KL- Divergence: Definition:

Kull back - Liebler Divergence is a distance that is not a metric.

Let X_a and X_b be two discrete distributions.

$$d_{KL}(X_a, X_b) = d_{KL}(a, b) = \sum_{i=1}^{d} a_i \ln(a_i / b_i)$$

Example Bag-of-Words

Consider D-dimensional space with D = 11.

For each co-ordinate list the corresponding word as.

(am, and, do, ham, I, jelly, like, not, Sam, them, zebra)

Use the text from the following 4 short documents.

D_1 : I am Sam

D_2 : Sam I am

D_3 : I do not like jelly and ham.

D_4 : I do not, do not, like them, Sam I am

For each of the above documents form the representative vectors.

$V_1 = (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0)$

$V_2 = (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0)$

$V_3 = (0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$

$V_4 = (1, 0, 2, 0, 2, 0, 1, 2, 1, 1, 0)$

Also find the distance between the (Euclid) documents, D_1, D_2, D_3 and D_4.

Find the L_2 distance or d_2 distance or Euclidean distance between the vectors.

Formula:

$$d_2(a,b) = \|a-b\| = \|a-b\|_2 = \sqrt{\sum_{i=1}^{d} (a_i - b_i)^2}$$

$$d_2(V_1,V_2) = \sqrt{\sum_{i=1}^{9} (1-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2 + (1-1)^2 + (0-0)^2 + (0-0)^2}$$

$$= \sqrt{0+0+\cdots+0} = 0$$

$$d_2(V_1,V_3) = \sqrt{(1-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (1-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2}$$

$$= \sqrt{1+1+1+1+0+1+1+1+1+0+0}$$

$$d_2(V_1,V_3) = \sqrt{8} = 2.83$$

$$d_2(V_1,V_4) = \sqrt{(1-1)^2 + (0-0)^2 + (0-2)^2 + (0-0)^2 + (1-2)^2 + (0-0)^2 + (0-1)^2 + (0-2)^2 + (1-1)^2 + (0-1)^2 + (0-0)^2}$$

$$= \sqrt{0+0+4+0+1+0+1+4+0+1+0}$$

$$= \sqrt{0+0+4+0+1+0+1+4+0+1+0}$$

$$= \sqrt{11}$$

$$d_2(V_1,V_4) = 3.32$$

$$d_2(V_2, V_3) = \sqrt{(1-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (1-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2}$$

$$= \sqrt{1+1+1+1+0+1+1+1+1+0+0}$$

$$= \sqrt{8} = 2.83$$

Vectors	1	2	3	4	5	6	7	8	9	10	11
V_1	1	0	0	0	1	0	0	0	1	0	0
V_2	1	0	0	0	1	0	0	0	1	0	0
V_3	0	1	1	1	1	1	1	1	0	0	0
V_4	1	0	2	0	2	0	1	2	1	1	0

$$d_2(V_2, V_4) = \sqrt{(1-1)^2 + (0-0)^2 + (0-2)^2 + (0-0)^2 + (1-2)^2 + (0-0)^2 + (0-1)^2 + (0-2)^2 + (1-1)^2 + (0-1)^2 + (0-0)^2}$$

$$= \sqrt{0+0+4+0+1+0+1+4+0+1+0}$$

$$= \sqrt{11} = 3.32$$

=

$$d_3(V_3, V_4) = \sqrt{(0-1)^2 + (1-0)^2 + (1-2)^2 + (1-0)^2 + (1-2)^2 + (1-0)^2 + (1-1)^2 + (1-2)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2}$$

$$= \sqrt{1+1+1+1+1+1+0+1+1+1+0}$$

$$= \sqrt{9}$$

$$\boxed{d_3(V_3, V_4) = 3}$$

Compute Jacard distance among the documents D_1, D_2, D_3 and D_4 for word 2-grms.

D_1: $G_1 = \{ [I\ am], [am\ Sam] \}$

D_2: $G_2 = \{ [Sam\ I], [I\ am] \}$

D_3: $G_3 = \{ [I\ do], [do\ not], [not\ like], [like, jelly], [jelly\ and], [and\ ham], [ham\ I] \}$

Solution:

Jacard distance between two sets A and B is

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

Here $|A \cap B|$ represents no of elements in $A \cap B$.

$|A \cup B|$ " No of elements in $A \cup B$.

$$d_J(D_1, D_2) = 1 - \frac{1}{3} = \frac{2}{3} = 0.667$$

=

~~$d_J(D_1, D_3) = 1 - 0$~~

Problem: Compute Jacard distance among the documents D_1, D_2, D_3, and D_4 for word 2-grams.

D_1 : $G_{1} = \{ [I\ am], [am, Sam] \}$

D_2 : $G_{2} = \{ [Sam\ I], [I\ am] \}$

D_3 : $G_{3} = \{ [I\ do], [do\ not], [not\ like], [like.\ jelly], [jelly\ and], [and\ ha$

D_4 : $G_{4} = \{ [I\ do], [do\ not], [not\ do], [not\ lik$
$[like\ them], [them\ Sam], [Sam\ I], [I\ $

Each gram is shown as a set of two words in square brackets.

Solution:

Formula : Jacard distance between any two sets A and B is

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

$|A \cap B|$ = Cardinality of $A \cap B$
= No of elements in $A \cap B$.

$|A \cup B|$ = No of elements in $A \cup B$.

$|D_1| = 2$, $|D_2| = 2$, $|D_3| = 6$

$|D_4| = 8$

$|D|$ represents no of grams in the docemat D.

$|D_1 \cap P_2| = 1$. $|D_1 \cap P_3| = 0$ $|D_1 \cap P_4| = 1$

$|D_1 \cup P_2| = 3$. $|D_1 \cup P_3| = 8$ $|D_1 \cup D_4| = 9$.

$|D_2 \cap D_3| = 0$ ~~$|D_2 \cap D_4| =$~~ $|D_2 \cap D_4| = 2$

$|D_2 \cup D_3| = 8$ $|D_2 \cup D_4| = 8$

$|D_3 \cup D_4| = 11$

$|D_1 \cap D_4| = 3$.

$$d_J(D_1, D_2) = 1 - \frac{|D_1 \cap P_2|}{|D_1 \cup P_2|} = 1 - \frac{1}{3} = 1 - 0.33 = 0.667$$

$$d_J(D_1, D_3) = 1 - \frac{|D_1 \cap D_3|}{|D_1 \cup D_2|} = 1 - \frac{0}{8} = 1 - 0 = 1$$

$$d_J(D_1, D_4) = 1 - \frac{|D_1 \cap D_4|}{|D_1 \cup D_4|} = 1 - \frac{1}{9} = 0.889$$

$$d_J(D_2, D_3) = 1 - \frac{|D_2 \cap D_3|}{|D_2 \cup D_3|} = 1 - \frac{0}{8} = 1$$

$$d_J(D_2, D_4) = 1 - \frac{|D_2 \cap D_4|}{|D_2 \cup D_4|} = 1 - \frac{2}{8} = 1 - 0.25 = 0.75$$

$$d_J(D_3, D_4) = 1 - \frac{|D_3 \cap D_4|}{|D_3 \cup D_4|} = 1 - \frac{3}{11} = 0.727.$$

— ++++ —

10 Marks:

Problem:

Construct character 3 grams for the document D_4.

$D_4 : G_4 = \{$ [I do], [do not], [not do], [not like], [like them], [them, Sam], [sam I], [I am]

Solution:

Characters 3 grams [ignore whitespace, punctuation and capitalization, repitions]

{ [ido], [don], [ono], [not], [not], [otd], [t
[not], [otl], [tli], [lik], [ike], [lik], [ket
[eth], [the], [hem], [the], [hem], [ems],
[msa], [sam], [sam], [ami], [mi], [iam]}

= { [ido], [don], [ono], [not], [otd], [tdo],
[otl], [tli], [lik], [ike], [ket],
[eth], [the], [hem], [ems], [msa],
[sam], [ami], [ma], [iam]}

Problem 6 Marks:

Construct characters 4-grams for the document D_4.

D_4: $G_4 = \{$[I do], [do not], [not do], [not like], [like them], [them Sam], [Sam I], [I am]$\}$

Solution:

character 4 grams for D_4

$= \{$ [idon], [dono], [onot], [notd], [otdo], [tdon], [otli], [tlik], [Like], [iket], [keth], [ethe], [them], [hems], [emsa], [msam], [sami], [amia], [miam]$\}$.

Problem 4: 4 Marks:

Find the Jacard distance between two sets $A = \{1, 2, 4, 8\}$, $B = \{1, 2, 3\}$

Solution:

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}.$$

Here $A \cap B = \{1, 2\} \Rightarrow |A \cap B| = 2$

$A \cup B = \{1, 2, 3, 4, 8\} \Rightarrow |A \cup B| = 5.$

$d_J(A, B) = 1 - \frac{2}{5} = 1 - 0.4 = 0.6.$

$\boxed{d_J(A, B) = 0.6.}$

6 Marks:

Construct ~~2-grams~~ characters 3-grams for the document D_3:

D_3 : G_3 = { [I do], [do not], [not like], [like, ~~j~~jelly.], [jelly and], [~~and~~ and ham] }

Solution:

Character 3-grams for D_3 is

{ [ido], [don], [ono], [not], [nOt], [otl], [tli], [lik], [ike], [kej] [eje], [jel], [ell], [lly], [~~and~~], [lya], [yan], [and], [ndh], [dha] [ham] }

= { [ido], [don], [ono], [not], [nOt], [otl], [tli], [lik], [ike], [kej], [eje], [jel], [ell], [lly], [lya], [yan], [and], [ndh], [dha], [ham] }

4 Marks:

Construct character 2-grams for the document D_2: $G_2 = \{[sam\ I], [I\ am]\}$

Solution:

Character 2-grams for the Document D_2 is

$\{[sa], [am], [mi], [ia], [am]\}$

$= \{[sa], [am], [mi], [ia]\}$ [∵ Repition not allowed]