Support decision system for diagnosing rare diseases using vector space model and medical text mining

DIKU Bachelorprojekt 2009 – 2010

Henrik G. Jensen og Michael Andersen

Vinter 2010

Motivation

Problemet

- Læger har stor viden, men ...
- Læger har begrænset tid per patient
- Sjældne sygdomme drukner i informations mængden.

Motivation

Problemet

- Læger har stor viden, men ...
- Læger har begrænset tid per patient
- Sjældne sygdomme drukner i informations mængden.

Behovet

- Der findes mange specifikke systemer, som bestemmer enkelt sygdomme, intet der rammer flere sygdomme.
- Giver foreslag til sygdomme, som hurtigt kan kontrolleres.

Midlet

Løsningen, systemet

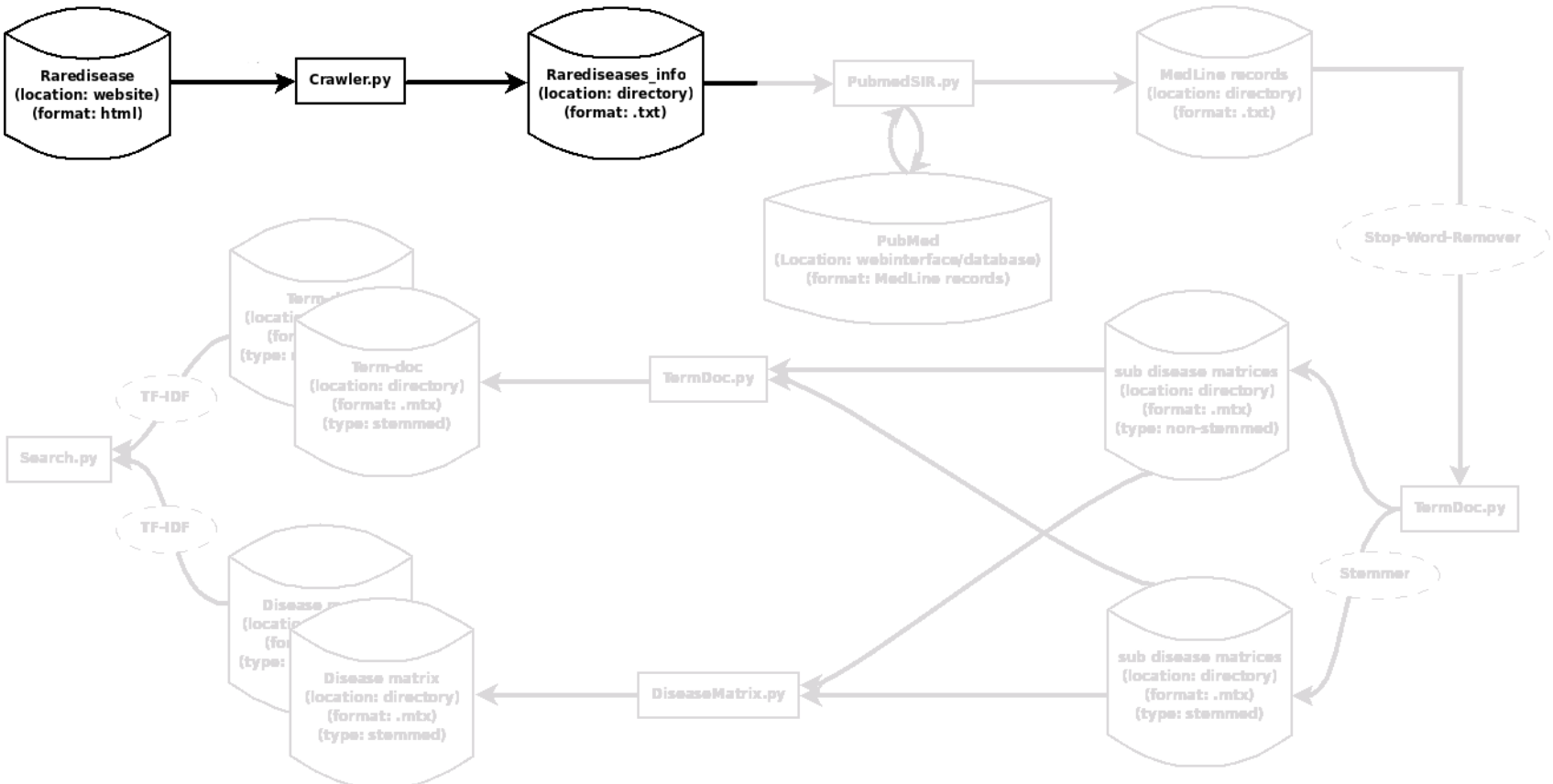
- Konstruer speciel database
- Vægtning af termer
- Vector space model
- Udregning af score for søgning
- Forslå top 20 sygdomme.

Resultater

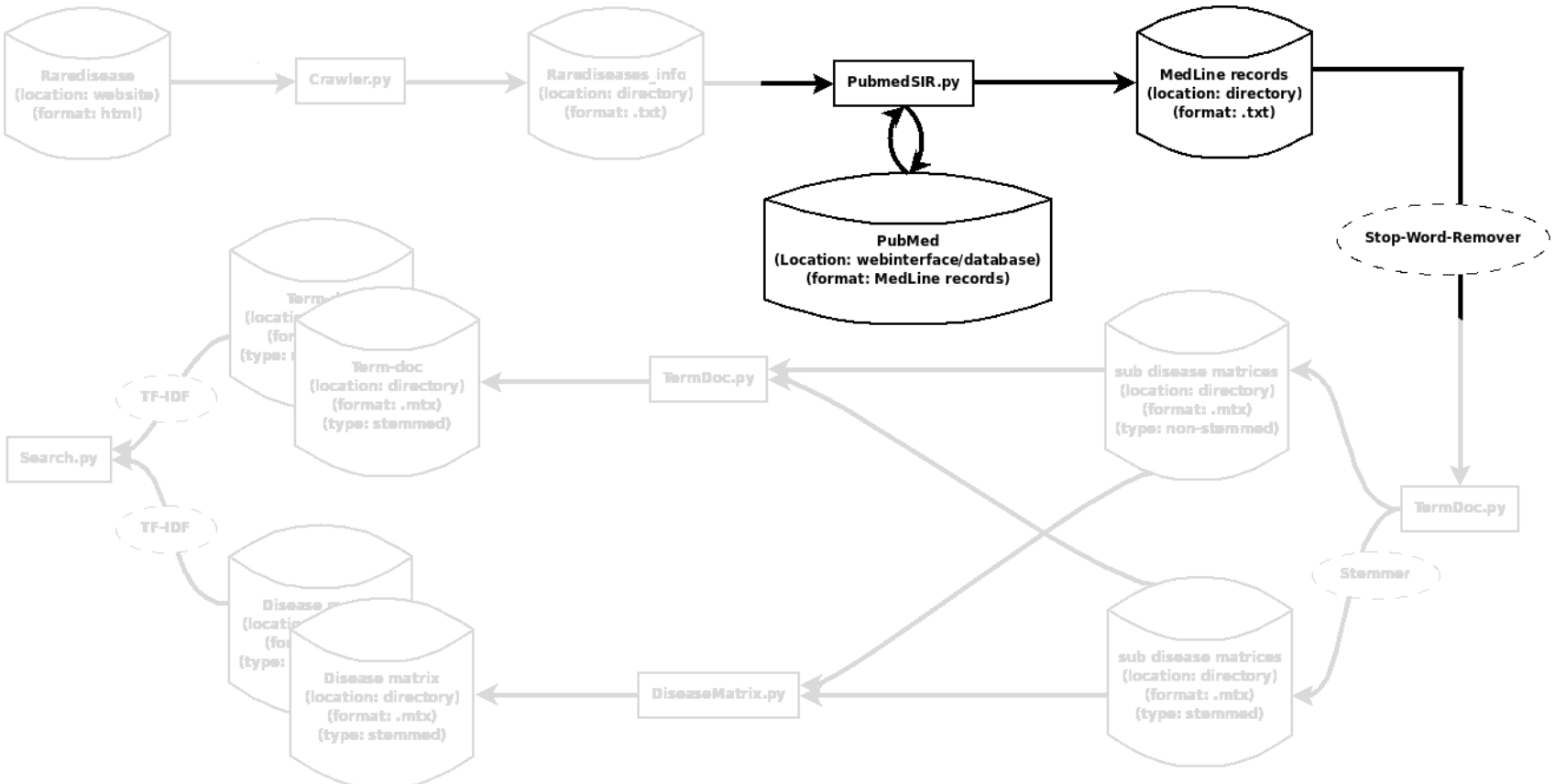
- BMJ test sager
- Orpha.net test sager
- Blind testen

- Fremtids muligheder

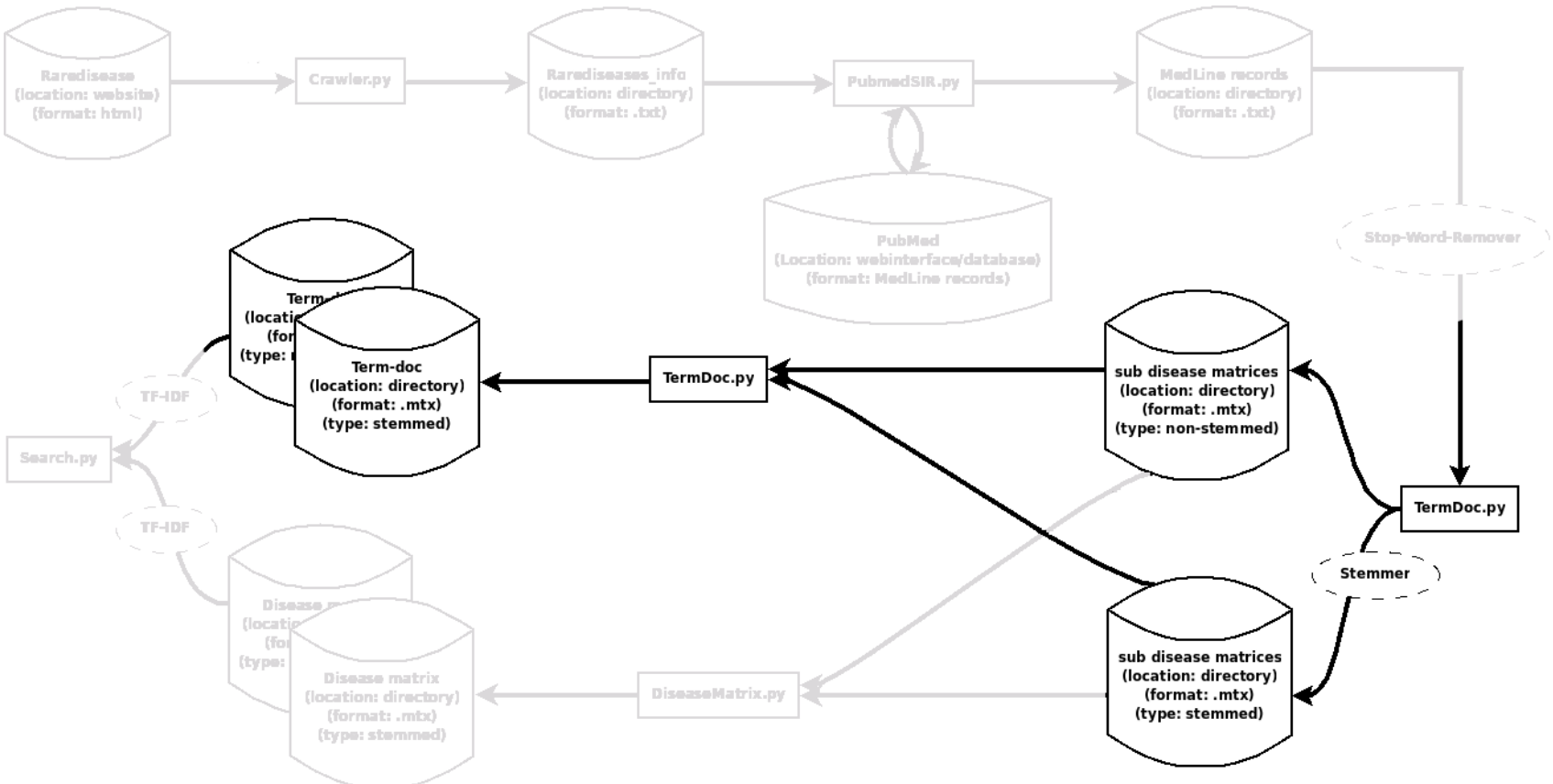
Oversigt over systemet: Crawler



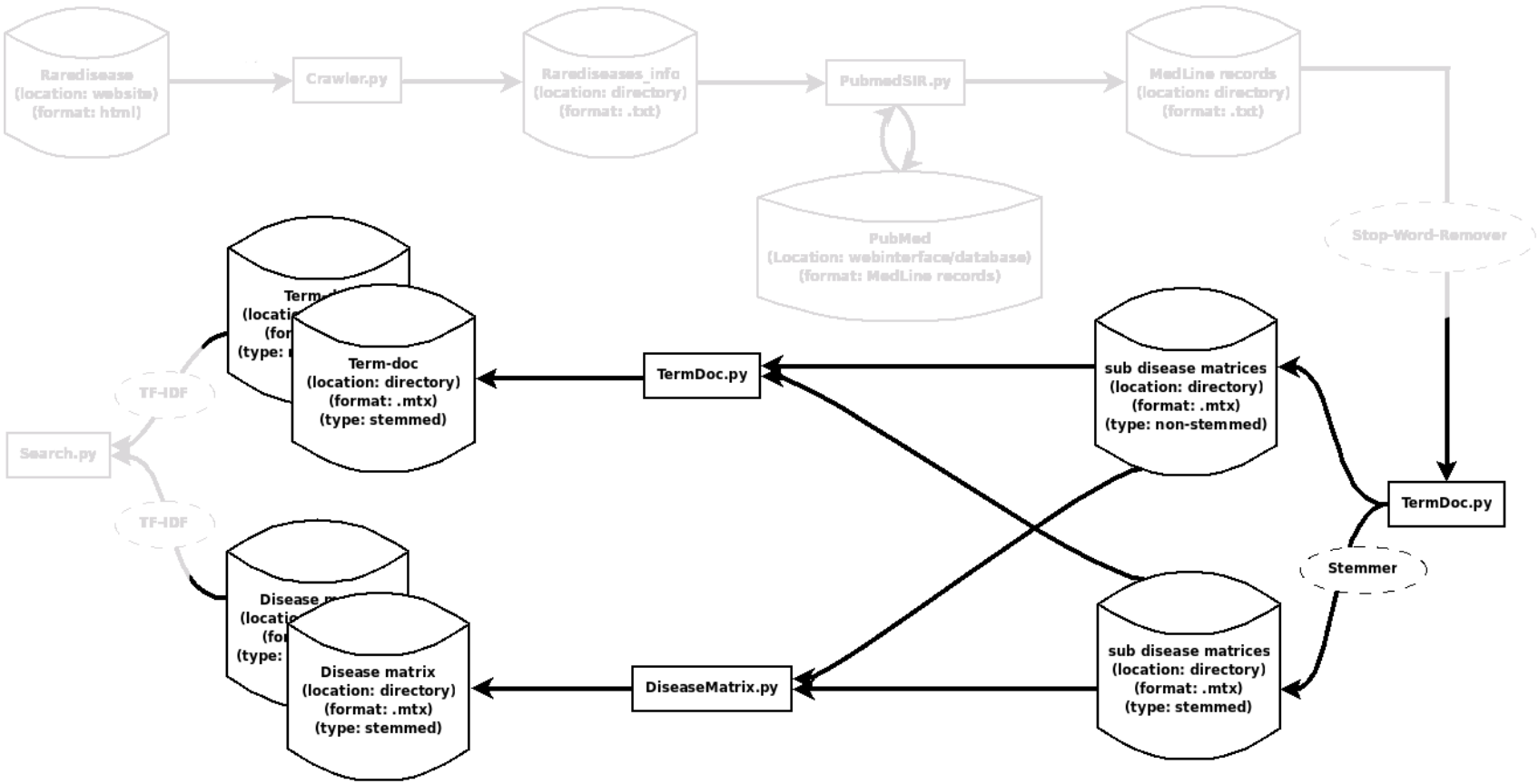
Oversigt over systemet: MedLine



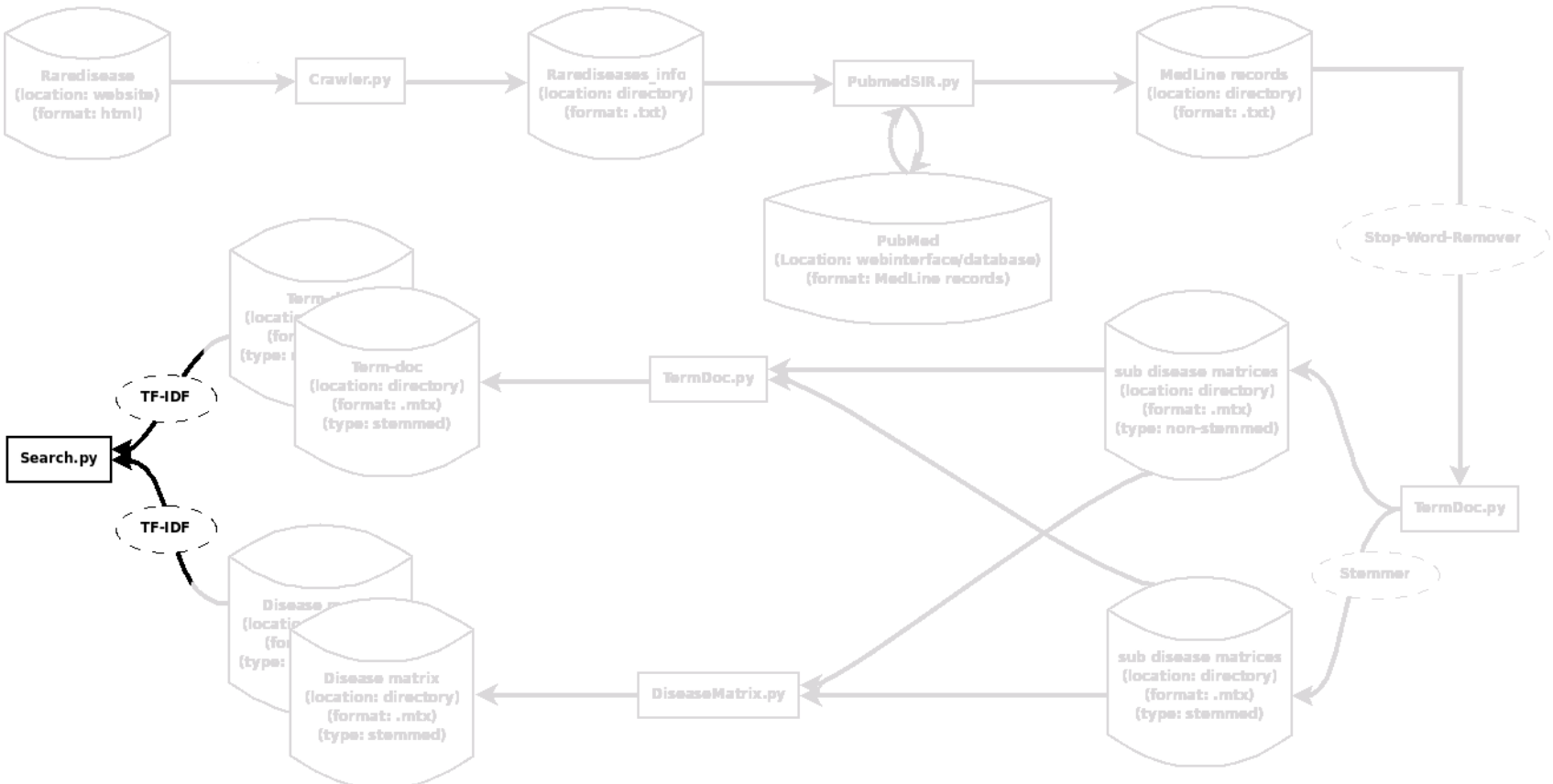
Oversigt over systemet: TermDoc



Oversigt over systemet: TermDoc / DiseaseMatrix



Oversigt over systemet: TF-IDF



Konstruktion af databasen

Indsæt tegning af interaktion mellem vores system og rarediseases, samt vores system og pubmed, endende med Medline Records. (evt. se figur 2.2 side 17)

Vægtning af term

Log-transformation af ord antallet:

$$x_{dw}^{log} = \log\left(1 + x_{dw}\right)$$

Der bruges TF-IDF til vægtning af termer, dette er for at fremhæve term som sjælendt optræder og nedvægte ofte forekommende ord:

$$x_{dw}^{tfidf} = x_{dw}^{log} \cdot \log \frac{D}{\sum_{d\prime=1}^{D} \delta_{d\prime w}}$$

Afslutningsvis normaliseres dokument vektoren for at sikre de har ens indflydelse på søgeresultatet:

$$x_{dw}^{norm} = \frac{x_{dw}^{tfidf}}{\sqrt{\sum_{w\prime=1}^{W} x_{dw}^{tfidf\,2}}}$$

Vector space model - Term doc matrix

Syntes denne skal op før TF-IDF, da det ikke mening at snakke om TF-IDF uden først at have en term-doc matrix.

$$\begin{array}{c} \\ \mathbf{d}_j \longrightarrow \end{array} \begin{array}{c} \mathbf{t}_i \\ \downarrow \\ \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} \end{array}$$

d_j hvor er en transponeret søjle vektor.

Vector space model - Cosine score

Udregning af vinkel mellem en symptom liste og et document:

$$\cos\theta_{D_j} = \frac{Q \circ D_j}{|Q| \cdot |D_j|}$$

Ovenstående svarer til

$$\text{score}_d = \frac{1}{|I|}\frac{1}{|x_d|}\sum_{i \in I} x_{dw}$$

Hvis der på forhånd er foretaget normalisering af dokument vektoren:

$$\propto \sum_{i \in I} \widehat{x_{dw}}$$

Outlier detection

Hvor fungere outlier detection??? Kan man goere det anderledes

Udregning af score for en sygdom

$$\text{disease}_j = \{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\}$$

- Cosine Mean
  - $\text{disease}_j = \text{mean}(\{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\})$
- Cosine Median
  - $\text{disease}_j = \{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\}$
- Cosine Max
  - $\text{disease}_j = \{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\}$
- Sum (På disease matrix)
  - final score: $\text{disease}_j \sum_{i \in I} t_{ji}$

Udregning af score for en sygdom

$$\text{disease}_j = \{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\}$$

- Cosine Mean
  - $\text{disease}_j = \text{mean}(\{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\})$
- Cosine Median
  - $\text{disease}_j = \{\text{d1}_{score}, \mathbf{d2_{score}}, \text{d3}_{score}, \text{d4}_{score}\}$
- Cosine Max
  - $\text{disease}_j = \{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\}$
- Sum (På disease matrix)
  - final score: $\text{disease}_j \sum_{i \in I} t_{ji}$

Udregning af score for en sygdom

$$\text{disease}_j = \{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\}$$

- Cosine Mean
  - $\text{disease}_j = \text{mean}(\{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \text{d4}_{score}\})$
- Cosine Median
  - $\text{disease}_j = \{\text{d1}_{score}, \mathbf{\text{d2}_{score}}, \text{d3}_{score}, \text{d4}_{score}\}$
- Cosine Max
  - $\text{disease}_j = \{\text{d1}_{score}, \text{d2}_{score}, \text{d3}_{score}, \mathbf{\text{d4}_{score}}\}$
- Sum (På disease matrix)
  - final score: $\text{disease}_j \sum_{i \in I} t_{ji}$

BMJ Resultater

Indsæt nogle af de gode resultater, evt. nogle af de dårlige.

Orpha.net Resultater

Indsæt nogle af de gode resultater, evt. nogle af de dårlige.

Blind tests Resultater

Indsæt nogle af de gode resultater, evt. nogle af de dårlige.

Forkerte forslag

Indsæt noget om keywords

Forkerte forslag 2

Vi har for lidt information om den enkelte sygdom, den “drukner” i symptomerne på andre sygdomme.

Clustering af resultater

Indsæt billede af clustering.
Hvorfor er det en god ide at cluster resultaterne, man kan se hvilke af de fundne sygdomme der minder om hinanden.

Udvidelser