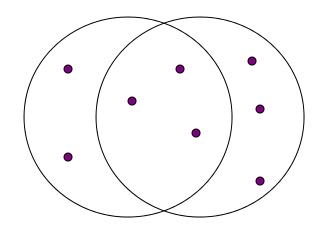
Minhashing

Jaccard Similarity Measure Constructing Signatures

Jaccard Similarity

- The Jaccard similarity of two sets is the size of their intersection divided by the size of their union.
- $Sim(C_1, C_2) = |C_1 \cap C_2|/|C_1 \cup C_2|$.

Example: Jaccard Similarity



3 in intersection.8 in union.Jaccard similarity= 3/8

From Sets to Boolean Matrices

- Rows = elements of the universal set.
 - Example: the set of all k-shingles.
- Columns = sets.
- 1 in row e and column S if and only if e is a member of S.
- Column similarity is the Jaccard similarity of the sets of their rows with 1.
- Typical matrix is sparse.

Example: Column Similarity

```
C_{1} C_{2}
0 1 *
1 0 *
1 1 * * Sim(C<sub>1</sub>, C<sub>2</sub>) =
0 0 2/5 = 0.4
1 1 * *
```

Four Types of Rows

• Given columns C_1 and C_2 , rows may be classified as:

$$\begin{array}{cccc}
 & C_1 & C_2 \\
 a & 1 & 1 \\
 b & 1 & 0 \\
 c & 0 & 1 \\
 d & 0 & 0
\end{array}$$

- Also, a = # rows of type a, etc.
- Note $Sim(C_1, C_2) = a/(a + b + c)$.

Minhashing

- Imagine the rows permuted randomly.
- Define minhash function h(C) = the number of the first (in the permuted order) row in which column C has 1.
- Use several (e.g., 100) independent hash functions to create a signature for each column.
- The signatures can be displayed in another matrix – the signature matrix – whose columns represent the sets and the rows represent the minhash values, in order for that column.

Minhashing Example

Input matrix

日本学 2 一条 表 身	4	
	+ L $+$	
0 00 00 00 00 00 00 0	0.00.00.00.00.00.00.00	
	2	
		887 48888
161		16
	13	
	6	
888 P 4686		
4	U	
	J	
_	J	
		_
	7	1
5	7	2
5	7	2
5	7	2
5	7	2
5	7	2
5	7	2
5	7	2
5	7	2
5	7	2
5	7	2
5	7	2 5
5 4	7	2 5

1	0	1	0		
1	0	O	1		
0	1	0	1		
О	1	О	1		
О	1	O	1		
1	0	1	0		
1	0	1	0		

Signature matrix M

		8 25 25 26 26 25 26 26 26 26 26	500000000000000000000000000000000000000
69 1000 1000 1000 1000 1000 1000 1000 10		10 TOO TOO DEE OOR TOO TOO TOO TOO TOO	100000000000000000000000000000000000000
N 100 90 - 91 100 100 100 100 100 100 100 100 100	000000=00000000000000000000000000000000	N 101 90 - 91 93 90 101 101 90 101 10	
	10000		10000
	<u></u>		<u></u>
	90001 100000000000000000000000000000000		9000
		S (2) 25 (2) (2) (2) (2) (2) (2) (2) (2) (2) (2)	
16元素的原金素的原来的8			
		(3) (CC CC) (CC (CC CC) (CC (CC CC) (C	
8 (8) 50 (8) (8) (8) (8) (8) (8) (8) (8)			
(原名名) 医多角质多角的		医医动性医多类医医多类医炎	
2		4	
	100000000000000000000000000000000000000		
2 100 100 101 101 100 100 101 101 101 10		0 20 20 20 20 20 20 20 20 20 20 20 20 20	
		3333 <u>-33333</u> 33	
CH 1005 COD 2001 TOX COD 2001 TOX COD 2001 TOX		CH 102 CO 203 102 CO 203 103 CO 203 103	
***********		**********	
187. 222222			
1646886681	2000		2000
		868888888888	
NO DOS TOOS COO DOS TOOS COO DOS TOOS COO DA	000000000000000000000000000000000000000	SC 100 300 000 000 100 000 000 100 100 000 0	



Surprising Property

- The probability (over all permutations of the rows) that $h(C_1) = h(C_2)$ is the same as $Sim(C_1, C_2)$.
- Both are a/(a+b+c)!
- Why?
 - Look down the permuted columns
 C₁ and C₂ until we see a 1.
 - If it's a type-a row, then $h(C_1) = h(C_2)$. If a type-b or type-c row, then not.

Similarity for Signatures

- The similarity of signatures is the fraction of the minhash functions in which they agree.
 - Thinking of signatures as columns of integers, the similarity of signatures is the fraction of rows in which they agree.
- Thus, the expected similarity of two signatures equals the Jaccard similarity of the columns or sets that the signatures represent.
 - And the longer the signatures, the smaller will be the expected error.

Min Hashing – Example

Input matrix

1 4 3 3 2 4 7 1 7 6 3 6 2 6 1 5 7 2 4 5 5			
3 2 4 7 1 7 6 3 6 2 6 1 5 7 2			
3 2 4 7 1 7 6 3 6 2 6 1 5 7 2			
3 2 4 7 1 7 6 3 6 2 6 1 5 7 2			
3 2 4 7 1 7 6 3 6 2 6 1 5 7 2			
3 2 4 7 1 7 6 3 6 2 6 1 5 7 2		4	HETEP ARE
7 1 7 6 3 6 2 6 1 5 7 2	0 100 105 00 1 205 00 0 0 100 0		
7 1 7 6 3 6 2 6 1 5 7 2			
7 1 7 6 3 6 2 6 1 5 7 2			
7 1 7 6 3 6 2 6 1 5 7 2			
7 1 7 6 3 6 2 6 1 5 7 2			
7 1 7 6 3 6 2 6 1 5 7 2			
7 1 7 6 3 6 2 6 1 5 7 2			8 55 toros s 10 50 5
2 6 1 5 7 2			
2 6 1 5 7 2			
2 6 1 5 7 2	0 0 0 0 0 0 0 0 0 0 0		5 (02 00 50) (02 00 50) (0
2 6 1 5 7 2			
2 6 1 5 7 2			
2 6 1 5 7 2			88887.48B
2 6 1 5 7 2			8887.488E
2 6 1 5 7 2			
2 6 1 5 7 2			
2 6 1 5 7 2			
2 6 1 5 7 2			
2 6 1 5 7 2			
2 6 1 5 7 2			
5 7 2		~	
5 7 2			
5 7 2			
5 7 2			
5 7 2			
5 7 2			
5 7 2			
5 7 2 4 5 5			
5 7 2 4 5 5			
57255	86600066		
5 7 2 4 5 5			
57255			
57255		+++++++	
5 / Z 4 5 5		-	
5 5 5			
5 5			
5 5			
5 5			
5 5			
4 5 5			
4 5 5			

1	0	1	0		
1	0	0	1		
0	1	0	1		
О	1	0	1		
О	1	0	1		
1	О	1	0		
1	0	1	0		

Signature matrix M

ĩŏ	100	88	dia.	10	de	d	88	100	0	di	ok.	80		0000	000	300	000	300	800	88	500	300	8	iii	88	88	333	-	100	i s	10	100	800	50	9	
18				18	88								81	10000									8	881											81	
Ñ		ø		П	Ħ	Ħ				Ħ	Ø.		8i	8888									88	88	ø					×				88	闘	100000000000000000000000000000000000000
福		88		t	Ŧ	ø				a	a		8	10000									88	81	8					m			m	to	8	
iii			ŧ	н	H	Ħ				٠	8		e e		*								8	н		ø	=	H		Ħ					ä	
ä		2	16	i	н	8	8		Н	٠	3	2	ä	10000		п							8		훓	24	6	п	1	ta	н				H	10000-01
쯦		æ	ŧ	и	ä	8	8		В	H	8	a	8	-		н							83	器	8		s	V.	a	ŧ	+	H	8	100	뿗	
엺		s	и	ĸ.	4	9	2		H	Ψ	9	2	8		9	4	9						22	93	2	9	7		1	æ	40	18	æ	æ	멾	<u></u>
œ		옾			H	9	2		H	4	9	2	8		<u>.</u>	=	_						9	ш	9	잎				£	10	P	8	P	뱶	200000000000000000000000000000000000000
		竪	4	+	4	4	星		н	4	4	噩	9										2	н	2				-	÷	+	н			럞	
題		2		4	4	4				4	4		틳										23	4						u					텴	
胡		88		Ш	ш	и				ø	Si.		8											翩	副					10				10	R	
Н	ø	20	46	Į.	46	d		pd.	ю	gil.	σĺ	zd.	4	1000		200	90	200	90	œů.	zó0				οÚ	οű					Į.		100	pů		
			ı																										ı	ı						
Н													H											н											H	
в													в											н											Н	
В													3										Э,	щ											я	
Н													4											щ											Ц	
Ш			×	٠	L								4		4									ш			7	,							Щ	
И				I)	ı.											н								21			•								ш	5000-65
М			ь	ı									8			н							\simeq	8		7	Λ	п							8	(000000)
И		r		۰									1			-								М		п									В	The second secon
В													3															П							М	
Ø													ī											丽											Ħ	
В																							8	т											ï	
ö	7	8	to	t	Ŧ	Ħ	8	в	Ħ	t	a	8	3		-	=	8	7	7	7	7	7		п	8	8		te	t	×	t	т	×	ю	ñ	
100	80	20	10	12	88	4	95	20	2	88	8	25	8	2000	880	888	200	800	880	880	000	800	8	100	2	盔	200	200	100	ķ.	6 500	100	88	50	i	
В				18	Ш								8	9888									88	翩											ø	
18				88									81										\approx												81	
				П	M								Ø	8000									80												N	
R		80	P	ı	H	ĸ.							80	00000	re.		70						100	闡		99	pri	100		80					M	
崩		ä	58	ā	ŔŔ				8				ā	00000	800	8	ı						883	顪		ы	ø	93			8 88	100			Ñ	000000000000000000000000000000000000000
Ю			88	á									Ø	0000		92	Já						800				10	88							8	
ã	8	Ŧ	22		н	8			8	×	a		ä	****	7	4	-						88	æ		æ	æ	3	я	×	18	te			笞	
蓄		8	i	i	f	d	붏		Ħ	ä	뷺	텷		00000	000	88	880							뷺		900	900	100	f		16			閪	뷻	
렱		뷻	r.	H	Ħ	ń	렲	台	Ħ	Ħ	計	젊	ii.	50000					**				200	器		體	蜡	H	н	-	16	10	100	16	펿	***************************************
В	8			Ю	н	8			ю	88	æ	æ	8	***	88	88	88	88	88	8	88	88	88	器			×	ю	t	×	æ	HB.			쨊	
텶	æ	嫠	46	10	4	ij.	폌		18	뫮	돢	텶	9	2000	200	2	20		2	嶷	8		2	의	뫮	鯔	蝎	IS.	18	46	18	10	100	歸	텶	

	1-3	2-4	1-2
Col/Col	0.75	0.75	0
Sig/Sig	0.67	1.00	0

Implementation of Minhashing

- Suppose 1 billion rows.
- Hard to pick a random permutation of 1...billion.
- Representing a random permutation requires
 1 billion entries.
- Accessing rows in permuted order leads to thrashing.

Implementation — (2)

- A good approximation to permuting rows: pick, say, 100 hash functions.
- For each column c and each hash function h_i, keep a "slot" M(i, c).
- Intent: M(i, c) will become the smallest value of $h_i(r)$ for which column c has 1 in row r.
 - I.e., $h_i(r)$ gives order of rows for i^{th} permutation.

Implementation – (3)

```
for each row r do begin
  for each hash function h<sub>i</sub> do
      compute h_i(r);
  for each column c
      if c has 1 in row r
        for each hash function h_i do
           if h_i(r) is smaller than M(i, c) then
              M(i, c) := h_i(r);
 end;
```

Example

Row	C1	C ₂
1	1	0
2	0	1
3	1	1
4	1	0
5	0	1

$$h(x) = x \mod 5$$
$$g(x) = (2x+1) \mod 5$$

Sig1 Sig2
$$h(1) = 1$$
 1 ∞ $g(1) = 3$ 3 ∞

$$h(2) = 2$$
 1 2 $g(2) = 0$ 3 ∞

$$h(3) = 3$$
 1 2 $g(3) = 2$ 2 ∞

$$h(4) = 4$$
 1 2 ∞

$$g(4) = 4$$
 2 ∞

0

g(5) = 1 2

Implementation – (4)

- Often, data is given by column, not row.
 - Example: columns = documents, rows = shingles.
- If so, sort matrix once so it is by row.