Search within a collection of documents

Mathematical Modelling

Nik Jenič, Tian Ključanin, Maša Uhan

Problem Introduction

- Finding relevant documents according to our search

Solution

- LSI - Latent Semantic Indexing

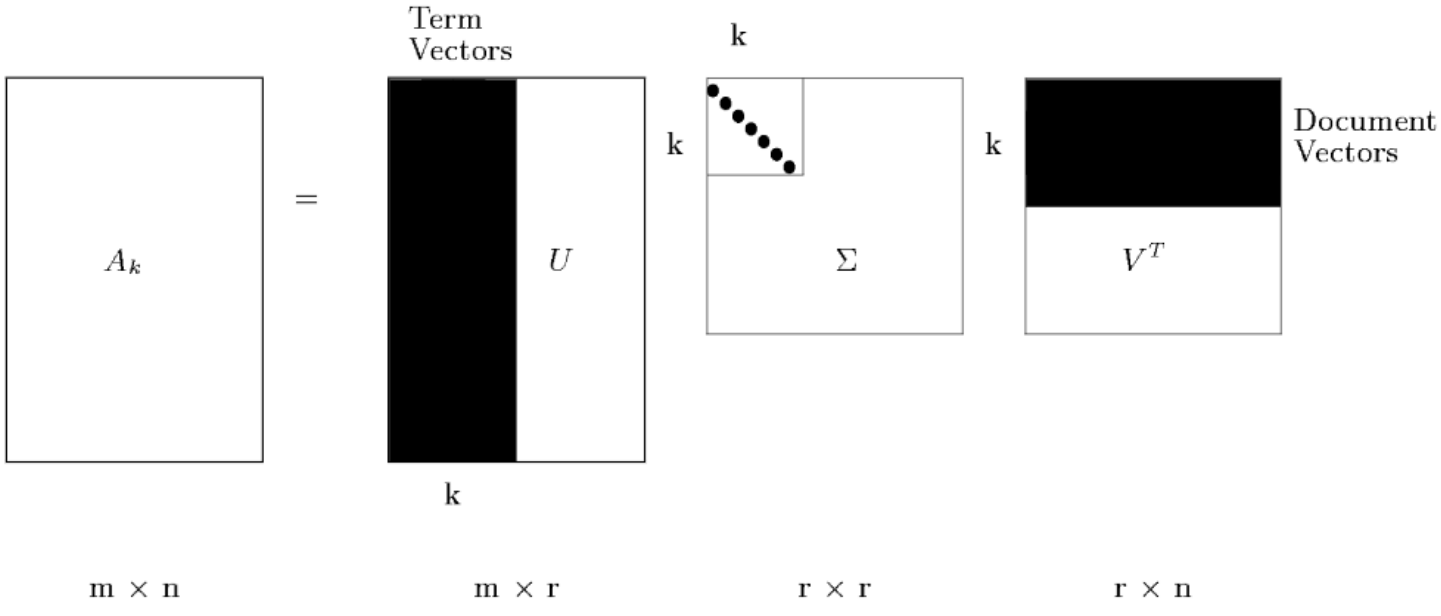

Figure: Mathematical representation of A_k

Solution

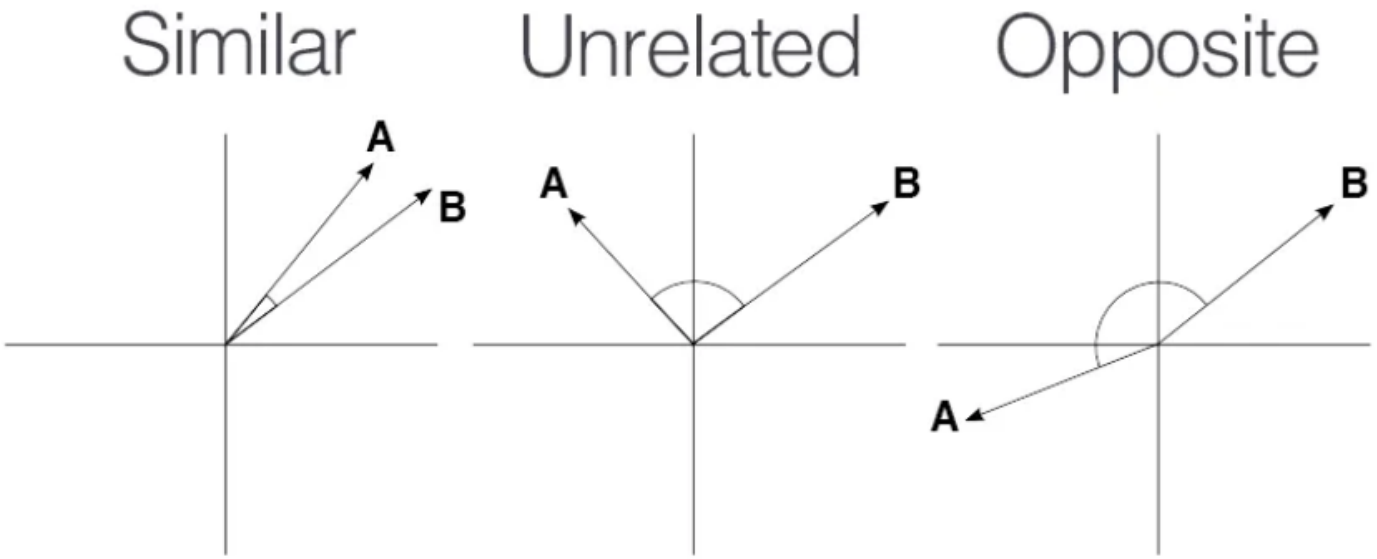

Figure: Cosine similarity in k=2

Optimization

- Giving words different weights
- Different ways of calculating the weights

$$a_{ij} = L_{ij} \cdot G_i$$

$$L_{ij} = \log(1 + f_{ij}), \quad G_i = 1 - \sum_j \frac{p_{ij} \log(p_{ij})}{\log n}, \quad p_{ij} = \frac{f_{ij}}{g_{f_i}}$$

Additional Improvements to the Solution

- Adding new documents without recalculation of SVD

$$\hat{q} = q^T U_k S_k^{-1}$$

- Adding new words without recalculation of SVD

$$\hat{q} = q^T V_k S_k$$

Results

- Frequency Solution:

	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
k=**10**	52.17	52.17	52.17	52.17	52.17	51.59	50.97	44.25	25.77
k=**50**	182.6	182.6	182.6	177.4	157.5	115.1	93.88	73.63	30
k=**100**	277.3	277.3	271.3	237.5	182.5	122.6	100.7	70	22
k=**250**	469.2	468.2	436.3	355.4	255.8	191.9	127.7	83	26
k=**500**	613.5	606.9	557.6	465.0	356.9	239.6	155	81	28
k=**750**	668.4	660.4	618.1	526.5	389.5	281.7	168	100	26
k=**1000**	641.9	627.2	580.6	486.4	385.2	295.6	213	119	37

Results

- Weighted Solution:

	0.10	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**	**0.90**
k=**10**	71.18	71.18	71.18	71.18	70.72	68.47	68.47	63.06	44.12
k=**50**	300.6	300.6	300.0	298.9	286.6	259.3	198.7	128.5	57.95
k=**100**	442.1	442.1	441.2	417.0	354.4	273.8	180.7	116.5	58
k=**250**	637.2	636.5	622.4	547.8	426.8	304.0	213.5	128.8	61
k=**500**	726.5	722.5	674.8	576.7	435.6	325.3	201.0	102	65
k=**750**	753.3	741.5	685.0	589.2	459.6	322.3	219	123	56
k=**1000**	673.5	657.3	604.0	499.4	409.7	316.6	232	141	70

Discussion

- Weighted approx. 10% better than Frequency solution
- Adding without recomputation overfitted, but functional

References

- Source for Figure 2: M. W. Berry, S.T. Dumais, G.W. O'Brien, Michael W. Berry, Susan T. Dumais, and Gavin. Using linear algebra for intelligent information retrieval. SIAM Review, 37:573–595, 1995