library(tm)

setwd("C:/Users/lohit/Desktop/course/KDD/FinalProject/diabetes/")

text\_corpus<-Corpus(DirSource("diabetes/"))

text\_corpus <- tm\_map(text\_corpus, stripWhitespace)

text\_corpus <- tm\_map(text\_corpus, content\_transformer(tolower))

text\_corpus <- tm\_map(text\_corpus, removeWords, stopwords("english"))

text\_corpus <- tm\_map(text\_corpus, removePunctuation)

dtm3 <-DocumentTermMatrix(text\_corpus, control=list(wordLengths=c(4, 20), bounds = list(global = c(1,Inf))))

dtm2 <- removeSparseTerms(dtm, sparse=0.95)

m3<-as.matrix(dtm3)

memory.limit(size=180000)

df3<-as.data.frame(m3)

entropy <- function(p) {

# Assumes: p is a numeric vector

if (sum(p) == 0) {

return(0)

}

p <- p/sum(p) # Normalize so it sums to 1

p <- p[p > 0] # Discard zero entries (because 0 log 0 = 0)

H = -sum(p\*log(p,base=2))

return(H)

}

|  |
| --- |
| maj<-df3[,"testicular"]  > entropy(maj)  [1] 2.641604  > entropy(table(maj))  [1] 0.01036369  > dib<-df3[,"diabetes"]  > entropy(dib)  [1] 12.787  > entropy(table(dib))  [1] 4.615899 |
|  |
| |  | | --- | |  | |

library(infotheo)

mutinformation(maj,dib,method="emp")

# identify those words with maximum MI with a particular word

find\_MI<-function(word,dataframe){

word\_dis<-dataframe[,word]

lexicon <- colnames(dataframe)

lexicon <- setdiff(lexicon,"label")

vocab.size = length(lexicon)

mis <- matrix(0,nrow=vocab.size,ncol=1)

rownames(mis) = lexicon

for (i in 1:vocab.size) {

word2\_dis<-dataframe[,lexicon[i]]

mi<-mutinformation(word\_dis,word2\_dis,method="emp")

mis[i,1]<-mi

}

return(mis)

}

Di<-find\_MI("diabetes",df3)

Di

write.csv(Di, file = "MyData.csv")

testicularrow\_numbers<-order(Di[,1],decreasing = TRUE)

head(Di[row\_numbers,],10)

tail(Di[row\_numbers,],10)

> Di<-find\_MI("diabetes",df3)

> row\_numbers<-order(Di[,1],decreasing = TRUE)

> head(Di[row\_numbers,],10)

diabetes type people care complications insulin million

3.19949705 0.31358421 0.16466932 0.10745795 0.10145380 0.09519288 0.09160672

association patients blood

0.09027007 0.08862078 0.08603939

> tail(Di[row\_numbers,],10)

1252 1293 1309 15261532 6210 auxology caranti centile maso

0.000214292 0.000214292 0.000214292 0.000214292 0.000214292 0.000214292 0.000214292 0.000214292 0.000214292

sartorio

0.000214292

[Jupyter Notebook](http://localhost:8888/tree)

LogoutUntitled Last Checkpoint: 3 hours ago (autosaved)

Python 3

Trusted

* [File](http://localhost:8888/notebooks/Desktop/course/KDD/Untitled.ipynb)
* [Edit](http://localhost:8888/notebooks/Desktop/course/KDD/Untitled.ipynb)
* [View](http://localhost:8888/notebooks/Desktop/course/KDD/Untitled.ipynb)
* [Insert](http://localhost:8888/notebooks/Desktop/course/KDD/Untitled.ipynb)
* [Cell](http://localhost:8888/notebooks/Desktop/course/KDD/Untitled.ipynb)
* [Kernel](http://localhost:8888/notebooks/Desktop/course/KDD/Untitled.ipynb)
* [Widgets](http://localhost:8888/notebooks/Desktop/course/KDD/Untitled.ipynb)
* [Help](http://localhost:8888/notebooks/Desktop/course/KDD/Untitled.ipynb)



In [237]:



*#Gensim accomplishes this by taking a corpus, a collection of text documents, and producing a vector representation of*

*#the text in the corpus. The vector representation can then be used to train a model, which is an algorithm to create*

*#different representations of the data, which are usually more semantic. These three concepts are key to understanding*

*#how Gensim works. At the same time, we'll work through a simple example that illustrates each of them.*

​

*#A corpus is a collection of digital documents. This corpus is fed to Gensim from which it will infer the structure of the*

*#documents and extract topics from the documents. Once the algorithm learns on how to infer topics from the training corpus,*

*#it can be used to assign topics to new documents which were not present in the training corpus.For this reason, we also refer*

*#to this collection as the training corpus. No human intervention is required - the topic classification is unsupervised.*

​

*# Reading whole corpus from the dibates and make a big corpus*

​

*## Using glob we can read and write*

In [ ]:



**import** io

**import** glob

​

a = open("output.txt",'a',encoding="utf-8")

counter = 0

*#var = ""*

list\_of\_files = glob.glob(r"C:\Users\lohit\Desktop\course\KDD\FinalProject\diabetes\diabetes\\*.txt") *# create the list of file*

**for** file\_name **in** list\_of\_files:

FI = open(file\_name, 'r',encoding="utf-8")

*#print(FI.read())*

**for** line **in** FI:

a.write(line)

*#print(line)*

counter = counter **+** 1

*#for line in FI:*

*# print(line)*

​

*#print(FI)*

FI.close()

*#FO.close()*

​

In [ ]:



*# We first remove all the words the commonly used English words - called stop words such as 'the', ‘a’, ‘we’, etc.) and words*

*# that occur only once in the corpus.*

*# Second we are counting number of different words inside the document and if the count is more than one it will keep on adding*

*# frequency and calcuate the total frequency of all words.*

*# Now we are take the words which are more than 0 time inside the corpus.*

*# After that we are printing the words.*

*# Using NLTK library we are calculating the operation*

In [ ]:



*#Using NLTK to appedning the whole in single file*

*# do not tun this part take a lot of time.*

**from** nltk.corpus **import** stopwords

**from** nltk.tokenize **import** word\_tokenize

*#word\_tokenize accepts a string as an input, not a file.*

stop\_words = set(stopwords.words('english'))

file1 = open(r"C:\Users\lohit\Desktop\course\KDD\output.txt",'r',encoding='utf-8')

line = file1.read()*# Use this to read file content as a stream:*

words = line.split()

**for** r **in** words:

**if** **not** r **in** stop\_words:

appendFile = open('filteredtext.txt','a',encoding='utf-8')

appendFile.write(" "**+**r)

In [1]:



*# Remove stop words from the corpus using NLTK*

*# Remove puncuations from the corpus using NLTK*

*# Remove lemma words from the Corpus usinf NLTK*

**from** nltk.corpus **import** stopwords

**from** nltk.stem.wordnet **import** WordNetLemmatizer

**import** string

stop = set(stopwords.words('english'))

exclude = set(string.punctuation)

lemma = WordNetLemmatizer()

**def** clean(doc):

*#print(doc)*

stop\_free = " ".join([i **for** i **in** doc.lower().split() **if** i **not** **in** stop])

punc\_free = ''.join(ch **for** ch **in** stop\_free **if** ch **not** **in** exclude)

normalized = " ".join(lemma.lemmatize(word) **for** word **in** punc\_free.split())

**return** normalized

​

document\_text = open(r"C:\Users\lohit\Desktop\course\KDD\output.txt",'r', encoding='utf-8')

x = document\_text.read()

doc\_clean = [clean(doc).split() **for** doc **in** x.split("\n")]

*#print(doc\_clean)*

In [4]:



*# Importing Gensim*

**import** gensim

**from** gensim **import** corpora

​

*# Creating the term dictionary of our courpus, where every unique term is assigned an index.*

dictionary = corpora.Dictionary(doc\_clean)

​

*# Converting list of documents (corpus) into Document Term Matrix using dictionary prepared above.*

doc\_term\_matrix = [dictionary.doc2bow(doc) **for** doc **in** doc\_clean]

In [2]:



*#Calculating the word frequency from the corpus and removing some of them*

**import** re

**import** string

dictionary\_arr=[]

document\_text = open(r"C:\Users\lohit\Desktop\course\KDD\output.txt",'r',encoding = 'utf-8')

x = document\_text.read()

​

*#text\_string = x.split()*

*# Count word frequencies*

**from** collections **import** defaultdict

frequency = defaultdict(int)

**for** text **in** x.split():

**for** token **in** text.split():

frequency[token] += 1

​

*#for words in x.lower().split('\n'):*

*# dictionary\_arr.append(words)*

processed\_corpus = [[token **for** token **in** text.split() **if** frequency[token] **>** 0] **for** text **in** x.split('\n')]

*#processed\_corpus*

*#processed\_corpus*

​

*#doc\_clean = [clean(doc).split() for doc in dictionary\_arr]*

In [ ]:



*#We now need to tokenize our data. This breaks the documents into words and assigns tokens: unique numbers to the words that*

*#have been repeated more than “x” times. Thus, we associate each word in the corpus with a unique integer ID. We can do this*

*#using the Gensim.corpora.Dictionary class. This dictionary defines the vocabulary of all words that our processing knows about.*

In [3]:



**from** gensim **import** corpora

dictionary = corpora.Dictionary(processed\_corpus)

dictionary.save('C:/Users/lohit/AppData/Local/Temp/den.dict')

*#print(dictionary)*

C:\Users\lohit\Anaconda3\lib\site-packages\gensim\utils.py:862: UserWarning: detected Windows; aliasing chunkize to chunkize\_serial

warnings.warn("detected Windows; aliasing chunkize to chunkize\_serial")

In [40]:



print(dictionary.token2id)

IOPub data rate exceeded.

The notebook server will temporarily stop sending output

to the client in order to avoid crashing it.

To change this limit, set the config variable

`--NotebookApp.iopub\_data\_rate\_limit`.

In [4]:



bow\_corpus = [dictionary.doc2bow(text) **for** text **in** processed\_corpus]

corpora.MmCorpus.serialize('C:/Users/lohit/AppData/Local/Temp/den.mm', bow\_corpus)

*#bow\_corpus*

In [ ]:



*#Next, we need to represent the documents mathematically to be able to continue further processing, so we represent each*

*#document as a vector. We use the bag-of-words model where each document is represented by a vector containing the frequency*

*#counts of each word in the dictionary. The length of the vector is the number of entries in the dictionary. One of the main*

*#properties of the bag-of-words model is that it completely ignores the order of the tokens in the document that is encoded,*

*#hence bag-of-words.*

*#The first entry in each tuple corresponds to the ID of the token in the dictionary, the second corresponds to the count*

*#of this token. Now changing the whole courpus into vector form.*

*#We are saving this into our temporary folder called with an extension of “.mm”. Note that while this list lives entirely*

*#in memory, in most applications you will want a more scalable solution.*

*#Now that we have vectorized our corpus we can begin to transform it using models. We use model as an abstract term*

*#referring to a transformation from one document representation to another. In Gensim documents are represented as vectors*

*#so a model can be thought of as a transformation between two vector spaces. The details of this transformation are learned*

*#from the training corpus. One simple example of a model is tf-idf. The tf-idf model transforms vectors from the bag-of-words*

*#representation to a vector space where the frequency counts are weighted according to the relative rarity of each word in*

*#the corpus.*

*#Term Frequency, Inverse Document Frequency (TF-IDF) TF-IDF is a way to score the importance of words (or "terms") in a*

*#document based on how frequently they appear across multiple documents.*

*#If a word appears frequently in a document, it's important and TF-IDF gives the word a high score.*

*#But if a word appears in many documents, then it's not a unique identifier and gives the word a low score.*

*#Therefore, stop words will be scaled down. Words that appear frequently in a single document will be scaled up.*

*#For a term t in a document d, the weight Wt, d of term t in document d is given by:*

*#Wt,d = TFt,d log (N/DFt )*

*#Where,*

*#TFt,d is the number of occurrences of t in document d.*

*#DFt is the number of documents containing term t.*

*#N is the total number of documents in corpus.*

In [5]:



**from** gensim **import** models

*# train the model*

tfidf = models.TfidfModel(bow\_corpus)

*# transform the "system minors" string*

*#top 10 frequrn*

print(tfidf[dictionary.doc2bow("diabetes patients type study insulin blood risk disease health research".lower().split())])

​

​

print(tfidf[dictionary.doc2bow("come opportunity supporting publication involving millions rapidly suggesting ".lower().split())])

[(53, 0.3492563305587527), (60, 0.22507556354736905), (62, 0.3512624990218378), (212, 0.30059628126992954), (394, 0.31066264090650114), (405, 0.3248702030107599), (431, 0.316902484435025), (557, 0.3622871041934056), (636, 0.3007640397348136), (712, 0.29904426990547595)]

[(204, 0.34994353712370985), (882, 0.34918760631019247), (3230, 0.3528768841850399), (3346, 0.34994353712370985), (4171, 0.35328037345900437), (4217, 0.3648741288092756), (4905, 0.358821439314328), (6213, 0.34918760631019247)]

In [6]:



putx = tfidf[dictionary.doc2bow("diabetes patients type study insulin blood risk disease health research people cells glucose said also levels researchers will treatment university care clinical medical title heart control high percent sugar found diabetic results studies body medicine obesity weight cell years group drug american first association data development information women professor center diseases increased time cardiovascular mice used associated children function help important including national year factors cancer million findings diet published metabolic well says system many journal compared effects lead based patient team human therapy school among trial higher developing institute http long complications according three googletag kidney beta cause drugs protein increase known company lower adults related need using however work participants life food significant world loss less program source potential resistance number management healthy develop hospital liver population better scientists pressure author exercise chronic improve metformin without activity phase likely authors changes even early term like reduce available called prevention trials individuals obese condition test states leading common role showed gene make prevent effect pancreas death part support provide effective cholesterol immune united normal conditions evidence reported groups metabolism include surgery whether developed treated public major different looking future genetic forward healthcare diagnosed statements level much daily quality events director report problems current safety shown lifestyle addition products americans within stroke take failure colleagues genes article significantly issue analysis scientific physical greater reduced placebo international brain currently treatments tissue technology primary taking might result additional treat often education bottom syndrome display dose oral small energy similar average response push months department able physicians form hormone combination risks today week possible impact 600xflex four linked online ability specific outcomes benefits rates foundation hba1c process large suggest release product proteins diabetics received medications around several target change show identified good reduction growth inflammation five provides improved society science just rate added factor college damage serious worldwide symptoms benefit countries causes controlled global family cases therapeutic production overweight produce needed involved made older since recent although times conducted period previous general another presented humans pancreatic community know every state european increasing weeks visit attack find therapies cost molecular programs producing monitoring pain living already severe services shows medication person least given either approximately middle suggests nutrition annual following glycemic total overall approach prevalence follow understanding screening sciences action incidence access muscle diagnosis become model hypoglycemia nearly included president half identify efficacy discovery 2007 designed 2008 senior across side costs disorders discovered novel self intake hypertension young subjects still stress working adverse increases estimated best anti treating delivery fatty receptor demonstrated onset previously active determine biology eating next approved finding adult mortality centers certain inflammatory manage vessels institutes mellitus amount needs link recently multiple standard range foods gain paper understand affect pharmaceutical coronary particularly baseline laboratory review reducing free mass project meeting announced funded together longer alone individual testing market receive fasting vision intervention caused 2010 survey class news guidelines administration practice index critical california systems tests types randomized affects measure white single autoimmune second billion molecules making differences lowering pharmaceuticals dietary month associate growing affected mechanism step especially chief days regulatory problem observed history present sensitivity enough nerve 2009 press mean mechanisms example studied blindness investigator doctors potentially lives interventions vascular cure experience elevated contact animals taken skin stage 2006 believe animal continue positive meal harvard produced earlier last mouse fact goal medicines difference progression seen injections order doctor recommended ways later point hope provided division hours aged expected funding appears advert poor case endocrinology largest third members therefore must improving biological expression improvement enzyme look complex includes pathway injection focused established decrease short limited play professionals models decreased considered receiving required reports physician plan focus achieve service rise areas cellular epidemic investigators full importance organization europe training tested preventing signaling area throughout others internal populations lack doses occurs regular contribute helps supported measured strategies targets june lipid molecule advanced almost clinic moderate chemical providing safe measures assistant leads start secretion tissues 2005 nature followed independent address examined efforts approval home place canada site collaboration light live excess whose past awareness boston despite thus particular natural issues means innovative open directly clear knowledge centre written keep deaths various double achieved companies grant recommendations unique whole improvements country experts offers genetics existing makes impaired offer york versus initial samples tolerance beneficial little post commonly explained researcher wide specifically appropriate works method field rather nation usually direct please latest member induced relationship attacks david regarding central strategy experienced highest design smoking appear main date occur responsible 2011 explains effectiveness simple america plans strong performance promote highly real twice call successful along environmental looked network difficult pathways implications require policy status regulate back leader effectively turn complete government 2004 profile resources maintain dependent thought environment controlling prior indicate assess right suggested approaches size line rapid executive stop protect chair managing combined advance took great controls progress allow regulation helping evaluated processes methods industry analyzed seven vice actual give clinicians value done noted promising burden normally investigation influence underlying fewer agents challenge reduces michael signs raise ongoing resulting eight officer completed evaluate explain notes activities essential performed success comprehensive improves generally larger adds monitor necessary want properly causing march suffer estimates create forms learn option instead course revealed variety indicated though contains getting actually defined april demonstrate reach toward john ensure continued differ principal note founded think amounts options presence respond examine concluded released majority fully allows reviewed assessed investigate plays basis continues avoid dedicated come opportunity supporting publication involving millions rapidly suggesting widely remains".lower().split())]

In [ ]:



​

In [7]:



**import** logging

​

logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)

In [9]:



**import** tempfile

**import** os.path

​

TEMP\_FOLDER = tempfile.gettempdir()

print('Folder "{}" will be used to save temporary dictionary and corpus.'.format(TEMP\_FOLDER))

Folder "C:\Users\lohit\AppData\Local\Temp" will be used to save temporary dictionary and corpus.

In [10]:



**from** gensim **import** corpora, models, similarities

**if** os.path.isfile(os.path.join(TEMP\_FOLDER, 'den.dict')):

dictionary = corpora.Dictionary.load(os.path.join(TEMP\_FOLDER, 'den.dict'))

corpus = corpora.MmCorpus(os.path.join(TEMP\_FOLDER, 'den.mm'))

print("Used files generated before ")

**else**:

print("Run again error")

2017-12-07 02:38:14,042 : INFO : loading Dictionary object from C:\Users\lohit\AppData\Local\Temp\den.dict

2017-12-07 02:38:14,154 : INFO : loaded C:\Users\lohit\AppData\Local\Temp\den.dict

2017-12-07 02:38:14,220 : INFO : loaded corpus index from C:\Users\lohit\AppData\Local\Temp\den.mm.index

2017-12-07 02:38:14,221 : INFO : initializing corpus reader from C:\Users\lohit\AppData\Local\Temp\den.mm

2017-12-07 02:38:14,222 : INFO : accepted corpus with 379126 documents, 205840 features, 5375083 non-zero entries

Used files generated before

In [11]:



print(dictionary[0])

print(dictionary[1])

print(dictionary[2])

Title:

The

Wyeth

In [12]:



tfidf = models.TfidfModel(corpus) *# step 1 -- initialize a model*

2017-12-07 02:38:20,307 : INFO : collecting document frequencies

2017-12-07 02:38:20,310 : INFO : PROGRESS: processing document #0

2017-12-07 02:38:20,631 : INFO : PROGRESS: processing document #10000

2017-12-07 02:38:20,889 : INFO : PROGRESS: processing document #20000

2017-12-07 02:38:21,189 : INFO : PROGRESS: processing document #30000

2017-12-07 02:38:21,512 : INFO : PROGRESS: processing document #40000

2017-12-07 02:38:21,859 : INFO : PROGRESS: processing document #50000

2017-12-07 02:38:22,199 : INFO : PROGRESS: processing document #60000

2017-12-07 02:38:22,518 : INFO : PROGRESS: processing document #70000

2017-12-07 02:38:22,828 : INFO : PROGRESS: processing document #80000

2017-12-07 02:38:23,149 : INFO : PROGRESS: processing document #90000

2017-12-07 02:38:23,448 : INFO : PROGRESS: processing document #100000

2017-12-07 02:38:23,768 : INFO : PROGRESS: processing document #110000

2017-12-07 02:38:24,067 : INFO : PROGRESS: processing document #120000

2017-12-07 02:38:24,386 : INFO : PROGRESS: processing document #130000

2017-12-07 02:38:24,670 : INFO : PROGRESS: processing document #140000

2017-12-07 02:38:25,024 : INFO : PROGRESS: processing document #150000

2017-12-07 02:38:25,352 : INFO : PROGRESS: processing document #160000

2017-12-07 02:38:25,692 : INFO : PROGRESS: processing document #170000

2017-12-07 02:38:26,061 : INFO : PROGRESS: processing document #180000

2017-12-07 02:38:26,476 : INFO : PROGRESS: processing document #190000

2017-12-07 02:38:26,889 : INFO : PROGRESS: processing document #200000

2017-12-07 02:38:27,309 : INFO : PROGRESS: processing document #210000

2017-12-07 02:38:27,724 : INFO : PROGRESS: processing document #220000

2017-12-07 02:38:28,160 : INFO : PROGRESS: processing document #230000

2017-12-07 02:38:28,514 : INFO : PROGRESS: processing document #240000

2017-12-07 02:38:28,787 : INFO : PROGRESS: processing document #250000

2017-12-07 02:38:29,027 : INFO : PROGRESS: processing document #260000

2017-12-07 02:38:29,273 : INFO : PROGRESS: processing document #270000

2017-12-07 02:38:29,517 : INFO : PROGRESS: processing document #280000

2017-12-07 02:38:29,771 : INFO : PROGRESS: processing document #290000

2017-12-07 02:38:30,021 : INFO : PROGRESS: processing document #300000

2017-12-07 02:38:30,271 : INFO : PROGRESS: processing document #310000

2017-12-07 02:38:30,533 : INFO : PROGRESS: processing document #320000

2017-12-07 02:38:30,803 : INFO : PROGRESS: processing document #330000

2017-12-07 02:38:31,054 : INFO : PROGRESS: processing document #340000

2017-12-07 02:38:31,332 : INFO : PROGRESS: processing document #350000

2017-12-07 02:38:31,622 : INFO : PROGRESS: processing document #360000

2017-12-07 02:38:31,880 : INFO : PROGRESS: processing document #370000

2017-12-07 02:38:32,175 : INFO : calculating IDF weights for 379126 documents and 205839 features (5375083 matrix non-zeros)

In [13]:



corpus\_tfidf = tfidf[corpus]

*#for doc in corpus\_tfidf:*

*#print(doc)*

In [14]:



​

lsi = models.LsiModel(corpus\_tfidf, id2word=dictionary, num\_topics=2) *# initialize an LSI transformation*

corpus\_lsi = lsi[corpus\_tfidf] *# create a double wrapper over the original corpus: bow->tfidf->fold-in-lsi*

2017-12-07 02:38:59,332 : INFO : using serial LSI version on this node

2017-12-07 02:38:59,333 : INFO : updating model with new documents

2017-12-07 02:39:00,134 : INFO : preparing a new chunk of documents

2017-12-07 02:39:00,226 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:00,228 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:00,345 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:04,021 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:04,484 : INFO : computing the final decomposition

2017-12-07 02:39:04,486 : INFO : keeping 2 factors (discarding 85.645% of energy spectrum)

2017-12-07 02:39:04,563 : INFO : processed documents up to #20000

2017-12-07 02:39:04,570 : INFO : topic #0(13.687): 0.390\*"the" + 0.318\*"of" + 0.297\*"and" + 0.243\*"to" + 0.215\*"in" + 0.184\*"diabetes" + 0.179\*"a" + 0.162\*"is" + 0.156\*"with" + 0.145\*"for"

2017-12-07 02:39:04,576 : INFO : topic #1(12.329): 0.641\*"googletag.cmd.push(function()" + 0.641\*"{" + -0.421\*"});" + 0.000\*"as" + 0.000\*"University" + 0.000\*"from" + 0.000\*"an" + 0.000\*"aspirin" + -0.000\*"," + -0.000\*"Institute"

2017-12-07 02:39:05,684 : INFO : preparing a new chunk of documents

2017-12-07 02:39:05,799 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:05,801 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:05,958 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:09,732 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:10,164 : INFO : computing the final decomposition

2017-12-07 02:39:10,164 : INFO : keeping 2 factors (discarding 83.163% of energy spectrum)

2017-12-07 02:39:10,206 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:10,267 : INFO : keeping 2 factors (discarding 9.450% of energy spectrum)

2017-12-07 02:39:10,281 : INFO : processed documents up to #40000

2017-12-07 02:39:10,291 : INFO : topic #0(19.446): 0.386\*"the" + 0.323\*"of" + 0.292\*"and" + 0.244\*"to" + 0.224\*"in" + 0.185\*"a" + 0.174\*"diabetes" + 0.160\*"with" + 0.157\*"is" + 0.144\*"for"

2017-12-07 02:39:10,297 : INFO : topic #1(15.981): -0.706\*"googletag.cmd.push(function()" + -0.706\*"{" + -0.055\*"});" + -0.000\*"as" + -0.000\*"from" + -0.000\*"an" + 0.000\*"was" + -0.000\*"aspirin" + 0.000\*"," + -0.000\*"University"

2017-12-07 02:39:11,587 : INFO : preparing a new chunk of documents

2017-12-07 02:39:11,712 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:11,712 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:11,866 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:15,849 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:16,256 : INFO : computing the final decomposition

2017-12-07 02:39:16,256 : INFO : keeping 2 factors (discarding 82.006% of energy spectrum)

2017-12-07 02:39:16,294 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:16,327 : INFO : keeping 2 factors (discarding 0.331% of energy spectrum)

2017-12-07 02:39:16,344 : INFO : processed documents up to #60000

2017-12-07 02:39:16,353 : INFO : topic #0(24.231): 0.388\*"the" + 0.322\*"of" + 0.288\*"and" + 0.248\*"to" + 0.225\*"in" + 0.190\*"a" + 0.165\*"diabetes" + 0.159\*"with" + 0.156\*"is" + 0.145\*"for"

2017-12-07 02:39:16,360 : INFO : topic #1(19.580): -0.706\*"{" + -0.706\*"googletag.cmd.push(function()" + -0.057\*"});" + -0.000\*"as" + -0.000\*"from" + -0.000\*"an" + 0.000\*"," + -0.000\*"aspirin" + -0.000\*"University" + -0.000\*"million"

2017-12-07 02:39:17,597 : INFO : preparing a new chunk of documents

2017-12-07 02:39:17,675 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:17,675 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:17,836 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:21,672 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:22,101 : INFO : computing the final decomposition

2017-12-07 02:39:22,103 : INFO : keeping 2 factors (discarding 80.937% of energy spectrum)

2017-12-07 02:39:22,148 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:22,187 : INFO : keeping 2 factors (discarding 13.575% of energy spectrum)

2017-12-07 02:39:22,200 : INFO : processed documents up to #80000

2017-12-07 02:39:22,216 : INFO : topic #0(28.122): 0.389\*"the" + 0.321\*"of" + 0.283\*"and" + 0.249\*"to" + 0.225\*"in" + 0.191\*"a" + 0.165\*"diabetes" + 0.158\*"with" + 0.154\*"is" + 0.146\*"for"

2017-12-07 02:39:22,225 : INFO : topic #1(19.580): -0.706\*"{" + -0.706\*"googletag.cmd.push(function()" + -0.057\*"});" + -0.000\*"as" + -0.000\*"from" + -0.000\*"an" + 0.000\*"," + -0.000\*"aspirin" + -0.000\*"University" + -0.000\*"million"

2017-12-07 02:39:23,336 : INFO : preparing a new chunk of documents

2017-12-07 02:39:23,418 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:23,418 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:23,550 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:27,463 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:27,834 : INFO : computing the final decomposition

2017-12-07 02:39:27,834 : INFO : keeping 2 factors (discarding 75.810% of energy spectrum)

2017-12-07 02:39:27,876 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:27,912 : INFO : keeping 2 factors (discarding 18.914% of energy spectrum)

2017-12-07 02:39:27,923 : INFO : processed documents up to #100000

2017-12-07 02:39:27,935 : INFO : topic #0(31.479): 0.388\*"the" + 0.320\*"of" + 0.282\*"and" + 0.247\*"to" + 0.227\*"in" + 0.191\*"a" + 0.165\*"diabetes" + 0.162\*"with" + 0.153\*"is" + 0.146\*"for"

2017-12-07 02:39:27,944 : INFO : topic #1(19.580): -0.706\*"googletag.cmd.push(function()" + -0.706\*"{" + -0.057\*"});" + -0.000\*"as" + -0.000\*"from" + -0.000\*"an" + 0.000\*"," + -0.000\*"aspirin" + -0.000\*"University" + -0.000\*"million"

2017-12-07 02:39:28,972 : INFO : preparing a new chunk of documents

2017-12-07 02:39:29,068 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:29,069 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:29,175 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:33,472 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:33,850 : INFO : computing the final decomposition

2017-12-07 02:39:33,850 : INFO : keeping 2 factors (discarding 74.566% of energy spectrum)

2017-12-07 02:39:33,900 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:33,935 : INFO : keeping 2 factors (discarding 17.968% of energy spectrum)

2017-12-07 02:39:33,944 : INFO : processed documents up to #120000

2017-12-07 02:39:33,955 : INFO : topic #0(34.530): 0.390\*"the" + 0.320\*"of" + 0.280\*"and" + 0.247\*"to" + 0.228\*"in" + 0.191\*"a" + 0.164\*"diabetes" + 0.161\*"with" + 0.153\*"is" + 0.147\*"that"

2017-12-07 02:39:33,968 : INFO : topic #1(19.580): -0.706\*"googletag.cmd.push(function()" + -0.706\*"{" + -0.057\*"});" + 0.000\*"Source:" + -0.000\*"as" + -0.000\*"from" + -0.000\*"an" + 0.000\*"," + -0.000\*"aspirin" + -0.000\*"million"

2017-12-07 02:39:34,920 : INFO : preparing a new chunk of documents

2017-12-07 02:39:35,004 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:35,004 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:35,135 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:39,226 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:39,617 : INFO : computing the final decomposition

2017-12-07 02:39:39,618 : INFO : keeping 2 factors (discarding 82.778% of energy spectrum)

2017-12-07 02:39:39,659 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:39,692 : INFO : keeping 2 factors (discarding 7.744% of energy spectrum)

2017-12-07 02:39:39,702 : INFO : processed documents up to #140000

2017-12-07 02:39:39,708 : INFO : topic #0(37.226): 0.389\*"the" + 0.320\*"of" + 0.277\*"and" + 0.247\*"to" + 0.230\*"in" + 0.192\*"a" + 0.163\*"diabetes" + 0.162\*"with" + 0.153\*"is" + 0.149\*"that"

2017-12-07 02:39:39,717 : INFO : topic #1(19.580): -0.706\*"{" + -0.706\*"googletag.cmd.push(function()" + -0.057\*"});" + 0.000\*"Source:" + -0.000\*"as" + -0.000\*"from" + -0.000\*"an" + 0.000\*"," + -0.000\*"aspirin" + -0.000\*"million"

2017-12-07 02:39:40,783 : INFO : preparing a new chunk of documents

2017-12-07 02:39:40,870 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:40,870 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:41,002 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:44,741 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:45,105 : INFO : computing the final decomposition

2017-12-07 02:39:45,105 : INFO : keeping 2 factors (discarding 80.499% of energy spectrum)

2017-12-07 02:39:45,144 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:45,179 : INFO : keeping 2 factors (discarding 1.372% of energy spectrum)

2017-12-07 02:39:45,188 : INFO : processed documents up to #160000

2017-12-07 02:39:45,195 : INFO : topic #0(40.073): 0.390\*"the" + 0.319\*"of" + 0.272\*"and" + 0.247\*"to" + 0.232\*"in" + 0.194\*"a" + 0.159\*"with" + 0.158\*"diabetes" + 0.155\*"that" + 0.153\*"is"

2017-12-07 02:39:45,205 : INFO : topic #1(21.258): -0.707\*"{" + -0.707\*"googletag.cmd.push(function()" + 0.038\*"});" + 0.000\*"Source:" + -0.000\*"an" + -0.000\*"as" + 0.000\*"diabetes" + -0.000\*"and" + -0.000\*"Americans" + -0.000\*"million"

2017-12-07 02:39:46,292 : INFO : preparing a new chunk of documents

2017-12-07 02:39:46,391 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:46,391 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:46,524 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:51,056 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:51,498 : INFO : computing the final decomposition

2017-12-07 02:39:51,498 : INFO : keeping 2 factors (discarding 80.050% of energy spectrum)

2017-12-07 02:39:51,548 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:51,579 : INFO : keeping 2 factors (discarding 0.482% of energy spectrum)

2017-12-07 02:39:51,591 : INFO : processed documents up to #180000

2017-12-07 02:39:51,598 : INFO : topic #0(43.079): 0.392\*"the" + 0.319\*"of" + 0.268\*"and" + 0.246\*"to" + 0.233\*"in" + 0.195\*"a" + 0.158\*"that" + 0.157\*"with" + 0.155\*"diabetes" + 0.151\*"is"

2017-12-07 02:39:51,607 : INFO : topic #1(22.295): -0.707\*"googletag.cmd.push(function()" + -0.707\*"{" + 0.007\*"});" + 0.000\*"Source:" + -0.000\*"concluded:" + 0.000\*"lower" + 0.000\*"one" + 0.000\*"high" + -0.000\*"women" + 0.000\*"therapy"

2017-12-07 02:39:53,061 : INFO : preparing a new chunk of documents

2017-12-07 02:39:53,192 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:39:53,194 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:39:53,399 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:39:58,178 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:39:58,508 : INFO : computing the final decomposition

2017-12-07 02:39:58,508 : INFO : keeping 2 factors (discarding 77.022% of energy spectrum)

2017-12-07 02:39:58,560 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:39:58,592 : INFO : keeping 2 factors (discarding 2.030% of energy spectrum)

2017-12-07 02:39:58,601 : INFO : processed documents up to #200000

2017-12-07 02:39:58,609 : INFO : topic #0(46.395): 0.393\*"the" + 0.320\*"of" + 0.265\*"and" + 0.246\*"to" + 0.236\*"in" + 0.196\*"a" + 0.162\*"that" + 0.155\*"with" + 0.152\*"diabetes" + 0.149\*"is"

2017-12-07 02:39:58,621 : INFO : topic #1(22.295): -0.707\*"googletag.cmd.push(function()" + -0.707\*"{" + 0.007\*"});" + 0.000\*"Source:" + -0.000\*"concluded:" + 0.000\*"lower" + -0.000\*"women" + 0.000\*"one" + 0.000\*"high" + 0.000\*"therapy"

2017-12-07 02:39:59,903 : INFO : preparing a new chunk of documents

2017-12-07 02:39:59,995 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:00,011 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:00,149 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:04,172 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:40:04,581 : INFO : computing the final decomposition

2017-12-07 02:40:04,581 : INFO : keeping 2 factors (discarding 77.345% of energy spectrum)

2017-12-07 02:40:04,625 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:04,655 : INFO : keeping 2 factors (discarding 1.711% of energy spectrum)

2017-12-07 02:40:04,667 : INFO : processed documents up to #220000

2017-12-07 02:40:04,673 : INFO : topic #0(49.452): 0.393\*"the" + 0.320\*"of" + 0.263\*"and" + 0.245\*"to" + 0.237\*"in" + 0.197\*"a" + 0.164\*"that" + 0.154\*"with" + 0.149\*"diabetes" + 0.147\*"is"

2017-12-07 02:40:04,682 : INFO : topic #1(22.295): -0.707\*"{" + -0.707\*"googletag.cmd.push(function()" + 0.007\*"});" + 0.000\*"Source:" + -0.000\*"concluded:" + 0.000\*"lower" + -0.000\*"women" + 0.000\*"one" + 0.000\*"high" + 0.000\*"therapy"

2017-12-07 02:40:05,926 : INFO : preparing a new chunk of documents

2017-12-07 02:40:06,026 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:06,028 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:06,215 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:11,042 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:40:11,625 : INFO : computing the final decomposition

2017-12-07 02:40:11,627 : INFO : keeping 2 factors (discarding 78.898% of energy spectrum)

2017-12-07 02:40:11,673 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:11,706 : INFO : keeping 2 factors (discarding 1.739% of energy spectrum)

2017-12-07 02:40:11,718 : INFO : processed documents up to #240000

2017-12-07 02:40:11,725 : INFO : topic #0(52.267): 0.393\*"the" + 0.319\*"of" + 0.261\*"and" + 0.246\*"to" + 0.235\*"in" + 0.198\*"a" + 0.165\*"that" + 0.153\*"with" + 0.147\*"is" + 0.146\*"diabetes"

2017-12-07 02:40:11,734 : INFO : topic #1(22.295): 0.707\*"googletag.cmd.push(function()" + 0.707\*"{" + -0.007\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + 0.000\*"women" + -0.000\*"one" + -0.000\*"high" + -0.000\*"therapy"

2017-12-07 02:40:12,872 : INFO : preparing a new chunk of documents

2017-12-07 02:40:12,966 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:12,968 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:13,111 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:17,510 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:40:17,988 : INFO : computing the final decomposition

2017-12-07 02:40:17,989 : INFO : keeping 2 factors (discarding 85.783% of energy spectrum)

2017-12-07 02:40:18,029 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:18,064 : INFO : keeping 2 factors (discarding 4.817% of energy spectrum)

2017-12-07 02:40:18,077 : INFO : processed documents up to #260000

2017-12-07 02:40:18,084 : INFO : topic #0(53.874): 0.393\*"the" + 0.320\*"of" + 0.261\*"and" + 0.246\*"to" + 0.236\*"in" + 0.197\*"a" + 0.164\*"that" + 0.154\*"with" + 0.148\*"is" + 0.148\*"diabetes"

2017-12-07 02:40:18,092 : INFO : topic #1(22.296): 0.707\*"{" + 0.707\*"googletag.cmd.push(function()" + -0.010\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + 0.000\*"women" + -0.000\*"one" + -0.000\*"high" + -0.000\*"therapy"

2017-12-07 02:40:18,863 : INFO : preparing a new chunk of documents

2017-12-07 02:40:18,946 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:18,946 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:19,060 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:22,812 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:40:23,161 : INFO : computing the final decomposition

2017-12-07 02:40:23,177 : INFO : keeping 2 factors (discarding 85.797% of energy spectrum)

2017-12-07 02:40:23,214 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:23,246 : INFO : keeping 2 factors (discarding 4.449% of energy spectrum)

2017-12-07 02:40:23,262 : INFO : processed documents up to #280000

2017-12-07 02:40:23,269 : INFO : topic #0(53.874): 0.393\*"the" + 0.320\*"of" + 0.261\*"and" + 0.246\*"to" + 0.236\*"in" + 0.197\*"a" + 0.164\*"that" + 0.154\*"with" + 0.148\*"is" + 0.148\*"diabetes"

2017-12-07 02:40:23,276 : INFO : topic #1(25.751): 0.707\*"{" + 0.707\*"googletag.cmd.push(function()" + -0.010\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + 0.000\*"women" + -0.000\*"lower" + -0.000\*"high" + -0.000\*"therapy" + 0.000\*"or"

2017-12-07 02:40:24,046 : INFO : preparing a new chunk of documents

2017-12-07 02:40:24,134 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:24,135 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:24,245 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:28,190 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:40:28,556 : INFO : computing the final decomposition

2017-12-07 02:40:28,556 : INFO : keeping 2 factors (discarding 84.645% of energy spectrum)

2017-12-07 02:40:28,601 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:28,631 : INFO : keeping 2 factors (discarding 4.655% of energy spectrum)

2017-12-07 02:40:28,640 : INFO : processed documents up to #300000

2017-12-07 02:40:28,649 : INFO : topic #0(53.874): 0.393\*"the" + 0.320\*"of" + 0.261\*"and" + 0.246\*"to" + 0.236\*"in" + 0.197\*"a" + 0.164\*"that" + 0.154\*"with" + 0.148\*"is" + 0.148\*"diabetes"

2017-12-07 02:40:28,660 : INFO : topic #1(29.088): 0.707\*"{" + 0.707\*"googletag.cmd.push(function()" + -0.010\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + 0.000\*"women" + -0.000\*"therapy" + -0.000\*"." + 0.000\*"may"

2017-12-07 02:40:30,061 : INFO : preparing a new chunk of documents

2017-12-07 02:40:30,148 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:30,148 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:30,248 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:34,066 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:40:34,415 : INFO : computing the final decomposition

2017-12-07 02:40:34,415 : INFO : keeping 2 factors (discarding 85.341% of energy spectrum)

2017-12-07 02:40:34,462 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:34,492 : INFO : keeping 2 factors (discarding 0.184% of energy spectrum)

2017-12-07 02:40:34,502 : INFO : processed documents up to #320000

2017-12-07 02:40:34,510 : INFO : topic #0(55.391): 0.394\*"the" + 0.320\*"of" + 0.262\*"and" + 0.246\*"to" + 0.235\*"in" + 0.197\*"a" + 0.163\*"that" + 0.154\*"with" + 0.149\*"is" + 0.148\*"diabetes"

2017-12-07 02:40:34,522 : INFO : topic #1(31.592): 0.707\*"googletag.cmd.push(function()" + 0.707\*"{" + -0.033\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + 0.000\*"women" + -0.000\*"therapy" + -0.000\*"our" + -0.000\*"high"

2017-12-07 02:40:35,362 : INFO : preparing a new chunk of documents

2017-12-07 02:40:35,425 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:35,425 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:35,548 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:39,358 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:40:39,723 : INFO : computing the final decomposition

2017-12-07 02:40:39,723 : INFO : keeping 2 factors (discarding 84.817% of energy spectrum)

2017-12-07 02:40:39,766 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:39,799 : INFO : keeping 2 factors (discarding 1.863% of energy spectrum)

2017-12-07 02:40:39,809 : INFO : processed documents up to #340000

2017-12-07 02:40:39,817 : INFO : topic #0(56.895): 0.394\*"the" + 0.321\*"of" + 0.264\*"and" + 0.246\*"to" + 0.235\*"in" + 0.196\*"a" + 0.162\*"that" + 0.154\*"with" + 0.150\*"is" + 0.149\*"diabetes"

2017-12-07 02:40:39,829 : INFO : topic #1(33.080): 0.702\*"googletag.cmd.push(function()" + 0.702\*"{" + -0.118\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + 0.000\*"women" + -0.000\*"our" + -0.000\*"therapy" + -0.000\*"high"

2017-12-07 02:40:40,721 : INFO : preparing a new chunk of documents

2017-12-07 02:40:40,785 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:40,785 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:40,899 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:45,028 : INFO : 2nd phase: running dense svd on (102, 20000) matrix

2017-12-07 02:40:45,378 : INFO : computing the final decomposition

2017-12-07 02:40:45,378 : INFO : keeping 2 factors (discarding 84.213% of energy spectrum)

2017-12-07 02:40:45,430 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:45,461 : INFO : keeping 2 factors (discarding 3.974% of energy spectrum)

2017-12-07 02:40:45,473 : INFO : processed documents up to #360000

2017-12-07 02:40:45,480 : INFO : topic #0(56.895): 0.394\*"the" + 0.321\*"of" + 0.264\*"and" + 0.246\*"to" + 0.235\*"in" + 0.196\*"a" + 0.162\*"that" + 0.154\*"with" + 0.150\*"is" + 0.149\*"diabetes"

2017-12-07 02:40:45,490 : INFO : topic #1(35.795): 0.702\*"googletag.cmd.push(function()" + 0.702\*"{" + -0.118\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + -0.000\*"our" + 0.000\*"women" + -0.000\*"therapy" + -0.000\*"one"

2017-12-07 02:40:46,316 : INFO : preparing a new chunk of documents

2017-12-07 02:40:46,408 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 02:40:46,409 : INFO : 1st phase: constructing (205840, 102) action matrix

2017-12-07 02:40:46,511 : INFO : orthonormalizing (205840, 102) action matrix

2017-12-07 02:40:50,324 : INFO : 2nd phase: running dense svd on (102, 19126) matrix

2017-12-07 02:40:50,662 : INFO : computing the final decomposition

2017-12-07 02:40:50,663 : INFO : keeping 2 factors (discarding 85.148% of energy spectrum)

2017-12-07 02:40:50,700 : INFO : merging projections: (205840, 2) + (205840, 2)

2017-12-07 02:40:50,730 : INFO : keeping 2 factors (discarding 0.926% of energy spectrum)

2017-12-07 02:40:50,743 : INFO : processed documents up to #379126

2017-12-07 02:40:50,750 : INFO : topic #0(58.360): 0.394\*"the" + 0.321\*"of" + 0.265\*"and" + 0.246\*"to" + 0.235\*"in" + 0.196\*"a" + 0.161\*"that" + 0.155\*"with" + 0.151\*"is" + 0.150\*"diabetes"

2017-12-07 02:40:50,756 : INFO : topic #1(37.328): 0.706\*"{" + 0.706\*"googletag.cmd.push(function()" + -0.067\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + -0.000\*"our" + -0.000\*"therapy" + 0.000\*"women" + 0.000\*"patients"

In [16]:



lsi.print\_topics(2)

2017-12-07 02:41:06,224 : INFO : topic #0(58.360): 0.394\*"the" + 0.321\*"of" + 0.265\*"and" + 0.246\*"to" + 0.235\*"in" + 0.196\*"a" + 0.161\*"that" + 0.155\*"with" + 0.151\*"is" + 0.150\*"diabetes"

2017-12-07 02:41:06,231 : INFO : topic #1(37.328): 0.706\*"{" + 0.706\*"googletag.cmd.push(function()" + -0.067\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + -0.000\*"our" + -0.000\*"therapy" + 0.000\*"women" + 0.000\*"patients"

Out[16]:

[(0,

'0.394\*"the" + 0.321\*"of" + 0.265\*"and" + 0.246\*"to" + 0.235\*"in" + 0.196\*"a" + 0.161\*"that" + 0.155\*"with" + 0.151\*"is" + 0.150\*"diabetes"'),

(1,

'0.706\*"{" + 0.706\*"googletag.cmd.push(function()" + -0.067\*"});" + -0.000\*"Source:" + 0.000\*"concluded:" + -0.000\*"lower" + -0.000\*"our" + -0.000\*"therapy" + 0.000\*"women" + 0.000\*"patients"')]

In [220]:



*#for doc in corpus\_lsi: # both bow->tfidf and tfidf->lsi transformations are actually executed here, on the fly*

*#print(doc)*

[(0, -1.0)]

[(1, 1.0)]

In [17]:



lsi.save(os.path.join(TEMP\_FOLDER, 'model.lsi')) *# same for tfidf, lda, ...*

*#lsi = models.LsiModel.load(os.path.join(TEMP\_FOLDER, 'model.lsi'))*

2017-12-07 02:41:27,994 : INFO : saving Projection object under C:\Users\lohit\AppData\Local\Temp\model.lsi.projection, separately None

2017-12-07 02:41:28,052 : INFO : saved C:\Users\lohit\AppData\Local\Temp\model.lsi.projection

2017-12-07 02:41:28,054 : INFO : saving LsiModel object under C:\Users\lohit\AppData\Local\Temp\model.lsi, separately None

2017-12-07 02:41:28,056 : INFO : not storing attribute projection

2017-12-07 02:41:28,057 : INFO : not storing attribute dispatcher

2017-12-07 02:41:28,213 : INFO : saved C:\Users\lohit\AppData\Local\Temp\model.lsi

In [18]:



**from** gensim **import** corpora, models, similarities

dictionary = corpora.Dictionary.load('C:/Users/lohit/AppData/Local/Temp/den.dict')

corpus = corpora.MmCorpus('C:/Users/lohit/AppData/Local/Temp/den.mm') *# comes from the first tutorial, "From strings to vectors"*

*#print(corpus)*

2017-12-07 02:41:41,397 : INFO : loading Dictionary object from C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 02:41:41,521 : INFO : loaded C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 02:41:41,567 : INFO : loaded corpus index from C:/Users/lohit/AppData/Local/Temp/den.mm.index

2017-12-07 02:41:41,568 : INFO : initializing corpus reader from C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 02:41:41,569 : INFO : accepted corpus with 379126 documents, 205840 features, 5375083 non-zero entries

In [19]:



doc = "diabetes research medicine"

vec\_bow = dictionary.doc2bow(doc.lower().split())

vec\_lsi = lsi[vec\_bow] *# convert the query to LSI space*

print(vec\_lsi)

​

[(0, 0.20682855845805731), (1, -1.7022889433835045e-05)]

In [20]:



index = similarities.MatrixSimilarity(lsi[corpus]) *# transform corpus to LSI space and index it*

2017-12-07 02:42:08,296 : WARNING : scanning corpus to determine the number of features (consider setting `num\_features` explicitly)

2017-12-07 02:42:23,292 : INFO : creating matrix with 379126 documents and 2 features

In [21]:



index.save('C:/Users/lohit/AppData/Local/Temp/den.index')

index = similarities.MatrixSimilarity.load('C:/Users/lohit/AppData/Local/Temp/den.index')

2017-12-07 02:43:04,791 : INFO : saving MatrixSimilarity object under C:/Users/lohit/AppData/Local/Temp/den.index, separately None

2017-12-07 02:43:04,839 : INFO : saved C:/Users/lohit/AppData/Local/Temp/den.index

2017-12-07 02:43:04,841 : INFO : loading MatrixSimilarity object from C:/Users/lohit/AppData/Local/Temp/den.index

2017-12-07 02:43:04,862 : INFO : loaded C:/Users/lohit/AppData/Local/Temp/den.index

In [22]:



sims = index[vec\_lsi] *# perform a similarity query against the corpus*

print(list(enumerate(sims))) *# print (document\_number, document\_similarity) 2-tuples*

IOPub data rate exceeded.

The notebook server will temporarily stop sending output

to the client in order to avoid crashing it.

To change this limit, set the config variable

`--NotebookApp.iopub\_data\_rate\_limit`.

In [23]:



sims = sorted(enumerate(sims), key=**lambda** item: **-**item[1])

print(sims) *# print sorted (document number, similarity score) 2-tuples*

IOPub data rate exceeded.

The notebook server will temporarily stop sending output

to the client in order to avoid crashing it.

To change this limit, set the config variable

`--NotebookApp.iopub\_data\_rate\_limit`.

In [ ]:



*# this time we are reading file one by one and apply tfIdf-> LSI*

*# calculating words from document to document to get the intersting topic*

In [ ]:



**import** glob

list\_of\_files = glob.glob(r"C:\Users\lohit\Desktop\course\KDD\FinalProject\diabetes\diabetes\\*.txt") *# create the list of file*

**for** file\_name **in** list\_of\_files:

FI = open(file\_name, 'r',encoding="utf-8")

x = FI.read()

​

*#text\_string = x.split()*

*# Count word frequencies*

**from** nltk.corpus **import** stopwords

**from** nltk.stem.wordnet **import** WordNetLemmatizer

**import** string

stop = set(stopwords.words('english'))

exclude = set(string.punctuation)

lemma = WordNetLemmatizer()

**def** clean(doc):

*#print(doc)*

stop\_free = " ".join([i **for** i **in** doc.lower().split() **if** i **not** **in** stop])

punc\_free = ''.join(ch **for** ch **in** stop\_free **if** ch **not** **in** exclude)

normalized = " ".join(lemma.lemmatize(word) **for** word **in** punc\_free.split())

**return** normalized

processed\_corpus = [clean(doc).split() **for** doc **in** x.split("\n")]

*#print(doc\_clean)*

**from** gensim **import** corpora, models, utils

dictionary = corpora.Dictionary(processed\_corpus)

dictionary.save('C:/Users/lohit/AppData/Local/Temp/den.dict')

*# print(dictionary)*

print(dictionary.token2id)

bow\_corpus = [dictionary.doc2bow(text) **for** text **in** processed\_corpus]

corpora.MmCorpus.serialize('C:/Users/lohit/AppData/Local/Temp/den.mm', bow\_corpus)

*#print(bow\_corpus)*

**from** gensim **import** models

*# train the model*

tfidf = models.TfidfModel(bow\_corpus)

*# transform the "system minors" string*

tfidf[dictionary.doc2bow("diabetes patients type study insulin blood risk disease health research ".lower().split())]

**import** logging

​

logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)

**import** tempfile

**import** os.path

​

TEMP\_FOLDER = tempfile.gettempdir()

print('Folder "{}" will be used to save temporary dictionary and corpus.'.format(TEMP\_FOLDER))

**from** gensim **import** corpora, models, similarities

**if** os.path.isfile(os.path.join(TEMP\_FOLDER, 'den.dict')):

dictionary = corpora.Dictionary.load(os.path.join(TEMP\_FOLDER, 'den.dict'))

corpus = corpora.MmCorpus(os.path.join(TEMP\_FOLDER, 'den.mm'))

print("Used files generated before ")

**else**:

print("Run again error")

tfidf = models.TfidfModel(corpus) *# step 1 -- initialize a model*

corpus\_tfidf = tfidf[corpus]

*#for doc in corpus\_tfidf:*

*# print(doc)*

lsi = models.LsiModel(corpus\_tfidf, id2word=dictionary, num\_topics=50) *# initialize an LSI transformation*

corpus\_lsi = lsi[corpus\_tfidf] *# create a double wrapper over the original corpus: bow->tfidf->fold-in-lsi*

lsi.print\_topics(50)

*#for doc in corpus\_lsi: # both bow->tfidf and tfidf->lsi transformations are actually executed here, on the fly*

*# print(doc)*

lsi.save(os.path.join(TEMP\_FOLDER, 'model.lsi')) *# same for tfidf, lda, ...*

*#lsi = models.LsiModel.load(os.path.join(TEMP\_FOLDER, 'model.lsi'))*

**from** gensim **import** corpora, models, similarities

dictionary = corpora.Dictionary.load('C:/Users/lohit/AppData/Local/Temp/den.dict')

corpus = corpora.MmCorpus('C:/Users/lohit/AppData/Local/Temp/den.mm') *# comes from the first tutorial, "From strings to vectors"*

*#print(corpus)*

doc = "diabetes research medicine discovert prevention cure "

vec\_bow = dictionary.doc2bow(doc.lower().split())

vec\_lsi = lsi[vec\_bow] *# convert the query to LSI space*

print(vec\_lsi)

index = similarities.MatrixSimilarity(lsi[corpus]) *# transform corpus to LSI space and index it*

index.save('C:/Users/lohit/AppData/Local/Temp/den.index')

index = similarities.MatrixSimilarity.load('C:/Users/lohit/AppData/Local/Temp/den.index')

sims = index[vec\_lsi] *# perform a similarity query against the corpus*

print(list(enumerate(sims))) *# print (document\_number, document\_similarity) 2-tuples*

sims = sorted(enumerate(sims), key=**lambda** item: **-**item[1])

print(sims) *# print sorted (document number, similarity score) 2-tuples*

​

2017-12-07 14:05:24,066 : INFO : adding document #0 to Dictionary(0 unique tokens: [])

2017-12-07 14:05:24,067 : INFO : built Dictionary(143 unique tokens: ['title', 'wyeth', 'symposium', 'metabolic', 'dysregulation']...) from 21 documents (total 232 corpus positions)

2017-12-07 14:05:24,079 : INFO : saving Dictionary object under C:/Users/lohit/AppData/Local/Temp/den.dict, separately None

2017-12-07 14:05:24,112 : INFO : saved C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 14:05:24,420 : INFO : storing corpus in Matrix Market format to C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 14:05:24,424 : INFO : saving sparse matrix to C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 14:05:24,425 : INFO : PROGRESS: saving document #0

2017-12-07 14:05:24,426 : INFO : saved 21x143 matrix, density=5.661% (170/3003)

2017-12-07 14:05:24,433 : INFO : saving MmCorpus index to C:/Users/lohit/AppData/Local/Temp/den.mm.index

2017-12-07 14:05:24,440 : INFO : collecting document frequencies

2017-12-07 14:05:24,440 : INFO : PROGRESS: processing document #0

2017-12-07 14:05:24,440 : INFO : calculating IDF weights for 21 documents and 142 features (170 matrix non-zeros)

2017-12-07 14:05:24,443 : INFO : loading Dictionary object from C:\Users\lohit\AppData\Local\Temp\den.dict

2017-12-07 14:05:24,445 : INFO : loaded C:\Users\lohit\AppData\Local\Temp\den.dict

2017-12-07 14:05:24,447 : INFO : loaded corpus index from C:\Users\lohit\AppData\Local\Temp\den.mm.index

2017-12-07 14:05:24,448 : INFO : initializing corpus reader from C:\Users\lohit\AppData\Local\Temp\den.mm

2017-12-07 14:05:24,449 : INFO : accepted corpus with 21 documents, 143 features, 170 non-zero entries

2017-12-07 14:05:24,456 : INFO : collecting document frequencies

2017-12-07 14:05:24,461 : INFO : PROGRESS: processing document #0

2017-12-07 14:05:24,464 : INFO : calculating IDF weights for 21 documents and 142 features (170 matrix non-zeros)

2017-12-07 14:05:24,466 : INFO : using serial LSI version on this node

2017-12-07 14:05:24,469 : INFO : updating model with new documents

2017-12-07 14:05:24,474 : INFO : preparing a new chunk of documents

2017-12-07 14:05:24,475 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 14:05:24,477 : INFO : 1st phase: constructing (143, 150) action matrix

2017-12-07 14:05:24,482 : INFO : orthonormalizing (143, 150) action matrix

2017-12-07 14:05:24,488 : INFO : 2nd phase: running dense svd on (143, 21) matrix

2017-12-07 14:05:24,492 : INFO : computing the final decomposition

2017-12-07 14:05:24,494 : INFO : keeping 8 factors (discarding 0.000% of energy spectrum)

2017-12-07 14:05:24,496 : INFO : processed documents up to #21

2017-12-07 14:05:24,497 : INFO : topic #0(1.398): 0.575\*"boston" + 0.522\*"university" + 0.190\*"wyeth" + 0.183\*"dysregulation" + 0.183\*"symposium" + 0.176\*"title" + 0.176\*"presented" + 0.162\*"metabolic" + 0.113\*"department" + 0.113\*"therapeutic"

2017-12-07 14:05:24,500 : INFO : topic #1(1.014): -0.305\*"chairman" + -0.305\*"hosted" + -0.305\*"h" + -0.305\*"farb" + -0.271\*"dr" + -0.248\*"experimental" + -0.248\*"therapeutic" + -0.248\*"department" + -0.241\*"david" + 0.234\*"university"

2017-12-07 14:05:24,502 : INFO : topic #2(1.001): -0.346\*"public" + -0.173\*"event" + -0.173\*"together" + -0.173\*"leading" + -0.173\*"am" + -0.173\*"obesity" + -0.173\*"impact" + -0.173\*"recent" + -0.173\*"inflammation" + -0.173\*"diabetes"

2017-12-07 14:05:24,504 : INFO : topic #3(1.000): 0.548\*"register" + 0.548\*"httpwwwbumcbuedubusmpmwyeth" + 0.548\*"go" + -0.183\*"digravio" + -0.183\*"source" + -0.183\*"gina" + 0.000\*"public" + 0.000\*"leading" + 0.000\*"am" + 0.000\*"500"

2017-12-07 14:05:24,506 : INFO : topic #4(1.000): 0.548\*"gina" + 0.548\*"source" + 0.548\*"digravio" + 0.183\*"httpwwwbumcbuedubusmpmwyeth" + 0.183\*"register" + 0.183\*"go" + 0.000\*"hosted" + 0.000\*"chairman" + 0.000\*"h" + 0.000\*"farb"

2017-12-07 14:05:24,540 : INFO : topic #0(1.398): 0.575\*"boston" + 0.522\*"university" + 0.190\*"wyeth" + 0.183\*"dysregulation" + 0.183\*"symposium" + 0.176\*"title" + 0.176\*"presented" + 0.162\*"metabolic" + 0.113\*"department" + 0.113\*"therapeutic"

2017-12-07 14:05:24,544 : INFO : topic #1(1.014): -0.305\*"chairman" + -0.305\*"hosted" + -0.305\*"h" + -0.305\*"farb" + -0.271\*"dr" + -0.248\*"experimental" + -0.248\*"therapeutic" + -0.248\*"department" + -0.241\*"david" + 0.234\*"university"

2017-12-07 14:05:24,546 : INFO : topic #2(1.001): -0.346\*"public" + -0.173\*"event" + -0.173\*"together" + -0.173\*"leading" + -0.173\*"am" + -0.173\*"obesity" + -0.173\*"impact" + -0.173\*"recent" + -0.173\*"inflammation" + -0.173\*"diabetes"

2017-12-07 14:05:24,547 : INFO : topic #3(1.000): 0.548\*"register" + 0.548\*"httpwwwbumcbuedubusmpmwyeth" + 0.548\*"go" + -0.183\*"digravio" + -0.183\*"source" + -0.183\*"gina" + 0.000\*"public" + 0.000\*"leading" + 0.000\*"am" + 0.000\*"500"

2017-12-07 14:05:24,549 : INFO : topic #4(1.000): 0.548\*"gina" + 0.548\*"source" + 0.548\*"digravio" + 0.183\*"httpwwwbumcbuedubusmpmwyeth" + 0.183\*"register" + 0.183\*"go" + 0.000\*"hosted" + 0.000\*"chairman" + 0.000\*"h" + 0.000\*"farb"

2017-12-07 14:05:24,550 : INFO : topic #5(0.957): -0.404\*"phd" + -0.354\*"professor" + -0.202\*"ma" + -0.202\*"md" + -0.191\*"dr" + -0.184\*"research" + -0.152\*"biology" + -0.152\*"medicine" + -0.152\*"school" + -0.152\*"neurology"

2017-12-07 14:05:24,552 : INFO : topic #6(0.801): 0.492\*"title" + 0.492\*"presented" + 0.237\*"wyeth" + -0.228\*"boston" + 0.204\*"symposium" + 0.200\*"dysregulation" + 0.175\*"metabolic" + -0.153\*"university" + -0.138\*"one" + -0.138\*"trustee"

2017-12-07 14:05:24,553 : INFO : topic #7(0.676): -0.390\*"university" + -0.265\*"boston" + 0.230\*"ballroom" + 0.230\*"street" + 0.230\*"october" + 0.230\*"present" + 0.230\*"mass" + 0.230\*"sherborn" + 0.230\*"trustee" + 0.230\*"22nd"

2017-12-07 14:05:24,554 : INFO : saving Projection object under C:\Users\lohit\AppData\Local\Temp\model.lsi.projection, separately None

2017-12-07 14:05:24,569 : INFO : saved C:\Users\lohit\AppData\Local\Temp\model.lsi.projection

2017-12-07 14:05:24,571 : INFO : saving LsiModel object under C:\Users\lohit\AppData\Local\Temp\model.lsi, separately None

2017-12-07 14:05:24,573 : INFO : not storing attribute projection

2017-12-07 14:05:24,575 : INFO : not storing attribute dispatcher

2017-12-07 14:05:24,584 : INFO : saved C:\Users\lohit\AppData\Local\Temp\model.lsi

2017-12-07 14:05:24,584 : INFO : loading Dictionary object from C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 14:05:24,588 : INFO : loaded C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 14:05:24,590 : INFO : loaded corpus index from C:/Users/lohit/AppData/Local/Temp/den.mm.index

2017-12-07 14:05:24,591 : INFO : initializing corpus reader from C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 14:05:24,593 : INFO : accepted corpus with 21 documents, 143 features, 170 non-zero entries

2017-12-07 14:05:24,598 : WARNING : scanning corpus to determine the number of features (consider setting `num\_features` explicitly)

2017-12-07 14:05:24,601 : INFO : creating matrix with 21 documents and 8 features

2017-12-07 14:05:24,605 : INFO : saving MatrixSimilarity object under C:/Users/lohit/AppData/Local/Temp/den.index, separately None

2017-12-07 14:05:24,612 : INFO : saved C:/Users/lohit/AppData/Local/Temp/den.index

2017-12-07 14:05:24,612 : INFO : loading MatrixSimilarity object from C:/Users/lohit/AppData/Local/Temp/den.index

2017-12-07 14:05:24,617 : INFO : loaded C:/Users/lohit/AppData/Local/Temp/den.index

{'title': 0, 'wyeth': 1, 'symposium': 2, 'metabolic': 3, 'dysregulation': 4, 'presented': 5, 'boston': 6, 'university': 7, 'department': 8, 'pharmacology': 9, 'experimental': 10, 'therapeutic': 11, 'present': 12, 'october': 13, '22nd': 14, 'trustee': 15, 'ballroom': 16, 'one': 17, 'sherborn': 18, 'street': 19, 'mass': 20, 'free': 21, 'event': 22, '830': 23, 'am': 24, '500': 25, 'pm': 26, 'bring': 27, 'together': 28, 'leading': 29, 'scientist': 30, 'pharmaceutical': 31, 'company': 32, 'executive': 33, 'public': 34, 'forum': 35, 'discus': 36, 'recent': 37, 'research': 38, 'impact': 39, 'obesity': 40, 'diabetes': 41, 'heart': 42, 'disease': 43, 'cancer': 44, 'inflammation': 45, 'open': 46, 'lecturer': 47, 'include': 48, 'dr': 49, 'steven': 50, 'a': 51, 'kliewer': 52, 'md': 53, 'nancy': 54, 'b': 55, 'jake': 56, 'l': 57, 'hamon': 58, 'distinguished': 59, 'chair': 60, 'basic': 61, 'professor': 62, 'molecular': 63, 'biology': 64, 'texas': 65, 'southwestern': 66, 'medical': 67, 'center': 68, 'dallas': 69, 'tx': 70, 'gary': 71, 'e': 72, 'landreth': 73, 'phd': 74, 'neuroscience': 75, 'neurology': 76, 'case': 77, 'western': 78, 'reserve': 79, 'cleveland': 80, 'oh': 81, 'timothy': 82, 'f': 83, 'osborne': 84, 'biochemistry': 85, 'california': 86, 'irvine': 87, 'ca': 88, 'giulio': 89, 'm': 90, 'pasinetti': 91, 'psychiatry': 92, 'geriatrics': 93, 'adult': 94, 'development': 95, 'mount': 96, 'sinai': 97, 'school': 98, 'medicine': 99, 'new': 100, 'york': 101, 'ny': 102, 'pere': 103, 'puigserver': 104, 'assistant': 105, 'cell': 106, 'dana': 107, 'farber': 108, 'institute': 109, 'ma': 110, 'david': 111, 'r': 112, 'riddell': 113, 'principal': 114, 'ii': 115, 'discovery': 116, 'princeton': 117, 'nj': 118, 'sudha': 119, 'seshadri': 120, 'associate': 121, 'andrei': 122, 'ruckenstein': 123, 'provost': 124, 'vice': 125, 'president': 126, 'james': 127, 'tobin': 128, 'cardiovascular': 129, 'cambridge': 130, 'benjamin': 131, 'wolozin': 132, 'hosted': 133, 'h': 134, 'farb': 135, 'chairman': 136, 'register': 137, 'go': 138, 'httpwwwbumcbuedubusmpmwyeth': 139, 'source': 140, 'gina': 141, 'digravio': 142}

Folder "C:\Users\lohit\AppData\Local\Temp" will be used to save temporary dictionary and corpus.

Used files generated before

[(0, 0.10438020609840562), (1, -0.15367473858110067), (2, -0.31691870729279276), (5, -0.32189135004841918), (6, -0.016264913126184739), (7, 0.065483330271116857)]

2017-12-07 14:05:24,700 : INFO : adding document #0 to Dictionary(0 unique tokens: [])

2017-12-07 14:05:24,705 : INFO : built Dictionary(301 unique tokens: ['title', 'fructose', 'hamper', 'hormone', 'control']...) from 43 documents (total 539 corpus positions)

2017-12-07 14:05:24,705 : INFO : saving Dictionary object under C:/Users/lohit/AppData/Local/Temp/den.dict, separately None

2017-12-07 14:05:24,713 : INFO : saved C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 14:05:24,713 : INFO : storing corpus in Matrix Market format to C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 14:05:24,717 : INFO : saving sparse matrix to C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 14:05:24,719 : INFO : PROGRESS: saving document #0

2017-12-07 14:05:24,720 : INFO : saved 43x301 matrix, density=3.554% (460/12943)

2017-12-07 14:05:24,734 : INFO : saving MmCorpus index to C:/Users/lohit/AppData/Local/Temp/den.mm.index

2017-12-07 14:05:24,740 : INFO : collecting document frequencies

2017-12-07 14:05:24,747 : INFO : PROGRESS: processing document #0

2017-12-07 14:05:24,749 : INFO : calculating IDF weights for 43 documents and 300 features (460 matrix non-zeros)

2017-12-07 14:05:24,751 : INFO : loading Dictionary object from C:\Users\lohit\AppData\Local\Temp\den.dict

2017-12-07 14:05:24,752 : INFO : loaded C:\Users\lohit\AppData\Local\Temp\den.dict

2017-12-07 14:05:24,754 : INFO : loaded corpus index from C:\Users\lohit\AppData\Local\Temp\den.mm.index

2017-12-07 14:05:24,756 : INFO : initializing corpus reader from C:\Users\lohit\AppData\Local\Temp\den.mm

2017-12-07 14:05:24,757 : INFO : accepted corpus with 43 documents, 301 features, 460 non-zero entries

2017-12-07 14:05:24,758 : INFO : collecting document frequencies

2017-12-07 14:05:24,760 : INFO : PROGRESS: processing document #0

2017-12-07 14:05:24,767 : INFO : calculating IDF weights for 43 documents and 300 features (460 matrix non-zeros)

2017-12-07 14:05:24,769 : INFO : using serial LSI version on this node

2017-12-07 14:05:24,772 : INFO : updating model with new documents

2017-12-07 14:05:24,779 : INFO : preparing a new chunk of documents

2017-12-07 14:05:24,784 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 14:05:24,790 : INFO : 1st phase: constructing (301, 150) action matrix

2017-12-07 14:05:24,791 : INFO : orthonormalizing (301, 150) action matrix

2017-12-07 14:05:24,804 : INFO : 2nd phase: running dense svd on (150, 43) matrix

2017-12-07 14:05:24,810 : INFO : computing the final decomposition

2017-12-07 14:05:24,811 : INFO : keeping 20 factors (discarding 0.000% of energy spectrum)

2017-12-07 14:05:24,814 : INFO : processed documents up to #43

2017-12-07 14:05:24,815 : INFO : topic #0(1.457): 0.261\*"rat" + 0.232\*"leptin" + 0.228\*"fructose" + 0.223\*"study" + 0.213\*"resistance" + 0.209\*"obesity" + 0.192\*"diet" + 0.172\*"group" + 0.165\*"finding" + 0.144\*"cause"

2017-12-07 14:05:24,816 : INFO : topic #1(1.248): 0.481\*"health" + 0.382\*"center" + 0.350\*"science" + 0.291\*"florida" + 0.257\*"university" + -0.155\*"rat" + -0.145\*"group" + 0.120\*"nursing" + 0.119\*"research" + 0.096\*"medicine"

2017-12-07 14:05:24,818 : INFO : topic #2(1.203): -0.330\*"rat" + -0.321\*"group" + -0.192\*"responded" + 0.178\*"study" + -0.177\*"health" + 0.168\*"resistance" + 0.164\*"cause" + 0.162\*"obesity" + -0.153\*"two" + -0.134\*"received"

2017-12-07 14:05:24,819 : INFO : topic #3(1.063): 0.192\*"author" + 0.177\*"johnson" + 0.172\*"shapiro" + 0.170\*"phd" + 0.150\*"lead" + 0.150\*"therapeutic" + 0.150\*"pharmacology" + 0.140\*"diet" + -0.134\*"appetite" + -0.129\*"hamper"

2017-12-07 14:05:24,820 : INFO : topic #4(1.028): 0.229\*"cause" + -0.217\*"appetite" + -0.209\*"title" + -0.209\*"find" + -0.209\*"control" + -0.209\*"hamper" + -0.188\*"hormone" + 0.174\*"way" + 0.167\*"work" + 0.161\*"alters"

2017-12-07 14:05:24,821 : INFO : topic #0(1.457): 0.261\*"rat" + 0.232\*"leptin" + 0.228\*"fructose" + 0.223\*"study" + 0.213\*"resistance" + 0.209\*"obesity" + 0.192\*"diet" + 0.172\*"group" + 0.165\*"finding" + 0.144\*"cause"

2017-12-07 14:05:24,822 : INFO : topic #1(1.248): 0.481\*"health" + 0.382\*"center" + 0.350\*"science" + 0.291\*"florida" + 0.257\*"university" + -0.155\*"rat" + -0.145\*"group" + 0.120\*"nursing" + 0.119\*"research" + 0.096\*"medicine"

2017-12-07 14:05:24,824 : INFO : topic #2(1.203): -0.330\*"rat" + -0.321\*"group" + -0.192\*"responded" + 0.178\*"study" + -0.177\*"health" + 0.168\*"resistance" + 0.164\*"cause" + 0.162\*"obesity" + -0.153\*"two" + -0.134\*"received"

2017-12-07 14:05:24,825 : INFO : topic #3(1.063): 0.192\*"author" + 0.177\*"johnson" + 0.172\*"shapiro" + 0.170\*"phd" + 0.150\*"lead" + 0.150\*"therapeutic" + 0.150\*"pharmacology" + 0.140\*"diet" + -0.134\*"appetite" + -0.129\*"hamper"

2017-12-07 14:05:24,826 : INFO : topic #4(1.028): 0.229\*"cause" + -0.217\*"appetite" + -0.209\*"title" + -0.209\*"find" + -0.209\*"control" + -0.209\*"hamper" + -0.188\*"hormone" + 0.174\*"way" + 0.167\*"work" + 0.161\*"alters"

2017-12-07 14:05:24,828 : INFO : topic #5(1.013): 0.294\*"cause" + 0.242\*"obesity" + 0.235\*"develops" + -0.193\*"vasselli" + -0.184\*"it" + -0.165\*"lot" + -0.149\*"happen" + -0.149\*"think" + -0.149\*"i" + -0.149\*"investigated"

2017-12-07 14:05:24,829 : INFO : topic #6(1.002): -0.203\*"show" + -0.139\*"block" + -0.136\*"new" + -0.133\*"chewing" + -0.133\*"future" + -0.133\*"setting" + -0.133\*"year" + -0.133\*"slurping" + -0.133\*"haunt" + -0.133\*"job"

2017-12-07 14:05:24,830 : INFO : topic #7(1.000): 1.000\*"source" + -0.000\*"body" + -0.000\*"could" + -0.000\*"show" + -0.000\*"new" + 0.000\*"health" + -0.000\*"chewing" + -0.000\*"year" + -0.000\*"come" + -0.000\*"blood"

2017-12-07 14:05:24,832 : INFO : topic #8(0.983): 0.194\*"title" + 0.194\*"hamper" + 0.194\*"find" + 0.194\*"control" + -0.177\*"johnson" + 0.161\*"amount" + 0.154\*"set" + 0.154\*"eat" + 0.153\*"diet" + 0.151\*"hormone"

2017-12-07 14:05:24,833 : INFO : topic #9(0.971): 0.210\*"johnson" + 0.207\*"way" + 0.190\*"control" + 0.190\*"title" + 0.190\*"hamper" + 0.190\*"find" + 0.160\*"alters" + 0.160\*"according" + 0.152\*"cause" + 0.143\*"appetite"

2017-12-07 14:05:24,834 : INFO : topic #10(0.965): 0.147\*"could" + 0.146\*"eat" + 0.146\*"set" + 0.144\*"blood" + -0.140\*"group" + 0.132\*"body" + -0.131\*"received" + -0.130\*"lead" + -0.130\*"therapeutic" + -0.130\*"pharmacology"

2017-12-07 14:05:24,836 : INFO : topic #11(0.945): -0.203\*"confirm" + -0.203\*"course" + -0.203\*"yet" + -0.203\*"current" + -0.203\*"apply" + 0.181\*"research" + -0.170\*"finding" + 0.148\*"associate" + -0.133\*"alters" + -0.133\*"according"

2017-12-07 14:05:24,837 : INFO : topic #12(0.931): -0.238\*"blood" + 0.234\*"research" + 0.201\*"associate" + 0.162\*"received" + -0.129\*"six" + -0.129\*"month" + 0.120\*"group" + -0.119\*"identical" + -0.119\*"glucose" + -0.119\*"level"

2017-12-07 14:05:24,838 : INFO : topic #13(0.918): -0.258\*"physiology" + 0.171\*"research" + 0.165\*"control" + 0.165\*"find" + 0.165\*"hamper" + 0.165\*"title" + 0.142\*"associate" + -0.129\*"everything" + -0.129\*"playing" + -0.129\*"published"

2017-12-07 14:05:24,839 : INFO : topic #14(0.899): -0.198\*"highfat" + -0.165\*"increasing" + 0.150\*"work" + 0.145\*"blood" + -0.138\*"diet" + 0.136\*"according" + 0.136\*"alters" + 0.131\*"scientist" + -0.129\*"resistance" + 0.127\*"action"

2017-12-07 14:05:24,841 : INFO : topic #15(0.880): -0.305\*"commentary" + -0.305\*"help" + -0.305\*"wrote" + -0.228\*"journal" + -0.208\*"also" + -0.201\*"researcher" + 0.174\*"current" + 0.174\*"apply" + 0.174\*"yet" + 0.174\*"confirm"

2017-12-07 14:05:24,842 : INFO : topic #16(0.867): 0.225\*"received" + -0.192\*"responded" + -0.183\*"resistant" + 0.155\*"blood" + -0.149\*"according" + -0.149\*"alters" + -0.133\*"work" + -0.128\*"occurring" + -0.128\*"sign" + -0.128\*"change"

2017-12-07 14:05:24,843 : INFO : topic #17(0.840): -0.347\*"develops" + 0.269\*"alters" + 0.269\*"according" + 0.187\*"way" + 0.185\*"work" + -0.179\*"responded" + -0.174\*"silent" + -0.174\*"introduced" + -0.174\*"undetected" + -0.174\*"expected"

2017-12-07 14:05:24,844 : INFO : topic #18(0.824): 0.383\*"responded" + -0.211\*"received" + 0.192\*"tested" + 0.192\*"normally" + 0.192\*"become" + 0.192\*"discovered" + 0.164\*"increasing" + 0.149\*"eating" + 0.126\*"amount" + -0.112\*"fructosefed"

2017-12-07 14:05:24,846 : INFO : topic #19(0.701): 0.306\*"science" + 0.301\*"university" + 0.290\*"florida" + -0.237\*"nursing" + 0.192\*"center" + -0.175\*"medicine" + -0.158\*"public" + -0.158\*"veterinary" + -0.158\*"patient" + -0.158\*"medical"

2017-12-07 14:05:24,847 : INFO : saving Projection object under C:\Users\lohit\AppData\Local\Temp\model.lsi.projection, separately None

2017-12-07 14:05:24,854 : INFO : saved C:\Users\lohit\AppData\Local\Temp\model.lsi.projection

2017-12-07 14:05:24,855 : INFO : saving LsiModel object under C:\Users\lohit\AppData\Local\Temp\model.lsi, separately None

2017-12-07 14:05:24,856 : INFO : not storing attribute projection

2017-12-07 14:05:24,857 : INFO : not storing attribute dispatcher

2017-12-07 14:05:24,866 : INFO : saved C:\Users\lohit\AppData\Local\Temp\model.lsi

2017-12-07 14:05:24,867 : INFO : loading Dictionary object from C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 14:05:24,869 : INFO : loaded C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 14:05:24,874 : INFO : loaded corpus index from C:/Users/lohit/AppData/Local/Temp/den.mm.index

2017-12-07 14:05:24,875 : INFO : initializing corpus reader from C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 14:05:24,878 : INFO : accepted corpus with 43 documents, 301 features, 460 non-zero entries

2017-12-07 14:05:24,881 : WARNING : scanning corpus to determine the number of features (consider setting `num\_features` explicitly)

2017-12-07 14:05:24,887 : INFO : creating matrix with 43 documents and 20 features

2017-12-07 14:05:24,897 : INFO : saving MatrixSimilarity object under C:/Users/lohit/AppData/Local/Temp/den.index, separately None

2017-12-07 14:05:24,904 : INFO : saved C:/Users/lohit/AppData/Local/Temp/den.index

2017-12-07 14:05:24,905 : INFO : loading MatrixSimilarity object from C:/Users/lohit/AppData/Local/Temp/den.index

2017-12-07 14:05:24,908 : INFO : loaded C:/Users/lohit/AppData/Local/Temp/den.index

[(0, 0.022665747), (1, 0.0), (2, 0.0), (3, 0.0), (4, 0.0), (5, 0.0), (6, 0.010256931), (7, 0.0), (8, 0.67596281), (9, 0.0), (10, 0.70037252), (11, 0.0), (12, 0.013728365), (13, 0.0), (14, 0.0), (15, 0.0), (16, 0.0), (17, 0.0), (18, 0.0), (19, 0.0), (20, 2.2351742e-08)]

[(10, 0.70037252), (8, 0.67596281), (0, 0.022665747), (12, 0.013728365), (6, 0.010256931), (20, 2.2351742e-08), (1, 0.0), (2, 0.0), (3, 0.0), (4, 0.0), (5, 0.0), (7, 0.0), (9, 0.0), (11, 0.0), (13, 0.0), (14, 0.0), (15, 0.0), (16, 0.0), (17, 0.0), (18, 0.0), (19, 0.0)]

{'title': 0, 'fructose': 1, 'hamper': 2, 'hormone': 3, 'control': 4, 'appetite': 5, 'uf': 6, 'study': 7, 'find': 8, 'could': 9, 'year': 10, 'chewing': 11, 'candy': 12, 'slurping': 13, 'sugary': 14, 'soda': 15, 'come': 16, 'back': 17, 'haunt': 18, 'you': 19, 'perhaps': 20, 'new': 21, 'university': 22, 'florida': 23, 'rat': 24, 'show': 25, 'fructosefilled': 26, 'diet': 27, 'block': 28, 'appetitecontrolling': 29, 'leptin': 30, 'job': 31, 'setting': 32, 'body': 33, 'future': 34, 'obesity': 35, 'critical': 36, 'controlling': 37, 'energy': 38, 'expenditure': 39, 'scientist': 40, 'long': 41, 'linked': 42, 'resistance': 43, 'several': 44, 'shown': 45, 'overconsumption': 46, 'sugar': 47, 'found': 48, 'everything': 49, 'apple': 50, 'cooky': 51, 'playing': 52, 'significant': 53, 'role': 54, 'epidemic': 55, 'recently': 56, 'published': 57, 'american': 58, 'journal': 59, 'physiology': 60, 'regulatory': 61, 'integrative': 62, 'comparative': 63, 'first': 64, 'link': 65, 'researcher': 66, 'became': 67, 'resistant': 68, 'fed': 69, 'high': 70, 'six': 71, 'month': 72, 'although': 73, 'visible': 74, 'sign': 75, 'change': 76, 'occurring': 77, 'fructosefed': 78, 'gained': 79, 'considerably': 80, 'weight': 81, 'never': 82, 'ate': 83, 'group': 84, 'switched': 85, 'highfat': 86, 'condition': 87, 'lead': 88, 'coupled': 89, 'surprising': 90, 'finding': 91, 'increasing': 92, 'amount': 93, 'without': 94, 'calorie': 95, 'led': 96, 'later': 97, 'exacerbated': 98, 'paired': 99, 'said': 100, 'philip': 101, 'j': 102, 'scarpace': 103, 'phd': 104, 'professor': 105, 'pharmacology': 106, 'therapeutic': 107, 'college': 108, 'medicine': 109, 'senior': 110, 'author': 111, 'according': 112, 'cause': 113, 'alters': 114, 'way': 115, 'work': 116, 'it': 117, 'action': 118, 'likely': 119, 'blocking': 120, 'entry': 121, 'brain': 122, 'alexandra': 123, 'shapiro': 124, 'assistant': 125, 'department': 126, 'test': 127, 'affect': 128, 'studied': 129, 'two': 130, 'received': 131, 'number': 132, 'day': 133, 'one': 134, 'chow': 135, 'containing': 136, '60': 137, 'percent': 138, 'kept': 139, 'fructosefree': 140, 'after': 141, 'detect': 142, 'difference': 143, 'exception': 144, 'elevation': 145, 'blood': 146, 'triglyceride': 147, 'they': 148, 'identical': 149, 'fat': 150, 'well': 151, 'level': 152, 'insulin': 153, 'glucose': 154, 'cholesterol': 155, 'tested': 156, 'responded': 157, 'discovered': 158, 'eating': 159, 'become': 160, 'normally': 161, 'from': 162, 'overall': 163, 'point': 164, 'view': 165, 'enough': 166, 'concentration': 167, 'induce': 168, 'implicate': 169, 'dietary': 170, 'potential': 171, 'risk': 172, 'factor': 173, 'human': 174, 'joseph': 175, 'vasselli': 176, 'research': 177, 'associate': 178, 'center': 179, 'st': 180, 'lukesroosevelt': 181, 'hospital': 182, 'york': 183, 'columbia': 184, 'typically': 185, 'develops': 186, 'showed': 187, 'silent': 188, 'undetected': 189, 'introduced': 190, 'greater': 191, 'expected': 192, 'set': 193, 'up': 194, 'if': 195, 'applicable': 196, 'consequence': 197, 'also': 198, 'consume': 199, 'excessive': 200, 'go': 201, 'trip': 202, 'attend': 203, 'celebration': 204, 'otherwise': 205, 'eat': 206, 'usually': 207, 'person': 208, 'consuming': 209, 'lowfructose': 210, 'may': 211, 'able': 212, 'handle': 213, 'individual': 214, 'longer': 215, 'unable': 216, 'burn': 217, 'extra': 218, 'gain': 219, 'lot': 220, 'current': 221, 'apply': 222, 'course': 223, 'yet': 224, 'confirm': 225, 'wrote': 226, 'commentary': 227, 'help': 228, 'i': 229, 'think': 230, 'important': 231, 'raise': 232, 'issue': 233, 'investigated': 234, 'happen': 235, 'collaborated': 236, 'wei': 237, 'mu': 238, 'carlos': 239, 'roncal': 240, 'kityan': 241, 'cheng': 242, 'richard': 243, 'johnson': 244, 'md': 245, 'rodale': 246, 'book': 247, 'the': 248, 'fix': 249, 'highfructose': 250, 'fallout': 251, 'making': 252, 'sick': 253, 'health': 254, 'science': 255, 'comprehensive': 256, 'academic': 257, 'southeast': 258, 'dedicated': 259, 'highquality': 260, 'program': 261, 'education': 262, 'patient': 263, 'care': 264, 'public': 265, 'service': 266, 'encompasses': 267, 'dentistry': 268, 'profession': 269, 'nursing': 270, 'pharmacy': 271, 'veterinary': 272, 'medical': 273, 'teaching': 274, 'campus': 275, 'jacksonville': 276, 'offering': 277, 'graduate': 278, 'activity': 279, 'banner': 280, 'ufshands': 281, 'provided': 282, 'network': 283, 'clinic': 284, 'gainesville': 285, 'statewide': 286, 'presence': 287, 'satellite': 288, 'dental': 289, 'staffed': 290, 'professional': 291, 'affiliation': 292, 'communitybased': 293, 'healthcare': 294, 'facility': 295, 'stretching': 296, 'hialeah': 297, 'miami': 298, 'panhandle': 299, 'source': 300}

Folder "C:\Users\lohit\AppData\Local\Temp" will be used to save temporary dictionary and corpus.

Used files generated before

[(0, 0.11723193089076037), (1, 0.21479615976407582), (2, 0.00094934305189918813), (3, 0.042060928900179283), (4, 0.010453043256257106), (5, 0.0071896165769442834), (6, 0.00099184298416972461), (8, -0.044923578184752878), (9, -0.07914237852480141), (10, -0.0084349730733776704), (11, 0.17369737984162453), (12, 0.21202924746566076), (13, 0.15888175844353661), (14, -0.029146460769247921), (15, 0.083143533656249061), (16, 0.02037499709209846), (17, 0.1330503389232166), (18, 0.10441096262517838), (19, -0.26882420769626342)]

[(0, 0.018363506), (1, 0.0), (2, 0.0), (3, 0.0), (4, 0.0), (5, 0.0), (6, 0.017424561), (7, 0.0), (8, 0.043857157), (9, 0.0), (10, 0.024567459), (11, 0.0), (12, 0.27098024), (13, 0.0), (14, 0.032384902), (15, 0.0), (16, 0.017319418), (17, 0.0), (18, 0.016137414), (19, 0.0), (20, 0.018494442), (21, 0.0), (22, 0.0054478981), (23, 0.0), (24, 0.74505937), (25, 0.0), (26, 0.05210327), (27, 0.0), (28, 0.029812388), (29, 0.0), (30, 0.024801657), (31, 0.0), (32, 0.023471013), (33, 0.0), (34, 0.018286549), (35, 0.0), (36, 0.006697949), (37, 0.0), (38, 0.0), (39, 0.48189354), (40, 0.0), (41, 0.0), (42, -0.0021772683)]

[(24, 0.74505937), (39, 0.48189354), (12, 0.27098024), (26, 0.05210327), (8, 0.043857157), (14, 0.032384902), (28, 0.029812388), (30, 0.024801657), (10, 0.024567459), (32, 0.023471013), (20, 0.018494442), (0, 0.018363506), (34, 0.018286549), (6, 0.017424561), (16, 0.017319418), (18, 0.016137414), (36, 0.006697949), (22, 0.0054478981), (1, 0.0), (2, 0.0), (3, 0.0), (4, 0.0), (5, 0.0), (7, 0.0), (9, 0.0), (11, 0.0), (13, 0.0), (15, 0.0), (17, 0.0), (19, 0.0), (21, 0.0), (23, 0.0), (25, 0.0), (27, 0.0), (29, 0.0), (31, 0.0), (33, 0.0), (35, 0.0), (37, 0.0), (38, 0.0), (40, 0.0), (41, 0.0), (42, -0.0021772683)]

2017-12-07 14:05:24,924 : INFO : adding document #0 to Dictionary(0 unique tokens: [])

2017-12-07 14:05:24,924 : INFO : built Dictionary(285 unique tokens: ['title', 'diabetes', 'pathfinder', 'award', 'go']...) from 27 documents (total 447 corpus positions)

2017-12-07 14:05:24,928 : INFO : saving Dictionary object under C:/Users/lohit/AppData/Local/Temp/den.dict, separately None

2017-12-07 14:05:24,932 : INFO : saved C:/Users/lohit/AppData/Local/Temp/den.dict

2017-12-07 14:05:24,937 : INFO : storing corpus in Matrix Market format to C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 14:05:24,939 : INFO : saving sparse matrix to C:/Users/lohit/AppData/Local/Temp/den.mm

2017-12-07 14:05:24,940 : INFO : PROGRESS: saving document #0

2017-12-07 14:05:24,942 : INFO : saved 27x285 matrix, density=4.964% (382/7695)

2017-12-07 14:05:24,951 : INFO : saving MmCorpus index to C:/Users/lohit/AppData/Local/Temp/den.mm.index

2017-12-07 14:05:24,958 : INFO : collecting document frequencies

2017-12-07 14:05:24,959 : INFO : PROGRESS: processing document #0

2017-12-07 14:05:24,966 : INFO : calculating IDF weights for 27 documents and 284 features (382 matrix non-zeros)

2017-12-07 14:05:24,968 : INFO : loading Dictionary object from C:\Users\lohit\AppData\Local\Temp\den.dict

2017-12-07 14:05:24,971 : INFO : loaded C:\Users\lohit\AppData\Local\Temp\den.dict

2017-12-07 14:05:24,974 : INFO : loaded corpus index from C:\Users\lohit\AppData\Local\Temp\den.mm.index

2017-12-07 14:05:24,979 : INFO : initializing corpus reader from C:\Users\lohit\AppData\Local\Temp\den.mm

2017-12-07 14:05:24,981 : INFO : accepted corpus with 27 documents, 285 features, 382 non-zero entries

2017-12-07 14:05:24,984 : INFO : collecting document frequencies

2017-12-07 14:05:24,986 : INFO : PROGRESS: processing document #0

2017-12-07 14:05:24,989 : INFO : calculating IDF weights for 27 documents and 284 features (382 matrix non-zeros)

2017-12-07 14:05:24,990 : INFO : using serial LSI version on this node

2017-12-07 14:05:24,991 : INFO : updating model with new documents

2017-12-07 14:05:24,995 : INFO : preparing a new chunk of documents

2017-12-07 14:05:24,999 : INFO : using 100 extra samples and 2 power iterations

2017-12-07 14:05:25,000 : INFO : 1st phase: constructing (285, 150) action matrix

2017-12-07 14:05:25,002 : INFO : orthonormalizing (285, 150) action matrix

2017-12-07 14:05:25,007 : INFO : 2nd phase: running dense svd on (150, 27) matrix

2017-12-07 14:05:25,012 : INFO : computing the final decomposition

2017-12-07 14:05:25,013 : INFO : keeping 11 factors (discarding 0.000% of energy spectrum)

2017-12-07 14:05:25,015 : INFO : processed documents up to #27

2017-12-07 14:05:25,017 : INFO : topic #0(1.330): -0.304\*"einstein" + -0.250\*"diabetes" + -0.230\*"college" + -0.215\*"center" + -0.212\*"medicine" + -0.194\*"albert" + -0.167\*"award" + -0.162\*"research" + -0.147\*"zang" + -0.145\*"pathfinder"

2017-12-07 14:05:25,018 : INFO : topic #1(1.154): 0.305\*"b7x" + -0.238\*"college" + -0.226\*"einstein" + -0.219\*"medicine" + -0.201\*"albert" + 0.189\*"lymphocyte" + 0.172\*"protein" + 0.166\*"dr" + 0.157\*"t1d" + 0.151\*"zang"

2017-12-07 14:05:25,020 : INFO : topic #2(1.039): 0.219\*"insulin" + -0.158\*"award" + 0.146\*"attack" + 0.146\*"injection" + 0.146\*"patient" + 0.146\*"cell" + 0.146\*"pancreatic" + -0.142\*"b7x" + -0.126\*"pathfinder" + -0.114\*"title"

2017-12-07 14:05:25,021 : INFO : topic #3(1.007): 0.300\*"title" + 0.300\*"go" + 0.289\*"award" + -0.259\*"b7x" + 0.223\*"investigator" + 0.217\*"pathfinder" + -0.149\*"protein" + 0.124\*"improving" + 0.124\*"promise" + 0.124\*"prevention"

2017-12-07 14:05:25,022 : INFO : topic #4(1.000): -0.577\*"michael" + -0.577\*"source" + -0.577\*"heller" + 0.000\*"fiveyear" + 0.000\*"diabetes" + -0.000\*"pathfinder" + 0.000\*"molecular" + 0.000\*"phd" + 0.000\*"15" + 0.000\*"protein"

2017-12-07 14:05:25,024 : INFO : topic #0(1.330): -0.304\*"einstein" + -0.250\*"diabetes" + -0.230\*"college" + -0.215\*"center" + -0.212\*"medicine" + -0.194\*"albert" + -0.167\*"award" + -0.162\*"research" + -0.147\*"zang" + -0.145\*"pathfinder"