ภาคผนวก 1

Math Tutorial

1. Summation $$s = \sum_{i=0}^{n-1} i$$

2. Program of that summation

```
public class Sumup {
    int n=10;
    int s = 0;
    int i = 0;
    while (i <= (n-1)) {
        s = s+i;
        i = i + 1;
    }
}
```

Or

```
public class multup2 {
    int n=10;
    int s = 0;
    int i;
    int a[] = {0,1,2,3,4,5,6,7,8,9,10,11};
    for (i = 1; i <= n; i++)
        s = s + (a[i] * a[i+1]);
}
```

3. Summation

$$s = \sum_{i=1}^{n} a_i \cdot a_{i+1}$$

```
public class multup {
    int n=10;
    int s = 0;
    int i = 1;
    int a[] = {0,1,2,3,4,5,6,7,8,9,10,11};
    while (i <= n) {
        s = s + (a[i] * a[i+1]);
        i = i + 1;
    }
}
or…
public class multup2 {
    int n=10;
    int s = 0;
    int i;
    int a[] = {0,1,2,3,4,5,6,7,8,9,10,11};
    for (i = 1; i <= n; i++)
        s = s + (a[i] * a[i+1]);
}
```

4. **Simple tf * idf**

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

T_k = term k in document D_i

tf_{ik} = frequency of term T_k in document D_i

idf_k = inverse document frequency of term T_k in C

N = total number of documents in the collection C

n_k = the number of documents in C that contain T_k

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

Inverse Document Frequency

For a collection of 10000 documents (N = 10000)

$$\log\left(\frac{10000}{10000}\right) = 0$$

$$\log\left(\frac{10000}{5000}\right) = 0.301$$

$$\log\left(\frac{10000}{20}\right) = 2.698$$

$$\log\left(\frac{10000}{1}\right) = 4$$

5. Similarity Measures

Simple matching (coordination level match)	$\lvert Q \cap D \rvert$
Dice's Coefficient	$2\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert + \lvert D \rvert}$
Jaccard's Coefficient	$\frac{\lvert Q \cap D \rvert}{\lvert Q \cup D \rvert}$
Cosine Coefficient	$\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert^{1/2} \times \lvert D \rvert^{1/2}}$
Overlap Coefficient	$\frac{\lvert Q \cap D \rvert}{\min(\lvert Q \rvert, \lvert D \rvert)}$

6. tf*idf Normalization

$$w_{ik} = \frac{\left(tf_{ik} * \log(N / n_k)\right)}{\sqrt{\left(\sum_{k=1}^{t}\left(((tf_{ik})^2) * ([\log(N / n_k)]^2)\right)\right)}}$$

7. Vector Space Similarity

$$sim(D_i, D_j) = \sum_{k=1}^{t} w_{ik} * w_{jk}$$

8. Vector Space Similarity Measure

$D_i = w_{d_{i1}}, w_{d_{i2}}, ..., w_{d_{it}}$

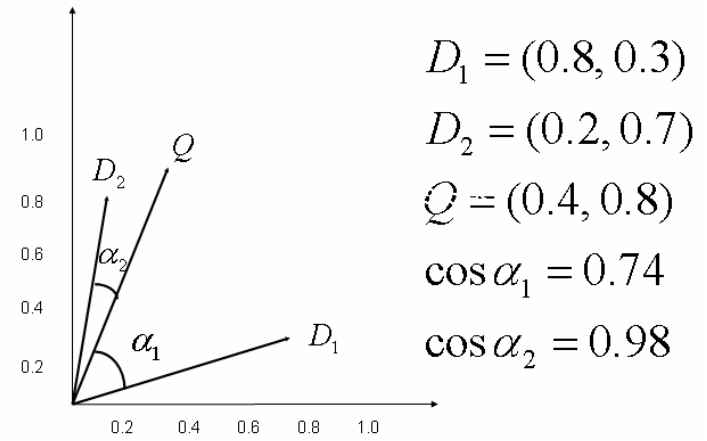
$Q = w_{q1}, w_{q2,}..., w_{qt}$ $\qquad w = 0$ if a term is absent

if term weights are normalized: $sim(Q, D_i) = \sum_{j=1}^{t} w_{qj} * w_{d_{ij}}$

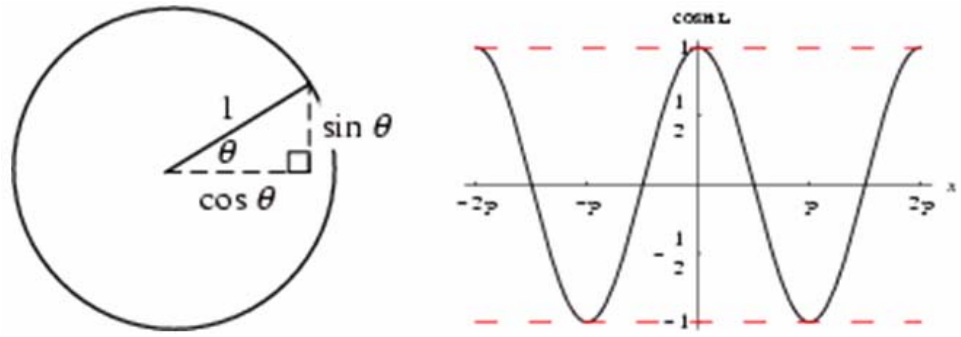
otherwise normalization and the similarity comparison are combined:

$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t}(w_{qj})^2 * \sum_{j=1}^{t}(w_{d_{ij}})^2}}$$

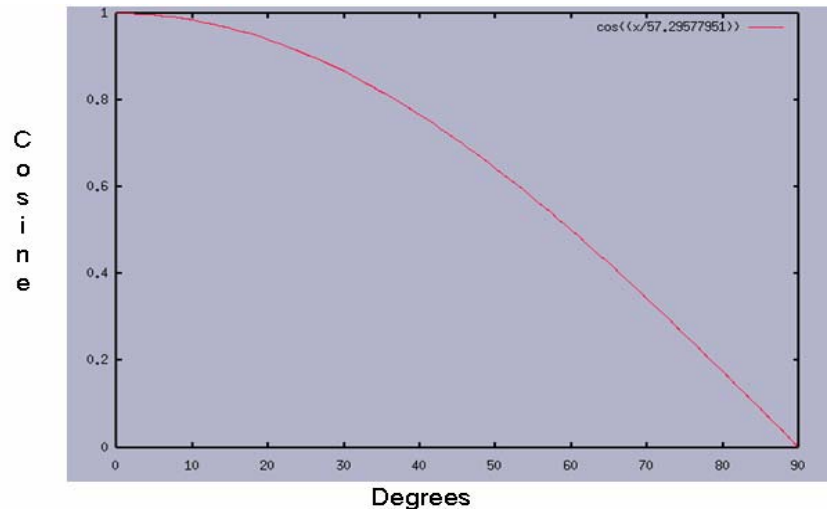
9. Computing Similarity Scores



What's Cosine Anyway?



Cosine vs. Degrees



Computing a Similarity Score

Say we have query vector $Q = (0.4,0.8)$

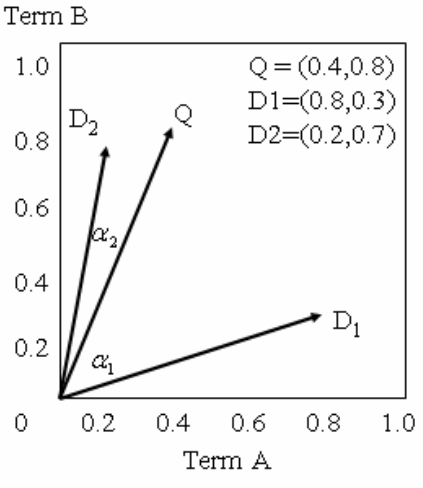
Also, document $D_2 = (0.2,0.7)$

What does their similarity comparison yield?

$$sim(Q, D_2) = \frac{(0.4*0.2)+(0.8*0.7)}{\sqrt{[(0.4)^2+(0.8)^2]*[(0.2)^2+(0.7)^2]}}$$

$$= \frac{0.64}{\sqrt{0.42}} = 0.98$$

Vector Space Matching



$$D_i=(d_{i1},w_{di1};d_{i2},\ w_{di2};\ \ldots;d_{it},\ w_{dit})$$
$$Q=(q_{i1},w_{qi1};q_{i2},\ w_{qi2};\ \ldots;q_{it},\ w_{qit})$$

$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{q_j} w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t}(w_{q_j})^2 \sum_{j=1}^{t}(w_{d_{ij}})^2}}$$

$$sim(Q, D2) = \frac{(0.4 \cdot 0.2)+(0.8 \cdot 0.7)}{\sqrt{[(0.4)^2+(0.8)^2]\cdot[(0.2)^2+(0.7)^2]}}$$

$$= \frac{0.64}{\sqrt{0.42}} = 0.98$$

$$sim(Q, D_1) = \frac{.56}{\sqrt{0.58}} = 0.74$$