CS5560 Knowledge Discovery and Management

Problem Set 3

June 19 (T), 2017

Name: SreeLakshmi Nandanamudi

Class ID: 17

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. a) **Stop words**: stop words are words which are filtered out before or after processing of natural language data
The stop words such as "and", "the", "an".

Removing stop words and punctuation:

Doc #1:
researchers focus computational phenotyping produce disease prediction models machine learning statistical tools.

Doc #2:
researchers develop tools Bayesian statistical information generate casual models large complex phenotyping datasets.

Doc #3:
researchers build computational information engine uses machine learning combine gene function gene interaction information disparate genomic data sources.

Detecting manually multi-word term using N-Gram:

N-Gram: An N-Gram is a contiguous sequence of n items from a given sequence of text or speech.
An n-gram of size 1 is referred to as a "unigram", Size 2 is a bigram", Size 3 is a "trigram". and so on.

N=2

Doc #1 :

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

Doc #2 :

researchers develop
develop tools
tools Bayesian
Bayesian statistical
statistical information
information generate
generate casual
casual models
models large

N=3

Doc #1 :

researchers focus computational
focus computational phenotyping
computational phenotyping produce
phenotyping produce disease
produce disease prediction
disease prediction models
prediction models machine
models machine learning
machine learning statistical
learning statistical tods.

Doc#2 :

researchers develop tools
develop tools Bayesian
tools Bayesian statistical
Bayesian statistical information
Statistical information generate
information generate casual
generate casual models
casual models large

N=2

large complex
complex phenotyping
phenotyping datasets.

Doc #3:

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

N=3

models large complex
large complex phenotyping
complex phenotyping datasets.

Doc #3:

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

Term - Document Matrix

Vocabulary (contains only terms that occur multiple times, no stop words)

Terms	Documents			Count in 3 documents
	D_1	D_2	D_3	
researchers	1	1	1	3
focus	1	0	0	1
computational	1	0	1	2
phenotyping	1	1	0	2
produce	1	0	0	1
disease	1	0	0	1
prediction	1	0	0	1
models	1	1	0	2
machine	1	0	1	2
learning	1	0	1	2
statistical	1	1	0	2
tools	1	1	0	2
develop	0	1	0	1

Terms	D_1	D_2	D_3	Count
Bayesian	0	1	0	1
information	0	1	1	2
generate	0	1	0	1
casual	0	1	0	1
large	0	1	0	1
complex	0	1	0	1
datasets	0	1	0	1
build	0	0	1	1
engine	0	0	1	1
uses	0	0	1	1
combine	0	0	1	1
gene	0	0	1	1
function	0	0	1	1
interaction	0	0	1	1
disparate	0	0	1	1
genomic	0	0	1	1
data	0	0	1	1
Sources	0	0	1	1

b) Term Frequency (TF): TF means which measures how frequently a term occurs in a document.

$$TF(t) = \frac{(\text{Number of times term } t \text{ appears in a document})}{(\text{Total number of terms in the document})}$$

Inverse Document Frequency (IDF): IDF means which measures how important a term is while computing TF, all terms are considered equally important.

$$IDF(t) = \log_{10}(\text{Total number of documents} / \text{Number of documents with term } t \text{ in it})$$

tf-idf: Tf-idf stands for term frequency - inverse document frequency, and the tf-idf weight is a weight often used in information retrieval and text mining.

Tf-idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

Terms	Doc 1	Doc 2	Doc 3
researchers	1	1	1
focus	1	0	0
computational	1	0	1
phenotyping	1	1	0

Terms	Doc1	Doc2	Doc3
produce	1	0	0
disease	1	0	0
prediction	1	0	0
models	1	1	0
machine	1	0	1
learning	1	0	1
statistical	1	1	0
tools	1	1	0
develop	0	1	0
Bayesian	0	1	0
information	0	1	2
generate	0	1	0
Casual	0	1	0
large	0	1	0
complex	0	1	0
datasets	0	1	0
build	0	0	1
engine	0	0	1

uses	0	0	1
Combine	0	0	1
gene	0	0	2
function	0	0	1
interaction	0	0	1
disparate	0	0	1
genomic	0	0	1
data	0	0	1
Sources	0	0	1

Computing the TF-IDF weights for terms:

Doc #1

researchers, $TF = \frac{1}{12}$, $IDF = \log_{10}\left(\frac{3}{3}\right) = \log_e(1) = 0$

$TF\text{-}IDF = TF \times IDF = \frac{1}{12} \times 0 = 0$

focus, $TF = \frac{1}{12}$, $IDF = \log_e\left(\frac{3}{1}\right) = \log_{10}(3) = 0.477$

$TF\text{-}IDF = \frac{1}{12} \times 0.477 = 0.0397$

Computational, $TF = \frac{1}{12}$, $IDF = \log_{10}\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = \frac{1}{12} \times 0.176 = 0.0146$

phenotyping, $TF = \frac{1}{12}$, $IDF = \log_{10}\left(\frac{3}{2}\right) = 0.176$; $TF\text{-}IDF = 0.0146$

produce , $TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{1}\right) = 0.477$, $TF\text{-}IDF = 0.0397$

disease , $TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{1}\right) = 0.477$, $TF\text{-}IDF = 0.0397$

prediction , $TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{1}\right) = 0.477$, $TF\text{-}IDF = 0.0397$

models , $TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = \frac{1}{12} \times 0.176 = 0.0146$

machine , $TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = \frac{1}{12} \times 0.176 = 0.0146$

learning , $TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = \frac{1}{12} \times 0.176 = 0.0146$

statistical , $TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = \frac{1}{12} \times 0.176 = 0.0146$

tools , $TF = \frac{1}{12}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = \frac{1}{12} \times 0.176 = 0.0146$

For the remaining terms, which are not present in the Doc 1

$TF = 0$ then $TF\text{-}IDF = 0$

Doc #2 :

researchers , $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{3}\right) = \log(1) = 0$, $TF\text{-}IDF = \frac{1}{13} \times 0 = 0$

develop , $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{1}\right) = \log(3) = 0.4771$, $TF\text{-}IDF = \frac{1}{13} \times 0.4771 = 0.0367$

tools , $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = \frac{1}{13} \times 0.176 = 0.0135$

Bayesian , $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{1}\right) = 0.4771$, $TF\text{-}IDF = 0.0367$

statistical, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = 0.0135$

information, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{3}\right) = 0$, $TF\text{-}IDF = 0$

generate, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{1}\right) = 0.4771$, $TF\text{-}IDF = 0.0367$

Casual, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{1}\right) = 0.4771$, $TF\text{-}IDF = 0.0367$

models, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = 0.0135$

large, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{1}\right) = 0.4771$, $TF\text{-}IDF = 0.0367$

Complex, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{1}\right) = 0.4771$, $TF\text{-}IDF = 0.0367$

phendyping, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = 0.0135$

datasets, $TF = \frac{1}{13}$, $IDF = \log\left(\frac{3}{1}\right) = 0.4771$, $TF\text{-}IDF = 0.0367$

Doc #3

researchers, $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{3}\right) = \log(1) = 0$, $TF\text{-}IDF = 0$

build, $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right) = \log\left(\frac{3}{1}\right) = 0.4771$, $TF\text{-}IDF = 0.0265$

Computational, $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = 0.0097$

information, $TF = \frac{2}{18}$, $IDF = \log\left(\frac{3}{2}\right) = 0.176$, $TF\text{-}IDF = 0.0195$

engine , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

uses , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

machine , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

learning , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{2}\right)$, $TF\text{-}IDF = 0.017$

Combine , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

gene , $TF = \frac{2}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0530$

function , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

interaction , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

disparate , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

genomic , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

data , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$

Sources , $TF = \frac{1}{18}$, $IDF = \log\left(\frac{3}{1}\right)$, $TF\text{-}IDF = 0.0265$