Fundamentals of Data Science

Semester B 20-21

Tutorial 2

1. There are a number of ways to compare two objects (vectors) **x** and **y** that consist of *n* binary attributes. The comparison of two such vectors leads to the following four quantities:
   f_{00}= the number of attributes with a value of 0 in both **x** and **y**.
   f_{01}= the number of attributes with a value of 0 in **x** and 1 in **y**.
   f_{10}= the number of attributes with a value of 1 in **x** and 0 in **y**.
   f_{11}= the number of attributes with a value of 1 in both **x** and **y**.

   Based on these quantities, we can define the following two measures:

   Simple Matching Coefficient (SMC):

   $$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

   Jaccard coefficient

   $$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

   a. Calculate the value of the Simple Matching Coefficient and the Jaccard coefficient for the two vectors **x**=(1,0,0,0,0,1,1,0) and **y**=(0,0,1,0,1,0,1,0).
   b. What is the main difference between these two measures?

2. The cosine similarity of two vectors **x** and **y** with continuous attributes is defined as follows:

   $$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

   where • indicates the dot product between two vectors, $\sum_{u=1}^{n} x_u y_u$ (x_u and y_u are the u-th attributes of **x** and **y** respectively), and $\|\mathbf{x}\|$ is the length of vector **x**, $\|\mathbf{x}\| = \sqrt{\sum_{u=1}^{n} x_u^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$.

We consider the following set of documents.

Article	Words
1	economics:3, market:5, bank:2
2	economics:12, market:15, bank:10, company:8
3	company:5, medicine:3, insurance:5

Calculate the value of the cosine similarity between (i) article 1 and article 2, and (ii) article 2 and article 3. What is your observation?

3. We consider the following set of documents.

Article	Words
1	dollar:1, industry:4, country:2, loan:3, deal:2, government:2, very:6
2	machinery:2, labor: 3, market:4, industry:2, government:3, very:5
3	job:5, inflation:3, company:2, market:3, country:2, index:3, very:8
4	domestic:3, forecast:2, government:1, market:2, sale:3, price:2, very:2
5	patient:4, symptom:2, drug:3, health:2, clinic:2, doctor:2, very:5
6	pharmaceutical:2, company:3, drug:2, vaccine:1, flu:3, very:3
7	death:2, cancer:4, drug:3, government:4, health:3, director:2, very:7
8	medical:2, cost:3, government:2, patient:2, health:3, care:1, very:5

Let t_{ij} be the count of the i-th word in the j-th document, and m be the number of documents. Consider the transformation defined by

$$t'_{ij} = t_{ij} \log_2 \frac{m}{n_i}$$

where n_i is the number of documents in which the i-th word appears.

a. Apply this transformation to (i) the count of the word “pharmaceutical” in article 6, (ii) the count of the word “government” in article 2, and (iii) the count of the word “very” in article 5. What is your observation?
b. What is the purpose of this transformation?