CS5560 Knowledge Discovery and Management

Problem Set 3

June 19 (T), 2017

Name: Chaitanya Kumar Peravalli

Class ID: 23

**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.

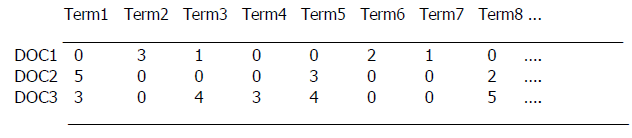
Doc #2:

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.

1. 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).
2. 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:



a)

After removing stop words and punctuation:

**Doc #1:**

researchers

researchers will

will focus

focus on

on computational

computational phenotyping

phenotyping and

and will

will produce

produce disease

disease prediction

prediction models

models from

from machine

machine learning

learning and

and statistical

statistical tools

tools

**Doc #2:**

researchers

researchers will

will develop

develop tools

tools that

that use

use Bayesian

Bayesian statistical

statistical information

information to

to generate

generate causal

causal models

models from

from large

large and

and complex

complex phenotyping

phenotyping datasets

datasets

**Doc #3**

researchers

researchers will

will build

build a

a computational

computational information

information engine

engine that

that uses

uses machine

machine learning

learning to

to combine

combine gene

gene function

function and

and gene

gene interaction

interaction information

information from

from disparate

disparate genomic

genomic data

data sources

**Dictionary of terms from 3 documents:**

|  |
| --- |
| Bayesian  information  build  interaction  causal  combine  large  complex  learning  computational  machine  data  models  datasets  phenotyping  develop  prediction  disease  produce  disparate  researchers  engine  sources  focus  statistical  tools  tools.  function  gene  generate  genomic  use  uses  will |

b)

Document vectors by computing TF-IDF weights:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Bayesian | information | build | interaction | causal | Combine | Large | Complex | Learning |
| Doc1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.287 |
| Doc2 | 0.693 | 0.575 | 0 | 0 | 0.693 | 0 | 0.693 | 0.693 | 0 |
| Doc3 | 0 | 0.575 | 0.693 | 0.693 | 0 | 0.693 | 0 | 0 | 0.287 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | computational | machine | data | datasets | develop | disease | disparate | engine | focus |
| Doc1 | 0.287 | 0.287 | 0 | 0 | 0 | 0.693 | 0 | 0 | 0.693 |
| Doc2 | 0 | 0 | 0 | 0.693 | 0.693 | 0 | 0 | 0 | 0 |
| Doc3 | 0.287 | 0.287 | 0.693 | 0 | 0 | 0 | 0.693 | 0.693 | 0 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | function | gene | generate | genomic | models | phenotyping | prediction | produce | researchers |
| Doc1 | 0 | 0 | 0 | 0 | 0.287 | 0.287 | 0.693 | 0.693 | 0 |
| Doc2 | 0 | 0 | 0.693 | 0 | 0.287 | 0 | 0 | 0 | 0 |
| Doc3 | 0.693 | 1.386 | 0 | 0.693 | 0 | 0 | 0 | 0 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | sources | statistical | tools | use | uses | will |
| Doc1 | 0 | 0.287 | 0.693 | 0 | 0 | 0 |
| Doc2 | 0 | 0 | 0 | 0.693 | 0 | 0 |
| Doc3 | 0.693 | 0 | 0 | 0 | 0.693 | 0 |

Calculation of TF-IDF:

T: term, D: document, F: frequency

**IDF (t, D) = (log (|D|+1/DF (t, D) +1))**

Example: for gene

= (Log (3 + 1 / 1+1))

= Log (2) = 0.693

|D| = No of documents in the corpse = 3

DF (t, D) = No of documents the term is present = 1 (Doc3)

TFIDF (t, d, D) =TF (t, d) ⋅IDF (t, D).

= 2 \* 0.693

= 1.386

TF (t, d) = No of times the term appears in a document.

The logarithmic base is e.

By Code:

(gene,2.1972245773362196)

(information,1.3862943611198906)

(dataset,1.0986122886681098)

(produce,1.0986122886681098)

(learning,1.0986122886681098)

(large,1.0986122886681098)

(focus,1.0986122886681098)

(develop,1.0986122886681098)

(generate,1.0986122886681098)

(bayesian,1.0986122886681098)

(genomic,1.0986122886681098)

(complex,1.0986122886681098)

(function,1.0986122886681098)

(source,1.0986122886681098)

(disease,1.0986122886681098)

(causal,1.0986122886681098)

(engine,1.0986122886681098)

(build,1.0986122886681098)

(prediction,1.0986122886681098)

(learn,1.0986122886681098)

(datum,1.0986122886681098)

(disparate,1.0986122886681098)

(interaction,1.0986122886681098)

(on,1.0986122886681098)

(combine,1.0986122886681098)

(computational,0.6931471805599453)

(phenotyping,0.6931471805599453)

(use,0.6931471805599453)

(machine,0.6931471805599453)

(statistical,0.6931471805599453)

(tool,0.6931471805599453)

(that,0.6931471805599453)

(,0.6931471805599453)

(model,0.6931471805599453)

(to,0.6931471805599453)

(from,0.4054651081081644)

(and,0.4054651081081644)

(researcher,0.4054651081081644)

(will,0.4054651081081644)