# CS5560 Knowledge Discovery and Management

Problem Set 3

June 19 (T), 2017

Name: Rashmi Tripathi

Class ID: 29

**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:

$$TF_t = \frac{\text{No of time } t \text{ appear in doc}}{\text{Total no of terms}}$$

$$IDF_t = \log_e \left( \frac{\text{No of Doc}}{\text{No of doc with term } t \text{ in it}} \right)$$

| | 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 | .... |

After stop words and punctuation removal :→

Doc 1 : Researchers focus computational phenotyping produce disease prediction models |machine learning| |stastical ~~tools~~| ← Using N Grams

Doc 2 : Researchers develop tools Bayesian stastical information generate casual models large complex phenotyping datasets.

Doc 3 : Researchers build computational information engine |machine learning| combine function gene interaction information disparate genomic |data sources|

create a dictionary of list of terms

$= TF_{t} * \log_e\left(\frac{3}{Doc1+Doc2+Doc3}\right)$

| S No | Word | Doc1 | Doc2 | Doc3 | TF Doc1 | Doc2 | Doc3 | TF-IDF Doc1 | Doc2 | Doc3 |
|---|---|---|---|---|---|---|---|---|---|---|
| Term 1 → | ~~Researchers~~ Bayesian | | 1 | | | =1/11 =0.09 | | 0 | 0.09 * log(3/1) 0.04 | 0 |
| Term 2 | build | | | 1 | | | =1/15 =0.07 | 0 | 0 | =0.07 * log(3/1) = 0.03 |
| Term 3 | casual | | 1 | | | 0.09 | | 0 | 0.04 | 0 |
| Term 4 | combine | | | 1 | | | 0.07 | 0 | 0 | 0.03 |
| 5 | complex | | 1 | | | 0.09 | | 0 | 0.04 | 0 |
| 6 | computational | 1 | | 1 | =1/10 =0.1 | ~~0.09~~ | 0.07 | 0.02 | 0 | 0.03 |
| 7 | data | | | 1 | | | 0.07 | 0 | 0 | 0.03 |
| 8 | datasets | | 1 | | | 0.09 | | 0 | 0.04 | 0 |
| 9 | disease | 1 | | | =1/10 =0.1 | 0.09 | 0.07 | 0.05 | 0 | 0 |
| 10 | engine | | | 1 | | | | 0 | 0 | 0.03 |
| 11 | gene | | | 2 | | | $\boxed{0.13}$ | 0 | 0 | 0.02 |
| 12 | genomic | | | 1 | | | 0.07 | 0 | 0 | 0.03 |
| 13 | information | | 1 | 2 | | 0.09 | 0.13 | 0 | 0 | 0.02 |
| 14 | interaction | | | 1 | | | 0.07 | 0 | 0 | 0.03 |
| 15 | large | | 1 | | | 0.09 | 0 | 0 | 0.04 | 0 |
| 16 | Machine Learning | 1 | | 1 | 0.1 | 9 | 0.07 | 0.62 | 0 | 0.03 |
| 17 | models | 1 | 1 | | 0.1 | 0.09 | | 0.02 | 0.02 | 0 |
| 18 | phenotypi1 | 1 | 1 | | 0.1 | 0.09 | | 0.02 | 0.02 | 0 |
| 19 | prediction | 1 | | | 0.1 | | | 0.05 | 0 | 0 |
| 20 | produce | 1 | | | 0.1 | | | 0.05 | 0 | 0.03 |
| 21 | Researchers | 1 | 1 | 1 | 0.1 | 0.09 | 0.07 | 0 | 0.02 | 0.03 |
| 22 | sources | | | 1 | | | 0.07 | 0 | 0 | 0.03 |
| 23 | statistical | 1 | 1 | | 0.1 | 0.09 | | 0.02 | 0.02 | 0 |
| 24 | tools | 1 | 1 | | 0.1 | 0.09 | | 0.02 | 0.02 | 0.03 |
| 25 | disparate | | | 1 | 0 | 0 | 0.07 | 0 | 0 | 0.03 |
| | SUM= | 10 | SUM=11 | SUM=15 | | | | | | |