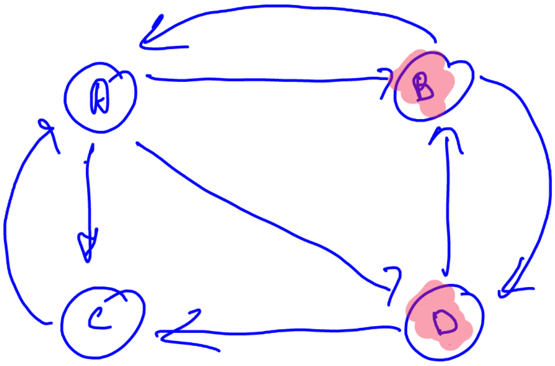
Topic-Sensitive PageRank



$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

$\beta = 0.8 \,/\, 0.85 \,/\, 0.9$

$$v' = \beta \underline{Mv} + (1-\beta)\, \boxed{\underline{e}}$$

$$e = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix} \longrightarrow e_S = \frac{1}{|S|}\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

sports

B and D

$$S = \{B, D\} \rightarrow e_S = \frac{1}{2}\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$

$$v' = \begin{bmatrix} 0 & 2/5 & 4/5 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} v + \underbrace{\begin{bmatrix} 0 \\ 1/10 \\ 0 \\ 1/10 \end{bmatrix}}$$

$\beta = 0.8$

$$e_S = \begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 1/2 \end{bmatrix}, \begin{bmatrix} 2/10 \\ 3/10 \\ 2/10 \\ 3/10 \end{bmatrix} \cdots \begin{bmatrix} 54/210 \\ 59/210 \\ 38/210 \\ 59/290 \end{bmatrix}$$

"Usual PageRank + Taxation"

$$\begin{bmatrix} 15/148 \\ 19/148 \\ 95/148 \\ 19/148 \end{bmatrix}$$

Importance of Words in Documents

Topics → special words

baseball → "ball", "bat", "run", "pitch"

Find significant words?

→ most frequent words? stop words ("the", "and"...)

Indicators of topics = relatively rare words

"notwithstanding" / "albeit"

$$\begin{tikzcd} \text{Polo} \arrow[r] & \text{"chukker"} \arrow[l, bend left] \end{tikzcd}$$

TF . IDF (Term Frequency times Inverse Document Frequence)

$$\boxed{TF_{\underset{\text{word}}{i}\,\underset{\text{document}}{j}} = \frac{f_{ij}}{\max_k f_{kj}}}$$

(frequence word i at doc j, exp. stop-words)

N documents in collection

term i appears in n_i documents

$$\boxed{IDF_i = \log_2 \left(N / n_i \right)}$$

For term i and doc j : $\boxed{TF_{ij} \times IDF_i}$

Example 20^{20} documents, word w appears in 2^{10}

$$IDF_w = \log_2 \left(2^{20} / 2^{10} \right) = \log_2 \left(2^{10} \right) = 10 .$$

If w appears in the doc j 20 times

$$TF_{wj} = 1 \rightarrow TF . IDF_{wj} = 10$$

document K, w appears only 1

$$TF_{wk} = 1/20 \rightarrow TF . IDF_{wj} = \frac{10}{20} = 1/2 ,$$

If the word "f" appears in each doc.

What will be IDF_f ? $IDF_f = 0$