Lister Hill National Center
for Biomedical Communications

A Method for Verifying a Vector-Based Text Classification System

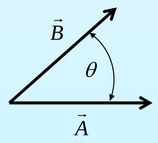
Text Categorization
http://specialist.nlm.nih.gov/tc

Objective:
Develop a methodology to compare two sets of vectors resulting in a single index measuring their similarity.

Methodology:

1). Comparing 2 Vectors

$$\vec{A}=(a_1,a_2,\ldots,a_n),\vec{B}=(b_1,b_2,\ldots,b_n)$$
$$\vec{A}\cdot\vec{B}=|\vec{A}|\cdot|\vec{B}|\cdot\cos(\theta)$$
$$\Rightarrow \cos(\theta)=\frac{\vec{A}\cdot\vec{B}}{|\vec{A}|\cdot|\vec{B}|}$$
$$\Rightarrow \cos(\theta)=\frac{(a_1\cdot b_1+a_2\cdot b_2+\ldots+a_n\cdot b_n)}{\sqrt{a_1{}^2+a_2{}^2+\ldots+a_n{}^2}\cdot\sqrt{b_1{}^2+b_2{}^2+\ldots+b_n{}^2}}$$
$$\Rightarrow \cos(\theta)=\frac{\sum_{i=0}^{n}a_i\cdot b_i}{\sqrt{\sum_{i=0}^{n}a_i{}^2}\sqrt{\sum_{i=0}^{n}b_i{}^2}}$$
$$\Rightarrow S(\vec{A},\vec{B})=\cos(\theta)$$



2). Comparing 2 sets of Vectors (Similarity Index)

Two sets of Vector	Similarity	Perfect Similarity
$\vec{A}_1=(a_{11},a_{12},\ldots,a_{1n}),\vec{B}_1=(b_{11},b_{12},\ldots,b_{1n})$	$S(\vec{A}_1,\vec{B}_1)=\cos(\theta_1)$	$S(\vec{A}_1,\vec{A}_1)=\cos(0)=1$
$\vec{A}_2=(a_{21},a_{22},\ldots,a_{2n}),\vec{B}_2=(b_{21},b_{22},\ldots,b_{2n})$	$S(\vec{A}_2,\vec{B}_2)=\cos(\theta_2)$	$S(\vec{A}_2,\vec{A}_2)=\cos(0)=1$
…	…	…
$\vec{A}_m=(a_{m1},a_{m2},\ldots,a_{mn}),\vec{B}_m=(b_{m1},b_{m2},\ldots,b_{mn})$	$S(\vec{A}_m,\vec{B}_m)=\cos(\theta_m)$	$S(\vec{A}_m,\vec{A}_m)=\cos(0)=1$
	$\vec{S}(A_i,B_i)_{i=0,\ldots m}$	$\vec{S}(A_i,A_i)_{i=0,\ldots m}$

SI

Applications (JDI):

1). Introduction (Word-JD Vectors):

- The words (400K) are from a multi-year collection of MEDLINE
- The JDs (about 120 biomedical disciplines, e.g., Cardiology, Genetics) are from about 4,000 records from NLM's Serials file representing journals indexed in MEDLINE.

Word	JDID	WC score	DC score
a00	JD036	0.5608018	0.3280480
…	…	…	…
heart	JD017	0.00180937	0.00524491
heart	JD018	0.04665726	0.09366356
heart	…	…	…
heart	JD114	0.00075555	0.00223503
heart	JD115	0.00665273	0.01562358
…	…	…	…

2). Approaches: Common vector components:

- Common words
- Common JDs

3). Results:

Three years of MEDLINE is a good increment for future release

No. of Years of MEDLINE	SI vs. 2005~07 Version
1 year: 2007~07	0.9803
2 years: 2006~07	0.9935
3 years: 2005~07	1.0000
4 years: 2004~07	0.9957
5 years: 2003~07	0.9920
6 years: 2002~07	0.9893

MEDLINE Versions	SI between Increments
1999~01 vs. 2000~02	0.9793
2000~02 vs. 2001~03	0.9772
2001~03 vs. 2002~04	0.9795
2002~04 vs. 2003~05	0.9808
2003~05 vs. 2004~06	0.9795
2004~06 vs. 2005~07	0.9797