CS5560 Knowledge Discovery and Management

Problem Set 3

June 19 (T), 2017

Name: SAITEJA MAKANI

Class ID: 12

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	....

(TF-IDF weights) $w_{t,d} = (1 + \log_{10} tf_{t,d}) \times \log_{10}(N / df_t)$

$tf_{t,d} = \dfrac{\text{highest frequency in doc (any term)}}{\text{absolute term frequency}}$ (1, 1, 2)

N = Number of Docs in database.

df_t = Number of Docs where term 't' appears.

a) Removing stop words and punctuations; Detecting Multi-word terms

Doc1: researchers, focus, computational, phenotyping, produce, disease, prediction, Models, Machine, learning, statistical, tools.

Doc2: researchers, develop, tools, Bayesian, statistical, information, generate, causal, Models, large, complex, phenotyping; datasets.

Doc3: researchers, build, computational, information, engine, uses, Machine, learning, combine, gene, function, interaction, information, disparate, genomic, data, Sources.

Dictionary: [researchers, focus, computational, phenotyping, Produce, disease, prediction, Models, Machine, learning, statistical, tools]^1, [develop, Bayesian, information, generate, causal, large, complex, datasets]^2, [build, engine, uses, combine, gene, function, interaction, disparate, genomic, data, Sources]^3

b)

word	Doc1 TF-IDF	Doc2 IF-IDF	Doc3 TF-IDF
interaction	0.4771	0.4771	0.6207
Researchers	0	0	0.6207
Focus	0.4771	0.4771	0.4771 0.6207
computational	0.1760	0.1760	0.0530
phenotyping	0.1760	0.1760	0.0309
produce	0.4771	0.4771	0.4771
disease	0.4771	0.4771	0.4771
Prediction	0.4771	0.4771	0.4771
Models	0.1760	0.1760	0.1760
Machine learning	0.1760	0.1760	0.6207
statistical	0.1760 0.4771	0.4771 0.1760	0.1760 0.4771
Disperate	0.4771	0.4771	0.6207
tools	0.1760	0.1760	0.1760
Develop	0.4771	0.4771	0.4771
Bayesian	0.4771	0.4771	0.4771
Information	0.1760	0.1760	0.1760 0.2291
Generate	0.4771	0.4771	0.4771
Causal	0.4771	0.4771	0.4771
Large	0.4771	0.4771	0.4771
Complex	0.4771	0.4771	0.4771
Datasets	0.4771	0.4771	0.4771
Build	0.4771	0.4771	0.6207
Engine	0.4771	0.4771	0.6207
uses	0.4771	0.4771	0.6207
Combine	0.4771	0.4771	0.6207
gene	0.4771	0.4771	0.4771
function	0.4771	0.4771	0.6207
genomic	0.4771	0.4771	0.6207
Datasources	0.4771	0.4771	0.6207

OBSERVATION: "Researchess" most common term in all Docs

so it got TF-IDF score as zero.