CS5560 Knowledge Discovery and Management

Problem Set 3

June 19 (T), 2017

Name: P.shankar

Class ID: 22

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).

A)

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

1. 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)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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:

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

Ex: 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.