Algorithms and Data Structures

Matrix Approximation Topic Extraction



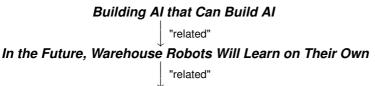
florence machines advertisement one people new company warehouse bins scott tolarge of jobs many one merchandise in 20 march one people arm now company warehouse bins scott tolarge of jobs many one merchandise in 20 march

Learning goals

Topic extraction

APPLICATION: TOPIC EXTRACTION

Often online articles refer to articles with similar content, e.g.



As Amazon Pushes Forward With Robots, Workers Find New Roles

The first two articles should definitely have one topic in common, just like the last two articles. We want to extract these two topics using a NMF.



APPLICATION: TOPIC EXTRACTION / 2

We set up the corresponding document-term matrix.

##		do c1	doc2	doc3
##	accelerate	1	0	0
##	accelerating	1	0	0
##	accurately	1	0	0
##	across	1	1	2
##	address	1	1	0
##	adiust	1	0	0



APPLICATION: TOPIC EXTRACTION / 3

We "search" two topics linking the articles.

```
set.seed(1)
res = nmf(tdm, 2, "Frobenius")
```

##		topicl	topic2
##	accelerate	0.0016	3.8e-11
##	accelerating	0.0016	3.8e-11
##	accurately	0.0016	3.8e-11
##	across	0.0023	4.1e-03
##	address	0.0026	6.3e-05
##	adiust	0.0016	3.8e-11



44

TOPIC EXTRACTION

For both topics, we print the 30 words with the largest values in the columns of matrix \mathbf{W} . The size of the word in the wordcloud is determined by the value of w_{ij} (placement of the word is completely random).



Duild algorithms algorithms algorithmdigital artificial companies one learn research experts eventually symphysical many project on two research ersearchers

florence machines advertisement one people new company warehouse being began around began around

(a) Topic 1

(b) Topic 2

TOPIC EXTRACTION: COEFFICIENT MATRIX H

H

##		topic 1	topic 2
##	doc1	4.5e+02	1.8e-09
##	doc2	2.9e+02	5.8e+01
##	doc3	1.6e-09	4.9e+02



The coefficient matrix shows: The first article clearly refers to the first extracted topic, article 3 clearly to the last. Article 2 addresses both topics.

Implementation in R: https://rpubs.com/JanpuHou/300168