# CS5560 Knowledge Discovery and Management

Problem Set 3

June 19 (T), 2017

Name: Sujitha Puthana

Class ID: 24

**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 .... |

a) Step1 - Removing Stop words & punctuation

Doc1 -

Researchers focus computational phenotyping produce disease prediction model from machine learning statistical tools.

Doc2 -

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

Doc3 -

Researches build computational information engine uses machine learning Combine gene function gene interaction information disparate genomic data sources.

Step 2 - Detect multi-word terms -

Here we are considering 2 closely related terms to perform multi-word terms.

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 causal
causal 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

interaction information

information from

from disparate

disparate genomic

genomic data

data sources

Step3 - Dictionary of terms

Research

focus

compute

phenotype

produce

predict

model

machine

learn

statistics

tools

develop

use

Bayesian

information

generate

causal

large

complex

dataset

build

engine

Combine

gene

function

interact

disparate

genome

data

source

| | | 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 |

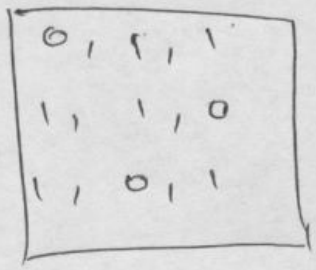
| | | | | | |
|---|---|---|---|---|---|
| Term37 | from | 1 | 1 | 1 | 0 |
| Term38 | and | 2 | 1 | 1 | 0 |
| Term39 | the | 1 | 1 | 1 | 0 |
| Term40 | researche | 1 | 1 | 1 | 0 |

## Calculation of TF-IDF

-> Here when the word appear in all documents then its

$$\log(3/3) = 0$$

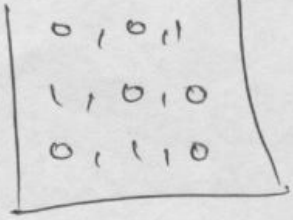
-> When the word appears in just 2 documents for 1 at a time i.e.,




then it is $\log(3/2) = 0.176$

$\Rightarrow 0.176 + 0.176 = 0.352.$

-> When the word appears only in one document i.e.,




$$\log(3/1) = 0.477$$

-> Here the exception case

① TF-IDF (gene) = $2 \times \log(3/1) = 0.95$

② TF-IDF (information) = $2 \times \log(3/2) + \log(3/2) = 0.522.$