CS5560 Knowledge Discovery and Management

Problem Set 3

June 19 (T), 2017

Name: Megha Nagabhushan

Class ID: 15

Information Retrieval (Text Mining) with TF-IDF

Consider the following three short documents

Doc #1:

The researchers will focus on computational phenotyping and will produce disease prediction models from machine learning and statistical tools.

Doc #2:

The researchers will develop tools that use Bayesian statistical information to generate causal models from large and complex phenotyping datasets.

Doc #3:

The researchers will build a computational information engine that uses machine learning to combine gene function and gene interaction information from disparate genomic data sources.

a) First remove stop words and punctuation; detect manually multi-word terms (using N-Gram or POS Tagging/Chunking); parse manually the documents and select the terms from the given 3 documents and created the dictionary (list of terms).

b) Create the document vectors by computing TF-IDF weights. Show how to compute the TF-IDF weights for terms. For each form of weighting list the document vectors in the following format:

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8 ...
DOC1	0	3	1	0	0	2	1	0 ....
DOC2	5	0	0	0	3	0	0	2 ....
DOC3	3	0	4	3	4	0	0	5 ....

① Removing stop words and punctuation

researchers focus computational phenotyping produce disease prediction models machine learning statistical tools ~~researc~~

researchers develop tools use Bayesian statistical information generate casual models large complex phenotyping datasets.

researchers build computational information engine uses machine learning combine gene function gene interaction information disparate genomic data sources.

~~Detect m~~ Commonly used stopwords:

a an and are as at be by for from has he in is it its of on that the to was were will with.

Detecting multi-word terms

Using n-gram technique and considering 2 closely related terms and performing chunking.

researchers focus
focus computational
computational phenotyping
phenotyping produce

produce disease
disease prediction
prediction models
models machine
machine learning
learning statistical
Statistical tools.
tools researchers
researchers develop
develop tools
tools use

use Bayesian
Bayesian statistical
statistical information
information generate
generate casual
casual models
models large
large complex
complex phenotyping
phenotyping datasets
datasets researchers.

researchers build.
build Computational
computational information
information engine
engine uses.
uses machine
machine learning
learning Combine
Combine gene
gene function.

function gene
gene interaction
Interfaction information
information disparate
disparate genomic.
genomic data
data sources.

Dictionary of terms

research
focus
compute
phenotype
produce
disease
predict
model
machine
learn
statistics
tool
develop
use
Bayesian
Information
generate
Casual
model
large
complex

dataset
build
engine
Combine
gene
function
interact
disparate
genome
data
source.

② N=3 (no of documents)

tf values.

	doc 1	doc 2	doc 3
the	1	1	1
researchers	1	1	1
will	2	1	1
focus	1	0	0
on	1	0	0
computational	1	0	1
phenotyping	1	1	0
and	2	1	1
with			
produce	1	0	0
disease	1	0	0
prediction	1	0	0
models	1	1	0
from	1	1	0
machine	1	0	1
learning	1	0	1
statistical	1	1	0
tools	1	1	0
develop	0	1	0

	doc1	doc2	doc3
that	0	1	1
use	0	1	1
Bayesian	0	1	0
information	0	1	1
to	0	1	0
generate	0	1	0
casual	0	1	0
large	0	1	0
complex	0	1	0
datasets	0	1	0
build	0	0	1
a	0	0	1
engine	0	0	1
combine	0	0	1
gene	0	0	1
function	0	0	1
interaction	0	0	1
disparate	0	0	1
genomic	0	0	1
data	0	0	1
sources	0	0	1

idf values

the → $\log_2(3/3) = 0$

researchers → $\log_2(3/3) = 0$

will → $\log_2(3/4) = 0$ (-ve).

focus → $\log_2(3/1) = 0.477$

on → $\log_2(3/1) = 0.477$

Computation → $\log_2(3/2) = 0.176$

phenotyping → $\log_2(3/2) = 0.176$

and → $\log_2(3/4) = 0$

produce → $\log_2(3/1) = 0.477$

disease → $\log_2(3/1) = 0.477$

prediction → $\log_2(3/1) = 0.477$

models → $\log_2(3/2) = 0.176$

from → $\log_2(3/2) = 0.176$

machine → $\log_2(3/2) = 0.176$

learning → $\log_2(3/2) = 0.176$

statistical → $\log_2(3/2) = 0.176$

tools → $\log_2(3/2) = 0.176$

develop → $\log_2(3/1) = 0.477$

that → $\log_2(3/2) = 0.176$

use → $\log_2(3/2) = 0.176$

Baysian → $\log_2(3/1) = 0.477$

information → $\log_2(3/2) = 0.176$

to → $\log_2(3/1) = 0.477$

generate → $\log_2(3/1) = 0.477$

casual → $\log_2(3/1) = 0.477$

large → $\log_2(3/1) = 0.477$

complex → $\log_2(3/1) = 0.477$

datasets → $\log_2(3/1) = 0.477$

build → $\log_2(3/1) = 0.477$

a → $\log_2(3/1) = 0.477$

engine → $\log_2(3/1) = 0.477$

combine → $\log_2(3/1) = 0.477$

gene → $\log_2(3/1) = 0.477$

function → $\log_2(3/1) = 0.477$

interaction → $\log_2(3/1) = 0.477$

disparate → $\log_2(3/1) = 0.477$

genomic → $\log_2(3/1) = 0.477$

data → $\log_2(3/1) = 0.477$

sources → $\log_2(3/1) = 0.477$.

		Doc1	Doc2	Doc3	TF-IDF
Term1	gene	0	1	2	1.386294361
Term2	develop	0	1	1	0.693147181
Term3	learn	1	0	1	0.693147181
Term4	source	0	0	1	0.693147181
Term5	interaction	0	0	1	0.693147181
Term6	learning	1	0	1	0.693147181
Term7	build	0	0	1	0.693147181
Term8	on	1	0	0	0.693147181
Term9	generate	0	1	0	0.693147181
Term10	engine	0	0	1	0.693147181
Term11	prediction	1	0	0	0.693147181
Term12	focus	1	0	0	0.693147181
Term13	causal	0	1	0	0.693147181
Term14	disease	1	0	0	0.693147181
Term15	large	0	1	0	0.693147181
Term16	data	0	0	1	0.693147181
Term17	bayesian	0	1	0	0.693147181
Term18	produce	1	0	0	0.693147181
Term19	complex	0	1	0	0.693147181
Term20	combine	0	0	1	0.693147181
Term21	a	0	0	1	0.693147181
Term22	dataset	0	1	0	0.693147181
Term23	disparate	0	0	1	0.693147181
Term24	genomic	0	0	1	0.693147181
Term25	function	0	0	1	0.693147181
Term26	information	0	1	2	0.575364145
Term27	phenotyping	1	1	0	0.287682072
Term28	computational	1	0	1	0.287682072
Term29	statistical	1	1	0	0.287682072
Term30	tool	1	1	0	0.287682072
Term31	model	1	1	0	0.287682072
Term32	that	0	1	1	0.287682072
Term33	to	0	1	1	0.287682072
Term34	machine	1	0	1	0.287682072
Term35	use	1	1	0	0.287682072
Term36	will	1	1	1	0

Term37	from	1	1	1	0
Term38	and	2	1	1	0
Term39	the	1	1	1	0
Term40	researcher	1	1	1	0