score': 0.6666534244750105},
{'doc_id': 466, 'score': 0.6666534244750105},
{'doc_id': 549, 'score': 0.6666534244750105},
{'doc_id': 550, 'score': 0.6666534244750105},
{'doc_id': 555, 'score': 0.6666534244750105},
{'doc_id': 556, 'score': 0.6666534244750105},
{'doc_id': 657, 'score': 0.6666534244750105},
{'doc_id': 662, 'score': 0.6666534244750105},
{'doc_id': 782, 'score': 0.6666534244750105},
{'doc_id': 801, 'score': 0.6666534244750105},
{'doc_id': 804, 'score': 0.6666534244750105},
{'doc_id': 805, 'score': 0.6666534244750105},
{'doc_id': 806, 'score': 0.6666534244750105},
{'doc_id': 778, 'score': 0.5391665330232535},
{'doc_id': 780, 'score': 0.5391665330232535},
```

```
            {'doc_id': 262, 'score': 0.44990913524539294},
            {'doc_id': 263, 'score': 0.44990913524539294},
            {'doc_id': 270, 'score': 0.44990913524539294},
            {'doc_id': 271, 'score': 0.44990913524539294},
            {'doc_id': 296, 'score': 0.44990913524539294},
            {'doc_id': 458, 'score': 0.44990913524539294},
            {'doc_id': 459, 'score': 0.44990913524539294},
            {'doc_id': 460, 'score': 0.44990913524539294},
            {'doc_id': 743, 'score': 0.44990913524539294},
            {'doc_id': 761, 'score': 0.44990913524539294},
            {'doc_id': 762, 'score': 0.44990913524539294},
            {'doc_id': 763, 'score': 0.44990913524539294},
            {'doc_id': 764, 'score': 0.44990913524539294},
            {'doc_id': 765, 'score': 0.44990913524539294},
            {'doc_id': 766, 'score': 0.44990913524539294},
            {'doc_id': 767, 'score': 0.44990913524539294},
            {'doc_id': 769, 'score': 0.44990913524539294},
            {'doc_id': 770, 'score': 0.44990913524539294},
            {'doc_id': 784, 'score': 0.44990913524539294},
            {'doc_id': 399, 'score': 0.43698676211708853},
            {'doc_id': 548, 'score': 0.43698676211708853},
            {'doc_id': 551, 'score': 0.43698676211708853},
            {'doc_id': 552, 'score': 0.43698676211708853},
            {'doc_id': 676, 'score': 0.43698676211708853},
            {'doc_id': 286, 'score': 0.43566057995314783},
            {'doc_id': 329, 'score': 0.43566057995314783},
            {'doc_id': 355, 'score': 0.43566057995314783},
            {'doc_id': 601, 'score': 0.43566057995314783},
            {'doc_id': 455, 'score': 0.34490764706053095}],
           'topic_2': [{'doc_id': 511, 'score': 0.723367759759252},
            {'doc_id': 516, 'score': 0.723367759759252},
            {'doc_id': 526, 'score': 0.723367759759252},
            {'doc_id': 528, 'score': 0.723367759759252},
            {'doc_id': 529, 'score': 0.723367759759252},
            {'doc_id': 687, 'score': 0.723367759759252},
            {'doc_id': 773, 'score': 0.723367759759252},
            {'doc_id': 783, 'score': 0.723367759759252},
            {'doc_id': 785, 'score': 0.723367759759252},
            {'doc_id': 611, 'score': 0.7182862983141161},
            {'doc_id': 612, 'score': 0.7182862983141161},
            {'doc_id': 613, 'score': 0.7182862983141161},
            {'doc_id': 615, 'score': 0.7182862983141161},
            {'doc_id': 616, 'score': 0.7182862983141161},
            {'doc_id': 181, 'score': 0.66662423509718},
            {'doc_id': 184, 'score': 0.66662423509718},
            {'doc_id': 189, 'score': 0.66662423509718},
            {'doc_id': 192, 'score': 0.66662423509718},
            {'doc_id': 194, 'score': 0.66662423509718},
            {'doc_id': 221, 'score': 0.66662423509718},
            {'doc_id': 242, 'score': 0.66662423509718},
            {'doc_id': 244, 'score': 0.66662423509718},
            {'doc_id': 246, 'score': 0.66662423509718},
            {'doc_id': 257, 'score': 0.66662423509718},
            {'doc_id': 258, 'score': 0.66662423509718},
            {'doc_id': 264, 'score': 0.66662423509718},
            {'doc_id': 274, 'score': 0.66662423509718},
            {'doc_id': 280, 'score': 0.66662423509718},
            {'doc_id': 298, 'score': 0.66662423509718},
            {'doc_id': 299, 'score': 0.66662423509718},
            {'doc_id': 301, 'score': 0.66662423509718},
            {'doc_id': 377, 'score': 0.66662423509718},
            {'doc_id': 390, 'score': 0.66662423509718},
            {'doc_id': 391, 'score': 0.66662423509718},
```

```
{'doc_id': 451, 'score': 0.66662423509718},
{'doc_id': 475, 'score': 0.66662423509718},
{'doc_id': 476, 'score': 0.66662423509718},
{'doc_id': 478, 'score': 0.66662423509718},
{'doc_id': 479, 'score': 0.66662423509718},
{'doc_id': 494, 'score': 0.66662423509718},
{'doc_id': 496, 'score': 0.66662423509718},
{'doc_id': 513, 'score': 0.66662423509718},
{'doc_id': 517, 'score': 0.66662423509718},
{'doc_id': 518, 'score': 0.66662423509718},
{'doc_id': 521, 'score': 0.66662423509718},
{'doc_id': 523, 'score': 0.66662423509718},
{'doc_id': 557, 'score': 0.66662423509718},
{'doc_id': 559, 'score': 0.66662423509718},
{'doc_id': 561, 'score': 0.66662423509718},
{'doc_id': 562, 'score': 0.66662423509718},
{'doc_id': 565, 'score': 0.66662423509718},
{'doc_id': 591, 'score': 0.66662423509718},
{'doc_id': 596, 'score': 0.66662423509718},
{'doc_id': 607, 'score': 0.66662423509718},
{'doc_id': 656, 'score': 0.66662423509718},
{'doc_id': 659, 'score': 0.66662423509718},
{'doc_id': 660, 'score': 0.66662423509718},
{'doc_id': 663, 'score': 0.66662423509718},
{'doc_id': 688, 'score': 0.66662423509718},
{'doc_id': 701, 'score': 0.66662423509718},
{'doc_id': 702, 'score': 0.66662423509718},
{'doc_id': 705, 'score': 0.66662423509718},
{'doc_id': 706, 'score': 0.66662423509718},
{'doc_id': 708, 'score': 0.66662423509718},
{'doc_id': 709, 'score': 0.66662423509718},
{'doc_id': 710, 'score': 0.66662423509718},
{'doc_id': 711, 'score': 0.66662423509718},
{'doc_id': 712, 'score': 0.66662423509718},
{'doc_id': 713, 'score': 0.66662423509718},
{'doc_id': 715, 'score': 0.66662423509718},
{'doc_id': 716, 'score': 0.66662423509718},
{'doc_id': 737, 'score': 0.66662423509718},
{'doc_id': 739, 'score': 0.66662423509718},
{'doc_id': 746, 'score': 0.66662423509718},
{'doc_id': 747, 'score': 0.66662423509718},
{'doc_id': 748, 'score': 0.66662423509718},
{'doc_id': 749, 'score': 0.66662423509718},
{'doc_id': 750, 'score': 0.66662423509718},
{'doc_id': 751, 'score': 0.66662423509718},
{'doc_id': 752, 'score': 0.66662423509718},
{'doc_id': 756, 'score': 0.66662423509718},
{'doc_id': 757, 'score': 0.66662423509718},
{'doc_id': 758, 'score': 0.66662423509718},
{'doc_id': 759, 'score': 0.66662423509718},
{'doc_id': 760, 'score': 0.66662423509718},
{'doc_id': 768, 'score': 0.66662423509718},
{'doc_id': 771, 'score': 0.66662423509718},
{'doc_id': 772, 'score': 0.66662423509718},
{'doc_id': 781, 'score': 0.66662423509718},
{'doc_id': 793, 'score': 0.66662423509718},
{'doc_id': 799, 'score': 0.66662423509718},
{'doc_id': 800, 'score': 0.66662423509718},
{'doc_id': 808, 'score': 0.66662423509718},
{'doc_id': 809, 'score': 0.66662423509718},
{'doc_id': 832, 'score': 0.66662423509718},
{'doc_id': 840, 'score': 0.66662423509718},
{'doc_id': 841, 'score': 0.66662423509718},
```

{'doc_id': 842, 'score': 0.6666242350971 8},
{'doc_id': 843, 'score': 0.66662423509718},
{'doc_id': 844, 'score': 0.66662423509718},
{'doc_id': 845, 'score': 0.66662423509718},
{'doc_id': 7, 'score': 0.6666003786695097},
{'doc_id': 13, 'score': 0.6666003786695097},
{'doc_id': 14, 'score': 0.6666003786695097},
{'doc_id': 15, 'score': 0.6666003786695097},
{'doc_id': 22, 'score': 0.6666003786695097},
{'doc_id': 23, 'score': 0.6666003786695097},
{'doc_id': 24, 'score': 0.6666003786695097},
{'doc_id': 25, 'score': 0.6666003786695097},
{'doc_id': 26, 'score': 0.6666003786695097},
{'doc_id': 27, 'score': 0.6666003786695097},
{'doc_id': 28, 'score': 0.6666003786695097},
{'doc_id': 29, 'score': 0.6666003786695097},
{'doc_id': 30, 'score': 0.6666003786695097},
{'doc_id': 31, 'score': 0.6666003786695097},
{'doc_id': 52, 'score': 0.6666003786695097},
{'doc_id': 67, 'score': 0.6666003786695097},
{'doc_id': 71, 'score': 0.6666003786695097},
{'doc_id': 72, 'score': 0.6666003786695097},
{'doc_id': 74, 'score': 0.6666003786695097},
{'doc_id': 75, 'score': 0.6666003786695097},
{'doc_id': 85, 'score': 0.6666003786695097},
{'doc_id': 101, 'score': 0.6666003786695097},
{'doc_id': 103, 'score': 0.6666003786695097},
{'doc_id': 104, 'score': 0.6666003786695097},
{'doc_id': 105, 'score': 0.6666003786695097},
{'doc_id': 106, 'score': 0.6666003786695097},
{'doc_id': 107, 'score': 0.6666003786695097},
{'doc_id': 108, 'score': 0.6666003786695097},
{'doc_id': 109, 'score': 0.6666003786695097},
{'doc_id': 110, 'score': 0.6666003786695097},
{'doc_id': 111, 'score': 0.6666003786695097},
{'doc_id': 112, 'score': 0.6666003786695097},
{'doc_id': 119, 'score': 0.6666003786695097},
{'doc_id': 120, 'score': 0.6666003786695097},
{'doc_id': 123, 'score': 0.6666003786695097},
{'doc_id': 124, 'score': 0.6666003786695097},
{'doc_id': 127, 'score': 0.6666003786695097},
{'doc_id': 136, 'score': 0.6666003786695097},
{'doc_id': 161, 'score': 0.6666003786695097},
{'doc_id': 165, 'score': 0.6666003786695097},
{'doc_id': 168, 'score': 0.6666003786695097},
{'doc_id': 169, 'score': 0.6666003786695097},
{'doc_id': 170, 'score': 0.6666003786695097},
{'doc_id': 171, 'score': 0.6666003786695097},
{'doc_id': 172, 'score': 0.6666003786695097},
{'doc_id': 213, 'score': 0.6666003786695097},
{'doc_id': 218, 'score': 0.6666003786695097},
{'doc_id': 222, 'score': 0.6666003786695097},
{'doc_id': 269, 'score': 0.6666003786695097},
{'doc_id': 514, 'score': 0.6666003786695097},
{'doc_id': 515, 'score': 0.6666003786695097},
{'doc_id': 527, 'score': 0.6666003786695097},
{'doc_id': 614, 'score': 0.6666003786695097},
{'doc_id': 620, 'score': 0.6666003786695097},
{'doc_id': 634, 'score': 0.6666003786695097},
{'doc_id': 641, 'score': 0.6666003786695097},
{'doc_id': 755, 'score': 0.6666003786695097},
{'doc_id': 774, 'score': 0.6666003786695097},
{'doc_id': 670, 'score': 0.43959889057783413},

```
{'doc_id': 671, 'score': 0.43959889057783413},
{'doc_id': 672, 'score': 0.43959889057783413},
{'doc_id': 677, 'score': 0.43959889057783413},
{'doc_id': 682, 'score': 0.43959889057783413}]}
```

In [10]:
```
# select only those sentences which have score more than 65%

sentences = {c:[] for c in categories}
selected_sentences = {c:[] for c in categories}
sentences_with_score = {c:[] for c in categories}


for each_topic in topic_leaders:
    for each_doc in topic_leaders[each_topic]:
        gt = ground_truth[each_doc['doc_id']]
        sentences[gt].append(each_doc['doc_id'])
        sentences_with_score[gt].append(each_doc['score'])
#         print(each_doc['score'])
        if each_doc['score']>0.66:
             selected_sentences[gt].append(each_doc['doc_id'])

selected_sentences
```

```
Out[10]: {'drug_gene': [226,
                        227,
                        228,
                        235,
                        32,
                        33,
                        34,
                        38,
                        39,
                        205,
                        206,
                        210,
                        216,
                        219,
                        0,
                        2,
                        3,
                        4,
                        5,
                        8,
                        9,
                        10,
                        11,
                        12,
                        134,
                        135,
                        147,
                        167,
                        181,
                        184,
                        189,
                        192,
                        194,
                        221,
                        242,
                        244,
                        246,
                        7,
                        13,
                        14,
                        15,
                        22,
                        23,
                        24,
                        25,
                        26,
                        27,
                        28,
                        29,
                        30,
                        31,
                        52,
                        67,
                        71,
                        72,
                        74,
                        75,
                        85,
                        101,
                        103,
                        104,
                        105,
                        106,
```

```
          107,
          108,
          109,
          110,
          111,
          112,
          119,
          120,
          123,
          124,
          127,
          136,
          161,
          165,
          168,
          169,
          170,
          171,
          172,
          213,
          218,
          222],
         'gene_disease': [260,
          261,
          273,
          277,
          278,
          322,
          325,
          326,
          327,
          328,
          346,
          347,
          359,
          362,
          367,
          471,
          472,
          503,
          558,
          570,
          248,
          249,
          283,
          285,
          293,
          294,
          318,
          319,
          320,
          338,
          357,
          360,
          363,
          364,
          366,
          371,
          372,
          375,
          381,
          383,
          384,
```

```
386,
387,
389,
415,
464,
483,
484,
487,
488,
489,
491,
500,
501,
502,
505,
507,
508,
519,
520,
524,
525,
589,
595,
288,
316,
317,
333,
337,
349,
350,
351,
352,
353,
376,
418,
419,
420,
421,
423,
446,
452,
454,
467,
470,
509,
510,
532,
534,
536,
538,
539,
540,
567,
573,
599,
265,
266,
267,
276,
354,
356,
394,
406,
```

```
          412,
          413,
          416,
          417,
          461,
          462,
          463,
          465,
          466,
          549,
          550,
          555,
          556,
          511,
          516,
          526,
          528,
          529,
          257,
          258,
          264,
          274,
          280,
          298,
          299,
          301,
          377,
          390,
          391,
          451,
          475,
          476,
          478,
          479,
          494,
          496,
          513,
          517,
          518,
          521,
          523,
          557,
          559,
          561,
          562,
          565,
          591,
          596,
          269,
          514,
          515,
          527],
         'drug_disease': [603,
          604,
          644,
          645,
          653,
          685,
          699,
          703,
          704,
          707,
          717,
```

```
791,
792,
657,
662,
782,
801,
804,
805,
806,
687,
773,
783,
785,
611,
612,
613,
615,
616,
607,
656,
659,
660,
663,
688,
701,
702,
705,
706,
708,
709,
710,
711,
712,
713,
715,
716,
737,
739,
746,
747,
748,
749,
750,
751,
752,
756,
757,
758,
759,
760,
768,
771,
772,
781,
793,
799,
800,
808,
809,
832,
840,
841,
842,
```

```
843,
844,
845,
614,
620,
634,
641,
755,
774]}
```

In [51]:
```
# Get average/mean of the sentence vectors that represent our topics
category_vecs = {}
for c in categories:
    vectors = np.asarray(list(df.loc[df.index.isin(selected_sentences
[c])].vector))
    category_vecs[c] = np.mean(vectors, axis=0)


category_vecs
```

```
Out[51]: {'drug_gene': array([-3.21358114e-02,  7.47584684e-02,  8.34615362e-02,  5.54770049e-02,
         1.22370782e-01,  6.47800698e-02,  1.48554523e-02, -4.29823855e-02,
        -5.56380577e-03, -1.21374275e-02,  1.48823383e-02, -7.69251622e-02,
         6.65993555e-02, -1.31682239e-03, -2.55387447e-02, -5.52268020e-02,
        -1.05548672e-01, -1.75949691e-01,  1.94938574e-02,  2.32216645e-02,
        -5.88295000e-02, -8.70603118e-02, -4.02607696e-02, -5.23781347e-02,
         3.46875427e-02, -6.53689495e-02,  3.71152185e-02, -6.13846829e-03,
        -2.33896747e-03, -7.99919579e-02,  2.72428694e-02, -1.00196057e-01,
        -1.26801660e-01, -1.43025047e-02,  6.74936439e-02, -9.01519123e-02,
        -3.06453595e-02,  3.62881264e-04, -8.61218259e-02, -9.98713983e-02,
         8.52346434e-03, -4.63008804e-02,  7.61586987e-02,  6.02405544e-02,
         1.49043512e-01, -4.74789195e-02, -1.12607301e-01, -8.31325829e-02,
         3.59362343e-02, -4.91476457e-02, -5.04195978e-02, -1.78867236e-02,
        -4.30926206e-02,  1.48498244e-03, -1.17466739e-01,  2.16820454e-02,
         1.07767951e-01, -9.93113626e-03, -1.24433138e-02,  2.56691025e-02,
         2.71850243e-02, -4.15987103e-02,  8.22673358e-04, -3.47494955e-02,
         1.32456623e-02,  1.60453196e-02, -4.28071777e-03,  3.00619219e-02,
        -3.07230631e-02, -7.02779287e-02, -6.12363811e-02, -4.29723483e-02,
        -6.42761764e-03, -1.09354332e-02,  5.83743663e-02,  9.97263251e-02,
        -8.77940536e-03,  4.00785255e-02, -1.16835386e-02,  5.95527576e-02,
        -8.76003605e-03, -6.34598044e-03,  1.07568391e-02,  4.66128902e-02,
        -4.02505743e-02,  4.01753399e-02, -8.93132851e-02, -5.00251547e-03,
         3.28462713e-02,  6.83463223e-02, -8.08918575e-02,  2.93257623e-02,
        -5.23161134e-03, -1.14727769e-01,  4.87972922e-02, -3.92898242e-02,
        -3.09750358e-02, -1.60451883e-02,  1.10024698e-01,  5.73426101e-03,
         8.07251200e-03,  7.61199947e-02, -5.18658055e-02, -5.38740028e-02,
         4.25488163e-02,  2.15148309e-02, -3.41121980e-02,  6.97964889e-02,
         4.62249754e-02,  4.63816486e-02,  3.17021479e-02, -2.74532656e-02,
         3.22619789e-02, -4.58491603e-02,  7.02673902e-02,  2.33391262e-02,
        -1.12527125e-04,  5.00852300e-03,  2.72571562e-02, -6.68687214e-02,
         4.99662371e-02,  7.79334784e-03,  1.01255842e-02, -5.51393342e-02,
         4.37017051e-02, -1.33691681e-02,  2.02518436e-02, -2.71089455e-
```

```
2e-02,
        -6.55510129e-02, -6.93302098e-02, -1.62063832e-02,  4.8118767
2e-02,
        -3.48948293e-02,  8.52577989e-02, -3.60349288e-02,  2.4698636
4e-02,
        -7.10419098e-02, -8.07612406e-02, -3.06760032e-02,  7.2486604
4e-02,
        -7.84599190e-02,  4.65381260e-02,  5.71827669e-02,  3.0278500
2e-02,
        -7.61449952e-02,  3.61775408e-02,  2.57776465e-02, -2.2741893
6e-02,
         5.49456848e-02,  5.22752115e-02,  9.56610620e-02,  2.3250120
6e-02,
         3.85851059e-02, -5.21761633e-03, -1.09777188e-02, -3.0703721
2e-02,
        -9.00035355e-02, -4.21000435e-02, -3.67912960e-02, -2.1536545
3e-02,
        -3.47972345e-03, -1.46597149e-02,  4.61616761e-02,  2.8926755
8e-02,
         4.50374549e-02, -1.60049456e-02,  7.57737199e-02,  5.2776023
4e-02,
         1.12110462e-01,  9.67839472e-04,  7.29464514e-03, -1.1062052
2e-01,
        -8.76695149e-02, -4.46648288e-02, -1.71999867e-01,  2.6458667
5e-02,
        -6.13840077e-02, -1.37182222e-01,  1.57664503e-02, -4.2387985
7e-02,
        -2.57846754e-02, -3.22512145e-02, -8.39264025e-02, -1.0153204
4e-01,
         4.18404134e-02,  7.59434879e-03, -1.05515737e-01, -2.3676074
0e-02,
         5.94890723e-02,  7.97219496e-02, -2.97200690e-02, -2.0964563
9e-02,
        -1.49659041e-02,  8.65183533e-02, -3.40028310e-02,  5.1889590
9e-02,
         9.15189238e-03,  1.50569177e-01, -1.44747186e-02, -1.3732551
1e-04]),
 'gene_disease': array([-0.02187848,  0.02261432,  0.04347951, -0.034
36503,  0.05994518,
         0.01016135, -0.04064447, -0.02884977, -0.02125299,  0.028787
05,
        -0.00961321, -0.11713743,  0.04926057,  0.01782062,  0.013883
86,
        -0.03751719, -0.06298924, -0.13242544,  0.02968876,  0.056578
15,
         0.02580438, -0.04663612, -0.03478725, -0.0656557 ,  0.038579
02,
        -0.04557844,  0.04125495, -0.06552493,  0.01678645, -0.063438
93,
         0.04707222, -0.0485334 , -0.04294874, -0.0002181 ,  0.057142
45,
        -0.04009597,  0.03610215,  0.02675375, -0.07729152, -0.028768
31,
        -0.00098729, -0.0614024 ,  0.05501022,  0.05027915,  0.097038
88,
        -0.07497141, -0.11520871, -0.09071015, -0.05484232, -0.088313
72,
         0.02790412,  0.00720131, -0.01327416,  0.037259  , -0.098706
81,
         0.03853304,  0.00272046, -0.00166644,  0.00517164,  0.017588
52,
         0.03860681,  0.00870928,  0.00118799, -0.02187852,  0.061098
,
```

```
        -0.0008905 , -0.02307511,  0.08415092,  0.00600624, -0.012047
02,
        -0.02262991, -0.04427554, -0.03507224, -0.03901289, -0.009809
61,
         0.06319557,  0.07922085,  0.00570427, -0.04413532, -0.012876
94,
         0.04755626, -0.02248873, -0.03658225, -0.02581983, -0.029596
72,
        -0.0559589 , -0.10173522, -0.01716438, -0.00368899,  0.028118
68,
         0.00583557,  0.02988015, -0.08527605, -0.06217288,  0.027981
27,
         0.0033972 ,  0.00375104, -0.04024579,  0.07685259, -0.053746
99,
         0.02228696,  0.04957056,  0.00663859, -0.02055714,  0.028205
57,
        -0.03466263,  0.02183046,  0.03147696,  0.01271996,  0.008499
64,
        -0.00071194, -0.02261991,  0.03998857, -0.02785565,  0.071842
61,
        -0.01712422,  0.01217643,  0.04677326,  0.02842865, -0.004455
77,
         0.06404803, -0.04052421,  0.0110433 , -0.03783824,  0.007866
23,
        -0.11159688,  0.03350807, -0.06536345, -0.01768567, -0.048658
84,
         0.01291716,  0.02250472, -0.06034084,  0.07633559, -0.026941
02,
         0.08186165, -0.0544764 , -0.08004978,  0.00317508,  0.010277
26,
        -0.04187827,  0.03649497,  0.05451602,  0.08189519,  0.004970
2 ,
        -0.01435066,  0.04395742, -0.03096783,  0.02236778,  0.073577
79,
         0.04618307,  0.03032329,  0.02596646, -0.01300092,  0.013862
76,
         0.00233598, -0.03143857, -0.06299815, -0.02081634, -0.000635
13,
        -0.02390152,  0.03012555, -0.00026283,  0.11429763,  0.030054
69,
         0.00265592, -0.0012079 ,  0.06557988,  0.04042032, -0.009574
21,
        -0.02174161, -0.1380254 ,  0.00837358, -0.06673517, -0.105917
6 ,
        -0.0392799 , -0.01971467, -0.10724238,  0.01637654,  0.002089
37,
        -0.08435808,  0.01913901, -0.08513355, -0.04749972,  0.092500
93,
         0.03824112, -0.08837281, -0.02147611,  0.03730662, -0.080697
29,
         0.0099313 , -0.01862315,  0.04886034,  0.07827532, -0.065298
06,
        -0.01439844, -0.01804423,  0.1041307 , -0.04506333,  0.006562
34]),
 'drug_disease': array([-8.86178220e-02,  1.23170130e-01,  4.62812525
e-02, -1.33085024e-02,
         2.56571521e-02, -2.16187252e-02,  3.85098258e-02, -2.4139740
9e-05,
         1.23486649e-02,  6.09282665e-02, -1.06062620e-02, -6.9389112
9e-02,
         1.23976207e-01, -1.01147045e-02, -4.99754701e-02, -9.4884857
1e-02,
        -5.98112924e-02, -1.13503379e-01,  2.87662765e-02,  3.7594017
```

```
4e-02,
         8.68969770e-02, -3.58414404e-02, -3.72036945e-02, -7.6340317
5e-02,
         2.36848877e-02, -2.84954570e-02,  8.44185425e-02, -2.4765699
9e-02,
        -5.30095734e-02, -2.50755140e-02,  1.48308931e-02, -8.6454199
1e-02,
        -2.89538332e-02, -9.05582890e-02,  8.01440288e-02, -1.8123613
9e-02,
        -2.28241684e-02,  2.92958882e-02, -8.88554189e-02, -6.8000589
1e-02,
        -7.51349604e-02, -4.16578529e-02, -7.53319021e-03,  4.7003241
2e-02,
         1.03828027e-01, -5.00640501e-02, -7.34445268e-02, -3.6720579
4e-02,
        -4.79209924e-02, -9.52058809e-02,  4.81136847e-02,  2.0077747
4e-02,
        -3.43602884e-03, -2.75044800e-02, -8.18629921e-02, -3.1468695
2e-02,
         5.05434394e-02, -6.31529410e-02,  3.06134730e-02,  2.1659933
6e-02,
         5.28917455e-02,  2.72663410e-02,  2.86331937e-02, -1.5595035
4e-02,
         4.16456857e-03,  6.28766710e-03, -1.96849027e-03,  3.2476955
8e-02,
         7.65568052e-02, -1.19039737e-03,  6.05962178e-02,  1.6243121
8e-02,
        -5.73209533e-02, -5.98052757e-03, -2.66848381e-02,  6.5410436
4e-02,
         1.10750260e-01, -4.98547778e-02, -5.95414548e-02,  4.2336627
6e-03,
        -4.56999248e-03, -2.10339651e-03,  1.15688143e-01, -2.5795106
0e-02,
        -1.21769063e-01,  5.59212175e-02, -6.42664748e-02,  2.2410501
0e-02,
         3.55095765e-02,  1.85836103e-02, -7.71755535e-02,  5.4757760
7e-02,
        -5.64399556e-02,  1.25472815e-02,  4.73831904e-02,  5.3961010
0e-02,
         8.09394126e-03, -4.45343608e-02,  4.04076181e-02,  1.8861880
9e-03,
         3.49743790e-03, -1.47724942e-02,  5.79980115e-02, -3.0807431
9e-02,
         5.00036306e-02,  1.35764345e-02,  1.52491129e-02, -3.3508232
2e-02,
        -4.80903336e-02,  3.29257034e-02,  6.45979581e-02, -6.4279498
8e-02,
         6.12040851e-03, -1.12536650e-01,  5.59143041e-02,  5.1324418
6e-02,
         2.51608981e-02, -1.07577571e-01,  5.42231490e-02,  8.4150181
5e-02,
         6.00497451e-02, -4.13553702e-02, -1.14999852e-02, -7.8732717
6e-02,
        -7.14465793e-02, -7.43784305e-02,  7.39431333e-02, -6.1815588
9e-02,
        -9.83908507e-03, -1.09549158e-02,  1.53343238e-02,  1.3255348
2e-02,
        -4.85962171e-02, -2.51257113e-02, -1.07426685e-01,  7.1947197
7e-02,
        -1.09830570e-01, -4.13533681e-02, -9.44798607e-02,  8.9254474
7e-02,
        -4.07701650e-02,  7.50666822e-03,  2.07753486e-02,  1.0819984
1e-01,
```

```
        -2.27786469e-02,  2.68488198e-02,  6.49099528e-02, -8.0006273
1e-02,
         1.25897158e-02,  5.03650786e-02,  7.19706704e-02, -1.3705830
8e-02,
         6.99947245e-02, -1.62304185e-02,  1.14290488e-01, -1.5905206
1e-02,
        -2.13328624e-02, -8.49158600e-02, -1.50532624e-03, -2.1788668
2e-02,
        -3.43258364e-02,  5.16449212e-02,  5.49266274e-02,  2.9988891
3e-02,
        -5.38857925e-02,  6.17925014e-02,  6.83476099e-02,  5.5455584
8e-02,
         8.73546982e-02, -2.37062424e-02, -1.40647141e-02, -1.0641354
4e-01,
        -5.39141071e-02, -3.56814369e-02, -1.07482799e-01,  2.2183890
4e-02,
        -6.14145386e-03, -7.60363625e-02,  1.74184539e-03, -8.2388111
6e-02,
         1.56537835e-02, -8.80889696e-03, -1.13468891e-01, -1.0054183
5e-01,
         6.55082564e-02,  3.24151194e-02, -8.83578373e-02, -3.2835205
1e-02,
        -4.57187398e-02, -1.94893242e-02, -3.80555925e-02, -1.2367087
6e-02,
         8.62890250e-03, -1.50413678e-02, -3.82473940e-02, -4.1833532
1e-02,
         3.13958102e-02,  1.38209840e-01, -6.45719768e-02,  1.6321967
2e-04])}
```

In [12]:
```
# Try to predict the label of unknown sentences

predictions = {}

selected_idx = [j for i in selected_sentences.values() for j in i]

for idx, row in df.iterrows():
    if idx in selected_idx:
        max_sim = 0
        winner = 'Unknown'
        for j in category_vecs:
            sim = cosine_similarity(row["vector"].reshape(1, -1), cat
egory_vecs[j].reshape(1, -1)).flatten()[0]
            if sim > max_sim:
                max_sim = sim
                winner = j
        predictions[idx] = winner
```

In [74]:
```
def get_accuracy_score(predictions, truth_dict):
    preds = []
    labels = []
    mis_classified = []
    mis_pred = []

    for k,v in predictions.items():
        preds.append(v)
        labels.append(truth_dict[k])
        if v!=truth_dict[k]:
            print(str(v) + '--x--' + str(truth_dict[k]))
            mis_pred.append(str(v))
            mis_classified.append(k)

    return f1_score(labels, preds, average='weighted'), mis_classifie
d, mis_pred


score, miss_classified_df, miss_pred = get_accuracy_score(predictions
, ground_truth)
```

```
drug_disease--x--drug_gene
drug_disease--x--drug_gene
drug_disease--x--drug_gene
drug_disease--x--drug_gene
drug_disease--x--drug_gene
drug_disease--x--drug_gene
drug_disease--x--drug_gene
drug_disease--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_disease--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
drug_disease--x--gene_disease
drug_disease--x--gene_disease
drug_gene--x--gene_disease
drug_gene--x--gene_disease
gene_disease--x--drug_disease
gene_disease--x--drug_disease
drug_gene--x--drug_disease
drug_gene--x--drug_disease
drug_gene--x--drug_disease
drug_gene--x--drug_disease
drug_gene--x--drug_disease
drug_gene--x--drug_disease
```

In [14]:
```
miss_calssified_df = df.iloc[miss_classified , [12,13]]
miss_calssified_df['Predicted-CLASS'] = miss_pred
miss_calssified_df
```

Out[14]:

	SENTENCE	CLASS	Predicted-CLASS
11	The expression of ABCG2 may explain in part th...	drug_gene	drug_disease
14	Oral administration of GSK221149A (5 mg/kg) wa...	drug_gene	drug_disease
15	Intravenous administration of GSK221149A produ...	drug_gene	drug_disease
189	Decreased plasma soluble RAGE in patients with...	drug_gene	drug_disease
242	A continuous epidural infusion of ropivacaine ...	drug_gene	drug_disease
244	Patients received 10 mg debrisoquine (a marker...	drug_gene	drug_disease
246	A continuous epidural infusion of ropivacaine ...	drug_gene	drug_disease
264	The prognostic value of the T393C SNP was eval...	gene_disease	drug_disease
265	However, the majority of colon cancer cells ha...	gene_disease	drug_gene
266	Lysophosphatidic acid facilitates proliferatio...	gene_disease	drug_gene
267	A recent study showed that LPA-mediated prolif...	gene_disease	drug_gene
269	The A6986G polymorphism of CYP3A5, which is in...	gene_disease	drug_disease
301	The results of a phase I clinical trial of the...	gene_disease	drug_gene
354	Regulation of hairy and enhancer of split homo...	gene_disease	drug_gene
356	Regulation of hairy and enhancer of split homo...	gene_disease	drug_gene
412	An anti-inflammatory and anticarcinogenic pote...	gene_disease	drug_gene
413	An anti-inflammatory and anticarcinogenic pote...	gene_disease	drug_gene
415	In contrast, enhanced COX-2 expression is cons...	gene_disease	drug_gene
416	An anti-inflammatory and anticarcinogenic pote...	gene_disease	drug_gene
417	An anti-inflammatory and anticarcinogenic pote...	gene_disease	drug_gene
511	In CD patients, TLR-induced GM-CSF secretion w...	gene_disease	drug_gene
514	Toll-like receptor-induced granulocyte-macroph...	gene_disease	drug_gene
515	Toll-like receptor-induced granulocyte-macroph...	gene_disease	drug_gene
516	Moreover, TNF-alpha production was induced by ...	gene_disease	drug_gene
517	CD patients have impaired GM-CSF secretion via...	gene_disease	drug_gene
518	CD patients have impaired GM-CSF secretion via...	gene_disease	drug_gene
521	CD patients with NOD2 mutations were able to s...	gene_disease	drug_gene
523	CD patients have impaired GM-CSF secretion via...	gene_disease	drug_gene
526	In CD patients, TLR-induced GM-CSF secretion w...	gene_disease	drug_gene
527	Toll-like receptor-induced granulocyte-macroph...	gene_disease	drug_gene
528	Moreover, TNF-alpha production was induced by ...	gene_disease	drug_gene
529	In CD patients, TLR-induced GM-CSF secretion w...	gene_disease	drug_gene
555	Vascular endothelial growth factor (VEGF) prom...	gene_disease	drug_gene
556	Vascular endothelial growth factor (VEGF) prom...	gene_disease	drug_gene

	SENTENCE	CLASS	Predicted-CLASS
591	However, IL-22R1 was only expressed in 55% of ...	gene_disease	drug_disease
595	Failure of medical and surgical therapy in CRS...	gene_disease	drug_disease
596	Fetuin-A is a calcification inhibitor and corr...	gene_disease	drug_gene
599	EGFR gene expression in pancreatic adenocarcin...	gene_disease	drug_gene
612	ET-1 mRNA expression was significantly higher ...	drug_disease	gene_disease
613	ET-1 mRNA expression was significantly higher ...	drug_disease	gene_disease
614	These results suggest that CsA can modulate th...	drug_disease	drug_gene
634	Another important finding was that the cytotox...	drug_disease	drug_gene
641	Another important finding was that the cytotox...	drug_disease	drug_gene
644	Indomethacin and indomethacin ethyl ester asso...	drug_disease	drug_gene
645	Indomethacin and indomethacin ethyl ester asso...	drug_disease	drug_gene
774	Absence of villin predisposes mice to dextran ...	drug_disease	drug_gene

In [15]:
```
result_path = os.path.abspath(os.path.join(os.path.dirname( '__file_
_' ), '..', 'Results'))+'/'
miss_calssified_df.to_csv(result_path+'miss_predictions_train.csv')
```

In [16]:
```
# Generalisation on the Unseen Dataset GAD


doc_t_d = codecs.open(data_path+'GAD_Corpus_IBIgroup/'+'GAD_Y_N'+'.cs
v','rU','UTF-8',errors='ignore')
GAD_target_disease = pd.read_csv(doc_t_d, sep='\t', na_filter = False
)
GAD_target_disease['CLASS'] = 'gene_disease'


GAD_target_disease
```

Out[16]:

	GAD_ID	GAD_ASSOC	GAD_GENE_SYMBOL	GAD_GENE_NAME	GAD_ENT
0	125111	N	HIF1A	Hypoxia-inducible factor 1, alpha subunit (bas...	3091
1	125062	N	HFE	Hemochromatosis	3077
2	125055	Y	HFE	Hemochromatosis	3077
3	125019	N	HFE	Hemochromatosis	3077
4	125015	Y	HFE	Hemochromatosis	3077
5	125009	N	HEXB	Hexosaminidase B (beta polypeptide)	3074
6	124975	Y	HCRTR2	Hypocretin (orexin) receptor 2	3062
7	124967	Y	HCCA2	HCCA2 protein	81532
8	124920	Y	HAVCR1	Hepatitis A virus cellular receptor 1	26762
9	124896	Y	GYS1	Glycogen synthase 1 (muscle)	2997
10	124890	N	GUCA1B	Guanylate cyclase activator 1B (retina)	2979
11	124828	Y	GSTT1	Glutathione S-transferase theta 1	2952
12	124820	Y	GSTT1	Glutathione S-transferase theta 1	2952
13	124814	Y	GSTT1	Glutathione S-transferase theta 1	2952

	GAD_ID	GAD_ASSOC	GAD_GENE_SYMBOL	GAD_GENE_NAME	GAD_ENT
14	124798	Y	GSTT1	Glutathione S-transferase theta 1	2952
15	124792	N	GSTT1	Glutathione S-transferase theta 1	2952
16	124778	Y	GSTT1	Glutathione S-transferase theta 1	2952
17	124778	Y	GSTT1	Glutathione S-transferase theta 1	2952
18	124776	N	GSTT1	Glutathione S-transferase theta 1	2952
19	124772	Y	GSTT1	Glutathione S-transferase theta 1	2952
20	124744	Y	GSTT1	Glutathione S-transferase theta 1	2952
21	124745	Y	GSTT1	Glutathione S-transferase theta 1	2952
22	124745	Y	GSTT1	Glutathione S-transferase theta 1	2952
23	124627	N	GSTP1	Glutathione S-transferase pi	2950
24	124626	N	GSTP1	Glutathione S-transferase pi	2950
25	124459	N	GSTM1	Glutathione S-transferase M1	2944
26	124454	Y	GSTM1	Glutathione S-transferase M1	2944
27	124452	Y	GSTM1	Glutathione S-transferase M1	2944

	GAD_ID	GAD_ASSOC	GAD_GENE_SYMBOL	GAD_GENE_NAME	GAD_ENT
28	124431	Y	GSTM1	Glutathione S-transferase M1	2944
29	124424	Y	GSTM1	Glutathione S-transferase M1	2944
...	...	...	...	...	...
2771	564629	Y	ADAM33	ADAM metallopeptidase domain 33	80332
2772	563212	N	XRCC1	X-ray repair complementing defective repair in...	7515
2773	560042	N	TGFB1	transforming growth factor, beta 1	7040
2774	559322	Y	STIP1	stress-induced-phosphoprotein 1	10963
2775	559277	N	STAT4	signal transducer and activator of transcripti...	6775
2776	558924	N	SOD2	superoxide dismutase 2, mitochondrial	6648
2777	558166	N	BRCA2	breast cancer 2, early onset	675
2778	557994	N	BRCA1	breast cancer 1, early onset	672
2779	557423	N	SHMT1	serine hydroxymethyltransferase 1 (soluble)	6470
2780	556700	N	BIK	BCL2-interacting killer (apoptosis-inducing)	638
2781	556671	N	RYR3	ryanodine receptor 3	6263
2782	707750	N	OGG1	8-oxoguanine DNA glycosylase	4968

	GAD_ID	GAD_ASSOC	GAD_GENE_SYMBOL	GAD_GENE_NAME	GAD_ENT
2783	705444	Y	NAT2	N-acetyltransferase 2 (arylamine N-acetyltrans...	10
2784	704807	N	MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH)	4524
2785	703633	N	LTC4S	leukotriene C4 synthase	4056
2786	692713	N	HSD3B1	hydroxy-delta-5-steroid dehydrogenase, 3 beta-...	3283
2787	691009	N	ALOX5AP	arachidonate 5-lipoxygenase-activating protein	241
2788	690996	N	ALOX5	arachidonate 5-lipoxygenase	240
2789	684878	Y	FCER2	Fc fragment of IgE, low affinity II, receptor ...	2208
2790	683138	Y	KIF21B		23046
2791	680619	N	CREG1	cellular repressor of E1A-stimulated genes 1	8804
2792	679932	N	KIF6	kinesin family member 6	221458
2793	679380	N	ADRB3	adrenergic, beta-3-, receptor	155
2794	678519	Y	ACE	angiotensin I converting enzyme (peptidyl-dipe...	1636
2795	677472	N	COMT	catechol-O-methyltransferase	1312
2796	675012	Y	CD14	CD14 molecule	929
2797	667271	Y	UGT2B17	UDP glucuronosyltransferase 2 family, polypept...	7367

	GAD_ID	GAD_ASSOC	GAD_GENE_SYMBOL	GAD_GENE_NAME	GAD_ENT
2798	667268	N	UGT2B17	UDP glucuronosyltransferase 2 family, polypept...	7367
2799	667058	N	TP53	tumor protein p53	7157
2800	666239	N	TNF	tumor necrosis factor (TNF superfamily, member 2)	7124

2801 rows × 12 columns

In [17]:
```
doc_embeddings = sif_embedding_wrapper.sentences2vecs(GAD_target_dise
ase["GAD_CONCLUSION"], embs, words, weight4ind)
GAD_target_disease["vector"] = pd.Series(list(doc_embeddings))
```

In [18]:
```
test_ground_truth = {}

for idx, row in GAD_target_disease.iterrows():
    test_ground_truth[idx] = row['CLASS']

# Try to predict the label of unknown sentences

test_predictions = {}

for idx, row in GAD_target_disease.iterrows():
    max_sim = 0.60
    winner = 'unknown'
    for j in category_vecs:
        sim = cosine_similarity(row["vector"].reshape(1, -1), categor
y_vecs[j].reshape(1, -1)).flatten()[0]
        if sim > max_sim:
            max_sim = sim
            winner = j
    test_predictions[idx] = winner
```

In [19]:
```
test_predictions
```

```
Out[19]: {0: 'gene_disease',
          1: 'gene_disease',
          2: 'gene_disease',
          3: 'gene_disease',
          4: 'gene_disease',
          5: 'gene_disease',
          6: 'gene_disease',
          7: 'drug_gene',
          8: 'gene_disease',
          9: 'gene_disease',
          10: 'unknown',
          11: 'gene_disease',
          12: 'gene_disease',
          13: 'gene_disease',
          14: 'gene_disease',
          15: 'gene_disease',
          16: 'gene_disease',
          17: 'gene_disease',
          18: 'gene_disease',
          19: 'gene_disease',
          20: 'gene_disease',
          21: 'gene_disease',
          22: 'gene_disease',
          23: 'gene_disease',
          24: 'gene_disease',
          25: 'gene_disease',
          26: 'gene_disease',
          27: 'gene_disease',
          28: 'gene_disease',
          29: 'gene_disease',
          30: 'gene_disease',
          31: 'gene_disease',
          32: 'gene_disease',
          33: 'gene_disease',
          34: 'gene_disease',
          35: 'gene_disease',
          36: 'gene_disease',
          37: 'gene_disease',
          38: 'gene_disease',
          39: 'unknown',
          40: 'gene_disease',
          41: 'gene_disease',
          42: 'gene_disease',
          43: 'gene_disease',
          44: 'gene_disease',
          45: 'gene_disease',
          46: 'gene_disease',
          47: 'gene_disease',
          48: 'gene_disease',
          49: 'gene_disease',
          50: 'gene_disease',
          51: 'gene_disease',
          52: 'unknown',
          53: 'unknown',
          54: 'unknown',
          55: 'unknown',
          56: 'gene_disease',
          57: 'gene_disease',
          58: 'gene_disease',
          59: 'gene_disease',
          60: 'unknown',
          61: 'gene_disease',
          62: 'unknown',
```

```
 63: 'gene_disease',
 64: 'gene_disease',
 65: 'gene_disease',
 66: 'gene_disease',
 67: 'gene_disease',
 68: 'gene_disease',
 69: 'gene_disease',
 70: 'gene_disease',
 71: 'gene_disease',
 72: 'gene_disease',
 73: 'gene_disease',
 74: 'gene_disease',
 75: 'gene_disease',
 76: 'gene_disease',
 77: 'gene_disease',
 78: 'gene_disease',
 79: 'gene_disease',
 80: 'gene_disease',
 81: 'unknown',
 82: 'gene_disease',
 83: 'gene_disease',
 84: 'gene_disease',
 85: 'gene_disease',
 86: 'gene_disease',
 87: 'drug_gene',
 88: 'gene_disease',
 89: 'gene_disease',
 90: 'gene_disease',
 91: 'gene_disease',
 92: 'gene_disease',
 93: 'gene_disease',
 94: 'gene_disease',
 95: 'gene_disease',
 96: 'gene_disease',
 97: 'gene_disease',
 98: 'gene_disease',
 99: 'gene_disease',
100: 'gene_disease',
101: 'gene_disease',
102: 'gene_disease',
103: 'gene_disease',
104: 'gene_disease',
105: 'gene_disease',
106: 'gene_disease',
107: 'gene_disease',
108: 'gene_disease',
109: 'gene_disease',
110: 'gene_disease',
111: 'gene_disease',
112: 'gene_disease',
113: 'gene_disease',
114: 'unknown',
115: 'gene_disease',
116: 'gene_disease',
117: 'gene_disease',
118: 'gene_disease',
119: 'gene_disease',
120: 'gene_disease',
121: 'gene_disease',
122: 'gene_disease',
123: 'gene_disease',
124: 'gene_disease',
125: 'gene_disease',
```

```
126: 'gene_disease',
127: 'gene_disease',
128: 'gene_disease',
129: 'gene_disease',
130: 'gene_disease',
131: 'gene_disease',
132: 'gene_disease',
133: 'gene_disease',
134: 'gene_disease',
135: 'gene_disease',
136: 'gene_disease',
137: 'gene_disease',
138: 'gene_disease',
139: 'gene_disease',
140: 'gene_disease',
141: 'gene_disease',
142: 'gene_disease',
143: 'gene_disease',
144: 'gene_disease',
145: 'gene_disease',
146: 'gene_disease',
147: 'gene_disease',
148: 'unknown',
149: 'unknown',
150: 'gene_disease',
151: 'gene_disease',
152: 'gene_disease',
153: 'gene_disease',
154: 'gene_disease',
155: 'gene_disease',
156: 'gene_disease',
157: 'gene_disease',
158: 'gene_disease',
159: 'gene_disease',
160: 'gene_disease',
161: 'gene_disease',
162: 'gene_disease',
163: 'gene_disease',
164: 'gene_disease',
165: 'gene_disease',
166: 'gene_disease',
167: 'gene_disease',
168: 'gene_disease',
169: 'gene_disease',
170: 'gene_disease',
171: 'gene_disease',
172: 'gene_disease',
173: 'gene_disease',
174: 'gene_disease',
175: 'gene_disease',
176: 'gene_disease',
177: 'gene_disease',
178: 'gene_disease',
179: 'unknown',
180: 'gene_disease',
181: 'gene_disease',
182: 'gene_disease',
183: 'drug_gene',
184: 'drug_gene',
185: 'unknown',
186: 'unknown',
187: 'gene_disease',
188: 'gene_disease',
```

```
189: 'gene_disease',
190: 'gene_disease',
191: 'gene_disease',
192: 'gene_disease',
193: 'gene_disease',
194: 'gene_disease',
195: 'gene_disease',
196: 'gene_disease',
197: 'gene_disease',
198: 'gene_disease',
199: 'gene_disease',
200: 'unknown',
201: 'unknown',
202: 'unknown',
203: 'unknown',
204: 'gene_disease',
205: 'gene_disease',
206: 'gene_disease',
207: 'gene_disease',
208: 'gene_disease',
209: 'gene_disease',
210: 'gene_disease',
211: 'gene_disease',
212: 'gene_disease',
213: 'gene_disease',
214: 'gene_disease',
215: 'gene_disease',
216: 'gene_disease',
217: 'gene_disease',
218: 'gene_disease',
219: 'gene_disease',
220: 'gene_disease',
221: 'gene_disease',
222: 'gene_disease',
223: 'gene_disease',
224: 'gene_disease',
225: 'gene_disease',
226: 'gene_disease',
227: 'gene_disease',
228: 'gene_disease',
229: 'gene_disease',
230: 'gene_disease',
231: 'gene_disease',
232: 'gene_disease',
233: 'gene_disease',
234: 'gene_disease',
235: 'gene_disease',
236: 'unknown',
237: 'gene_disease',
238: 'gene_disease',
239: 'gene_disease',
240: 'gene_disease',
241: 'gene_disease',
242: 'gene_disease',
243: 'gene_disease',
244: 'gene_disease',
245: 'gene_disease',
246: 'gene_disease',
247: 'gene_disease',
248: 'gene_disease',
249: 'gene_disease',
250: 'gene_disease',
251: 'gene_disease',
```

```
252: 'gene_disease',
253: 'unknown',
254: 'gene_disease',
255: 'gene_disease',
256: 'gene_disease',
257: 'gene_disease',
258: 'unknown',
259: 'gene_disease',
260: 'gene_disease',
261: 'gene_disease',
262: 'gene_disease',
263: 'gene_disease',
264: 'gene_disease',
265: 'gene_disease',
266: 'gene_disease',
267: 'gene_disease',
268: 'gene_disease',
269: 'gene_disease',
270: 'unknown',
271: 'unknown',
272: 'gene_disease',
273: 'gene_disease',
274: 'gene_disease',
275: 'gene_disease',
276: 'gene_disease',
277: 'gene_disease',
278: 'unknown',
279: 'unknown',
280: 'gene_disease',
281: 'gene_disease',
282: 'gene_disease',
283: 'gene_disease',
284: 'gene_disease',
285: 'gene_disease',
286: 'gene_disease',
287: 'gene_disease',
288: 'gene_disease',
289: 'gene_disease',
290: 'gene_disease',
291: 'gene_disease',
292: 'gene_disease',
293: 'gene_disease',
294: 'gene_disease',
295: 'gene_disease',
296: 'gene_disease',
297: 'gene_disease',
298: 'gene_disease',
299: 'gene_disease',
300: 'gene_disease',
301: 'gene_disease',
302: 'gene_disease',
303: 'gene_disease',
304: 'gene_disease',
305: 'gene_disease',
306: 'gene_disease',
307: 'gene_disease',
308: 'gene_disease',
309: 'gene_disease',
310: 'gene_disease',
311: 'gene_disease',
312: 'gene_disease',
313: 'gene_disease',
314: 'gene_disease',
```

```
315: 'gene_disease',
316: 'gene_disease',
317: 'gene_disease',
318: 'gene_disease',
319: 'gene_disease',
320: 'gene_disease',
321: 'gene_disease',
322: 'gene_disease',
323: 'gene_disease',
324: 'gene_disease',
325: 'gene_disease',
326: 'gene_disease',
327: 'gene_disease',
328: 'gene_disease',
329: 'gene_disease',
330: 'gene_disease',
331: 'gene_disease',
332: 'gene_disease',
333: 'gene_disease',
334: 'gene_disease',
335: 'gene_disease',
336: 'gene_disease',
337: 'gene_disease',
338: 'gene_disease',
339: 'gene_disease',
340: 'gene_disease',
341: 'gene_disease',
342: 'gene_disease',
343: 'gene_disease',
344: 'gene_disease',
345: 'gene_disease',
346: 'gene_disease',
347: 'gene_disease',
348: 'gene_disease',
349: 'gene_disease',
350: 'gene_disease',
351: 'gene_disease',
352: 'gene_disease',
353: 'gene_disease',
354: 'gene_disease',
355: 'gene_disease',
356: 'gene_disease',
357: 'gene_disease',
358: 'gene_disease',
359: 'gene_disease',
360: 'gene_disease',
361: 'gene_disease',
362: 'gene_disease',
363: 'gene_disease',
364: 'unknown',
365: 'gene_disease',
366: 'gene_disease',
367: 'gene_disease',
368: 'gene_disease',
369: 'gene_disease',
370: 'gene_disease',
371: 'gene_disease',
372: 'gene_disease',
373: 'gene_disease',
374: 'gene_disease',
375: 'gene_disease',
376: 'gene_disease',
377: 'gene_disease',
```

```
 378: 'gene_disease',
 379: 'gene_disease',
 380: 'gene_disease',
 381: 'gene_disease',
 382: 'gene_disease',
 383: 'gene_disease',
 384: 'gene_disease',
 385: 'gene_disease',
 386: 'gene_disease',
 387: 'gene_disease',
 388: 'gene_disease',
 389: 'gene_disease',
 390: 'gene_disease',
 391: 'gene_disease',
 392: 'gene_disease',
 393: 'gene_disease',
 394: 'gene_disease',
 395: 'gene_disease',
 396: 'gene_disease',
 397: 'gene_disease',
 398: 'gene_disease',
 399: 'drug_gene',
 400: 'gene_disease',
 401: 'gene_disease',
 402: 'gene_disease',
 403: 'gene_disease',
 404: 'gene_disease',
 405: 'gene_disease',
 406: 'unknown',
 407: 'gene_disease',
 408: 'gene_disease',
 409: 'gene_disease',
 410: 'gene_disease',
 411: 'gene_disease',
 412: 'gene_disease',
 413: 'gene_disease',
 414: 'gene_disease',
 415: 'gene_disease',
 416: 'gene_disease',
 417: 'gene_disease',
 418: 'gene_disease',
 419: 'gene_disease',
 420: 'gene_disease',
 421: 'gene_disease',
 422: 'gene_disease',
 423: 'gene_disease',
 424: 'gene_disease',
 425: 'gene_disease',
 426: 'gene_disease',
 427: 'gene_disease',
 428: 'gene_disease',
 429: 'gene_disease',
 430: 'gene_disease',
 431: 'gene_disease',
 432: 'gene_disease',
 433: 'gene_disease',
 434: 'gene_disease',
 435: 'gene_disease',
 436: 'gene_disease',
 437: 'gene_disease',
 438: 'gene_disease',
 439: 'gene_disease',
 440: 'gene_disease',
```

```
441: 'gene_disease',
442: 'drug_gene',
443: 'gene_disease',
444: 'gene_disease',
445: 'gene_disease',
446: 'gene_disease',
447: 'gene_disease',
448: 'gene_disease',
449: 'gene_disease',
450: 'gene_disease',
451: 'gene_disease',
452: 'gene_disease',
453: 'gene_disease',
454: 'gene_disease',
455: 'gene_disease',
456: 'gene_disease',
457: 'gene_disease',
458: 'gene_disease',
459: 'gene_disease',
460: 'gene_disease',
461: 'gene_disease',
462: 'gene_disease',
463: 'gene_disease',
464: 'gene_disease',
465: 'gene_disease',
466: 'unknown',
467: 'gene_disease',
468: 'gene_disease',
469: 'gene_disease',
470: 'gene_disease',
471: 'gene_disease',
472: 'gene_disease',
473: 'gene_disease',
474: 'gene_disease',
475: 'gene_disease',
476: 'gene_disease',
477: 'gene_disease',
478: 'gene_disease',
479: 'gene_disease',
480: 'gene_disease',
481: 'gene_disease',
482: 'gene_disease',
483: 'gene_disease',
484: 'gene_disease',
485: 'gene_disease',
486: 'drug_gene',
487: 'gene_disease',
488: 'gene_disease',
489: 'gene_disease',
490: 'gene_disease',
491: 'gene_disease',
492: 'gene_disease',
493: 'gene_disease',
494: 'gene_disease',
495: 'gene_disease',
496: 'gene_disease',
497: 'gene_disease',
498: 'gene_disease',
499: 'gene_disease',
500: 'gene_disease',
501: 'drug_gene',
502: 'gene_disease',
503: 'unknown',
```

```
504: 'gene_disease',
505: 'gene_disease',
506: 'gene_disease',
507: 'gene_disease',
508: 'gene_disease',
509: 'gene_disease',
510: 'gene_disease',
511: 'gene_disease',
512: 'gene_disease',
513: 'gene_disease',
514: 'gene_disease',
515: 'gene_disease',
516: 'gene_disease',
517: 'unknown',
518: 'gene_disease',
519: 'gene_disease',
520: 'gene_disease',
521: 'gene_disease',
522: 'gene_disease',
523: 'gene_disease',
524: 'gene_disease',
525: 'gene_disease',
526: 'gene_disease',
527: 'gene_disease',
528: 'gene_disease',
529: 'unknown',
530: 'gene_disease',
531: 'unknown',
532: 'gene_disease',
533: 'gene_disease',
534: 'gene_disease',
535: 'unknown',
536: 'gene_disease',
537: 'gene_disease',
538: 'gene_disease',
539: 'gene_disease',
540: 'gene_disease',
541: 'gene_disease',
542: 'gene_disease',
543: 'gene_disease',
544: 'gene_disease',
545: 'gene_disease',
546: 'gene_disease',
547: 'gene_disease',
548: 'gene_disease',
549: 'gene_disease',
550: 'gene_disease',
551: 'unknown',
552: 'unknown',
553: 'unknown',
554: 'gene_disease',
555: 'unknown',
556: 'gene_disease',
557: 'unknown',
558: 'gene_disease',
559: 'gene_disease',
560: 'gene_disease',
561: 'gene_disease',
562: 'gene_disease',
563: 'gene_disease',
564: 'gene_disease',
565: 'gene_disease',
566: 'gene_disease',
```

```
567: 'unknown',
568: 'gene_disease',
569: 'gene_disease',
570: 'gene_disease',
571: 'gene_disease',
572: 'gene_disease',
573: 'gene_disease',
574: 'gene_disease',
575: 'unknown',
576: 'unknown',
577: 'gene_disease',
578: 'unknown',
579: 'gene_disease',
580: 'gene_disease',
581: 'gene_disease',
582: 'gene_disease',
583: 'gene_disease',
584: 'gene_disease',
585: 'gene_disease',
586: 'unknown',
587: 'gene_disease',
588: 'gene_disease',
589: 'gene_disease',
590: 'gene_disease',
591: 'gene_disease',
592: 'gene_disease',
593: 'gene_disease',
594: 'gene_disease',
595: 'gene_disease',
596: 'gene_disease',
597: 'gene_disease',
598: 'gene_disease',
599: 'unknown',
600: 'unknown',
601: 'unknown',
602: 'unknown',
603: 'unknown',
604: 'gene_disease',
605: 'gene_disease',
606: 'gene_disease',
607: 'gene_disease',
608: 'gene_disease',
609: 'gene_disease',
610: 'gene_disease',
611: 'gene_disease',
612: 'gene_disease',
613: 'gene_disease',
614: 'gene_disease',
615: 'gene_disease',
616: 'gene_disease',
617: 'gene_disease',
618: 'gene_disease',
619: 'gene_disease',
620: 'gene_disease',
621: 'unknown',
622: 'gene_disease',
623: 'gene_disease',
624: 'gene_disease',
625: 'gene_disease',
626: 'gene_disease',
627: 'gene_disease',
628: 'gene_disease',
629: 'gene_disease',
```

```
630: 'gene_disease',
631: 'gene_disease',
632: 'gene_disease',
633: 'gene_disease',
634: 'gene_disease',
635: 'gene_disease',
636: 'gene_disease',
637: 'unknown',
638: 'gene_disease',
639: 'gene_disease',
640: 'gene_disease',
641: 'gene_disease',
642: 'gene_disease',
643: 'gene_disease',
644: 'gene_disease',
645: 'gene_disease',
646: 'gene_disease',
647: 'unknown',
648: 'gene_disease',
649: 'gene_disease',
650: 'gene_disease',
651: 'gene_disease',
652: 'gene_disease',
653: 'gene_disease',
654: 'gene_disease',
655: 'gene_disease',
656: 'gene_disease',
657: 'gene_disease',
658: 'gene_disease',
659: 'gene_disease',
660: 'gene_disease',
661: 'gene_disease',
662: 'gene_disease',
663: 'gene_disease',
664: 'gene_disease',
665: 'gene_disease',
666: 'gene_disease',
667: 'gene_disease',
668: 'gene_disease',
669: 'gene_disease',
670: 'gene_disease',
671: 'gene_disease',
672: 'gene_disease',
673: 'gene_disease',
674: 'gene_disease',
675: 'gene_disease',
676: 'gene_disease',
677: 'gene_disease',
678: 'unknown',
679: 'unknown',
680: 'gene_disease',
681: 'gene_disease',
682: 'gene_disease',
683: 'gene_disease',
684: 'gene_disease',
685: 'gene_disease',
686: 'gene_disease',
687: 'gene_disease',
688: 'unknown',
689: 'unknown',
690: 'gene_disease',
691: 'gene_disease',
692: 'gene_disease',
```

```
693: 'gene_disease',
694: 'unknown',
695: 'gene_disease',
696: 'gene_disease',
697: 'gene_disease',
698: 'gene_disease',
699: 'gene_disease',
700: 'gene_disease',
701: 'gene_disease',
702: 'gene_disease',
703: 'gene_disease',
704: 'gene_disease',
705: 'unknown',
706: 'gene_disease',
707: 'gene_disease',
708: 'gene_disease',
709: 'gene_disease',
710: 'gene_disease',
711: 'gene_disease',
712: 'gene_disease',
713: 'gene_disease',
714: 'gene_disease',
715: 'gene_disease',
716: 'gene_disease',
717: 'gene_disease',
718: 'unknown',
719: 'gene_disease',
720: 'gene_disease',
721: 'gene_disease',
722: 'gene_disease',
723: 'gene_disease',
724: 'gene_disease',
725: 'gene_disease',
726: 'gene_disease',
727: 'gene_disease',
728: 'gene_disease',
729: 'gene_disease',
730: 'gene_disease',
731: 'unknown',
732: 'gene_disease',
733: 'gene_disease',
734: 'gene_disease',
735: 'gene_disease',
736: 'gene_disease',
737: 'gene_disease',
738: 'unknown',
739: 'gene_disease',
740: 'gene_disease',
741: 'unknown',
742: 'gene_disease',
743: 'gene_disease',
744: 'unknown',
745: 'gene_disease',
746: 'gene_disease',
747: 'gene_disease',
748: 'gene_disease',
749: 'gene_disease',
750: 'gene_disease',
751: 'gene_disease',
752: 'gene_disease',
753: 'gene_disease',
754: 'gene_disease',
755: 'gene_disease',
```

```
756: 'unknown',
757: 'gene_disease',
758: 'gene_disease',
759: 'gene_disease',
760: 'gene_disease',
761: 'gene_disease',
762: 'gene_disease',
763: 'gene_disease',
764: 'gene_disease',
765: 'gene_disease',
766: 'gene_disease',
767: 'gene_disease',
768: 'gene_disease',
769: 'gene_disease',
770: 'gene_disease',
771: 'gene_disease',
772: 'gene_disease',
773: 'unknown',
774: 'gene_disease',
775: 'gene_disease',
776: 'gene_disease',
777: 'gene_disease',
778: 'gene_disease',
779: 'gene_disease',
780: 'gene_disease',
781: 'gene_disease',
782: 'gene_disease',
783: 'gene_disease',
784: 'gene_disease',
785: 'gene_disease',
786: 'gene_disease',
787: 'gene_disease',
788: 'unknown',
789: 'gene_disease',
790: 'gene_disease',
791: 'gene_disease',
792: 'gene_disease',
793: 'gene_disease',
794: 'unknown',
795: 'gene_disease',
796: 'unknown',
797: 'unknown',
798: 'gene_disease',
799: 'gene_disease',
800: 'gene_disease',
801: 'gene_disease',
802: 'gene_disease',
803: 'gene_disease',
804: 'gene_disease',
805: 'gene_disease',
806: 'gene_disease',
807: 'gene_disease',
808: 'gene_disease',
809: 'gene_disease',
810: 'gene_disease',
811: 'gene_disease',
812: 'gene_disease',
813: 'gene_disease',
814: 'gene_disease',
815: 'gene_disease',
816: 'gene_disease',
817: 'gene_disease',
818: 'gene_disease',
```

```
819: 'gene_disease',
820: 'gene_disease',
821: 'gene_disease',
822: 'gene_disease',
823: 'gene_disease',
824: 'gene_disease',
825: 'gene_disease',
826: 'gene_disease',
827: 'gene_disease',
828: 'gene_disease',
829: 'gene_disease',
830: 'gene_disease',
831: 'gene_disease',
832: 'gene_disease',
833: 'gene_disease',
834: 'gene_disease',
835: 'unknown',
836: 'gene_disease',
837: 'gene_disease',
838: 'gene_disease',
839: 'gene_disease',
840: 'gene_disease',
841: 'gene_disease',
842: 'gene_disease',
843: 'gene_disease',
844: 'gene_disease',
845: 'gene_disease',
846: 'gene_disease',
847: 'gene_disease',
848: 'gene_disease',
849: 'gene_disease',
850: 'gene_disease',
851: 'gene_disease',
852: 'gene_disease',
853: 'gene_disease',
854: 'unknown',
855: 'gene_disease',
856: 'gene_disease',
857: 'gene_disease',
858: 'unknown',
859: 'gene_disease',
860: 'gene_disease',
861: 'gene_disease',
862: 'gene_disease',
863: 'unknown',
864: 'drug_gene',
865: 'drug_gene',
866: 'gene_disease',
867: 'gene_disease',
868: 'unknown',
869: 'gene_disease',
870: 'gene_disease',
871: 'gene_disease',
872: 'gene_disease',
873: 'gene_disease',
874: 'gene_disease',
875: 'gene_disease',
876: 'gene_disease',
877: 'gene_disease',
878: 'gene_disease',
879: 'gene_disease',
880: 'gene_disease',
881: 'gene_disease',
```

```
882: 'gene_disease',
883: 'gene_disease',
884: 'gene_disease',
885: 'gene_disease',
886: 'gene_disease',
887: 'gene_disease',
888: 'gene_disease',
889: 'gene_disease',
890: 'gene_disease',
891: 'gene_disease',
892: 'gene_disease',
893: 'gene_disease',
894: 'gene_disease',
895: 'gene_disease',
896: 'gene_disease',
897: 'gene_disease',
898: 'gene_disease',
899: 'gene_disease',
900: 'gene_disease',
901: 'gene_disease',
902: 'unknown',
903: 'unknown',
904: 'gene_disease',
905: 'gene_disease',
906: 'gene_disease',
907: 'unknown',
908: 'unknown',
909: 'unknown',
910: 'unknown',
911: 'gene_disease',
912: 'gene_disease',
913: 'gene_disease',
914: 'gene_disease',
915: 'gene_disease',
916: 'gene_disease',
917: 'gene_disease',
918: 'gene_disease',
919: 'gene_disease',
920: 'gene_disease',
921: 'gene_disease',
922: 'unknown',
923: 'gene_disease',
924: 'gene_disease',
925: 'gene_disease',
926: 'gene_disease',
927: 'gene_disease',
928: 'gene_disease',
929: 'unknown',
930: 'unknown',
931: 'gene_disease',
932: 'unknown',
933: 'unknown',
934: 'gene_disease',
935: 'gene_disease',
936: 'gene_disease',
937: 'gene_disease',
938: 'gene_disease',
939: 'gene_disease',
940: 'gene_disease',
941: 'unknown',
942: 'gene_disease',
943: 'gene_disease',
944: 'gene_disease',
```

```
945: 'unknown',
946: 'gene_disease',
947: 'gene_disease',
948: 'gene_disease',
949: 'gene_disease',
950: 'gene_disease',
951: 'gene_disease',
952: 'gene_disease',
953: 'gene_disease',
954: 'gene_disease',
955: 'gene_disease',
956: 'gene_disease',
957: 'gene_disease',
958: 'gene_disease',
959: 'gene_disease',
960: 'gene_disease',
961: 'gene_disease',
962: 'gene_disease',
963: 'gene_disease',
964: 'gene_disease',
965: 'unknown',
966: 'unknown',
967: 'gene_disease',
968: 'gene_disease',
969: 'gene_disease',
970: 'gene_disease',
971: 'gene_disease',
972: 'gene_disease',
973: 'gene_disease',
974: 'drug_gene',
975: 'drug_gene',
976: 'drug_gene',
977: 'drug_gene',
978: 'drug_gene',
979: 'drug_gene',
980: 'gene_disease',
981: 'gene_disease',
982: 'gene_disease',
983: 'gene_disease',
984: 'gene_disease',
985: 'gene_disease',
986: 'gene_disease',
987: 'gene_disease',
988: 'gene_disease',
989: 'gene_disease',
990: 'gene_disease',
991: 'gene_disease',
992: 'unknown',
993: 'gene_disease',
994: 'gene_disease',
995: 'gene_disease',
996: 'gene_disease',
997: 'gene_disease',
998: 'gene_disease',
999: 'gene_disease',
...}
```

In [20]:
```
def get_accuracy_score(predictions, truth_dict):
    preds = []
    labels = []
    mis_classified = []
    mis_pred = []

    for k,v in predictions.items():
        preds.append(v)
        labels.append(truth_dict[k])
        if v!=truth_dict[k]:
#             print(str(v) + '--x--' + str(truth_dict[k]))
            mis_classified.append(k)
            mis_pred.append(str(v))

    return f1_score(labels, preds, average='weighted'), mis_classifie
d, mis_pred


score, miss_classified, miss_pred = get_accuracy_score(test_predictio
ns, test_ground_truth)
score
```

```
/home/stirunag/anaconda3/lib/python3.7/site-packages/sklearn/metrics/
classification.py:1145: UndefinedMetricWarning: F-score is ill-define
d and being set to 0.0 in labels with no true samples.
  'recall', 'true', average, warn_for)
```

Out[20]: 0.9487708763370238

In [21]:
```
miss_calssified_df_test = GAD_target_disease.iloc[miss_classified , [
10,11]]
miss_calssified_df_test['Predicted-CLASS'] = miss_pred
miss_calssified_df_test
```

Out[21]:

	GAD_CONCLUSION	CLASS	Predicted-CLASS
7	The novel gene HCCA2 may be related with the i...	gene_disease	drug_gene
10	We found no evidence that mutation in GUCA1B,G...	gene_disease	unknown
39	Because our samples provided quite high power,...	gene_disease	unknown
52	We conclude that GRIK2 does not play a major ...	gene_disease	unknown
53	We conclude that GRIK2 does not play a major ...	gene_disease	unknown
54	We conclude that GRIK1 does not play a major ...	gene_disease	unknown
55	We conclude that GRIK1 does not play a major ...	gene_disease	unknown
60	Determination of VNTR of the GPIba gene may pr...	gene_disease	unknown
62	We found no evidence that mutation in GUCA1B,G...	gene_disease	unknown
81	Our results suggest that GAD2 does not play a ...	gene_disease	unknown
87	These results suggest the GABRA3 gene may be a...	gene_disease	drug_gene
114	The results showed that the FGF1 gene is assoc...	gene_disease	unknown
148	In conclusion, there was no association betwee...	gene_disease	unknown
149	In conclusion, there was no association betwee...	gene_disease	unknown
179	In conclusion, our findings support an etiolog...	gene_disease	unknown
183	This study suggests that high EGF production m...	gene_disease	drug_gene
184	This study suggests that high EGF production m...	gene_disease	drug_gene
185	All our results indicate that the presence of ...	gene_disease	unknown
186	All our results indicate that the presence of ...	gene_disease	unknown
200	In children with ADHD, possession of the DRD4 ...	gene_disease	unknown
201	In children with ADHD, possession of the DRD4 ...	gene_disease	unknown
202	In children with ADHD, possession of the DRD4 ...	gene_disease	unknown
203	In children with ADHD, possession of the DRD4 ...	gene_disease	unknown
236	These results suggest that DBP does not contri...	gene_disease	unknown
253	CYP46 influences brain beta-amyloid load, cere...	gene_disease	unknown
258	CYP46 influences brain beta-amyloid load, cere...	gene_disease	unknown
270	Our results suggest that rate of gastric empty...	gene_disease	unknown
271	Our data suggest that deficient CYP2A6 activit...	gene_disease	unknown
278	Our results suggested that the Val CYP1B1 alle...	gene_disease	unknown
279	These results do not support a favoring role o...	gene_disease	unknown
...	...	...	...
2482	there is no main effect of APOE in our autism ...	gene_disease	unknown
2485	we could not find evidence of association betw...	gene_disease	unknown

	GAD_CONCLUSION	CLASS	Predicted-CLASS
2487	Patients with X-linked retinitis pigmentosa du...	gene_disease	unknown
2489	Patients with X-linked retinitis pigmentosa du...	gene_disease	unknown
2492	Large genomic rearrangements in SCN5A are not ...	gene_disease	unknown
2515	These findings provide initial evidence that e...	gene_disease	unknown
2516	These findings provide initial evidence that e...	gene_disease	unknown
2522	These findings provide initial support for gen...	gene_disease	unknown
2523	These findings provide initial support for gen...	gene_disease	unknown
2524	These findings provide initial support for gen...	gene_disease	unknown
2549	These findings suggest that transforming growt...	gene_disease	drug_gene
2608	These preliminary findings are suggestive of a...	gene_disease	unknown
2612	According to this study, FGFR4 Arg388 genotype...	gene_disease	drug_disease
2627	we could not find evidence of association betw...	gene_disease	unknown
2647	The His645Asp polymorphism of the histamine me...	gene_disease	unknown
2649	we could not find evidence of association betw...	gene_disease	unknown
2657	The results of the present study, which is muc...	gene_disease	unknown
2667	These findings bring further evidence on the r...	gene_disease	unknown
2672	We found evidence for association between GABR...	gene_disease	unknown
2682	this study has found that the IL-9 gene was sl...	gene_disease	unknown
2688	it is unlikely that common variants in MLH1, M...	gene_disease	unknown
2689	it is unlikely that common variants in MLH1, M...	gene_disease	unknown
2692	it is unlikely that common variants in MLH1, M...	gene_disease	unknown
2699	PPP2R1B genes may not play a role in the carci...	gene_disease	unknown
2700	The present study provides further evidence in...	gene_disease	unknown
2703	These findings provide replication of the asso...	gene_disease	unknown
2771	Our findings suggest a relevant role of ADAM33...	gene_disease	unknown
2789	FCER2 predicts the likelihood of treatment pro...	gene_disease	unknown
2793	the Trp(64)Arg mutation of ADRB3 has little or...	gene_disease	unknown
2794	The ACE deletion allele may protect against hy...	gene_disease	unknown

273 rows × 3 columns

In [22]:
```
miss_calssified_df_test.to_csv(result_path+'miss_predictions_test.csv')
```

In [76]:
```
#Final model

final_data_1 = df[['SENTENCE', 'CLASS']][~df.index.isin(miss_classifi
ed_df)]
final_data_2 = GAD_target_disease[['GAD_CONCLUSION', 'CLASS']][~GAD_t
arget_disease.index.isin(miss_classified)]

final_data_2.rename(columns={"GAD_CONCLUSION": "SENTENCE"}, inplace =
True)

final_data = final_data_1.append(final_data_2)

final_data.reset_index(inplace=True)


doc_embeddings = sif_embedding_wrapper.sentences2vecs(final_data["SEN
TENCE"], embs, words, weight4ind)
final_data["vector"] = pd.Series(list(doc_embeddings))
```

In [77]:
```
ground_truth = {}

for idx, row in final_data.iterrows():
    ground_truth[idx] = row['CLASS']

# ground_truth

inv_map = {}
for k, v in ground_truth.items():
    inv_map[v] = inv_map.get(v, [])
    inv_map[v].append(k)

inv_map
```

```
Out[77]: {'drug_gene': [0,
                        1,
                        2,
                        3,
                        4,
                        5,
                        6,
                        7,
                        8,
                        9,
                        10,
                        11,
                        12,
                        13,
                        14,
                        15,
                        16,
                        17,
                        18,
                        19,
                        20,
                        21,
                        22,
                        23,
                        24,
                        25,
                        26,
                        27,
                        28,
                        29,
                        30,
                        31,
                        32,
                        33,
                        34,
                        35,
                        36,
                        37,
                        38,
                        39,
                        40,
                        41,
                        42,
                        43,
                        44,
                        45,
                        46,
                        47,
                        48,
                        49,
                        50,
                        51,
                        52,
                        53,
                        54,
                        55,
                        56,
                        57,
                        58,
                        59,
                        60,
                        61,
                        62,
```

```
63,
64,
65,
66,
67,
68,
69,
70,
71,
72,
73,
74,
75,
76,
77,
78,
79,
80,
81,
82,
83,
84,
85,
86,
87,
88,
89,
90,
91,
92,
93,
94,
95,
96,
97,
98,
99,
100,
101,
102,
103,
104,
105,
106,
107,
108,
109,
110,
111,
112,
113,
114,
115,
116,
117,
118,
119,
120,
121,
122,
123,
124,
125,
```

```
126,
127,
128,
129,
130,
131,
132,
133,
134,
135,
136,
137,
138,
139,
140,
141,
142,
143,
144,
145,
146,
147,
148,
149,
150,
151,
152,
153,
154,
155,
156,
157,
158,
159,
160,
161,
162,
163,
164,
165,
166,
167,
168,
169,
170,
171,
172,
173,
174,
175,
176,
177,
178,
179,
180,
181,
182,
183,
184,
185,
186,
187,
188,
```

```
          189,
          190,
          191,
          192,
          193,
          194,
          195,
          196,
          197,
          198,
          199,
          200,
          201,
          202,
          203,
          204,
          205,
          206,
          207,
          208,
          209,
          210,
          211,
          212,
          213,
          214,
          215,
          216,
          217,
          218,
          219,
          220,
          221,
          222,
          223,
          224,
          225,
          226,
          227,
          228,
          229,
          230,
          231,
          232,
          233,
          234,
          235,
          236,
          237,
          238,
          239],
         'gene_disease': [240,
          241,
          242,
          243,
          244,
          245,
          246,
          247,
          248,
          249,
          250,
          251,
```

```
252,
253,
254,
255,
256,
257,
258,
259,
260,
261,
262,
263,
264,
265,
266,
267,
268,
269,
270,
271,
272,
273,
274,
275,
276,
277,
278,
279,
280,
281,
282,
283,
284,
285,
286,
287,
288,
289,
290,
291,
292,
293,
294,
295,
296,
297,
298,
299,
300,
301,
302,
303,
304,
305,
306,
307,
308,
309,
310,
311,
312,
313,
314,
```

```
315,
316,
317,
318,
319,
320,
321,
322,
323,
324,
325,
326,
327,
328,
329,
330,
331,
332,
333,
334,
335,
336,
337,
338,
339,
340,
341,
342,
343,
344,
345,
346,
347,
348,
349,
350,
351,
352,
353,
354,
355,
356,
357,
358,
359,
360,
361,
362,
363,
364,
365,
366,
367,
368,
369,
370,
371,
372,
373,
374,
375,
376,
377,
```

```
378,
379,
380,
381,
382,
383,
384,
385,
386,
387,
388,
389,
390,
391,
392,
393,
394,
395,
396,
397,
398,
399,
400,
401,
402,
403,
404,
405,
406,
407,
408,
409,
410,
411,
412,
413,
414,
415,
416,
417,
418,
419,
420,
421,
422,
423,
424,
425,
426,
427,
428,
429,
430,
431,
432,
433,
434,
435,
436,
437,
438,
439,
440,
```

```
441,
442,
443,
444,
445,
446,
447,
448,
449,
450,
451,
452,
453,
454,
455,
456,
457,
458,
459,
460,
461,
462,
463,
464,
465,
466,
467,
468,
469,
470,
471,
472,
473,
474,
475,
476,
477,
478,
479,
480,
481,
482,
483,
484,
485,
486,
487,
488,
489,
490,
491,
492,
493,
494,
495,
496,
497,
498,
499,
500,
501,
502,
503,
```

```
504,
505,
506,
507,
508,
509,
510,
511,
512,
513,
514,
515,
516,
517,
518,
519,
520,
521,
522,
523,
524,
525,
526,
527,
528,
529,
530,
531,
532,
533,
534,
535,
536,
537,
538,
539,
540,
541,
542,
543,
544,
545,
546,
547,
548,
549,
550,
551,
552,
553,
554,
555,
556,
557,
558,
559,
560,
561,
562,
563,
800,
801,
802,
```

```
803,
804,
805,
806,
807,
808,
809,
810,
811,
812,
813,
814,
815,
816,
817,
818,
819,
820,
821,
822,
823,
824,
825,
826,
827,
828,
829,
830,
831,
832,
833,
834,
835,
836,
837,
838,
839,
840,
841,
842,
843,
844,
845,
846,
847,
848,
849,
850,
851,
852,
853,
854,
855,
856,
857,
858,
859,
860,
861,
862,
863,
864,
865,
```

```
866,
867,
868,
869,
870,
871,
872,
873,
874,
875,
876,
877,
878,
879,
880,
881,
882,
883,
884,
885,
886,
887,
888,
889,
890,
891,
892,
893,
894,
895,
896,
897,
898,
899,
900,
901,
902,
903,
904,
905,
906,
907,
908,
909,
910,
911,
912,
913,
914,
915,
916,
917,
918,
919,
920,
921,
922,
923,
924,
925,
926,
927,
928,
```

```
929,
930,
931,
932,
933,
934,
935,
936,
937,
938,
939,
940,
941,
942,
943,
944,
945,
946,
947,
948,
949,
950,
951,
952,
953,
954,
955,
956,
957,
958,
959,
960,
961,
962,
963,
964,
965,
966,
967,
968,
969,
970,
971,
972,
973,
974,
975,
976,
977,
978,
979,
980,
981,
982,
983,
984,
985,
986,
987,
988,
989,
990,
991,
```

```
992,
993,
994,
995,
996,
997,
998,
999,
1000,
1001,
1002,
1003,
1004,
1005,
1006,
1007,
1008,
1009,
1010,
1011,
1012,
1013,
1014,
1015,
1016,
1017,
1018,
1019,
1020,
1021,
1022,
1023,
1024,
1025,
1026,
1027,
1028,
1029,
1030,
1031,
1032,
1033,
1034,
1035,
1036,
1037,
1038,
1039,
1040,
1041,
1042,
1043,
1044,
1045,
1046,
1047,
1048,
1049,
1050,
1051,
1052,
1053,
1054,
```

```
1055,
1056,
1057,
1058,
1059,
1060,
1061,
1062,
1063,
1064,
1065,
1066,
1067,
1068,
1069,
1070,
1071,
1072,
1073,
1074,
1075,
1076,
1077,
1078,
1079,
1080,
1081,
1082,
1083,
1084,
1085,
1086,
1087,
1088,
1089,
1090,
1091,
1092,
1093,
1094,
1095,
1096,
1097,
1098,
1099,
1100,
1101,
1102,
1103,
1104,
1105,
1106,
1107,
1108,
1109,
1110,
1111,
1112,
1113,
1114,
1115,
1116,
1117,
```

```
1118,
1119,
1120,
1121,
1122,
1123,
1124,
1125,
1126,
1127,
1128,
1129,
1130,
1131,
1132,
1133,
1134,
1135,
1136,
1137,
1138,
1139,
1140,
1141,
1142,
1143,
1144,
1145,
1146,
1147,
1148,
1149,
1150,
1151,
1152,
1153,
1154,
1155,
1156,
1157,
1158,
1159,
1160,
1161,
1162,
1163,
1164,
1165,
1166,
1167,
1168,
1169,
1170,
1171,
1172,
1173,
1174,
1175,
1176,
1177,
1178,
1179,
1180,
```

1181,
1182,
1183,
1184,
1185,
1186,
1187,
1188,
1189,
1190,
1191,
1192,
1193,
1194,
1195,
1196,
1197,
1198,
1199,
1200,
1201,
1202,
1203,
1204,
1205,
1206,
1207,
1208,
1209,
1210,
1211,
1212,
1213,
1214,
1215,
1216,
1217,
1218,
1219,
1220,
1221,
1222,
1223,
1224,
1225,
1226,
1227,
1228,
1229,
1230,
1231,
1232,
1233,
1234,
1235,
1236,
1237,
1238,
1239,
1240,
1241,
1242,
1243,

```
1244,
1245,
1246,
1247,
1248,
1249,
1250,
1251,
1252,
1253,
1254,
1255,
1256,
1257,
1258,
1259,
1260,
1261,
1262,
1263,
1264,
1265,
1266,
1267,
1268,
1269,
1270,
1271,
1272,
1273,
1274,
1275,
1276,
1277,
1278,
1279,
1280,
1281,
1282,
1283,
1284,
1285,
1286,
1287,
1288,
1289,
1290,
1291,
1292,
1293,
1294,
1295,
1296,
1297,
1298,
1299,
1300,
1301,
1302,
1303,
1304,
1305,
1306,
```

```
1307,
1308,
1309,
1310,
1311,
1312,
1313,
1314,
1315,
1316,
1317,
1318,
1319,
1320,
1321,
1322,
1323,
1324,
1325,
1326,
1327,
1328,
1329,
1330,
1331,
1332,
1333,
1334,
1335,
1336,
1337,
1338,
1339,
1340,
1341,
1342,
1343,
1344,
1345,
1346,
1347,
1348,
1349,
1350,
1351,
1352,
1353,
1354,
1355,
1356,
1357,
1358,
1359,
1360,
1361,
1362,
1363,
1364,
1365,
1366,
1367,
1368,
1369,
```

```
1370,
1371,
1372,
1373,
1374,
1375,
1376,
1377,
1378,
1379,
1380,
1381,
1382,
1383,
1384,
1385,
1386,
1387,
1388,
1389,
1390,
1391,
1392,
1393,
1394,
1395,
1396,
1397,
1398,
1399,
1400,
1401,
1402,
1403,
1404,
1405,
1406,
1407,
1408,
1409,
1410,
1411,
1412,
1413,
1414,
1415,
1416,
1417,
1418,
1419,
1420,
1421,
1422,
1423,
1424,
1425,
1426,
1427,
1428,
1429,
1430,
1431,
1432,
```

```
        1433,
        1434,
        1435,
        1436,
        1437,
        1438,
        1439,
        1440,
        1441,
        1442,
        1443,
        1444,
        1445,
        1446,
        1447,
        1448,
        1449,
        1450,
        1451,
        1452,
        1453,
        1454,
        1455,
        1456,
        1457,
        1458,
        1459,
        1460,
        1461,
        1462,
        1463,
        1464,
        1465,
        1466,
        1467,
        1468,
        1469,
        1470,
        1471,
        1472,
        1473,
        1474,
        1475,
        ...],
       'drug_disease': [564,
        565,
        566,
        567,
        568,
        569,
        570,
        571,
        572,
        573,
        574,
        575,
        576,
        577,
        578,
        579,
        580,
        581,
        582,
```

```
583,
584,
585,
586,
587,
588,
589,
590,
591,
592,
593,
594,
595,
596,
597,
598,
599,
600,
601,
602,
603,
604,
605,
606,
607,
608,
609,
610,
611,
612,
613,
614,
615,
616,
617,
618,
619,
620,
621,
622,
623,
624,
625,
626,
627,
628,
629,
630,
631,
632,
633,
634,
635,
636,
637,
638,
639,
640,
641,
642,
643,
644,
645,
```

```
646,
647,
648,
649,
650,
651,
652,
653,
654,
655,
656,
657,
658,
659,
660,
661,
662,
663,
664,
665,
666,
667,
668,
669,
670,
671,
672,
673,
674,
675,
676,
677,
678,
679,
680,
681,
682,
683,
684,
685,
686,
687,
688,
689,
690,
691,
692,
693,
694,
695,
696,
697,
698,
699,
700,
701,
702,
703,
704,
705,
706,
707,
708,
```

```
709,
710,
711,
712,
713,
714,
715,
716,
717,
718,
719,
720,
721,
722,
723,
724,
725,
726,
727,
728,
729,
730,
731,
732,
733,
734,
735,
736,
737,
738,
739,
740,
741,
742,
743,
744,
745,
746,
747,
748,
749,
750,
751,
752,
753,
754,
755,
756,
757,
758,
759,
760,
761,
762,
763,
764,
765,
766,
767,
768,
769,
770,
771,
```

772,
773,
774,
775,
776,
777,
778,
779,
780,
781,
782,
783,
784,
785,
786,
787,
788,
789,
790,
791,
792,
793,
794,
795,
796,
797,
798,
799]}

In [78]:
```
# Get average/mean of the sentence vectors that represent our topics

categories = list(final_data["CLASS"].unique())
print(categories)


category_vecs_ = {}
for c in categories:
    vectors = np.asarray(list(final_data.loc[final_data.index.isin(in
v_map[c])].vector))
    category_vecs_[c] = np.mean(vectors, axis=0)


category_vecs_
```

```
['drug_gene', 'gene_disease', 'drug_disease']
```

```
Out[78]: {'drug_gene': array([-0.02259501,  0.06900093,  0.08034494,  0.051286
         72,  0.109773  ,
                0.07043651,  0.01890659, -0.04447399,  0.00232815, -0.031831
         96,
                0.02219459, -0.06560161,  0.05480214, -0.0008662 , -0.036251
         38,
               -0.03070597, -0.10438451, -0.16020964,  0.00480927,  0.025631
         61,
               -0.06444762, -0.0747943 , -0.02560701, -0.05884123,  0.037142
         24,
               -0.07155826,  0.03770538, -0.0045408 ,  0.00113782, -0.075157
         98,
                0.04877341, -0.08112333, -0.10605722, -0.007455  ,  0.061097
         01,
               -0.08524513, -0.01665239, -0.00193276, -0.08086548, -0.091132
         11,
                0.01160008, -0.04256436,  0.06923264,  0.05370258,  0.139539
         33,
               -0.04828128, -0.10773625, -0.08710372,  0.03560877, -0.044034
         7 ,
               -0.04703573, -0.00654404, -0.05748476,  0.00903911, -0.090228
         5 ,
                0.02511453,  0.08825264,  0.00662434, -0.02267997,  0.029066
         21,
                0.01720275, -0.04489324, -0.01539939, -0.0211833 ,  0.008635
         89,
                0.01029012, -0.00341215,  0.02266545, -0.04107853, -0.065592
         95,
               -0.0492788 , -0.04646665, -0.00425594, -0.00879286,  0.072807
         27,
                0.09293385, -0.02217183,  0.04853517, -0.01381857,  0.071499
         28,
               -0.0046454 , -0.02435077,  0.01293399,  0.04367903, -0.030009
         52,
                0.03377294, -0.07752994,  0.00459051,  0.01960824,  0.053857
         29,
               -0.06998995,  0.02772035, -0.0112851 , -0.1046491 ,  0.051055
         56,
               -0.01107914, -0.03322789, -0.014381  ,  0.11331822,  0.013468
         26,
                0.01423458,  0.09652433, -0.04998231, -0.03524452,  0.035954
         67,
                0.01047966, -0.04607962,  0.07103991,  0.03767436,  0.051907
         8 ,
                0.02106631, -0.01602907,  0.03807558, -0.02678066,  0.067820
         98,
                0.01417979, -0.01665879,  0.00436057,  0.02230283, -0.063107
         81,
                0.04569304,  0.03064912,  0.00540505, -0.04279056,  0.040786
         83,
               -0.0097842 ,  0.02130368, -0.04016707, -0.06234886, -0.055506
         53,
               -0.028472  ,  0.03840448, -0.03265039,  0.07322495, -0.021719
         42,
                0.0219988 , -0.07528665, -0.09371537, -0.02687253,  0.057981
         7 ,
               -0.06420033,  0.04318409,  0.05758439,  0.04212966, -0.069765
         53,
                0.03406704,  0.0216992 , -0.01751402,  0.05846941,  0.055000
         27,
                0.08659255,  0.01035483,  0.03605185,  0.01029625, -0.004406
         26,
               -0.01686907, -0.08494996, -0.0204646 , -0.02754893, -0.023140
```

```
06,
        -0.00833148, -0.01588855,  0.03042198,  0.03391189,  0.048127
59,
        -0.03573181,  0.06215374,  0.06567121,  0.08501771,  0.005723
01,
        -0.00544842, -0.10782876, -0.0821683 , -0.0439988 , -0.155123
49,
         0.02804725, -0.06658659, -0.11911979,  0.00269219, -0.037008
09,
        -0.02821007, -0.03748677, -0.07513077, -0.10859924,  0.027878
82,
         0.00974063, -0.09650454, -0.02150409,  0.05049756,  0.064454
23,
        -0.02594272, -0.00940941, -0.01749841,  0.10001498, -0.028754
55,
         0.05609788, -0.00061992,  0.14480043, -0.00793043, -0.000522
9 ]),
 'gene_disease': array([-0.01894094,  0.0026824 ,  0.05278864, -0.049
94545,  0.06973205,
        -0.00715203, -0.03732716, -0.028285  , -0.03021973,  0.017963
67,
        -0.00866856, -0.13118232,  0.02993769,  0.01045083,  0.002905
35,
        -0.06039446, -0.07724169, -0.13998248,  0.02163892,  0.076546
87,
         0.03099466, -0.06496584, -0.04123766, -0.08287935,  0.039962
92,
        -0.03143508,  0.04435205, -0.10256732,  0.0232578 , -0.058746
06,
         0.0219973 , -0.0486413 , -0.04402548, -0.00741007,  0.076764
83,
        -0.02208246,  0.04607483,  0.03614996, -0.05852792, -0.025431
51,
        -0.0027273 , -0.06236734,  0.0708188 ,  0.04669061,  0.105552
5 ,
        -0.07922249, -0.10981007, -0.11103149, -0.04820119, -0.131596
94,
         0.0110788 , -0.00328608, -0.01818729,  0.04317896, -0.087124
44,
         0.04692777, -0.02968445,  0.0111203 ,  0.01402305, -0.004438
84,
         0.03805884,  0.00645043, -0.01333097,  0.00830524,  0.059119
98,
        -0.00370814, -0.03271034,  0.08991915,  0.01743721,  0.003759
78,
        -0.01801892, -0.04478382, -0.03234207, -0.04723298, -0.035319
08,
         0.03023364,  0.10519775, -0.00099013, -0.0343597 , -0.007468
95,
         0.07082928, -0.04647901, -0.03605074, -0.01985432, -0.049707
24,
        -0.03663353, -0.11370036, -0.05383802,  0.00295939,  0.025941
26,
         0.01381267,  0.01257951, -0.1044039 , -0.06473194,  0.016136
96,
        -0.01962321,  0.0079649 , -0.03765243,  0.10690313, -0.064533
64,
         0.01880765,  0.05319942, -0.01443602, -0.02166466,  0.032254
19,
        -0.03917902,  0.01484661,  0.04468778,  0.01132086, -0.004426
88,
         0.01135866, -0.04061285,  0.0470459 , -0.01480132,  0.053582
81,
```

```
        -0.02430859, -0.00370396,  0.00644308,  0.0112515 , -0.009823
82,
         0.06629728, -0.07685832,  0.01149585, -0.03483788,  0.014352
51,
        -0.12354892,  0.03485231, -0.06347502, -0.00572563, -0.067933
55,
         0.0144394 ,  0.01704359, -0.07134108,  0.05815698, -0.010918
13,
         0.07398397, -0.07420843, -0.07630211,  0.00325303, -0.002278
87,
        -0.0525499 ,  0.02622819,  0.05983656,  0.07826244,  0.000243
85,
        -0.02631064,  0.0424064 , -0.05613592,  0.03854308,  0.079826
39,
         0.04066342,  0.0155632 ,  0.0522846 , -0.01499126,  0.009874
8 ,
        -0.0091032 ,  0.00238626, -0.06941868, -0.03876348, -0.027028
36,
        -0.01528063,  0.04179048, -0.00912749,  0.11076091,  0.025021
58,
         0.00408332,  0.00341762,  0.04338575,  0.05023738,  0.006381
38,
        -0.01582713, -0.12511575,  0.01775416, -0.07794587, -0.115324
02,
        -0.03153813,  0.00726716, -0.16100741,  0.01333724,  0.014803
09,
        -0.1028725 ,  0.04394676, -0.08594431, -0.03324341,  0.101837
84,
         0.05323173, -0.05624724, -0.02914044,  0.03809479, -0.098374
05,
         0.03651038, -0.01889545,  0.02978605,  0.07177132, -0.069190
33,
        -0.0336774 ,  0.00215992,  0.10554505, -0.05418101,  0.004824
24]),
 'drug_disease': array([-8.08392393e-02,  1.07544154e-01,  5.07261586
e-02, -7.63855879e-03,
        4.20294581e-02, -1.86339118e-02,  3.45490900e-02,  1.5968065
5e-03,
        4.71229241e-03,  4.93887052e-02, -1.51885675e-02, -7.5271002
8e-02,
        1.05176153e-01, -2.11550538e-02, -4.61599000e-02, -7.7286075
8e-02,
        -8.08744149e-02, -1.05410263e-01,  2.52544196e-02,  3.9965722
4e-02,
        8.52303592e-02, -2.81857952e-02, -2.63347009e-02, -6.6486591
3e-02,
        6.74278522e-03, -3.47732691e-02,  8.11143022e-02, -3.5568367
4e-02,
        -6.12660033e-02, -8.83450442e-03,  2.71653893e-02, -9.9585942
1e-02,
        -3.90031168e-02, -6.72858266e-02,  7.12616571e-02, -2.3815741
4e-02,
        -2.84257743e-02,  1.57893809e-02, -9.70614031e-02, -6.4664937
0e-02,
        -7.91901072e-02, -2.72769947e-02,  6.38239988e-03,  4.7651791
7e-02,
        1.10229683e-01, -5.45760261e-02, -8.71110026e-02, -3.0709936
2e-02,
        -4.07317947e-02, -1.01426498e-01,  4.75029037e-02,  7.6334963
3e-03,
        4.97566438e-05, -1.10346348e-02, -8.06670487e-02, -3.6846895
3e-02,
        5.93816519e-02, -4.34332763e-02,  4.27593094e-02,  4.4832464
```

```
3e-03,
         4.11583412e-02,  1.36420227e-02,  2.91610699e-02, -2.5080944
9e-02,
        -6.82629255e-04,  1.56183206e-02, -1.07726300e-02,  2.2636439
9e-02,
         6.84676222e-02,  1.87075333e-03,  5.19577105e-02,  1.7809738
0e-02,
        -4.85580615e-02,  1.60969578e-02, -2.92258248e-02,  5.8543472
0e-02,
         1.17887182e-01, -4.11336059e-02, -5.40964933e-02,  1.4502668
0e-02,
         7.74309828e-03, -1.87204061e-02,  1.12390597e-01, -2.0105929
8e-02,
        -1.12989185e-01,  5.04171514e-02, -6.25617620e-02,  2.5012505
3e-02,
         4.11678011e-02,  3.02352421e-02, -5.43229013e-02,  4.9704823
4e-02,
        -4.87038539e-02,  7.65902119e-03,  4.42130840e-02,  3.5862417
5e-02,
         1.47823361e-02, -3.77691528e-02,  4.34341766e-02, -1.2229202
9e-02,
        -2.75046446e-02, -5.87096100e-03,  3.81385351e-02, -4.2619629
7e-02,
         3.15735410e-02,  1.13913980e-02,  1.62923440e-02, -2.4147986
5e-02,
        -2.83084175e-02,  4.03391589e-02,  7.36034811e-02, -4.6791143
4e-02,
         1.19673244e-02, -9.78305302e-02,  3.54012787e-02,  5.3309452
3e-02,
         2.76526361e-02, -8.02144147e-02,  3.65307779e-02,  7.1878540
2e-02,
         5.65107900e-02, -5.25904331e-02, -2.00148826e-02, -8.4448027
8e-02,
        -6.08861982e-02, -6.45479330e-02,  5.63616355e-02, -7.3794890
1e-02,
        -1.23510253e-02, -1.37231388e-03,  1.17334497e-02,  8.0796253
9e-03,
        -3.20168359e-02, -2.72035957e-02, -9.13218078e-02,  4.1837185
7e-02,
        -1.20814114e-01, -4.35413689e-02, -7.38197109e-02,  6.9802263
7e-02,
        -2.02128357e-02,  1.55208874e-02,  8.12812275e-03,  9.2753581
3e-02,
        -2.69912140e-02,  2.65074761e-02,  5.08721041e-02, -7.2461008
9e-02,
         2.58917041e-02,  5.47822620e-02,  4.45604219e-02, -4.9611432
5e-03,
         7.12214357e-02, -7.05062708e-03,  1.15441222e-01, -1.7689600
1e-02,
        -7.03170819e-03, -6.00978234e-02, -2.82504615e-03, -2.7545736
3e-02,
        -3.60451418e-02,  3.93191636e-02,  3.93019914e-02,  2.4697148
9e-02,
        -4.63719877e-02,  4.45779242e-02,  5.14349063e-02,  5.1546710
4e-02,
         7.38075400e-02, -1.61509765e-02, -1.39308377e-02, -1.0366696
0e-01,
        -4.38634953e-02, -3.09122887e-02, -9.40224995e-02,  2.6571560
2e-02,
        -2.34101962e-03, -8.26211852e-02, -9.82162232e-03, -7.0738699
7e-02,
         6.68611038e-03, -1.01406407e-02, -9.43302041e-02, -9.1638647
7e-02,
```

```
         5.29445651e-02,  2.56046891e-02, -6.58245566e-02, -1.3110060
8e-02,
        -4.95847725e-02, -2.56343192e-02, -2.02271077e-02, -6.4641612
2e-03,
        -1.80915229e-03,  1.09901095e-03, -2.56411728e-02, -5.5804179
3e-02,
         2.72605923e-02,  1.20513499e-01, -5.31299967e-02, -7.3606906
2e-03])}
```

In [84]:
```
# Test new sentence

test_sample = 'This study assessed associations between the CYP4F2 ge
ne and myocardial infarction (MI), using a haplotype-based case-contr
ol study of 234 MI patients and 248 controls genotyped for 5 single-n
ucleotide polymorphisms (rs3093105, rs3093135, rs1558139, rs2108622,
 rs3093200).'
# test_sample = 'Assessment of 1177 human immunodeficiency virus (HI
V) resistance genotypes at an HIV/AIDS clinic showed a decrease in th
e incidence of the K65R mutation, from 15.2% of isolates during the p
eriod 2002-2004 to 2.7% of isolates during the period 2005-2006 (P <
 .001), despite elevated and stable rates of tenofovir use.'
# test_sample = 'Doxorubicin-induced DNA damage was also specifically
abolished by the proteasome inhibitors bortezomib and MG132 and much
 reduced in top2beta(-/-) mouse embryonic fibroblasts (MEF) compared
 with TOP2beta(+/+) MEFs, suggesting the involvement of proteasome an
d DNA topoisomerase IIbeta (Top2beta).'
# test_sample = 'SLC9A6 at Xq26.3 (Gilfillan et al., 2008)X-linked me
ntal retardation'
# test_sample = 'DLBCL was identified by a microenvironment gene expr
ession signature and is associated with increased expression of infla
mmatory mediators, such as multiple components of the T-cell receptor
(TCR), molecules associated with T/NK-cell activation and the complem
ent cascade, downstream targets of IFNγ'

test_embedding = sif_embedding_wrapper.sentences2vecs([test_sample],
embs, words, weight4ind)

sim = {}
for j in category_vecs:
    sim[j] = cosine_similarity(test_embedding.reshape(1, -1), categor
y_vecs[j].reshape(1, -1)).flatten()[0]

sim
```

Out[84]:
```
{'drug_gene': 0.27992935769402444,
 'gene_disease': 0.6860943582066192,
 'drug_disease': 0.36788278692274523}
```

In [83]:
```
# Test new sentence

test_sample = 'This study assessed associations between the CYP4F2 ge
ne and myocardial infarction (MI), using a haplotype-based case-contr
ol study of 234 MI patients and 248 controls genotyped for 5 single-n
ucleotide polymorphisms (rs3093105, rs3093135, rs1558139, rs2108622,
 rs3093200).'
# test_sample = 'Assessment of 1177 human immunodeficiency virus (HI
V) resistance genotypes at an HIV/AIDS clinic showed a decrease in th
e incidence of the K65R mutation, from 15.2% of isolates during the p
eriod 2002-2004 to 2.7% of isolates during the period 2005-2006 (P <
 .001), despite elevated and stable rates of tenofovir use.'
# test_sample = 'Doxorubicin-induced DNA damage was also specifically
abolished by the proteasome inhibitors bortezomib and MG132 and much
 reduced in top2beta(-/-) mouse embryonic fibroblasts (MEF) compared
 with TOP2beta(+/+) MEFs, suggesting the involvement of proteasome an
d DNA topoisomerase IIbeta (Top2beta).'
# test_sample = 'SLC9A6 at Xq26.3 (Gilfillan et al., 2008)X-linked me
ntal retardation'
# test_sample = 'DLBCL was identified by a microenvironment gene expr
ession signature and is associated with increased expression of infla
mmatory mediators, such as multiple components of the T-cell receptor
(TCR), molecules associated with T/NK-cell activation and the complem
ent cascade, downstream targets of IFNγ'

test_embedding = sif_embedding_wrapper.sentences2vecs([test_sample],
embs, words, weight4ind)

sim = {}
for j in category_vecs_:
    sim[j] = cosine_similarity(test_embedding.reshape(1, -1), categor
y_vecs_[j].reshape(1, -1)).flatten()[0]

sim
```

Out[83]:
```
{'drug_gene': 0.27115843650125687,
 'gene_disease': 0.6521062059914209,
 'drug_disease': 0.3627070600955426}
```