Project

IFT6758 - Data Science

Mila

Université de Montréal

Announcements …

- 

Announcements …

- The Ugly: We have been forced to shut down part of the project! We are not allowed to use the raw image/text anymore.

- The Bad: We can still use some features (next slides), so the project remains the same! it gets much easier… less fun…

- The Good: We can still use the relational data!

LIWC features

- LIWC -> Linguistic Inquiry and Word Count
- LIWC is the gold standard in computerized text analysis.

I. STANDARD LINGUISTIC DIMENSIONS	
Pronouns	I, them, itself
Articles	a, an, the
Past tense	walked, were, had
Present tense	Is, does, hear
Future tense	will, gonna
Prepositions	with, above
Negations	no, never, not
Numbers	one, thirty, million
Swear words	*****
II. PSYCHOLOGICAL PROCESSES	
Social Processes	talk, us, friend
Friends	pal, buddy, coworker
Family	mom, brother, cousin
Humans	boy, woman, group
Affective Processes	happy, ugly, bitter
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Anger	hate, kill, pissed
Sadness	grief, cry, sad
Cognitive Processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain
Inclusive	with, and, include
Exclusive	but, except, without
Perceptual Processes	see, touch, listen
Seeing	view, saw, look
Hearing	heard, listen, sound
Feeling	touch, hold, felt
Biological Processes	eat, blood, pain
Body	ache, heart, cough
Sexuality	horny, love, incest
Relativity	area, bend, exit, stop
Motion	walk, move, go

- How to interpret LIWC features? http://liwc.wpengine.com/interpreting-liwc-output/
- Read this paper: https://www.cs.cmu.edu/~ylataus/files/TausczikPennebaker2010.pdf

How does it look like?

```
liwc
```

	userId	WC	WPS	SixItr	Dic	Numerals	funct	pronoun	ppron	i	...	Colon	SemiC	QMark	Exclam	Dash	Quo
0	1c1bb692d7765344d418c0247962e7f8	156	17.33	25.00	76.92	1.92	45.51	9.62	6.41	3.21	...	1.92	0.00	5.77	5.77	2.56	0.
1	2eba17f4ec950f23af8d31a1d1db4518	176	11.00	21.02	84.66	0.00	53.41	15.34	12.50	7.39	...	0.57	0.00	0.00	13.07	0.00	0.
2	4f3fc35de4f026bbe96d6aead80a011c	179	17.90	17.88	90.50	0.00	54.19	11.73	7.82	6.15	...	0.00	0.56	0.56	1.68	1.12	0.
3	5b670721d060e5063273aefefd0e0731	179	179.00	20.67	77.09	0.00	51.40	12.29	9.50	4.47	...	0.00	0.00	1.68	0.00	1.12	0.
4	5ca01e48cdf661e932cfe58588360a7f	18	18.00	44.44	55.56	0.00	27.78	0.00	0.00	0.00	...	22.22	0.00	3661.11	0.00	0.00	11.
5	7b561be60d859de59fec1929522643a8	164	18.22	23.78	73.17	0.61	40.85	12.80	10.37	9.76	...	8.54	0.00	0.61	13.41	0.61	0.
6	7eaa36752bccce478ab5ee651fcc4d68	125	20.83	29.60	71.20	1.60	38.40	13.60	11.20	7.20	...	2.40	0.00	0.00	15.20	1.60	0.
7	8bbb62b38592bcbf869e384f66d64383	155	10.33	20.65	41.29	1.29	25.16	5.81	4.52	3.87	...	0.65	2.58	4.52	3.23	1.94	1.
8	23bb3332905ec10bf36b0e463a38bd5e	186	186.00	18.82	69.35	1.08	38.17	12.37	9.68	7.53	...	0.54	0.54	1.08	5.38	0.00	0.
9	35e30200663adc2848fce13563497eed	168	21.00	20.24	74.40	0.00	48.81	14.29	9.52	4.17	...	2.38	0.00	9.52	24.40	0.00	0.
10	80a59cefb87607b116d8252cc4f1b802	171	19.00	20.47	90.06	0.58	60.23	19.30	12.28	9.36	...	1.17	0.00	0.00	5.26	0.00	0.
11	739c217bf88c96c5149349fafab0eb1c	178	25.43	19.66	75.28	2.25	46.63	8.99	5.06	3.93	...	0.00	0.00	0.56	19.66	0.00	0.
12	740aa055e9dccaee874ab7ec7e499d8e	45	22.50	26.67	84.44	0.00	62.22	22.22	13.33	13.33	...	0.00	0.00	0.00	0.00	4.44	0.
13	6773ab780fbccf3500386dcadfb0d4fe	188	94.00	14.89	87.23	0.00	57.45	16.49	11.70	7.45	...	0.00	0.00	0.00	16.49	0.00	0.
14	353022f5d05661c517f5829c96909b8a	175	35.00	20.00	78.29	0.00	46.86	16.00	12.57	9.71	...	0.57	0.57	0.57	6.29	2.29	0.

LIWC features

```
liwc.columns
```

```
Index([u'userId', u'WC', u'WPS', u'Sixltr', u'Dic', u'Numerals', u'funct',
       u'pronoun', u'ppron', u'i', u'we', u'you', u'shehe', u'they', u'ipron',
       u'article', u'verb', u'auxverb', u'past', u'present', u'future',
       u'adverb', u'preps', u'conj', u'negate', u'quant', u'number', u'swear',
       u'social', u'family', u'friend', u'humans', u'affect', u'posemo',
       u'negemo', u'anx', u'anger', u'sad', u'cogmech', u'insight', u'cause',
       u'discrep', u'tentat', u'certain', u'inhib', u'incl', u'excl',
       u'percept', u'see', u'hear', u'feel', u'bio', u'body', u'health',
       u'sexual', u'ingest', u'relativ', u'motion', u'space', u'time', u'work',
       u'achieve', u'leisure', u'home', u'money', u'relig', u'death',
       u'assent', u'nonfl', u'filler', u'Period', u'Comma', u'Colon', u'SemiC',
       u'QMark', u'Exclam', u'Dash', u'Quote', u'Apostro', u'Parenth',
       u'OtherP', u'AllPct'],
      dtype='object')
```

Full details:

https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

NRC features

- NRC Emotion Lexicon -> Nuances of Emotion

- NRC can be used to identify personality, take a look at this paper: https://pdfs.semanticscholar.org/45d0/660b7bbf60f53be75e4c263bd7c135b66a1d.pdf

How does it look like?

```
nrc
```

	userId	positive	negative	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
0	0000e06e07496624211632e8e264126c	0.578947	0.421053	0.107143	0.107143	0.071429	0.214286	0.142857	0.178571	0.035714	0.142857
1	000235a2ba2f48231b7d24e1f08d7878	0.450000	0.550000	0.156250	0.125000	0.125000	0.093750	0.187500	0.093750	0.000000	0.218750
2	000c4b6e2468f7d528876fd1a6dffd4c	0.617647	0.382353	0.127273	0.163636	0.090909	0.127273	0.163636	0.090909	0.036364	0.200000
3	001187432d2a247562082cd0000dec40	0.730769	0.269231	0.084746	0.186441	0.033898	0.084746	0.254237	0.101695	0.101695	0.152542
4	001494c3b74f124a2e3435fff17f376b	0.875000	0.125000	0.019231	0.134615	0.019231	0.019231	0.384615	0.019231	0.057692	0.346154
5	0016684d01925c6e96ba9895801207c8	0.857143	0.142857	0.055556	0.166667	0.055556	0.111111	0.194444	0.111111	0.083333	0.222222
6	001d4fa530e67f4fa73374472a59ef5d	0.818182	0.181818	0.032258	0.193548	0.000000	0.064516	0.258065	0.032258	0.129032	0.290323
7	00200961a099690696e2854d1ac7f186	0.520000	0.480000	0.135135	0.108108	0.081081	0.162162	0.081081	0.189189	0.081081	0.162162
8	0028c7db14b7a987bac0f3163239eaa9	0.920000	0.080000	0.023256	0.255814	0.023256	0.046512	0.325581	0.023256	0.069767	0.232558
9	002efabe296f426d60b815c26517749a	0.375000	0.625000	0.072727	0.090909	0.090909	0.109091	0.163636	0.236364	0.109091	0.127273
10	002fe3062d7faa7e05b40610bd256a4c	0.740741	0.259259	0.142857	0.228571	0.028571	0.028571	0.171429	0.057143	0.085714	0.257143
11	00318d5dcc34668f68fba3af27e95c4d	0.551724	0.448276	0.097561	0.000000	0.146341	0.146341	0.195122	0.121951	0.024390	0.268293
12	00365f2d35f94b2a814e15194701c702	0.607143	0.392857	0.088235	0.264706	0.147059	0.058824	0.147059	0.147059	0.000000	0.147059
13	0037fce9ff25e2363a94561c7d7c1fd2	0.640000	0.360000	0.035714	0.071429	0.000000	0.107143	0.035714	0.214286	0.107143	0.428571
14	003a80b077f815e91c31e0b216727b50	0.656250	0.343750	0.081081	0.135135	0.108108	0.081081	0.243243	0.108108	0.054054	0.189189
15	003bd172967ca5c2b4f2b3f7ad7f45cf	0.750000	0.250000	0.108108	0.243243	0.000000	0.054054	0.189189	0.108108	0.081081	0.216216

NRC features

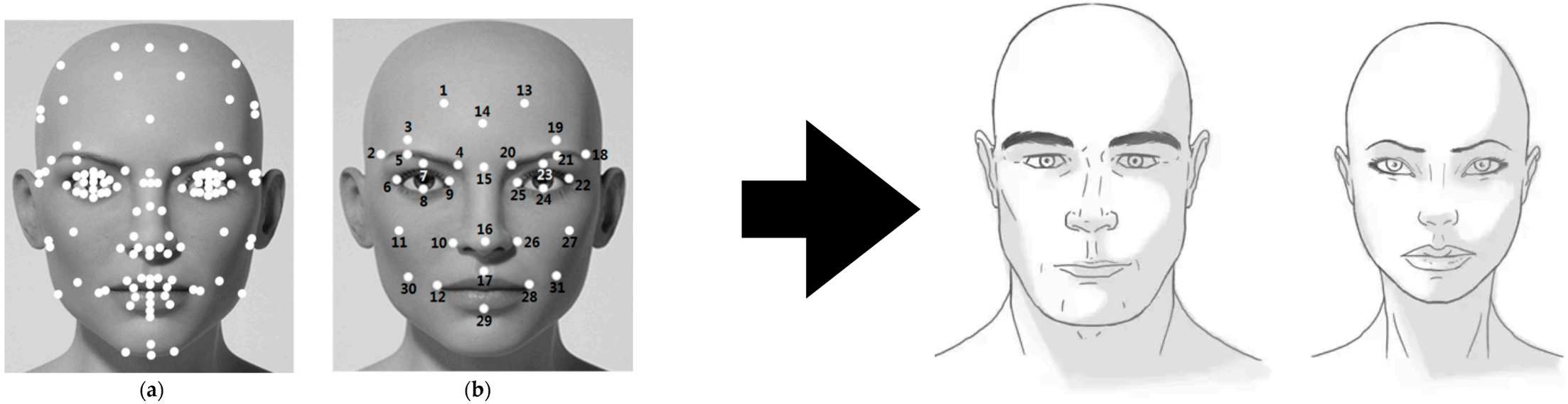
```
nrc.columns
```

```
Index([u'userId', u'positive', u'negative', u'anger', u'anticipation',
       u'disgust', u'fear', u'joy', u'sadness', u'surprise', u'trust'],
      dtype='object')
```

We use a bag of words approach and do not consider any misspellings (e.g., hapy or haaaappy), negation (e.g., not good), strength of the emotions using adjectives or adverbs (e.g., very happy vs. happy) or combined words (e.g., long-awaited vs. long awaited). Moreover, any emotions expressed with words that are not present in the NRC lexicon will remain undetected.

Oxford features

- Facial point features



(a) (b)

- More info: https://azure.microsoft.com/en-ca/services/cognitive-services/face/

How does it look like?

```
oxford
```

	userId	faceID	faceRectangle_width	faceRectangle_height	faceRectangle_left	faceRectangle_top	pupilLeft_x	pupilL
0	0000e06e07496624211632e8e264126c	f7e072db-8532-4686-9074-27e83fee5e94	61	61	15	49	32.7	
1	000235a2ba2f48231b7d24e1f08d7878	934b5179-acec-4dea-a348-feae87767c2d	83	83	91	95	114.4	
2	000235a2ba2f48231b7d24e1f08d7878	118c1f96-b32a-4021-a993-8f60e9859517	76	76	22	50	49.6	
3	000c4b6e2468f7d528876fd1a6dffd4c	1eb367c8-9467-411f-9689-fd1affa95654	121	121	10	21	45.8	

You may have an image without a face
You may have an image with multiple faces

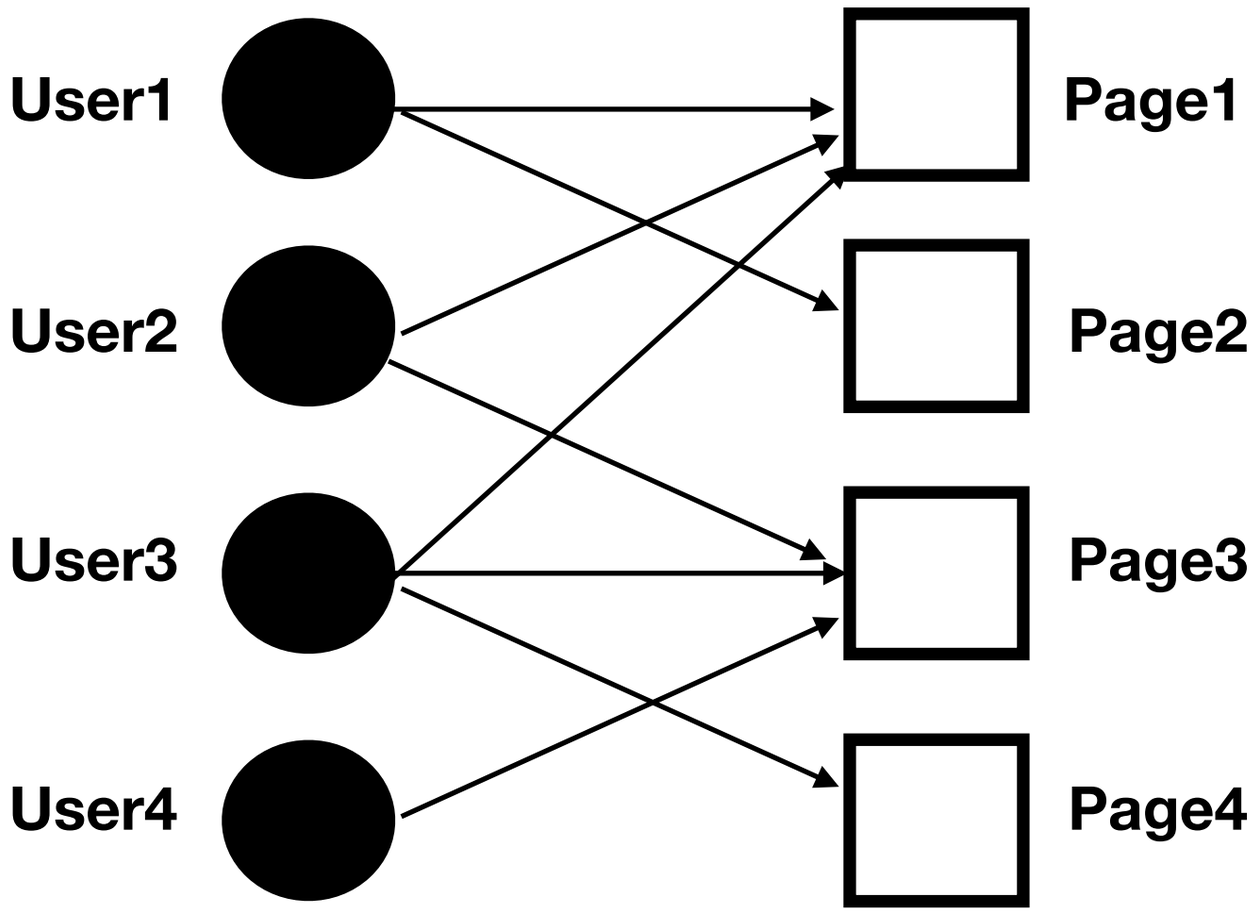
Oxford features

```
oxford.columns
```

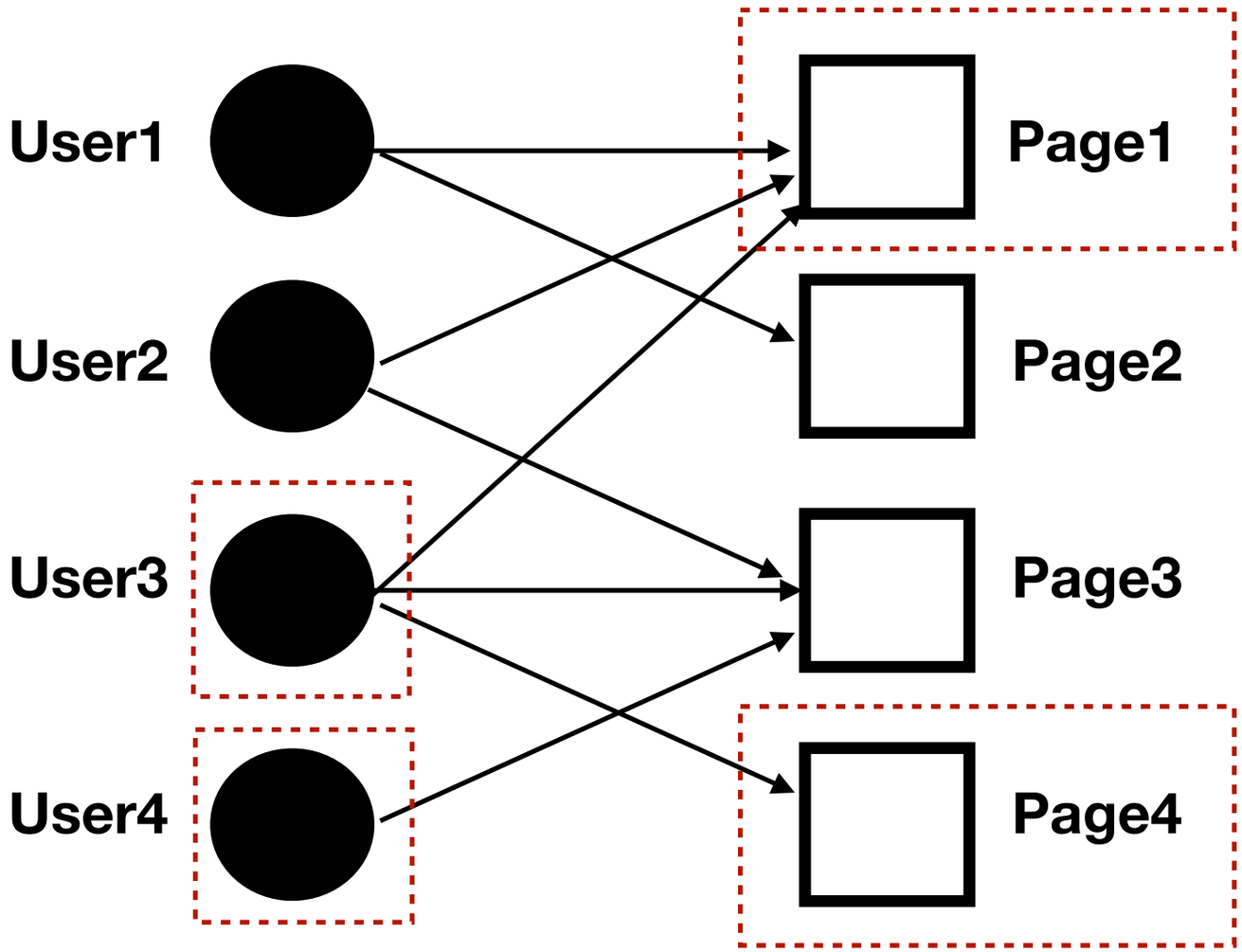
```
Index([u'userId', u'faceID', u'faceRectangle_width', u'faceRectangle_height',
       u'faceRectangle_left', u'faceRectangle_top', u'pupilLeft_x',
       u'pupilLeft_y', u'pupilRight_x', u'pupilRight_y', u'noseTip_x',
       u'noseTip_y', u'mouthLeft_x', u'mouthLeft_y', u'mouthRight_x',
       u'mouthRight_y', u'eyebrowLeftOuter_x', u'eyebrowLeftOuter_y',
       u'eyebrowLeftInner_x', u'eyebrowLeftInner_y', u'eyeLeftOuter_x',
       u'eyeLeftOuter_y', u'eyeLeftTop_x', u'eyeLeftTop_y', u'eyeLeftBottom_x',
       u'eyeLeftBottom_y', u'eyeLeftInner_x', u'eyeLeftInner_y',
       u'eyebrowRightInner_x', u'eyebrowRightInner_y', u'eyebrowRightOuter_x',
       u'eyebrowRightOuter_y', u'eyeRightInner_x', u'eyeRightInner_y',
       u'eyeRightTop_x', u'eyeRightTop_y', u'eyeRightBottom_x',
       u'eyeRightBottom_y', u'eyeRightOuter_x', u'eyeRightOuter_y',
       u'noseRootLeft_x', u'noseRootLeft_y', u'noseRootRight_x',
       u'noseRootRight_y', u'noseLeftAlarTop_x', u'noseLeftAlarTop_y',
       u'noseRightAlarTop_x', u'noseRightAlarTop_y', u'noseLeftAlarOutTip_x',
       u'noseLeftAlarOutTip_y', u'noseRightAlarOutTip_x',
       u'noseRightAlarOutTip_y', u'upperLipTop_x', u'upperLipTop_y',
       u'upperLipBottom_x', u'upperLipBottom_y', u'underLipTop_x',
       u'underLipTop_y', u'underLipBottom_x', u'underLipBottom_y',
       u'facialHair_mustache', u'facialHair_beard', u'facialHair_sideburns',
       u'headPose_roll', u'headPose_yaw', u'headPose_pitch'],
      dtype='object')
```

How to use page likes?

Page likes



Page likes

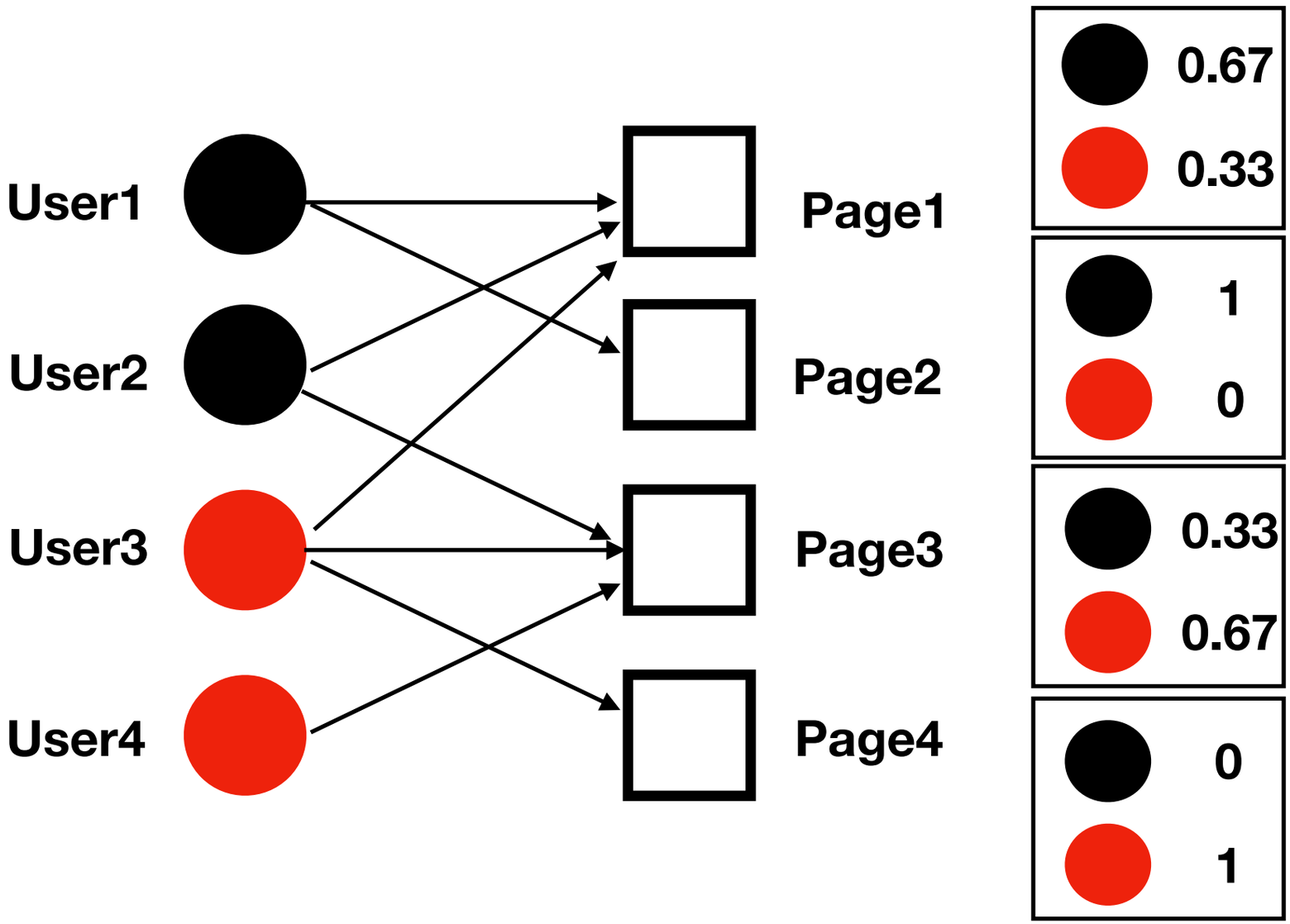


Matrix based

	Page1	Page2	Page3	Page4
User1	1	1	0	0
User2	1	0	1	0
User3	1	0	1	1
User4	0	0	1	0

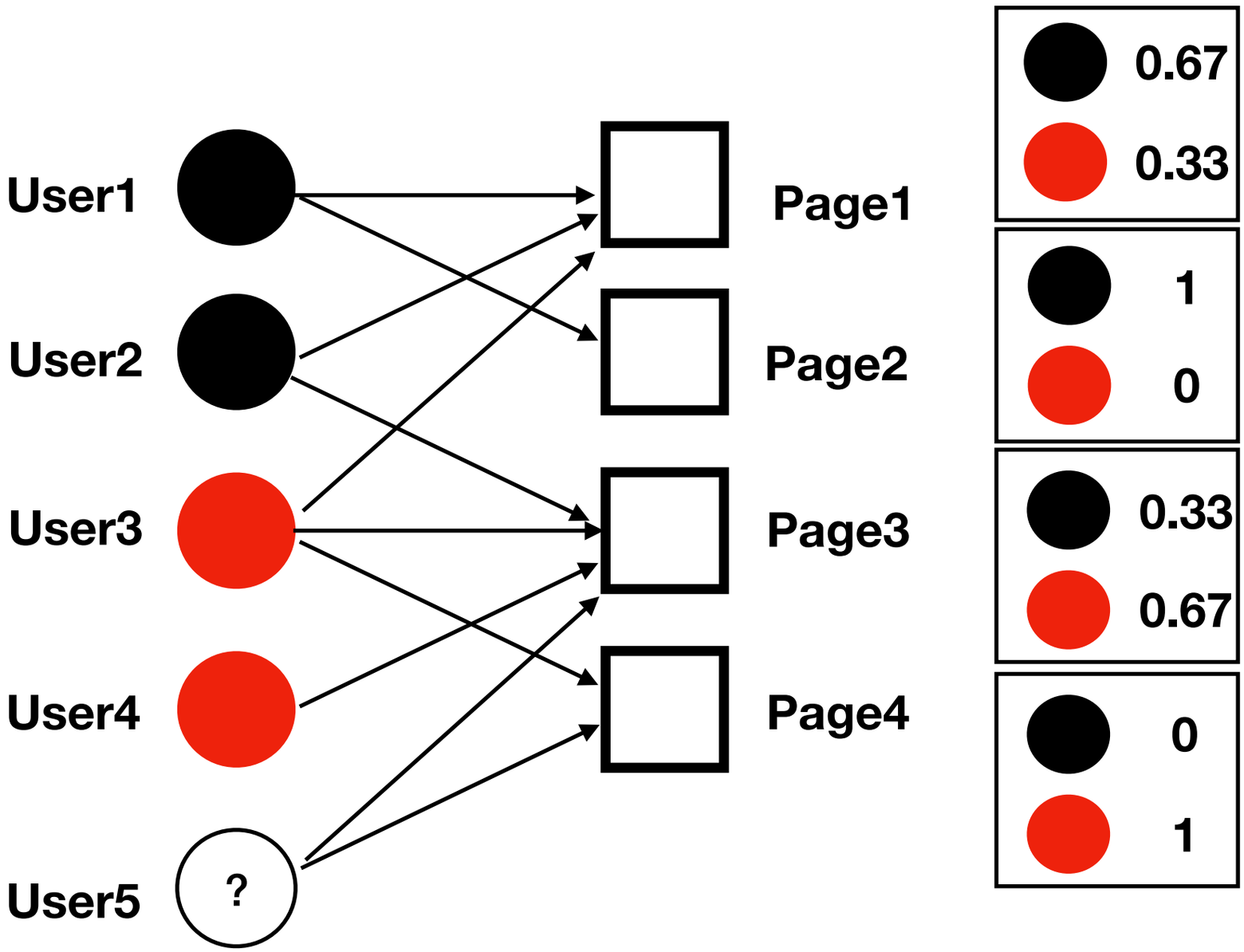
User-Page-Profiling

1- page profiling



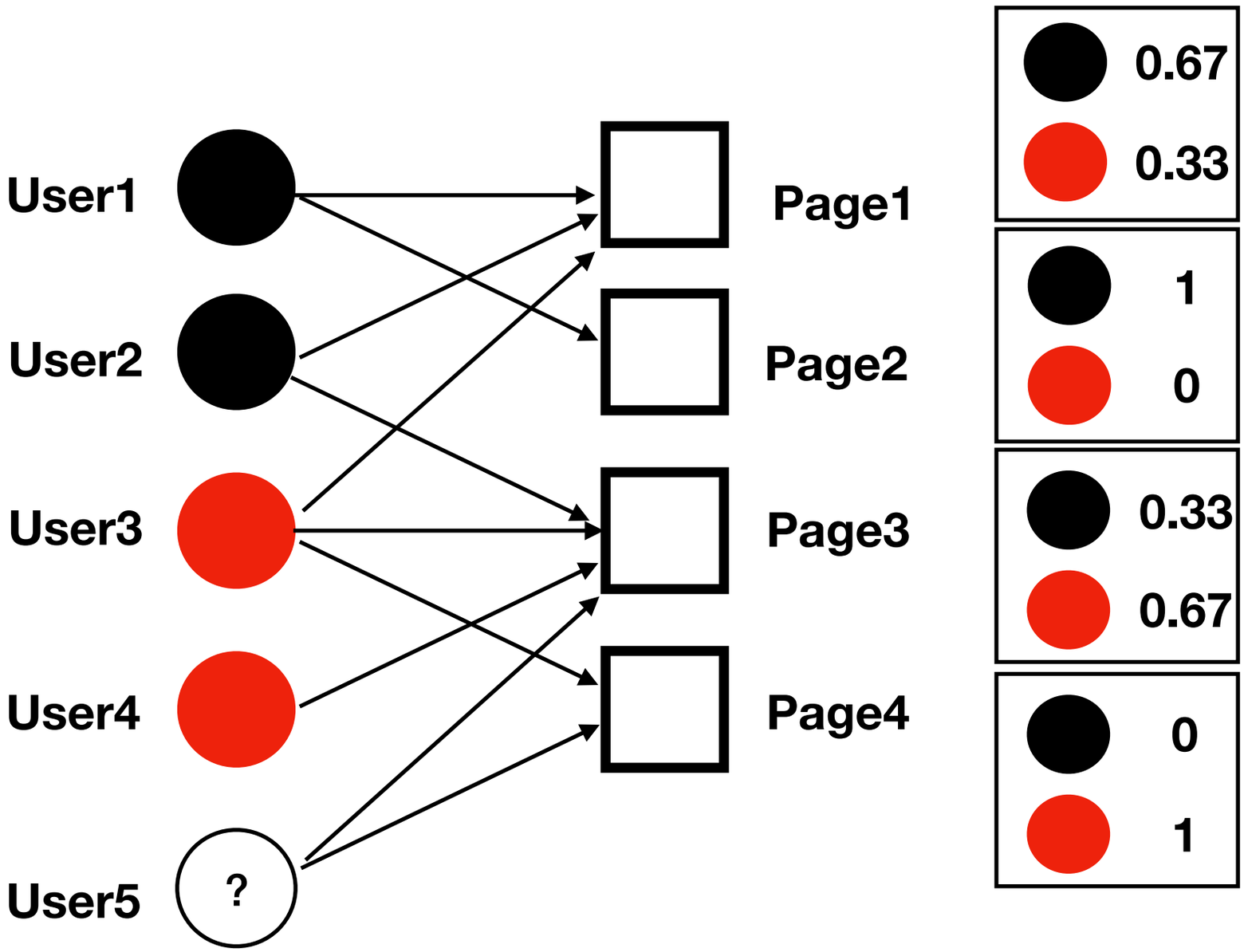
User-Page-Profiling

2- user profiling



User-Page-Profiling

2- user profiling



Next steps?

- Try feature selection (**Week 6**)
- Try dimension reduction techniques e.g., PCA (**Week 3**)
- Use different classification/regression techniques (**Week 3**)
- Measure the performance of your model on the train data, i.e., try model selection, cross validation (**Week 4**)
- Measure the importance of each page/feature (**Week 6**)
- Combine features, use clustering (**Week 3**), ….
- …
- Do you want to know more about feature engineering in graphs? Graph mining and recommender systems on **Week 12**