Machine learning based patient diagnosis (CMB-100)

CMB-100 DEMO: Machine Learning In Diagnosis (Case-3)

Code

Jinsong Zhang

2018-08-17

A. Introduction

1. Machine learning is a powerful way to discover underlying relationship(s) embedded within a large number of objects (words, patient samples,genes, etc.).

2. Text mining is now a very popular area of study, used in studies such as topics finding, natural language precessing.

3. Deep learning / neuronal network is a specific type of machinery learning and has been very successful in image processing.

4. Data Science Bowl 2017 is featured by lung cancer diagnosis using the deep learning technology (https://www.kaggle.com/c/data-science-bowl-2017 (https://www.kaggle.com/c/data-science-bowl-2017)).

Questions to answer:

1. How many topics are embedded in the case report?
2. Can we get the correct diagnosis by examining these topics?

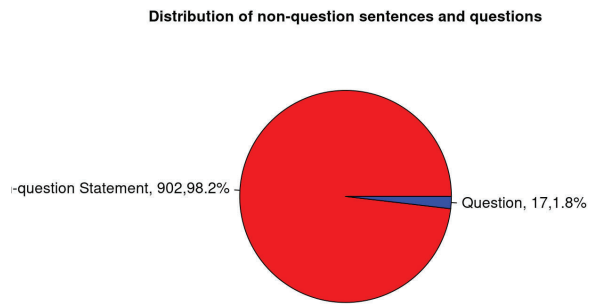
B. Dataset and pre-processing

1. The entire report downloaded from Google Docs was converted into a text file

2. All sentences were extracted from the file.

3. Sentences containing questions were also extracted into a separate group.

4. Distribution of non-question sentences and questions in the document.



5. In total, there are 932 sentences to be analyzed.

Representative sentences/text:


```
##  [1] "* Tenderness in the deltoids."
##  [2] "* Plethysmography: An airtight box in which the patient sits/stands."
##  [3] "* Joint pain in ankles, wrists, fingers, back."
##  [4] "Age 12: Femoral anteversion."
##  [5] "-widespread joint pain/inflammation."
##  [6] "Most cases described in the literature have significant heart problems, such as mitral
valve insufficiency and regurgitation, which require surgery."
##  [7] "* Osteoarthritis - cartilage that cushions the ends of bones in your joints gradually d
eteriorates; risk factors: old age, women more likely to be affected, obesity injuries from spor
ts or accidents increase risk, can be inherited ."
##  [8] "-       typically sporadic, occurring in people with no family history of the conditio
n."
##  [9] "Regular ultrasound cannot detect blood flow, but Doppler estimates the rate of change i
n pitch coming from the blood vessel."
## [10] "Ehlers-Danlos syndrome with progressive kyphoscoliosis, myopathy, and hearing loss; inb
orn genetic issues."
## [11] "* Negative ANA, c-ANCA, and p-ANCA."
## [12] "Essentially, a disease appears more severe with each succeeding generation."
## [13] "American Journal of Sports Medicine 8, 3 (1980)."
## [14] "Symptoms: most common, but not seen in every case (no case of lupus is the same)."
## [15] "* Method: Blood is placed into a tall, thin tube and red blood cells are allowed to set
tle for 1 hour."
## [16] "There is nothing that suggests any connection between femoral anteversion and any other
symptoms that Susan experienced, so that is rather an unrelated incident that was worth investig
ating."
## [17] "gov/pmc/articles/PMC4026000/."
## [18] "* There are common growth abnormalities, including leg length discrepancy."
## [19] "* Bizzare parosteal osteochondromatous proliferation: no bone tumor."
## [20] "Craniofacial dysmorphism (ex: strabismus microretrognathia, low-set and posteriorly rot
ated ears, etc."
```

C. Building the model using Latent Dirichlet Allocation Algorithm

C.1. Convert text corpus into a document matrix

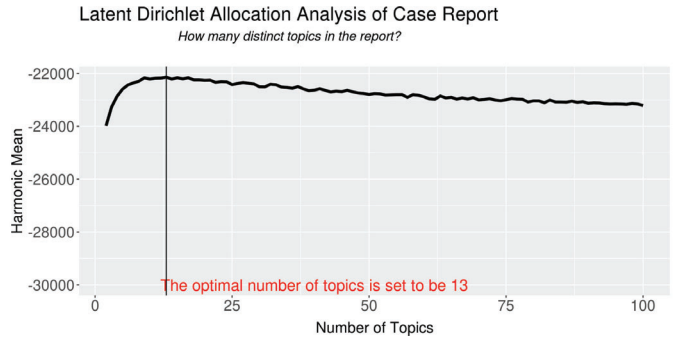
1. A text corpus was created to contain all sentences.

2. Further processing the corpus to remove punctuation, non-text symbols was conducted

3. Finally, the corpus was convected to a document matrix with rows corresponding to sentences and columns corresponding to terms/words.


```
## List of 6
##  $ i       : int [1:3216] 1 1 1 1 1 1 2 2 2 2 ...
##  $ j       : int [1:3216] 1 2 3 4 5 6 7 8 9 10 ...
##  $ v       : num [1:3216] 1 1 1 1 1 1 1 1 1 1 ...
##  $ nrow    : int 544
##  $ ncol    : int 1440
##  $ dimnames:List of 2
##   ..$ Docs : chr [1:544] "3" "4" "6" "8" ...
##   ..$ Terms: chr [1:1440] "disorders" "father" "genetic" "information" ...
##  - attr(*, "class")= chr [1:2] "DocumentTermMatrix" "simple_triplet_matrix"
##  - attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

C.2. Determining the optimal number of topics using gradient ascent



C.3. Building the model using 10 topics

```
## A LDA_Gibbs topic model with 13 topics.
```

```
##    Length     Class      Mode
##         1 LDA_Gibbs        S4
```

D. Exploring Results to find topics-specific tags and top questions within each topic

D.1. Representative terms within each topic

Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topi
skin	muscle	bone	pain	eds	femoral	tissue	severe	syndrome	disorders	blo
genetic	pain	cola	eds	gene	anteversion	type	hip	tissue	inflammation	an
toimmune	joints	strength	tests	vascular	recessive	collagen	disorder	rarediseases	result	lup
positive	diseases	arthritis	imperfecta	age	hypermobility	autosomal	feet	blood	crp	shou
type	knee	joints	medical	inflammatory	heterozygous	impingement	condition	skin	susan	eleva
physical	susan	fatigue	bone	polyarthritis	diseases	used	anticipation	abnormal	esr	syste
body	age	upper	ultrasound	including	knees	susan	early	weakness	purpose	shoul
fingers	jia	motion	polyarthritis	therapy	connective	antibodies	thin	decreased	swelling	ner
lung	others	abnormalities	increased	genes	chronic	disorder	missense	idiopathic	tendon	err
back	crepitus	dislocations	plod	arterial	jia	arms	rate	impingement	rheumatoid	has
toes	hypermobility	joint	susan	lupus	often	complications	fractures	benign	surgery	oligoar
yndrome	resulting	cytoplasmic	age	tendonitis	skeletal	muscular	connective	juvenile	physical	kne
mozygous	musculoskeletal	leg	rotation	blue	ages	rupture	stiffness	issues	method	bilate
lisorders	lower	eds	arm	william	arthritis	cardiac	accompanied	spinal	response	arthi
low	occurring	improves	evaluate	regular	typically	loss	asthma	recessive	prednisone	mani
rash	therapy	collagen	neer	ana	forms	antineutrophil	inward	com	rotated	go
fragile	scoliosis	positive	onset	cytoplasmic	appearance	resistance	volume	cracking	increase	la
nportant	immediate	clinical	affects	internal	extracellular	injuries	onset	proteins	autosomal	resu
hips	fingers	results	myasthenic	cuff	positive	appears	density	evaluator	muscular	antinu
tests	types	panca	recurrent	rotator	incident	sports	deficiency	subacromial	variants	lo
bilateral	typically	patellar	iii	vessel	correction	presentation	tenascin	motor	congenital	rese
romyalgia	women	leads	uterine	capacity	neurologic	area	loss	studies	hypotonia	exter
ndicated	learn	form	missense	clumps	assess	mild	swelling	affects	protein	hea
muscle	tendons	idiopathic	fragility	values	hand	ligaments	around	evidence	known	elect
canca	asthma	appear	diagnose	pediatric	juvenile	hypotonia	weakness	depression	arthritis	cel
dditional	activity	rare	nervous	touch	imaging	white	patterns	bones	indication	ran
arms	electrode	characterized	osteoarthritis	structure	rashes	bilateral	functional	acid	michael	redn
sound	range	ankles	autoimmune	swollen	bones	swelling	findings	difficulty	weakness	fev
ndicative	numbness	childhood	supraspinatus	lower	nrs	young	worse	myopathy	american	oft
umbness	management	mutation	around	tenderness	absence	immune	negative	trait	lesions	enoi

2. Topic-specific Labels

Appendix_I

Topic	Label	# of Terms
1	MUTATIONS, ORG, LOSS	72
2	SKIN, GENETIC, AUTOIMMUNE	58
3	MUSCLE, PAIN, JOINTS	59
4	BONE, COLA, STRENGTH	60
5	PAIN, EDS, TESTS	48
6	EDS, GENE, VASCULAR	47
7	FEMORAL, ANTEVERSION, RECESSIVE	37
8	TISSUE, TYPE, COLLAGEN	29
9	SEVERE, HIP, DISORDER	34
10	SYNDROME, TISSUE, RAREDISEASES	28
11	DISORDERS, INFLAMMATION, RESULT	27
12	BLOOD, ANA, LUPUS	21
13	JOINT, EHLERSDANLOS, PAIN	24

```
## png
##   2
```

3. Top 10 sentences within each of topics

op 10 Questions from Topic 1 (MUTATIONS, ORG, LOSS)
d intractable periodontitis with early onset.
IST14 dermatan 4-O-sulfotransferase-1 D4ST1 , which is responsible for the biosynthesis of dermatan sulfate.
IST14 dermatan 4-O-sulfotransferase-1 D4ST1 , which is responsible for the biosynthesis of dermatan sulfate.
sess the health of muscles and motor neurons.
Spondylodysplalstic Carson .
, numbness also, and trembling hyper mobility related to Parkinson , loss of sensation.
ionnaires and reported 7 or more symptoms at the Small Fiber Neuropathy Symptoms Inventory Questionnaire.
ations, including increased risk for arterial ruptures, uterine ruptures, colonic perforations.
-negative autoimmune testing.
wollen joint and fever; caused by penetrating injury to joint or bacteria that travels from another part of the body; infants and older adults

op 10 Questions from Topic 2 (KIN, GENETIC, AUTOIMMUNE)
supraspinatus and infraspinatus shoulders ; positive Neer and Hawkins tests indicative of issues w superspinatus and other structure
mal wound healing, fragile blood vessels, osteopenia.
ive kyphoscoliosis, myopathy, and hearing loss; inborn genetic issues.
in and silvery scales, more likely to cause swollen fingers and toes, foot and lower back pain; genetic environmental causes.
, and p-ANCA not indicative of auto-immune disorder .
important for skin, bones, connective tissues.
disorders, as a number of other disorders have been discovered through research on it.
vers the cheeks and bridge of the nose or rashes elsewhere on the body.
gile X Syndrome and Huntington Disease often exhibit genetic anticipation.
umothorax collection of air between the lung and chest wall, impairing proper lung inflation are commonly experienced.

p 10 Questions from Topic 3 (MUSCLE, PAIN, JOINTS)	Topic	
typically spreads from one part of body to others bilateral! * What are the common types of arthritis pain?	3	and in women m pa
e pain + shoulder pain + numbness in both hands.	3	
2 or more limbs, recurring daily for at least 3 months.	3	M
phy, or muscle biopsy to test for muscle abnormalities.	3	Ultra
ral valve prolapse, recurrent hernias, musculoskeletal pain, and recurrent joint dislocations.	3	Pt typica
erted directly into a muscle to record electrical activity in that muscle.	3	* Needle E
f the nerve-rich membrane in the back of the eye retina are common and can result in vision loss.	3	Microcornea, n
and memory loss Susan nervous system symptoms?	3	* H
VI extracellular matrix of skeletal muscle and XII component of skeletal muscle .	3	Not very ur
praspinatus muscle and tendons.	3	

Appendix_I

p 10 Questions from Topic 4 (BONE, COLA, STRENGTH)
ch as heart surgery, pain management and minimizing pressure on joints.
ng childhood and adolescence that often result from minor trauma.
r two or more atraumatic dislocations in two different joints occurring at different times.
mothorax collection of air between the lung and chest wall, impairing proper lung inflation are commonly experienced.
plasmic Antibodies p-ANCA , Cytoplasmic Anti-neutrophil Cytoplasmic Antibodies c-ANCA .
l range of hip motion and normal strength in hips.
se swelling of joins; no joint deformities or limited range of motion; crepitus cracking popping in knees, shoulders, ankles and wrists.
rature , arthritis; heart, lung and skin abnormalities; kidney disease; muscle weakness, and dysfunction of the esophagus.
OL1A2 genes that cause complete or partial loss of exon 6 of the gene.
lerness of patella, knee, patellar tendon, and tibia; good strength in hips.

p 10 Questions from Topic 5 (PAIN, EDS, TESTS)
adolescence can stand and walk unaided, but over time, walking and climbing stairs may become increasingly difficult.
supraspinatus and infraspinatus shoulders ; positive Neer and Hawkins tests indicative of issues w superspinatus and other structure
early adolescent in RF-negative polyarthritis, and 9-11 years in RF-positive polyarthritis.
hronic pain, stiffness, joint pain and limited range of motion.
emory, mood issues depression ; tingling or numbness in hands and feet more sensitive to pain; unknown cause.
her risk of developing spondyloarthritis found that around 70% of individuals with SpA carry the HLA-B27 gene.
mon is intestinal rupture - leads to acute pain in the abdominal area and requires immediate medical attention.
mplications - arterial rupture, uterine rupture, and intestinal perforation.
require immediate hospitalisation, observation in an intensive care unit.
Hypermobility Syndrome HMS Benign Joint Hypermobility Syndrome BJHS and EhlersDanlos Syndrome EDS , compared with the n

Top 10 Questions from Topic 6 (EDS, GENE, VASCULAR)
ered by emotions, too much movement, temp changes, touch, etc.
es her risk of developing spondyloarthritis found that around 70% of individuals with SpA carry the HLA-B27 gene.
e Investigation of Peripheral Nerve Injuries.
atients with inflammation, RBCs form clumps.
ggest that RF, ANA, or CCP values were NOT validated enough to differentiate between juvenile idiopathic polyarthritis and pediatric sy
et syndrome TOS and rotator cuff tendonitis .
rome TOS and rotator cuff tendonitis PT prescribed.
ld-moderate vascular and neurological impingement bilaterally.
e, hearing loss, respiratory problems, and a disorder of tooth development .
ns of patient that are consistent with EDS.

Top 10 Questions from Topic 7 ORAL, ANTEVERSION, RECESSIVE)
nd feet to turn inward, or have what is also known as a pigeon-toed appearance.
any other symptoms that Susan experienced, so that is rather an unrelated incident that was worth investigating.
n VI extracellular matrix of skeletal muscle and XII component of skeletal muscle .
noral anteversion correction surgery.
ptoms including joint pain, swelling, stiffness, fever, swollen lymph nodes, rash subtypes include systemic, oligoarticular, polyarticular
Femoral anteversion describes the inward rotation of the femur bone in the upper leg.
and premature ovarian failure: nothing to indicate ovarian failure.
ndant in our bodies - it forms scar tissue, ligaments, dentin, etc.
vers the cheeks and bridge of the nose or rashes elsewhere on the body.
What is femoral anteversion?

Appendix_I

op 10 Questions from Topic 8 (TISSUE, TYPE, COLLAGEN)	Topic
mune response, which may lead to the inflammatory symptoms of redness, warmth, and swelling in the affected area.	8
hic features when young but as you get older they become less distinct.	8
ndant in our bodies - it forms scar tissue, ligaments, dentin, etc.	8
ysical therapy, bracing, surgery, medication - bisphosphonates used to slow loss of existing bone.	8
d to moderate vascular & neurologic impingement bilaterally.	8
ndant in our bodies - it forms scar tissue, ligaments, dentin, etc.	8
unning, stairs, burpees diffuse, peripatellar, worse over patellar tendon .	8
Cardiac Phenotype predisposition to cardiac issues .	8
quences… A two-nucleotide deletion underlined was found in codon 1184.	8
isk factors: old age, women more likely to be affected, obesity injuries from sports or accidents increase risk, can be inherited .	8

Top 10 Questions from Topic 9 (SEVERE, HIP, DISORDER)
e hypotonia,Joint hypermobility, hyperextensible skin, bowed limbs.
ng thin lips and thin nose, prominent eyes due to abnormally decreased levels of fatty tissue under skin layers.
tall, thin tube and red blood cells are allowed to settle for 1 hour.
bone, also known as the femur the bone that is located between the hip and the knee .
0s and it becomes important to watch that their scoliosis does not begin to impede normal breathing patterns.
Age of onset: teen or early adult.
ia, hyperextensible thin skin with easy bruisability and atrophic scarring, wrinkled palms, joint hypermobility, and ocular involvement.
d early adolescent in RF-negative polyarthritis, and 9-11 years in RF-positive polyarthritis.
al disorders, immune disorders, connective tissue disorder, anticipation.
such as braces, wheelchairs, or scooters, surgery for hip dislocation.

Top 10 Questions from Topic 10 NDROME, TISSUE, RAREDISEASES)
he diagnosis, although the presence of antinuclear antibody ANA and rheumatoid factor RF can help classify JIA.
causes abnormal growth of new bone tissue on top of existing bones.
r weakness hypotonia or abnormal spinal rotations and curvatures scoliosis .
o identify subacromial impingement syndrome.
evelops in some people who have high levels of uric acid in the blood.
kin due to decreased levels of fatty tissue under skin resulting in characteristic facial appearance, easy bruising.
ns of both Benign Joint Hypermobility Syndrome and Ehlers-Danlos Syndrome, and 2 is potentially protective against heart disease.
e tissue and touch sensation across various parts of the body.
* Responds to glucocorticoids.
rmal shape of the cranium, short metacarpals typically from osteoclastic overactivity.

Interactive presentation of topic-term association

```
## <<TermDocumentMatrix (terms: 1440, documents: 544)>>
## Non-/sparse entries: 3216/780144
## Sparsity           : 100%
## Maximal term length: 71
## Weighting          : term frequency (tf)
```

```
##                       word freq
## pain                   pain   48
## joint                 joint   37
## eds                     eds   36
## skin                   skin   25
## syndrome           syndrome   24
## susan               susan   22
## muscle             muscle   22
## type                   type   21
## disorders         disorders   20
## tissue              tissue   18
## joints              joints   18
## autoimmune       autoimmune   18
## bone                   bone   17
## result              result   16
## gene                   gene   16
## positive           positive   15
## age                     age   15
## femoral             femoral   15
## hypermobility hypermobility   15
## loss                   loss   14
## arthritis         arthritis   14
## inflammation   inflammation   14
## anteversion     anteversion   14
## blood                 blood   14
## severe               severe   14
## cola                   cola   14
## diseases           diseases   13
## genetic             genetic   12
## tests                 tests   12
## disorder           disorder   12
## physical           physical   12
## recessive         recessive   12
## collagen           collagen   12
## connective       connective   11
## weakness           weakness   11
## polyarthritis polyarthritis   11
## swelling           swelling   10
## lupus                 lupus   10
## purpose             purpose   10
## strength           strength   10
## ehlersdanlos   ehlersdanlos   10
## autosomal         autosomal   10
## mutations         mutations   10
## missense           missense   10
## esr                     esr    9
## ana                     ana    9
## surgery             surgery    9
## hip                     hip    9
## vascular           vascular    9
## therapy             therapy    9
## crp                     crp    8
```


Appendix_I

```
## impingement     impingement    8
## abnormal           abnormal    8
## org                     org    8
## including         including    8
## idiopathic       idiopathic    8
## jia                     jia    8
## onset                onset    8
## knees                knees    8
## often                often    8
## fatigue             fatigue    8
## back                   back    8
## heterozygous   heterozygous    8
## imperfecta       imperfecta    8
## fingers             fingers    7
## body                   body    7
## typically         typically    7
## antibodies       antibodies    7
## asthma               asthma    7
## inflammatory   inflammatory    7
## feet                   feet    7
## rarediseases   rarediseases    7
## numbness           numbness    7
## complications complications    7
## arterial           arterial    7
## muscular          muscular    7
## rupture            rupture    7
## hypotonia         hypotonia    7
## mutation           mutation    7
## osteogenesis   osteogenesis    7
## rheumatoid       rheumatoid    6
## anticipation   anticipation    6
## clinical           clinical    6
## small                small    6
## knee                   knee    6
## medical             medical    6
## results             results    6
## cytoplasmic     cytoplasmic    6
## upper                 upper    6
## testing             testing    6
## method               method    6
## low                     low    6
## nerve                nerve    6
## arms                  arms    6
## shoulder          shoulder    6
## lung                  lung    6
## thin                  thin    6
## juvenile          juvenile    6
## negative          negative    6
## affects            affects    6
##  [ reached 'max' / getOption("max.print") -- omitted 1340 rows ]
```