huew_task

June 22, 2018

1   System Configuration

RAM -> 8GB
Processor -> AMD A6-5350M
OS -> Ubuntu 16.04
OS Type -> 64-bit

2   # Approach

The original dataset contains 4,057,189 rows. Out of these it was given in the problem statement, that only those value has to be selected which has subcategories "Tunics". After selecting those values, with the help of Pandas, total rows become 644,661. Since this was still too big for the system to handle, I randomly selected 40,000 data entries from the entire dataset. Still no label was given to us, it was a unsupervised problem. Out of all 32 feature points, I primarily selected title, description, productFamily, keySpecsStr, detailedSpecsStr because title and description describe the product in layman language; productFamily describes a class of product; keySpecsStr and detailedSpecsStr are describing the product in detail. All these 5 features were string type, so features were extracted from these by converting textual data to word2vec vectors. After that cosine similarity was calculated for each features using these word2vec vectors. These cosine similarity will tell how two products are similar for a particular feature. After calculating for all 5 features, a score was calculated using those.

```
In [1]: import pandas as pd
        import numpy as np
        from gensim.models import doc2vec
        from collections import namedtuple
        from scipy import spatial
```

```
/home/naman/.local/lib/python2.7/site-packages/requests/__init__.py:83: RequestsDependencyWarnin
  warnings.warn(warning, RequestsDependencyWarning)
```

```
In [2]: # data1 = pd.read_csv('2oq-c1r.csv') # This is the original dataset's dataframe
        # index_list = []
        # for i in range(len(df)):
        #     if 'unic' in df['categories'][i]:
        #         index_list.append(i)
        # data2 = data1.iloc[index_list, :] # This dataframe contains values having only Tunics
```

```
In [3]: # data2 = data2.sample(frac=1).reset_index(drop=True)
        # df = data2[:40000] # This dataframe contains randomly selected 40000 rows containing 1

In [4]: df = pd.read_csv('tunic40k.csv')

In [5]: df.shape

Out[5]: (40000, 34)

In [6]: df.columns

Out[6]: Index([u'Unnamed: 0', u'Unnamed: 0.1', u'productId', u'title', u'description',
               u'imageUrlStr', u'mrp', u'sellingPrice', u'specialPrice', u'productUrl',
               u'categories', u'productBrand', u'productFamily', u'inStock',
               u'codAvailable', u'offers', u'discount', u'shippingCharges',
               u'deliveryTime', u'size', u'color', u'sizeUnit', u'storage',
               u'displaySize', u'keySpecsStr', u'detailedSpecsStr',
               u'specificationList', u'sellerName', u'sellerAverageRating',
               u'sellerNoOfRatings', u'sellerNoOfReviews', u'sleeve', u'neck',
               u'idealFor'],
              dtype='object')

In [7]: df.neck = df.neck.fillna('')
        df.sizeUnit = df.sizeUnit.fillna('')
        df['size'] = df['size'].fillna('')
        df.keySpecsStr = df.keySpecsStr.fillna('')
        df.detailedSpecsStr = df.detailedSpecsStr.fillna('')
        df.description = df.description.fillna('')
        df.title = df.title.fillna('')
        df.productFamily = df.productFamily.fillna('')

In [8]: df = df.drop(['Unnamed: 0', 'Unnamed: 0.1', 'imageUrlStr', 'productUrl', 'inStock', 'cod

In [9]: df.shape

Out[9]: (40000, 17)

In [10]: def return_similarity_score(doc1):
             docs = []
             analyzedDocument = namedtuple('AnalyzedDocument', 'words tags')
             for i, text in enumerate(doc1):
                 words = text.lower().split()
                 tags = [i]
                 docs.append(analyzedDocument(words, tags))
             model = doc2vec.Doc2Vec(docs, size = 100, window = 300, min_count = 1, workers = 4)
             sim_score = []
             for i in range(len(model.docvecs)):
                 sim_score.append(1-(spatial.distance.cosine(model.docvecs[0], model.docvecs[i])
             return sim_score
```

2

```
In [11]: doc_title = df['title'].tolist()
         sim_score_title = return_similarity_score(doc_title)

/home/naman/.local/lib/python2.7/site-packages/gensim/models/doc2vec.py:366: UserWarning: The pa
  warnings.warn("The parameter `size` is deprecated, will be removed in 4.0.0, use `vector_size`


In [12]: sorted(sim_score_title, reverse=True)

Out[12]: [1.0,
          0.9580023884773254,
          0.9572941064834595,
          0.9560819864273071,
          0.9532979130744934,
          0.9523184895515442,
          0.9522967338562012,
          0.9522823691368103,
          0.9508616924285889,
          0.9507434368133545,
          0.9503090381622314,
          0.9481085538864136,
          0.9478361010551453,
          0.9473828673362732,
          0.9472674131393433,
          0.947053074836731,
          0.9460029006004333,
          0.9445982575416565,
          0.9435232877731323,
          0.9435194134712219,
          0.9434640407562256,
          0.9430810809135437,
          0.9427804350852966,
          0.9427653551101685,
          0.9424615502357483,
          0.9421488642692566,
          0.9421411752700806,
          0.9420633316040039,
          0.9419798851013184,
          0.9419723153114319,
          0.9409561157226562,
          0.9409391283988953,
          0.9409373998641968,
          0.9408707022666931,
          0.9407768249511719,
          0.9407556056976318,
          0.940752387046814,
          0.9405441880226135,
          0.9405302405357361,
```

0.9401859045028687,
0.9401128888130188,
0.9399386048316956,
0.9398705363273621,
0.9397174119949341,
0.9396159052848816,
0.9395330548286438,
0.939530611038208,
0.9393796324729919,
0.9392148852348328,
0.9391650557518005,
0.9390748739242554,
0.939039409160614,
0.9387760162353516,
0.9385697841644287,
0.9385076761245728,
0.9384440183639526,
0.9384400844573975,
0.9380615949630737,
0.9379655718803406,
0.9378096461296082,
0.9376811385154724,
0.9376200437545776,
0.9375274181365967,
0.9372771382331848,
0.937240481376648,
0.9371634125709534,
0.9369938969612122,
0.9369491934776306,
0.9366521835327148,
0.9365747570991516,
0.9363778829574585,
0.9362097382545471,
0.9361670613288879,
0.9360546469688416,
0.9360296726226807,
0.9358532428741455,
0.93573397397995,
0.9357234239578247,
0.935718297958374,
0.9356980323791504,
0.9355965852737427,
0.9355705976486206,
0.9354432225227356,
0.9353780746459961,
0.9353705048561096,
0.9350839257240295,
0.9349771738052368,

0.9348522424697876,
0.9348034858703613,
0.934689462184906,
0.934463620185852,
0.9344353079795837,
0.9343814253807068,
0.9343054294586182,
0.9342015385627747,
0.9341400861740112,
0.9341288805007935,
0.9340789914131165,
0.9338177442550659,
0.933718204498291,
0.9336870312690735,
0.9335954785346985,
0.9335799813270569,
0.9335373640060425,
0.9334658980369568,
0.933415412902832,
0.9332342743873596,
0.9332325458526611,
0.9331343173980713,
0.9330700039863586,
0.9330360293388367,
0.9330193400382996,
0.932949960231781,
0.9329098463058472,
0.932775616645813,
0.9327297210693359,
0.9326911568641663,
0.9326115250587463,
0.9325599670410156,
0.9323965311050415,
0.9321737885475159,
0.9321134090423584,
0.9321133494377136,
0.9320469498634338,
0.9320351481437683,
0.9320273399353027,
0.9320156574249268,
0.9320144653320312,
0.9319814443588257,
0.9319718480110168,
0.9318414330482483,
0.9317208528518677,
0.9316912293434143,
0.9316836595535278,
0.9316720962524414,

0.9315611720085144,
0.931489884853363,
0.931474506855011,
0.9314236044883728,
0.931403398513794,
0.931268572807312,
0.9312143921852112,
0.9311304092407227,
0.9310915470123291,
0.9310237765312195,
0.9310221672058105,
0.9310219287872314,
0.9308419823646545,
0.9308158755302429,
0.930773913860321,
0.9307703971862793,
0.9307268261909485,
0.9305217266082764,
0.9304553270339966,
0.9303675293922424,
0.9303340315818787,
0.9303038120269775,
0.9302549958229065,
0.930080235004425,
0.9298834204673767,
0.9298639893531799,
0.9298552870750427,
0.9298527240753174,
0.9296844005584717,
0.9296714067459106,
0.9296620488166809,
0.9296331405639648,
0.9296225905418396,
0.9295110702514648,
0.9295027852058411,
0.9293761253356934,
0.9293083548545837,
0.9292798042297363,
0.929261326789856,
0.9290882349014282,
0.9290338158607483,
0.9289442300796509,
0.9289401769638062,
0.9289211630821228,
0.9289103150367737,
0.928848147392273,
0.9288450479507446,
0.9288159012794495,

0.9288057684898376,
0.9287737607955933,
0.9287537336349487,
0.928719699382782,
0.9286646246910095,
0.9284791350364685,
0.9284300804138184,
0.9284201264381409,
0.9284091591835022,
0.9283953905105591,
0.928270697593689,
0.928153395652771,
0.9280552268028259,
0.9280075430870056,
0.9280012249946594,
0.9279949069023132,
0.9279174208641052,
0.9279142618179321,
0.9278568029403687,
0.9278436899185181,
0.9278056025505066,
0.9277932047843933,
0.9275931119918823,
0.9274988174438477,
0.9274575710296631,
0.9273489713668823,
0.9273051619529724,
0.9271460771560669,
0.9270957112312317,
0.9270898699760437,
0.9270778894424438,
0.9268966317176819,
0.9268698692321777,
0.926855206489563,
0.9268397092819214,
0.9267942309379578,
0.9267541766166687,
0.9267425537109375,
0.9266989231109619,
0.9265049695968628,
0.9264785051345825,
0.9264401197433472,
0.9263765215873718,
0.9263690114021301,
0.926345705986023,
0.9263427257537842,
0.9260028600692749,
0.9260013699531555,

0.9259555339813232,
0.92594975233078,
0.925757884979248,
0.9256116151809692,
0.9255028963088989,
0.9254962205886841,
0.9254539608955383,
0.9254034161567688,
0.9253906011581421,
0.9253884553909302,
0.9253568649291992,
0.9253431558609009,
0.9253414869308472,
0.9253316521644592,
0.9252766966819763,
0.9252074360847473,
0.9252025485038757,
0.9251710176467896,
0.9251264929771423,
0.9251022934913635,
0.9250966310501099,
0.9250643849372864,
0.9250370264053345,
0.9250257611274719,
0.9250078797340393,
0.9249587655067444,
0.9247440099716187,
0.924697756767273,
0.9246461391448975,
0.9246131777763367,
0.9246065616607666,
0.9245980381965637,
0.9245668649673462,
0.9245641827583313,
0.9245408177375793,
0.924475371837616,
0.9244427680969238,
0.9243579506874084,
0.924205482006073,
0.9241894483566284,
0.9241834878921509,
0.9241525530815125,
0.9240929484367371,
0.9240909218788147,
0.9240237474441528,
0.923838198184967,
0.9238144755363464,
0.9237198829650879,

0.9236658811569214,
0.9235804677009583,
0.9235729575157166,
0.9235551357269287,
0.9235341548919678,
0.9235249161720276,
0.9234713912010193,
0.9234043955802917,
0.9233912825584412,
0.9233514070510864,
0.9232305884361267,
0.9231811761856079,
0.9231598973274231,
0.9230563044548035,
0.9230352640151978,
0.9229908585548401,
0.9229472875595093,
0.9229177832603455,
0.9228684902191162,
0.9228259325027466,
0.9227514863014221,
0.9226789474487305,
0.9225978851318359,
0.9225651621818542,
0.9225610494613647,
0.9224446415901184,
0.9223560690879822,
0.9222796559333801,
0.922269344329834,
0.9221946597099304,
0.9221908450126648,
0.9221163392066956,
0.922102153301239,
0.9220790863037109,
0.9219652414321899,
0.9219170808792114,
0.9219169020652771,
0.9217974543571472,
0.921782910823822,
0.9217455983161926,
0.9217384457588196,
0.9217349290847778,
0.9216665625572205,
0.9216262698173523,
0.9215947389602661,
0.921534538269043,
0.9215080142021179,
0.9214996695518494,

0.9214856624603271,
0.9214792251586914,
0.9214028120040894,
0.9213878512382507,
0.9213868975639343,
0.9213266372680664,
0.921286940574646,
0.9211958646774292,
0.9211093783378601,
0.9211006760597229,
0.921063244342804,
0.9210608601570129,
0.9210104942321777,
0.9209688305854797,
0.9208980202674866,
0.9208519458770752,
0.9208351969718933,
0.9207877516746521,
0.9207338690757751,
0.9207229614257812,
0.9207099676132202,
0.9207059741020203,
0.9206984639167786,
0.9206829071044922,
0.9206339716911316,
0.9205787777900696,
0.9205688834190369,
0.9205634593963623,
0.9205591678619385,
0.9205331802368164,
0.920529305934906,
0.9205031394958496,
0.9204839468002319,
0.9204730987548828,
0.9204532504081726,
0.9204283356666565,
0.9203289747238159,
0.9202134609222412,
0.920199990272522,
0.9201131463050842,
0.9200845956802368,
0.9199159145355225,
0.9198248982429504,
0.9198151230812073,
0.9198023676872253,
0.9197850227355957,
0.9197512865066528,
0.9196933507919312,

0.9196693301200867,
0.9196401238441467,
0.919623851776123,
0.9195604920387268,
0.9195430278778076,
0.9195250272750854,
0.919451117515564,
0.9194411635398865,
0.9194409251213074,
0.9194083213806152,
0.9193999767303467,
0.9193601608276367,
0.9193499088287354,
0.9193276166915894,
0.9192925691604614,
0.919291079044342,
0.9192756414413452,
0.919154942035675,
0.9190661311149597,
0.9190317392349243,
0.9190284609794617,
0.9189488887786865,
0.9189131259918213,
0.9188954830169678,
0.9187474250793457,
0.9187209606170654,
0.9187138080596924,
0.9186027646064758,
0.9185158610343933,
0.9185128808021545,
0.9185081720352173,
0.9184698462486267,
0.9184619784355164,
0.9184491634368896,
0.9184480309486389,
0.9184130430221558,
0.9183931946754456,
0.918359100818634,
0.9183480143547058,
0.9183396697044373,
0.9182206988334656,
0.9182018041610718,
0.9181824326515198,
0.918146014213562,
0.9181159138679504,
0.9181126952171326,
0.9180654287338257,
0.9180471897125244,

0.9180249571800232,
0.9180167317390442,
0.9180039763450623,
0.9179664254188538,
0.9179098606109619,
0.9178860187530518,
0.9178756475448608,
0.9178721904754639,
0.9177818298339844,
0.9177794456481934,
0.917770504951477,
0.9177607893943787,
0.9177201390266418,
0.9176483750343323,
0.9176467657089233,
0.9175834655761719,
0.917454719543457,
0.9174502491950989,
0.9174428582191467,
0.917441725730896,
0.9173579216003418,
0.9173381328582764,
0.9172970056533813,
0.9172862768173218,
0.9172587394714355,
0.9171479344367981,
0.9171013236045837,
0.9170429110527039,
0.9170243740081787,
0.9169710874557495,
0.9169021844863892,
0.9169010519981384,
0.9168840050697327,
0.9168540835380554,
0.9168412685394287,
0.9168158769607544,
0.9168003797531128,
0.9167928695678711,
0.9167906045913696,
0.9167207479476929,
0.9167184829711914,
0.916671097278595,
0.9166518449783325,
0.916638970375061,
0.9166381359100342,
0.916593074798584,
0.9165914058685303,
0.9165846109390259,

0.9165791273117065,
0.9165524840354919,
0.9165521264076233,
0.9165346026420593,
0.9165228009223938,
0.9164940714836121,
0.9163748621940613,
0.9163637161254883,
0.9163591265678406,
0.9163229465484619,
0.9162914156913757,
0.9162881374359131,
0.9162121415138245,
0.91618812084198,
0.9161759614944458,
0.9161199927330017,
0.9160631895065308,
0.9160612225532532,
0.9159698486328125,
0.9159523248672485,
0.9159500002861023,
0.9159089922904968,
0.9158985614776611,
0.9158011674880981,
0.9157735109329224,
0.9157727360725403,
0.915758490562439,
0.9157502055168152,
0.9157386422157288,
0.9157143235206604,
0.9157072901725769,
0.9156869053840637,
0.9156859517097473,
0.9156774282455444,
0.9156445264816284,
0.9155827760696411,
0.9155786633491516,
0.9155616164207458,
0.9154855012893677,
0.9154790043830872,
0.9154697060585022,
0.9154613018035889,
0.9154564738273621,
0.9154142737388611,
0.9154142737388611,
0.9153160452842712,
0.9153013229370117,
0.9152613878250122,

0.9152576923370361,
0.9152511954307556,
0.915237545967102,
0.9152117371559143,
0.9151976704597473,
0.9151811003684998,
0.9151553511619568,
0.9151485562324524,
0.9151442050933838,
0.9151293039321899,
0.915103554725647,
0.9150760173797607,
0.9150597453117371,
0.9150475859642029,
0.9150266051292419,
0.9150206446647644,
0.9149779081344604,
0.914970338344574,
0.9149686098098755,
0.9149268865585327,
0.9149141907691956,
0.9149021506309509,
0.9148754477500916,
0.9148513078689575,
0.914840817451477,
0.9148305654525757,
0.9148034453392029,
0.9148033261299133,
0.9147956371307373,
0.914793848991394,
0.9147657155990601,
0.9147613048553467,
0.9147542715072632,
0.9146382212638855,
0.9146372675895691,
0.9146144390106201,
0.9146100282669067,
0.9146087765693665,
0.9145926833152771,
0.9145740270614624,
0.9145533442497253,
0.9145106077194214,
0.9144635200500488,
0.9144351482391357,
0.9144333600997925,
0.9143936038017273,
0.9143744707107544,
0.9143511056900024,

0.9143227934837341,
0.9142239093780518,
0.914203405380249,
0.9141926169395447,
0.9141920804977417,
0.914161205291748,
0.9141045212745667,
0.914043664932251,
0.9140249490737915,
0.9139856100082397,
0.9139633178710938,
0.9139339923858643,
0.9139216542243958,
0.9139102101325989,
0.9138795137405396,
0.9138534665107727,
0.9138267636299133,
0.9138244986534119,
0.9138187170028687,
0.9137557744979858,
0.9137535691261292,
0.9137272834777832,
0.913683295249939,
0.9136425852775574,
0.9136293530464172,
0.9136266112327576,
0.9136118292808533,
0.9135927557945251,
0.9135161638259888,
0.9135098457336426,
0.9134860038757324,
0.9134413599967957,
0.9134408831596375,
0.9134262800216675,
0.9134026765823364,
0.913382351398468,
0.913377046585083,
0.913338303565979,
0.9133324027061462,
0.9133238792419434,
0.9132998585700989,
0.9132940769195557,
0.9132688045501709,
0.9132587313652039,
0.9132530093193054,
0.9132437109947205,
0.913227915763855,
0.9132193326950073,

```
0.9131938219070435,
0.9131395816802979,
0.9131287336349487,
0.9131216406822205,
0.9131177663803101,
0.913098156452179,
0.9130975604057312,
0.913090169429779,
0.9130610823631287,
0.9130181074142456,
0.9129996299743652,
0.9129974246025085,
0.9129669070243835,
0.9129036068916321,
0.9128545522689819,
0.9128498435020447,
0.9127857685089111,
0.9127610325813293,
0.9127529263496399,
0.912743091583252,
0.9126443862915039,
0.9126036763191223,
0.9126020073890686,
0.9125484228134155,
0.9125345349311829,
0.9125222563743591,
0.9125215411186218,
0.9125190377235413,
0.9125039577484131,
0.912485659122467,
0.9124698638916016,
0.912413477897644,
0.9123974442481995,
0.9123923778533936,
0.9123713374137878,
0.9123631715774536,
0.9123563766479492,
0.9123353362083435,
0.9123021364212036,
0.9123002886772156,
0.9122843146324158,
0.9122517704963684,
0.9122142791748047,
0.912211000919342,
0.9121875762939453,
0.9121819138526917,
0.9121787548065186,
0.9120728373527527,
```

0.9120311737060547,
0.9120224118232727,
0.9120211005210876,
0.9120147824287415,
0.9119557738304138,
0.9119470715522766,
0.9119279980659485,
0.9119183421134949,
0.9119064211845398,
0.9118569493293762,
0.9118459224700928,
0.9117839932441711,
0.9117429256439209,
0.9116831421852112,
0.91164231300354,
0.9116168022155762,
0.9115534424781799,
0.9115515947341919,
0.9115383625030518,
0.9115312099456787,
0.9115304946899414,
0.9114693999290466,
0.9114231467247009,
0.9114139676094055,
0.9113931655883789,
0.9113894701004028,
0.9113240242004395,
0.9113222360610962,
0.9112992286682129,
0.9112969040870667,
0.911281943321228,
0.9112769365310669,
0.9112711548805237,
0.9112065434455872,
0.9111975431442261,
0.9111973643302917,
0.9111760258674622,
0.9111447930335999,
0.9111227989196777,
0.9110607504844666,
0.9110472202301025,
0.9110439419746399,
0.9110315442085266,
0.9109907746315002,
0.9109795689582825,
0.9109712243080139,
0.9109658002853394,
0.9109312891960144,

0.9109306931495667,
0.9109158515930176,
0.9108520746231079,
0.9108455181121826,
0.910816490650177,
0.9108135104179382,
0.9108116030693054,
0.9107810854911804,
0.9107710719108582,
0.9107588529586792,
0.9107470512390137,
0.9107386469841003,
0.9106753468513489,
0.9106715321540833,
0.9106548428535461,
0.9106125831604004,
0.9106017351150513,
0.9105846881866455,
0.9105732440948486,
0.91053307056427,
0.9105024337768555,
0.9104858040809631,
0.9104759097099304,
0.9104626178741455,
0.9103752970695496,
0.9103671312332153,
0.9103298783302307,
0.9103102684020996,
0.9102795720100403,
0.9102351069450378,
0.9102253317832947,
0.9101972579956055,
0.9101812243461609,
0.910133957862854,
0.9101228713989258,
0.9100795984268188,
0.9100674986839294,
0.9100671410560608,
0.9100283980369568,
0.9100187420845032,
0.9099968075752258,
0.9099831581115723,
0.9099786877632141,
0.9099640250205994,
0.9099572896957397,
0.9099557399749756,
0.9099423885345459,
0.9098937511444092,

0.90988689661026,
0.9098471999168396,
0.909838080406189,
0.9098339080810547,
0.909791111946106,
0.9097381234169006,
0.9097256064414978,
0.9096971750259399,
0.9096361994743347,
0.9095558524131775,
0.9095457792282104,
0.9094967246055603,
0.9094830751419067,
0.909475564956665,
0.909461498260498,
0.9094608426094055,
0.9094405174255371,
0.9093964695930481,
0.9093308448791504,
0.9093306064605713,
0.9092991352081299,
0.9092892408370972,
0.9092884659767151,
0.9092410206794739,
0.9091969132423401,
0.9091893434524536,
0.9091811180114746,
0.9091587066650391,
0.9091477394104004,
0.909145176410675,
0.9091301560401917,
0.9091214537620544,
0.9091127514839172,
0.9091054201126099,
0.9090896248817444,
0.9090861678123474,
0.9090449213981628,
0.9089995622634888,
0.9089981913566589,
0.9089719653129578,
0.9089412093162537,
0.9088950157165527,
0.9088403582572937,
0.9088262319564819,
0.908791184425354,
0.9087241291999817,
0.9087075591087341,
0.9086801409721375,

0.9086757898330688,
0.9086664915084839,
0.9086013436317444,
0.9085448980331421,
0.9085193276405334,
0.9085109233856201,
0.908463716506958,
0.9084558486938477,
0.9084052443504333,
0.9084019064903259,
0.9083937406539917,
0.9083852767944336,
0.9083685874938965,
0.908366858959198,
0.9083648324012756,
0.9083535671234131,
0.9083373546600342,
0.9083101749420166,
0.9082757234573364,
0.9082605242729187,
0.9082473516464233,
0.9082422852516174,
0.9082329273223877,
0.9082115292549133,
0.9082009792327881,
0.9081886410713196,
0.9081819653511047,
0.9081500768661499,
0.9081394672393799,
0.9081325531005859,
0.908123254776001,
0.9081203937530518,
0.908096194267273,
0.9080935120582581,
0.908092737197876,
0.9080673456192017,
0.9080559015274048,
0.9080156087875366,
0.9080103039741516,
0.9080085158348083,
0.9079779982566833,
0.9079715609550476,
0.9079399704933167,
0.907912015914917,
0.9078512191772461,
0.9078295230865479,
0.9078261256217957,
0.9078136682510376,

0.9078133702278137,
0.9077813029289246,
0.9077781438827515,
0.9077754616737366,
0.9077499508857727,
0.9077021479606628,
0.907689094543457,
0.9076604247093201,
0.9076290130615234,
0.9076123237609863,
0.9076106548309326,
0.9075859189033508,
0.9075731635093689,
0.907565176486969,
0.9075432419776917,
0.9075383543968201,
0.9075363874435425,
0.9075266718864441,
0.9075065851211548,
0.9074768424034119,
0.9074752330780029,
0.9074426889419556,
0.9074405431747437,
0.9074395298957825,
0.9074143171310425,
0.9073944687843323,
0.9073932766914368,
0.9073821902275085,
0.9073553085327148,
0.9072999358177185,
0.9072898626327515,
0.9072878956794739,
0.9072826504707336,
0.9072537422180176,
0.9072529673576355,
0.9072436094284058,
0.9072336554527283,
0.9072285294532776,
0.9071798324584961,
0.9070970416069031,
0.9070694446563721,
0.9070581793785095,
0.907042920589447,
0.9069456458091736,
0.9069395661354065,
0.9069361686706543,
0.9068624377250671,
0.906837522983551,

0.9068349599838257,
0.9067639112472534,
0.9066954851150513,
0.9066653251647949,
0.9066064357757568,
0.9065494537353516,
0.9065083265304565,
0.9065003395080566,
0.9064364433288574,
0.906427264213562,
0.9064124226570129,
0.9063889980316162,
0.9063791632652283,
0.9063729047775269,
0.906336784362793,
0.906312108039856,
0.9063088297843933,
0.9063042402267456,
0.9062886834144592,
0.9062639474868774,
0.9062449336051941,
0.9062238335609436,
0.9062143564224243,
0.9062024354934692,
0.9061718583106995,
0.9061501026153564,
0.9061465263366699,
0.9061225652694702,
0.9060785174369812,
0.9059544205665588,
0.9058605432510376,
0.9058056473731995,
0.9057973623275757,
0.9057948589324951,
0.9057757258415222,
0.9057525396347046,
0.9057475924491882,
0.9057087302207947,
0.9056482911109924,
0.9056476354598999,
0.905577540397644,
0.9055572748184204,
0.9055512547492981,
0.9055467247962952,
0.9055318236351013,
0.905519425868988,
0.9054822325706482,
0.9054752588272095,

0.9054647088050842,
0.9054409265518188,
0.905377984046936,
0.9053544998168945,
0.9053537845611572,
0.9053310751914978,
0.9052688479423523,
0.9052523374557495,
0.9052416086196899,
0.9052246809005737,
0.905209481716156,
0.905174970626831,
0.9051586985588074,
0.9051254987716675,
0.9051119685173035,
0.9051089286804199,
0.9050997495651245,
0.9050673842430115,
0.905056893825531,
0.9050372242927551,
0.9050306677818298,
0.9050127863883972,
0.9049676060676575,
0.9049610495567322,
0.9049397110939026,
0.9048972129821777,
0.9048826098442078,
0.9048821330070496,
0.9048692584037781,
0.9048022627830505,
0.9048011898994446,
0.9047881960868835,
0.9047663807868958,
0.9047538042068481,
0.9047517776489258,
0.9047389030456543,
0.9047288298606873,
0.9047268629074097,
0.9047085046768188,
0.9047048687934875,
0.904689371585846,
0.9046809077262878,
0.9046330451965332,
0.9046145677566528,
0.9045920968055725,
0.9045885801315308,
0.9045868515968323,
0.9045842885971069,

```
         0.9045805931091309,
         ...]

In [13]: doc_description = df['description'].tolist()
         sim_score_description = return_similarity_score(doc_description)
         doc_keyspec = df['keySpecsStr'].tolist()
         sim_score_keyspec = return_similarity_score(doc_keyspec)
         doc_detailspec = df['detailedSpecsStr'].tolist()
         sim_score_detailspec = return_similarity_score(doc_detailspec)
         doc_prodfam = df['productFamily'].tolist()
         sim_score_prodfam = return_similarity_score(doc_prodfam)

In [14]: avg_score_list = []
         for i in range(len(df)):
             sum_score = 0
             sum_score = sim_score_title[i] + sim_score_description[i] + sim_score_keyspec[i] +
             avg_score = sum_score / 5
             avg_score_list.append(avg_score)

In [15]: avg_score_list

Out[15]: [1.0,
          0.19913675151765348,
          0.12857295498251914,
          0.24324544370174409,
          0.291514178365469,
          0.2309822678565979,
          0.3559082821011543,
          0.17346120178699492,
          -0.07107783854007721,
          0.2975344300270081,
          0.25993516221642493,
          0.11861620284616947,
          0.36114214099943637,
          0.38421919345855715,
          0.015791203081607818,
          -0.15120927318930627,
          0.3144567418843508,
          0.3376820407807827,
          0.28393520414829254,
          0.18876098841428757,
          0.33565987199544906,
          0.3609116733074188,
          0.04419657401740551,
          0.1991051932796836,
          0.30343054942786696,
          0.39893912672996523,
          -0.058128141611814496,
          0.3418563693761826,
```

0.3148974165320396,
0.1370036207139492,
0.09239943586289882,
0.2136879812926054,
0.40556347388774155,
0.2590924009680748,
0.1299200788140297,
0.17243934869766236,
0.27137675806879996,
0.12001860151067376,
0.09272409677505493,
0.19870373010635375,
0.16775750182569027,
0.3101510763168335,
0.018730311840772628,
-0.10449664462357759,
0.24982584714889527,
0.3001817863434553,
0.18173222243785858,
0.39828359335660934,
0.2418445959687233,
-0.02193716838955879,
0.25784298926591875,
0.4963927701115608,
0.1716344863176346,
0.3724038615822792,
0.45609653419815005,
0.24533959925174714,
0.27810213482007384,
0.23628193326294422,
0.38584492907393725,
0.2059544265270233,
0.12130662053823471,
0.3954315036535263,
0.26980709955096244,
0.3801268197596073,
-0.09342062324285508,
0.21503434479236602,
0.31910487413406374,
0.15027576684951782,
0.2095564876217395,
0.31572669595479963,
0.20048188865184785,
0.3102779470384121,
-0.0423806831240654,
0.09797882437705993,
0.3033389613032341,
-0.0050334900617599486,

0.26393361687660216,
0.4134491704404354,
0.4458282008767128,
0.5275585647672415,
0.26666365638375283,
0.3290563106536865,
0.31802023500204085,
0.16467879191040993,
0.2125329539179802,
0.11183370817452669,
0.4603152900934219,
0.5154373943805695,
0.3304747447371483,
0.28395502716302873,
0.45514692589640615,
-0.14654012897517532,
-0.10418398603796959,
0.2695125237107277,
0.30115709602832796,
0.2822379507124424,
0.36934633143246176,
0.25628963112831116,
0.24243744760751723,
0.20904318988323212,
0.082955102622509,
-0.1666610401123762,
0.2748139724135399,
0.30703980177640916,
0.12688829824328424,
0.43398471027612684,
0.21880769804120065,
0.30086979568004607,
0.15441424278542398,
0.1729903608560562,
0.191536657884717,
0.26955424547195433,
0.12868901938199998,
0.2753732919692993,
0.07293345034122467,
0.11540439203381539,
0.12984811812639235,
0.003726760298013687,
0.14528195892926304,
0.4185972176492214,
0.12024370729923248,
0.21386129409074783,
0.16659774407744407,
0.30085557103157046,

0.16726185381412506,
-0.18044372983276843,
-0.14900858979672194,
0.28934362828731536,
0.12883732989430427,
0.09948769509792328,
0.11840068101882935,
0.16660062000155448,
0.409086287766669503,
0.4532257951796055,
0.3179732780903578,
0.1574319706647657,
0.11821730136871338,
0.10299918204545974,
0.27867211997509,
0.22064028829336166,
0.07515078485012054,
0.13313997015357018,
0.22220698595046998,
0.2133908063173294,
0.014206371060572565,
0.5026415344327688,
0.2891395315527916,
0.15975541137158872,
0.016193927108542994,
0.37640998885035515,
0.20683229938149453,
0.037135059013962746,
0.3318168604746461,
0.15244727432727814,
0.47796734916046263,
0.2564009092748165,
0.2413947841152549,
0.06923681639600546,
0.19757965877652167,
0.1091587096452713,
0.2102738119661808,
0.2987751178443432,
0.01661422662436962,
0.6132114961743355,
0.25556244775652887,
0.34001090824604036,
0.24248546827584505,
0.007765452563762665,
0.09025073535740376,
0.46263532526791096,
0.4714390397071838,
0.02541269138455391,

```
0.3194301245734096,
0.12783229053020478,
-0.21764302700757981,
0.31352812200784685,
0.45997163392603396,
0.39617414176464083,
0.11582090817391873,
0.12493790239095688,
0.3752253159880638,
0.2589384578168392,
0.4660921812057495,
0.23338648602366446,
-0.11361076347529889,
0.23099144063889981,
0.5007645159959793,
-0.09197647646069526,
0.24226887971162797,
0.18447392527014017,
0.20127000212669371,
-0.014058231934905052,
0.20035878121852874,
0.13351453915238382,
0.2529911160469055,
0.31411479115486146,
0.19923263490200044,
0.04446790218353271,
0.2502944961190224,
0.016311022639274596,
0.31398393101990224,
0.3580167431384325,
0.02269575744867325,
0.2593538288027048,
0.4725270189344883,
0.27694427967071533,
0.14739927644841372,
0.3223603626713157,
0.2042074203491211,
0.3281174188479781,
0.15706857196055352,
0.2944061689078808,
0.08474428802728654,
0.4726571589708328,
0.23580883741378783,
0.2973944524303079,
0.17332826182246208,
0.1638457365334034,
0.23797702491283418,
0.2986506339162588,
```

0.2538488134741783,
0.20992183592170477,
0.2659642405807972,
0.1719648467376828,
0.08538785800337792,
0.516775780916214,
0.16953496634960175,
0.185397769510746,
-0.03182473480701446,
0.24994323402643204,
0.01946440488100052,
0.46965244598686695,
0.2506353497505188,
0.323425629734993,
0.5398602768778801,
0.20551178976893425,
0.1518582402728498,
0.13583901897072792,
0.19058478623628616,
0.2639256536960602,
0.42439757883548734,
0.37798679396510126,
0.24975161999464035,
0.45731380246579645,
0.3076880142092705,
0.3745664071291685,
0.2787241511046886,
0.2944809928536415,
0.23579020649194718,
0.17177917025983333,
0.2216307818889618,
0.23820743672549724,
0.11783003583550453,
0.18735568039119244,
0.38170815706253053,
0.29122588112950326,
0.02310015484690666,
0.32397247701883314,
0.48175342017784717,
0.23902415484189987,
-0.12456174790859223,
0.12038919478654861,
0.08675627708435059,
0.28446077965199945,
0.1254773408174515,
0.1713464565575123,
0.07282816767692565,
0.34191477447748186,

0.054948809742927554,
0.1456108011305332,
0.16986779496073723,
0.33039484918117523,
0.29843097254633905,
0.12642333507537842,
0.5324118800461293,
0.3079554557800293,
0.2763350196182728,
-0.023572185635566713,
0.193035451695323,
0.38678216338157656,
-0.010901418328285218,
0.1701043765991926,
0.3434653222560883,
0.49352007657289504,
0.06504548117518424,
0.31344434767961504,
0.16858696490526198,
0.25067590475082396,
0.2941296145319939,
0.07572659850120544,
0.24848915804177524,
-0.1320868283510208,
0.09231055825948715,
0.2765450018458068,
0.4041174784302711,
0.48361451700329783,
-0.003763715922832489,
0.0839573860168457,
0.2538516469299793,
-0.10714685916900635,
0.11339122019708156,
0.39496003091335297,
0.28580953925848007,
0.31127038449048994,
-0.10930895507335663,
0.2208692230284214,
0.18433509171009063,
0.4676122710108757,
0.3271852761507034,
-0.03113679736852646,
0.5316492527723312,
0.08043301310390234,
0.40848452746868136,
0.24532470703125,
0.4223471138626337,
0.03492995202541351,

```
0.2908198952674866,
0.18746754191815854,
0.49591285437345506,
0.512407181598246,
0.03377473503351212,
0.18656983710825442,
0.4621160373091698,
0.26375741040101275,
0.5366614993195981,
0.3709502339363098,
0.3654743280261755,
0.2250356527045369,
0.4041674127802253,
0.2630366124212742,
0.11180728860199451,
0.3229347500950098,
0.12654227614402772,
0.20289764357730747,
0.3169640928506851,
0.32785936295986173,
0.13136006051208823,
0.3841835618019104,
-0.024800701439380644,
0.1511538341641426,
0.24352358281612396,
0.48944675233215096,
0.24205513373017312,
0.05752951733302325,
0.16096094399690627,
0.44004314839839936,
0.27880155146121977,
0.424661860242486,
0.2683677840977907,
0.250798861682415,
0.2951378580182791,
0.37142764776945114,
-0.07477244138717651,
0.18086830899119377,
0.21525496542453765,
0.2223628457635641,
0.24160640146583318,
0.17388578951358796,
0.41252539604902266,
0.440320660173893,
0.19572151750326156,
0.2587583899497986,
0.23350099958479403,
0.2919198237359524,
```

```
0.5190992474555969,
0.5142935141921043,
0.2422141835093498,
0.11048953086137772,
0.17334821373224257,
0.16122352331876755,
0.27429221719503405,
0.48808740079402924,
0.08802015855908393,
0.3835040658712387,
0.133095396310091,
0.16087833195924758,
0.31682249158620834,
0.13145280466414988,
0.11342756450176239,
0.2611101850867271,
-0.04170744102448225,
0.31151740774512293,
-0.009573417343199253,
0.4356417268514633,
0.3088370367884636,
0.29555292762815955,
0.15250681936740876,
0.2587291955947876,
0.17864602580666541,
0.5412735879421234,
0.3057147413492203,
0.2338263601064682,
0.2708068147301674,
0.09662016779184342,
0.5101632624864578,
0.27162343114614484,
0.3971876010298729,
0.23241190761327743,
0.21923334952443838,
0.14226463586091995,
0.33569849490304476,
0.3661908507347107,
0.024347954988479616,
0.021974488347768783,
0.22725395262241363,
0.2697632139548659,
0.26756596118211745,
0.338502349704504,
0.04940065033733846,
0.3459766566753387,
0.18629901260137557,
-0.028999101743102072,
```

0.2705601967871189,
0.09206187073141336,
-0.01333445981144905,
0.14063862711191177,
0.4614193938672543,
0.24229488223791124,
0.2766347900032997,
-0.0457403264939785,
0.2081921875476837,
0.28103465735912325,
0.47138278409838674,
0.14709700644016266,
0.3416199222207069,
0.13262753784656525,
0.3394358068704605,
0.08953337036073208,
0.1612906173802912,
0.38903399407863615,
0.19549489540513604,
0.300839776545763,
0.19907813146710396,
-0.16643818616867065,
0.26397968605160715,
0.21192058064043523,
0.09433678686618804,
0.20624006018042565,
0.26456473469734193,
0.49161956608295443,
0.04306457638740539,
0.38082313537597656,
0.3763814002275467,
0.41676075661089274,
0.027799104899168016,
0.19948662519454957,
0.296858243457973,
0.3348547473549843,
0.3260485500097275,
0.4073137910454534,
0.06024307367042638,
0.42180445715785025,
0.2133688010275364,
0.17814458161592484,
0.3633943647146225,
0.11405256688594818,
0.34995184019207953,
0.35746739283204076,
0.2960520401597023,
0.5882265865802765,

```
0.3909260492771864,
-0.058336713910102846,
0.13780149444937706,
0.5835372984409333,
0.3762403279542923,
0.27213563583791256,
0.2954206317663193,
0.23680890444666147,
0.25434756129980085,
0.4120491147041321,
0.31412757784128187,
0.28335938006639483,
0.47062764316797256,
0.41630774885416033,
0.41035648062825203,
0.3729512669146061,
0.21349132359027861,
0.29527326226234435,
0.40460666120052335,
0.32788021191954614,
0.23417479498311877,
0.381354851461947,
0.39094214215874673,
0.32610267363488676,
0.40381753370165824,
0.25460513681173325,
0.45199990570545195,
0.4880660682916641,
0.41591191589832305,
0.18921000547707081,
0.358243003487587,
0.12119080722332001,
0.43573653846979143,
0.22595964213833214,
0.39695952348411084,
0.2883284881711006,
0.30306842774152754,
0.3232507824897766,
0.29676816016435625,
0.40664413571357727,
0.37794584929943087,
0.5175892129540444,
0.25742766773328185,
0.07446999251842498,
0.4251264877617359,
0.3165603634901345,
0.2411881074309349,
0.34898578822612764,
```

```
0.2958631794899702,
0.13066391311585904,
0.20971569567918777,
0.43689636029303075,
0.3839064136147499,
0.425798943429254,
0.3280686289072037,
0.5717117629945279,
0.08058438301086426,
0.2416948825120926,
0.5910181906074286,
0.4605680324137211,
0.487580743432045,
0.24106912249699236,
0.4654033824801445,
0.3849650837481022,
0.08214296698570252,
0.12388753890991211,
0.2484528750181198,
0.37450483739376067,
0.3707695215009153,
0.2742236077785492,
0.3965903133153915,
0.1283604197204113,
0.3554527997970581,
0.09232774674892426,
0.3625123739242554,
0.3221516191959381,
0.232353812456131,
0.056509120762348174,
0.4978400141000748,
0.11244948692619801,
0.105010139197111114,
0.4608433127403259,
0.3819480836391449,
0.41447146479040387,
0.5421996712684631,
0.042155236657708886,
0.38713412582874296,
0.3728835374116898,
0.26726412400603294,
0.43059664964675903,
0.44163842499256134,
0.4980440676212311,
0.4462211951613426,
0.47124897725880144,
0.11478526555001736,
0.21536660343408584,
```

0.33728061318397523,
0.29682046100497245,
0.3053234666585922,
0.30157585218548777,
0.31340207941830156,
0.29975806176662445,
0.5437451347708702,
0.2819452464580536,
0.31531585231423376,
0.392506193369627,
0.31135825738310813,
0.30028539448976516,
0.27145847789943217,
0.47115197628736494,
0.48514789938926695,
0.07531520277261734,
0.3812735378742218,
0.3267027661204338,
0.5045906594023108,
0.44246949851512907,
0.3625956892967224,
0.27900235503911974,
0.21276965588331223,
0.4936488509178162,
0.4092737875878811,
0.30490435026586055,
0.1491052806377411,
0.15803986489772798,
0.41931881457567216,
0.26822039783000945,
0.19611361473798752,
0.29822735786437987,
0.42868283540010454,
0.5069260910153389,
0.21596418768167497,
0.37327165007591245,
0.539303581416607,
0.5512010529637337,
0.23545323507860302,
0.44221668012100457,
0.10722513124346733,
0.510884191095829,
0.3999801414087415,
0.36224065348505974,
-0.0933161661028862,
0.27620361745357513,
0.4297589886933565,
0.42394666075706483,

0.3432230852544308,
0.34582240767776967,
0.10707534104585648,
0.5041838705539703,
0.4291947349905968,
0.3059893697500229,
0.336315780878067,
0.1868813470005989,
0.1803976982831955,
0.4794867470860481,
0.25503038354218005,
0.4003763157874346,
0.26175096035003664,
0.3540259301662445,
0.4202642798423767,
0.4993043467402458,
0.504356899112463,
0.03357137888669968,
0.4211313173174858,
0.3874327950179577,
0.38834119141101836,
0.18420207239687442,
0.3637408211827278,
0.5165135076269507,
0.2961579531431198,
0.21306165531277657,
0.17998599112033845,
0.38646352253854277,
0.32788882665336133,
0.34976314175873996,
0.4654170244932175,
0.19957820177078248,
0.31841666251420975,
0.3143704215064645,
0.24390481263399125,
0.14671863615512848,
0.29162505865097044,
0.465476905554533,
0.3645729884505272,
0.39544016420841216,
0.45105685740709306,
0.14802867844700812,
0.4620740432292223,
0.24631503373384475,
0.5170058816671371,
0.15822397097945212,
0.16674095969647168,
0.2871920123696327,

```
0.5551360279321671,
0.2219694998115301,
0.3825542524456978,
0.5006986021995544,
0.39425877928733827,
0.3567972034215927,
0.5270757377147675,
0.35734503213316204,
0.1270160975633189,
0.6089782666414976,
0.21841152794659138,
0.301883402466774,
0.43017353117465973,
-0.13406797125935555,
0.5542268499732017,
0.44971439689397813,
0.2800620751455426,
0.302654168009758,
0.38682558536529543,
0.229554183781147,
0.5433880478143692,
0.38649145513772964,
0.377520246617496,
0.17337820939719678,
0.2672402017749846,
0.3236890811473131,
0.16411082930862902,
0.24656708240509034,
0.20467095971107482,
0.460347056388855,
0.4596671048551798,
0.34916461147367955,
0.42459039725363257,
0.5775296330451966,
0.44382154047489164,
0.19388314336538315,
0.613528487086296,
0.4883438564836979,
0.4401578851044178,
0.41155912578105924,
0.08182897530496121,
0.13689967393875122,
0.323658518306911,
0.5025135464966297,
0.42924417555332184,
0.24612710713408886,
0.13733540773391723,
0.2737923016771674,
```

```
0.3735650047659874,
0.21137070953845977,
-0.08120909035205841,
0.41687509417533875,
0.47750569358468054,
0.15949537567794322,
0.21686351746320726,
0.21490555442869663,
0.44179933247651204,
0.3303563825786114,
0.16316613499075175,
0.5485476464033127,
0.42789407819509506,
0.3251831393688917,
0.44218413531780243,
0.3338965490460396,
0.3726653873920441,
0.39412225484848024,
0.45345921590924265,
0.38347976729273797,
0.3018071368336678,
0.2553625889122486,
0.20682992991060017,
0.4657163016498089,
0.29460245221853254,
0.371944010025772095,
0.05442440807819367,
0.4095657140016556,
0.286845950782299,
0.3621041439473629,
0.16491157785058022,
0.23034125566482544,
0.42258172333240507,
0.6062886215746403,
0.2855738289654255,
0.32170806676149366,
0.560184794664383,
0.5277637660503387,
0.4642560213804245,
0.34419227540493014,
0.33889293372631074,
0.42805148363113404,
0.382175150513649,
0.3424639978446066,
0.18114826530218125,
0.23394685946404933,
0.6262893617153168,
0.30855868831276895,
```

0.4458391427993774,
0.21227820813655854,
0.2672091191634536,
0.31833242485299706,
0.17041270984336734,
0.4766113795340061,
0.3994746521115303,
0.3230139374732971,
0.5455854773521424,
0.1983478520065546,
0.31076186150312424,
0.04992940947413445,
0.13619330562651158,
0.25662637427449225,
0.2889593843370676,
0.39502236172556876,
0.2854122593998909,
0.5365651473402977,
0.17331977486610411,
0.39261443465948104,
-0.17334943413734435,
0.4904042512178421,
0.39896996282041075,
0.42298159301280974,
0.34306780695915223,
0.3710155636072159,
0.4370101767592132,
0.343238091468811,
0.3898003689944744,
0.563551265001297,
0.4121967479586601,
0.2716474112123251,
0.11537655144929886,
0.4270157949998975,
0.34263926148414614,
0.3297884237021208,
0.12197951227426529,
0.3837335668504238,
0.3376162819564342,
0.5623391836881637,
0.2351723112165928,
0.4541350953280926,
0.410426814109087,
0.3996345944702625,
0.41267878711223605,
0.3134020671248436,
0.36354036927223204,
0.5004033774137497,

0.4180078461766243,
0.16427032873034478,
0.41321050226688383,
0.20248538148589432,
0.26583756376057865,
0.44345848709344865,
0.3625927772372961,
0.2783931016921997,
0.34795658886432645,
0.2557650238275528,
0.21353365033864974,
0.150784283131361,
0.4234680965542793,
0.36502342261374,
-0.03531100861728191,
0.5133937269449234,
0.2437156058847904,
0.23911560121923686,
-0.015602895617485046,
0.4570554792881012,
0.08956515491008758,
0.5217036575078964,
0.4551774986088276,
0.3362005054950714,
0.31846603453159333,
0.14130393117666246,
0.4879369601607323,
0.30842492915689945,
0.34256661087274554,
0.21983017176389694,
0.4484895788133144,
0.37262679934501647,
0.3169735597446561,
0.358855427056551,
0.5098572127521038,
0.2875761419534683,
0.0768539659678936,
0.22596026211977005,
0.40733070150017736,
0.5645564839243888,
0.4896609663963318,
0.3626783177256584,
-0.2587976653128862,
0.22521358542144299,
0.3268108174204826,
0.060369497537612914,
0.04065467119216919,
0.09937111716717481,

-0.006749372184276581,
0.37168197631835936,
0.13311538994312286,
0.5676994651556015,
0.19729661121964454,
0.2091612622141838,
0.4686545252799988,
0.22115673199296,
0.36470712497830393,
0.043819546699523926,
0.2957609124481678,
0.335616309940815,
0.20083045400679111,
0.36957547552883624,
0.3426987253129482,
0.3078825056552887,
0.35367412865161896,
0.3959581792354584,
0.09400088042020797,
0.3997384667396545,
0.47196958437561987,
0.28950536660850046,
0.327714154869318,
0.410525231808424,
0.013030168414115906,
0.296468049287796,
0.5493047326803208,
0.26353254467248916,
0.2668274898082018,
0.2522467762231827,
0.4406699866056442,
0.0811478391289711,
0.3958885818719864,
0.27498829886317255,
0.4687125638127327,
0.20476676076650618,
-0.0013720318675041199,
0.3438017715699971,
0.26506808325648307,
0.1902645096182823,
0.25879150032997134,
0.17392450012266636,
0.3348876442760229,
0.20787022169679403,
0.4222490020096302,
0.5113186150789261,
0.4281776774674654,
0.0854182243347168,

```
0.4547800192609429,
0.26505443304777143,
0.1795368582010269,
0.5563403159379959,
0.46289356220513583,
0.1797214314341545,
0.054234142601490024,
0.3975692528299987,
0.5377436727285385,
0.4727750382386148,
0.3037240780889988,
0.4892115067690611,
0.31805808544158937,
0.31719285640865563,
0.4111985430121422,
0.13158652633428575,
0.22440101057291031,
0.4715019389986992,
0.39387743175029755,
0.11870383769273758,
0.30805050656199456,
0.033772974833846094,
0.40896979104727504,
0.089715570025146,
0.5621355965733528,
0.38047391921281815,
0.3731765434145927,
0.1447601795196533,
0.23194478303194047,
0.31074223294854164,
0.11450013518333435,
0.22581266313791276,
0.403629707545042,
0.3260642122477293,
0.1561926607042551,
0.37800917997956274,
0.4069990985095501,
0.309877073764801,
0.2587543301284313,
0.28688941076397895,
0.08548336140811444,
0.4003660866059363,
0.45698135029524567,
0.383116140589118,
0.31609964836388826,
0.3295352332293987,
0.43736740564927457,
0.08578020296990871,
```

```
0.27783431112766266,
0.3603551104664803,
0.4564946010708809,
0.14293716996908187,
0.3519990175962448,
0.25179566740989684,
0.2114116832613945,
0.6497033791150898,
0.20927879195660354,
0.14173804111778737,
0.3857002711854875,
0.1136973237618804,
0.5485957257449627,
0.19437639117240907,
0.3750911369919777,
0.45045451521873475,
0.4756422214210033,
0.3088750779628754,
0.40423843301832674,
0.5501191213726997,
0.4591722421348095,
0.07897912096232176,
0.2946118667721748,
0.24916223101317883,
0.38902320712804794,
0.5177069500088691,
0.2503184570930898,
0.13334644436836243,
0.2641759693622589,
0.29409386664628984,
0.3624126374721527,
0.3485656308475882,
0.4016115993261337,
0.17958231903612615,
0.44833843410015106,
-0.06800356805324555,
0.233811704069376,
0.5976150691509247,
0.38851225227117536,
0.5068334192037582,
0.24375160336494445,
0.1586824268102646,
0.4113643883960322,
0.423284912109375,
0.39430648684501646,
0.2645902572199702,
0.4908867634832859,
0.4593303516507149,
```

```
0.28320488668978216,
0.14898979812860488,
0.5255038847681135,
-0.05825780928134918,
0.5675791352987289,
0.10507981777191162,
0.18630860820412637,
0.41451741605997083,
0.45020828396081924,
0.5082612411119044,
0.20658182874321937,
0.01581605914980173,
...]
```

In the above code snipet, I have equally weighted the all five features, but it can be weighted according to importance. Also I have calculated all these 5 feature vectors cosinie similarity for only the zeroth product. It should be the covariance matrix. The reason for not calculating matrix is the time taken to calculate 40k once is nearly 20 seconds. To calculate matrix, it will be taking 20* 40000 = 800000 seconds.

The model is not performing much well as I have neglected more than nearly 94% dataset because of system issues. System was not able to handle such a huge dataset. If model is run on a good system, it can give promising output.

2.1 Additional Work

If it was a supervised problem, i.e., if we have labelled dataset, we could train a classification model, like neural network on these word2vec feature vectors separately and can combine their result. Apart from these five features, prices also play a very significant role in determining the similar products. On a very particular subcategory, generally prices form a normal distribution, so we can try to find the outliers available in the dataset using various statistical testing, like chi-square test and after finding the outliers can remove them. Features like colors, size also play a significant role, and they are part of categorical data. If two products fall into same category can help in narrowing our search and if different categories can help in reject them. Apart from all these features, one very important feature was the image of product. Since, these imageUrls are showing error 400, if it was accessible one can train the model to identify similar looking products.