9주차 과제

Latent semantic analysis

담당교수님: 윤장혁 교수님

과목명: Data Analytics

이름: 박민성

학번: 201611145

전공: 산업공학과

제출일: 2020.05.15

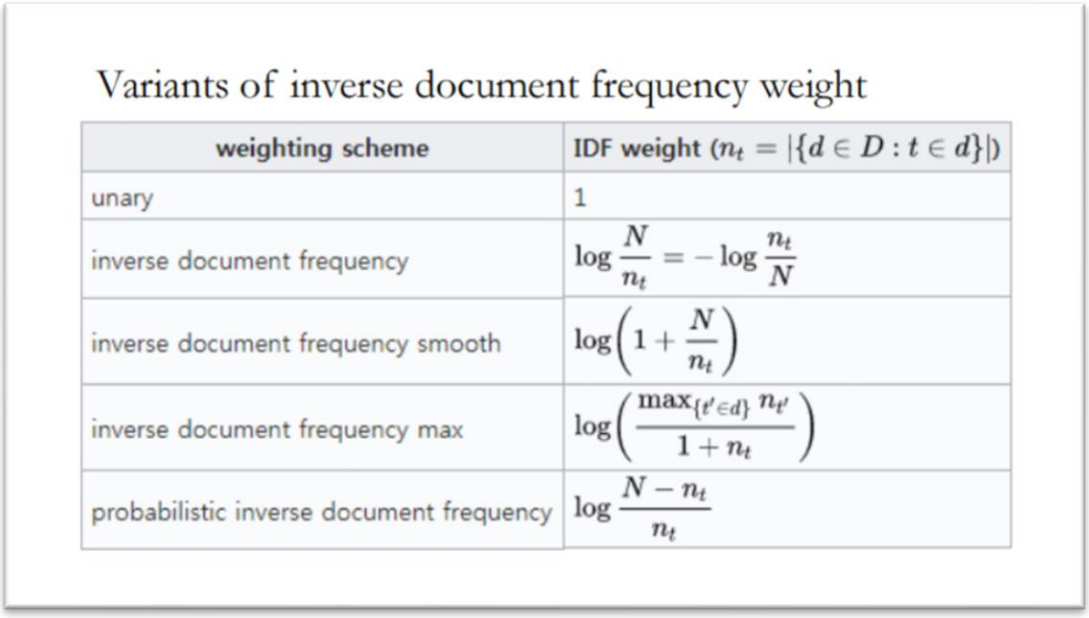
Object

- 주어진 Term-document matrix(data_week9.csv)를 이용하여 LSA를 수행한 후, 다음 요구사항에 대한 분석 수행
    - 1) 단어 'database'와 가장 유사한 단어 탐색
    - 2) 문서 'D6'과 가장 유사한 문서 탐색
- 주의 사항
    - Term-document matrix를 TF-IDF matrix로 변환한 후 LSA를 수행할 것
        - TF(raw count), IDF(inverse document frequency smooth)
    - 2개의 singular values를 활용
    - Cosine 유사도를 활용

Term-document matrix

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
database	11	21	12	0	37	2	0	3	1	6
SQL	21	10	0	7	31	0	0	21	5	0
index	9	0	5	2	20	0	1	0	11	0
regression	3	5	2	2	0	18	32	11	21	8
likelihood	0	3	0	0	3	7	12	4	27	4
linear	3	0	0	4	0	16	21	2	16	15

Lecture note p. 261 idf 자료 활용



Variants of inverse document frequency weight

weighting scheme	IDF weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(1 + \frac{N}{n_t}\right)$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1 + n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N - n_t}{n_t}$

Weighting scheme로 inverse document frequency smooth 사용

TF-IDF Matrix

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
database	3.874008	7.395833	4.22619	0	13.03075	0.704365	0	1.056548	0.352183	2.113095
SQL	8.945343	4.259687	0	2.981781	13.20503	0	0	8.945343	2.129844	0
index	3.468158	0	1.926754	0.770702	7.707018	0	0.385351	0	4.23886	0
regression	0.973533	1.622555	0.649022	0.649022	0	5.8412	10.38435	3.569622	6.814733	2.596089
likelihood	0	1.156053	0	0	1.156053	2.697456	4.624211	1.541404	10.40447	1.541404
linear	1.156053	0	0	1.541404	0	6.165614	8.092369	0.770702	6.165614	5.780263

SVD 결과(순서대로 U, Σ, Vt)

Ui j i \ j	1	2	3	4	5	6	7	8	9	10
1	-0.50349	0.328264	0.684444	0.312753	0	0.25236	0.093662	0	0	0
2	-0.63171	0.342577	-0.67504	0.061554	0	-0.00897	0.155049	0	0	0
3	-0.34387	0.099611	0.218471	-0.57493	0	-0.52054	-0.47181	0	0	0
4	-0.30911	-0.56141	-0.11636	0.312402	0	0.277101	-0.63353	0	0	0
5	-0.26288	-0.4069	0.06217	-0.62814	0	0.475259	0.375549	0	0	0
6	-0.25405	-0.53312	0.103568	0.275146	0	-0.60218	0.449655	0	0	0

Wj \ j	1	2	3	4	5	6	7	8	9	10
	26.82261	20.60061	7.939011	6.962247	0	4.197844	2.671985	0	0	0

Vti j i \ j	1	2	3	4	5	6	7	8	9	10
1	-0.35471	-0.26918	-0.11411	-0.10323	-0.67615	-0.16537	-0.2471	-0.29405	-0.35574	-0.13944
2	0.172577	0.121635	0.059954	-0.00387	0.445593	-0.3608	-0.58169	0.017923	-0.48709	-0.21711
3	-0.3203	0.260695	0.41345	-0.21949	0.244115	0.076671	0.001305	-0.69971	0.04024	0.231603
4	0.025902	0.338396	0.04309	0.046049	-0.10571	0.294038	0.333389	0.178112	-0.74153	0.30078
5	0.778344	0.281949	0.090096	-0.26389	-0.39138	-0.03989	-0.03576	-0.26714	0.085051	-0.04778
6	-0.36319	0.673493	0.032802	-0.29029	-0.1704	-0.15113	-0.00466	0.343977	0.178924	-0.35626
7	-0.05834	0.284205	-0.38182	0.128107	-0.1188	0.056452	-0.52559	0.056099	0.192727	0.647916
8	0	-0.32762	0.461684	-0.58034	-0.04974	0.174573	-0.25528	0.420661	0.009297	0.259929
9	0	-0.06778	-0.65581	-0.6103	0.268944	0.273693	0.069807	-0.14076	-0.08628	-0.11612
10	0	-0.06698	-0.11124	-0.24382	0.048711	-0.78558	0.376623	0.047906	-0.02791	0.39908

Singular Value는 (26.82261, 20.60061)로 결정했다.

1) 단어 'database'와 가장 유사한 단어 탐색

행렬 A

-0.503495	0.328264
-0.631708	0.342577
-0.343865	0.099611
-0.309106	-0.561406
-0.262878	-0.406898
-0.254055	-0.533122

행렬 B

26.822608	0
0	20.600609

행렬 A * B

-13.505042	6.762429
-16.944050	7.057300
-9.223363	2.052044
-8.291019	-11.5653
-7.051079	-8.382341
-6.814409	-10.982643

	DIM2	DIM2
database	**-13.505042**	**6.762429**
SQL	**-16.944050**	**7.057300**
index	**-9.223363**	**2.052044**
regression	**-8.291019**	**-11.5653**
likelihood	**-7.051079**	**-8.382341**
linear	**-6.814409**	**-10.982643**

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

	database 와 cosine 유사도
SQL	0.99758
index	0.97006
regression	0.157083
likelihood	0.232958
linear	0.090975

따라서, database와 가장 유사한 단어는 SQL이다.

2) 문서 'D6'과 가장 유사한 문서 탐색

행렬 A

26.822608	0
0	20.600609

행렬 B

-0.354712	-0.269179	-0.114115	-0.103226	-0.676148	-0.165372	-0.247099	-0.294050	-0.355744	-0.139438
0.172577	0.121635	0.059954	-0.003872	0.445593	-0.360799	-0.581692	0.017923	-0.487094	-0.217109

행렬 A * B

-9.514293	-7.220083	-3.060860	-2.768783	-18.136054	-4.435697	-6.627847	-7.887201	-9.54198	-3.740102
3.555182	2.505747	1.235095	-0.079769	9.179492	-7.432681	-11.983208	0.369219	-10.034426	-4.472588

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
DIM1	**-9.51429**	**-7.22008**	**-3.06086**	**-2.76878**	**-18.1361**	**-4.4357**	**-6.62785**	**-7.8872**	**-9.54198**	**-3.7401**
DIM2	**3.555182**	**2.505747**	**1.235095**	**-0.07977**	**9.179492**	**-7.43268**	**-11.9832**	**0.369219**	**-10.0344**	**-4.47259**

	D1	D2	D3	D4	D5	D7	D8	D9	D10
D6 와 cosine 유사도	**0.179471**	**0.202592**	**0.153906**	**0.53698**	**0.069442**	**0.999461**	**0.471748**	**0.975415**	**0.987482**

따라서 D6와 가장 유사한 문서는 D7이다.