ม

CS5560 Knowledge Discovery and Management

Problem Set 3

June 19 (T), 2017

Name: Revanth Chakilam

Class ID: 02

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 ....

Revanth Chakilam

a) Stop words:

Basically the stop words are those which do not contain important significance to be used in search queries. Usually these words are filtered out from queries because they return huge amount of data.

→ Removal of stop words / punctuation:

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

Doc #2: researchers develop tools Bayesian statistical information generate causal 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.

→ N-gram approach:

In this scenario, I am considering N-gram size as 3 (N=3)

Doc #1: The researchers will
researchers will focus
will focus on
focus on computational
on computational phenotyping
computational phenotyping and
phenotyping and will
and will produce.
.... so on
and statistical tools.

The output for N=3 after removal stop words for doc#1 is (2)

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

The N-gram Value - 3

Doc #2:

researchers develop tools
develop tools ~~to~~ bayesian
tool bayesian statistical
Bayesian statistical information
statistical information generate
generate casual models
causal models large
models large complex
large complex phenotyping
complex phenotyping datasets.

N-gram - 3

Doc #3:

recearchers build computational
build computational information
computational information engine
information engine uses
engine uses machine
uses machine learning
machine learning combine
learning combine gene
combine gene function
gene function gene
function gene interactions

N-gram Value - 3

Revanth Chakilam

interaction information disparate
information disparate genomic
disparate genomic data
genomic data sources

→ Parse manually the documents cy select the terms from given 3 docs:

	#doc 1	doc#2	doc#3	Count in all docs
researchers	1	1	1	3
focus	1	0	0	1
Computational	1	0	1	2
phenotyping	1	1	0	2
problem	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
Bayesian	0	1	0	1
information	0	1	1	2
generate	0	1	0	1
causal	0	1	0	1
large	0	1	0	1
Complex	0	1	0	1
datasets	0	1	0	1
build	0	0	1	1
sources	0	0	1	1
engine	0	0	1	1
⋮				
			1	1
		0		
genomic	0			

b) Term Frequency: It means how frequently a term occurs in a document since every document is difference in length It is possible that a term would appear much more times

$$T(t)/t = \frac{\text{No. of times 't' appears in doc}}{\text{Total no. of terms in the doc}}$$

Inverse Document Frequency:

$$IDF(t) = \log_e(\text{Total no. of document} / \text{No. of documents with term 't' on it})$$

TFIDF: Its weight often used in information retrieval and text mining.

★ The matrix shows how many times a term occur in document is already shown in previous page-③ ★
Based on the matrix, each word count in all docs

TF-IDF values for each term in Documents:

Doc #1:

Researchers:-

$TF = 1/12$
$= 0.083$

$IDF = \log_e\left(\frac{3}{32}\right)$ $\quad TF\text{-}IDF = \frac{1}{12} \times \log_e\left(\frac{3}{32}\right)$
$= 0.0146$

Focus:

$TF = 1/12$ $\quad IDF = \log_e\left(\frac{3}{1}\right) = 0.477$ $\quad TF\text{-}IDF = \frac{1}{12} \times 0.477 = 0.039$

Computational:

$TF = 1/12$, $IDF = \log_e\left(\frac{3}{2}\right) = 0.176$ $\quad TF\text{-}IDF = 0.0146$

Phenotyping:

$TF = 1/12$, $IDF = \log_e\left(\frac{3}{2}\right) = 0.176$ $\quad TF\text{-}IDF = 0.0146$

Produce:

$TF = 1/12$, $IDF = \log_e(3/1) = 3$ $\quad TF\text{-}IDF = \frac{1}{12} \times 0.477 = 0.039$

Revanth Chakilam

Disease:

$TF = 1/12$ $IDF = \log_e(3/1) = 3$ $TF\text{-}IDF = \frac{1}{12} \times 0.477 = 0.039$

Prediction:

$TF = 1/12$, $IDF = \log(3/1) = 0.477$, $TF\text{-}IDF = \frac{1}{12} \times 0.477 = 0.039$

Models:

$TF = 1/12$ $IDF = 0.176$ $TF\text{-}IDF = 0.0146$

Machine:

$TF = 1/12$ $IDF = \log(3/2) = 0.176$, $TF\text{-}IDF = 0.0146$

For learning:

$TF = 1/12$, $IDF = \log(3/2) = 0.176$, $TF\text{-}IDF = 0.0146$

Statistical:

$1/12$, 0.176 , 0.0146

tools:

$1/12$, 0.176 , 0.0146.

The other words in Doc #1 have $TF = 0$, $TF\text{-}IDF = 0$.

Doc #2:

Develop:

$TF = 1/13$, $IDF = \log(3/1) = 0.477$, $TF\text{-}IDF = 0.036$

Bayesian: $1/13$, 0.477 , 0.036

Information: $1/13$, $\log(3/2) = 0.176$, 0.0135

Researchers: $1/13$, $\log(3/3)$, 0

Tools: $1/13$ $\log(3/2)$, 0.0135

Statistical: $1/13$, $\log(3/2)$, 0.0135

Generate, large, datasets, causal, complex :

$TF = 1/13$, $IDF = \log(3/1)$ $TF\text{-}IDF = 0.036$

phenotyping, models

$TF = 1/13$, $IDF = \log(3/2)$ $TF\text{-}IDF = 0.0135$

Doc #3 :-

Researchers : $TF = 1/18$, $IDF = \log(3/3)$, $TF\text{-}IDF = 0$
Build : $1/18$, $\log(3/1)$, 0.026
Computational : $1/18$, $\log(3/2)$, 0.978
Information : $2/18$, $\log(3/2)$, 0.0195

Engine, users, machine, combine, function, interaction, data, sources, genomic, disparate :

$TF = 1/18$, $IDF = \log(3/1)$, $TF\text{-}IDF = 0.026$

Gene : $TF = 2/18$, $IDF = \log(3/1)$, $TF\text{-}IDF = 0.0529$

Learning : $TF = 1/18$, $IDF = \log(3/2)$, $TF\text{-}IDF = 0.978$

— END —