Data Mining
Lecture 12: Finding Independent Features

Jo Grundy

ECS Southampton

22^{nd} March 2022

Finding Features - Introduction

LSA found *concepts* that were linear mixtures of words associated in different documents.

- Weightings were unconstrained, could be negative
- Difficult to interpret, couldn't give *meaning* to the concept
- Assumes each document will have one topic only

Want to find different *themes* or **topics** for a corpus.

Finding Features - Topic Modelling

Uncover hidden thematic structure in a collection of documents
Helps with

- Searching
- Browsing
- Summarising

A single document will often have more than one topic

Finding Features - Introduction

There are a number of ways to do 'Topic Modelling'
Using probabilistic models:

- Probabilistic LSA
- Latent Dirichlet Allocation (LDA) (covered in AML)
- Pachinko Allocation (PAM)

Finding Features - NMF

Topic modelling is like clustering as we group documents into similar sets

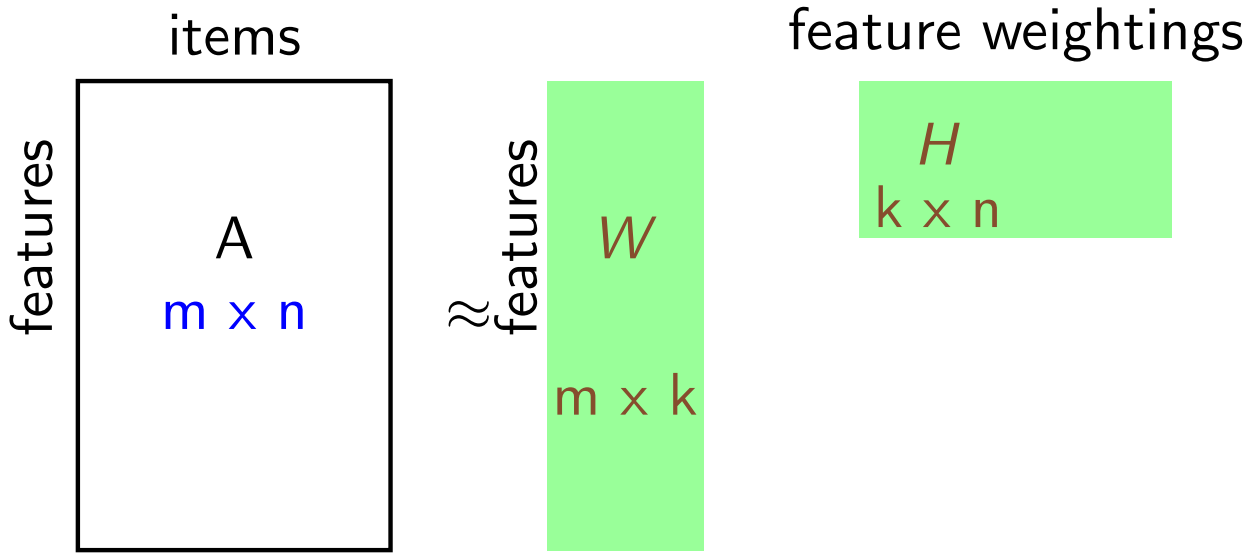
However we want *soft* clusters
a document should be a weighted mixture of topics

Non-Negative Matrix Factorisation achieves this via a different matrix decomposition

$$A \approx WH$$

With PCA and vector quantisation.

Finding Features - NMF



W contains a k dimensional feature vector for each of the items

The weightings for each document are in the columns of H

Each basis vector (row of W) can be interpreted as a cluster. Membership of each cluster is encoded in H

Finding Features - NMF Algorithm

W and H are found by an iterative Expectation Maximisation process
A cost function is minimised:

- Euclidean Norm $||M - WH||^2$
- KL divergence $\sum_{ij}(-M_{ij}\log(WH)_{ij} + (WH)_{ij})$

Lee and Seung 1999, Nature

Finding Features - NMF Algorithm

Algorithm 1: NMF Algorithm with Euclidean distance

Data: A ($m \times n$ non-negative matrix),d dimensions to use
Initialise W with $m \times k$ random values;
Initialise H with $d \times n$ random values;
while *not converged* **do**

$$W_{ij} = W_{ij} \frac{(AH^T)_{ij}}{(WHH^T)_{ij}};$$

$$W_{ij} = \frac{W_{ij}}{\sum_k W_{ik}};$$

$$H_{ij} = H_{ij} \frac{W^T A_{ij}}{(WHH^T)_{ij}};$$

end

This has the effect of minimising the norm $||V - WH||_F^2$ subject to $W \geq 0$, $H \geq 0$

Finding Features - NMF Algorithm

Algorithm 2: NMF Algorithm with KL-Divergence

Data: A ($m \times n$ non-negative matrix), d dimensions to use

Initialise W with $m \times d$ random values;

Initialise H with $d \times n$ random values;

while *not converged* **do**

$$W_{ij} = W_{ij} \sum_k \frac{A_{ik}}{(WH)_{ik}} H_{jk};$$

$$W_{ij} = \frac{W_{ij}}{\sum_k W_{ik}};$$

$$H_{ij} = H_{ij} \sum_k W_{ki} \frac{A_{kj}}{(WH)_{kj}};$$

end

This has the effect of minimising the generalized KL Divergence $\sum_{ij}(-M_{ij} \log(WH)_{ij} + (WH)_{ij})$ subject to $W \geq 0$, $H \geq 0$

Finding Features - NMF Algorithm

Initialisation is usually random.
Different random initialisations can lead to **instability**
i.e. different results for different runs with the same data and d value.
Improvement was reported using SVD initialisation (Boutsidis and Gallopoulos 2008)
Where:

- W is initialised as $U_d\sqrt{\Sigma_d}$
- H is initialised as $\sqrt{\Sigma_d}V_d^T$

However, a further study reported that random initialisation was better (Utsumi 2010)

Finding Features - NMF Algorithm

Other variants can involve:

- For distance: Use of Bregman divergence (Li *et al* 2012)
- For optimisation: alternating least squares with projected gradient method for sub-problems (Lin 2007)
- For constraints:
  - Enforcing Sparseness (Hoyer 2004)
  - Using background information (Semi-NMF)
- Inputs: Symmetric matrices, e.g. Document - Documents cosine similarity matrix (Ding & He, 2005)

Finding Features - NMF Algorithm

Example:
a set of strings:

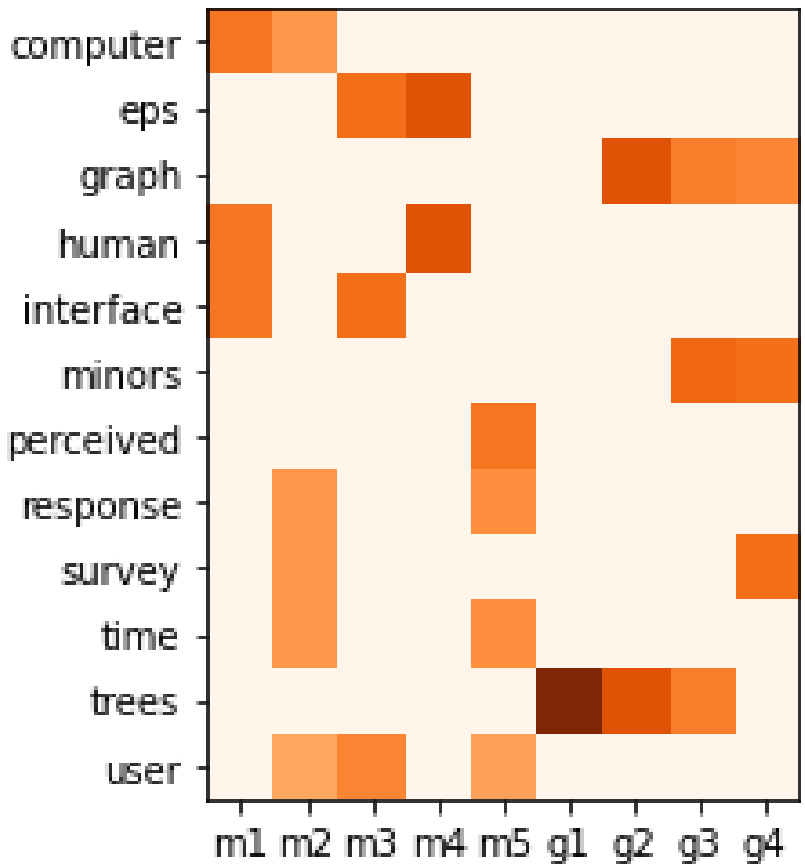
m1	"Human machine interface for ABC computer applications"
m2	"A survey of user opinion of computer system response time"
m3	"The EPS user interface management system"
m4	"System and human system engineering testing of EPS"
m5	"Relation of user perceived response time to error measurement"
g1	"The generation of random, binary, ordered trees"
g2	"The intersection graph of paths in trees"
g3	"Graph minors IV: Widths of trees and well-quasi-ordering"
g4	"Graph minors: A survey"

`http://lsa.colorado.edu/papers/dp1.LSAintro.pdf`

Finding Features - NMF Algorithm

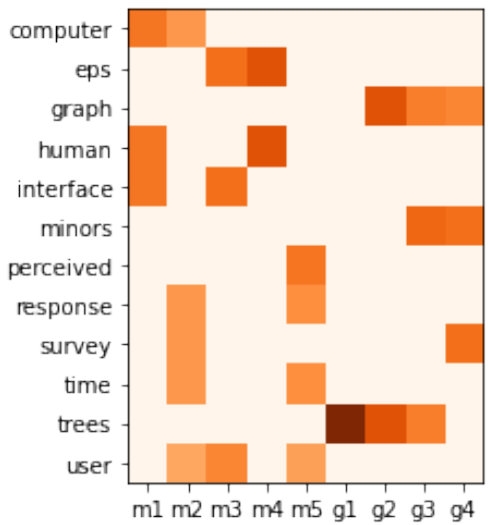
calculate TF.IDF

0.58	0.46	0.	0.	0.	0.	0.	0.	0.
0.	0.	0.6	0.71	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.71	0.55	0.52
0.58	0.	0.	0.71	0.	0.	0.	0.	0.
0.58	0.	0.6	0.	0.	0.	0.	0.	0.
0.	0.	0.	0.	0.	0.	0.	0.63	0.6
0.	0.	0.	0.	0.58	0.	0.	0.	0.
0.	0.46	0.	0.	0.49	0.	0.	0.	0.
0.	0.46	0.	0.	0.	0.	0.	0.	0.6
0.	0.46	0.	0.	0.49	0.	0.	0.	0.
0.	0.	0.	0.	0.	1.	0.71	0.55	0.
0.	0.4	0.52	0.	0.43	0.	0.	0.	0.

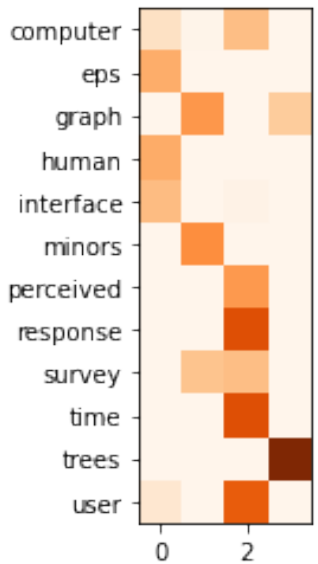


Finding Features - NMF Algorithm

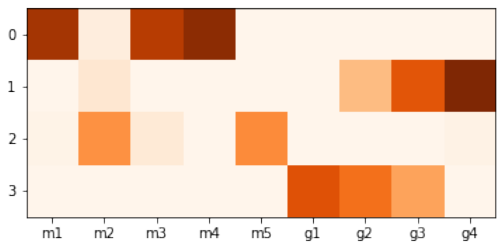
NMF: $A \approx WH$, $d = 4$



A



W



H

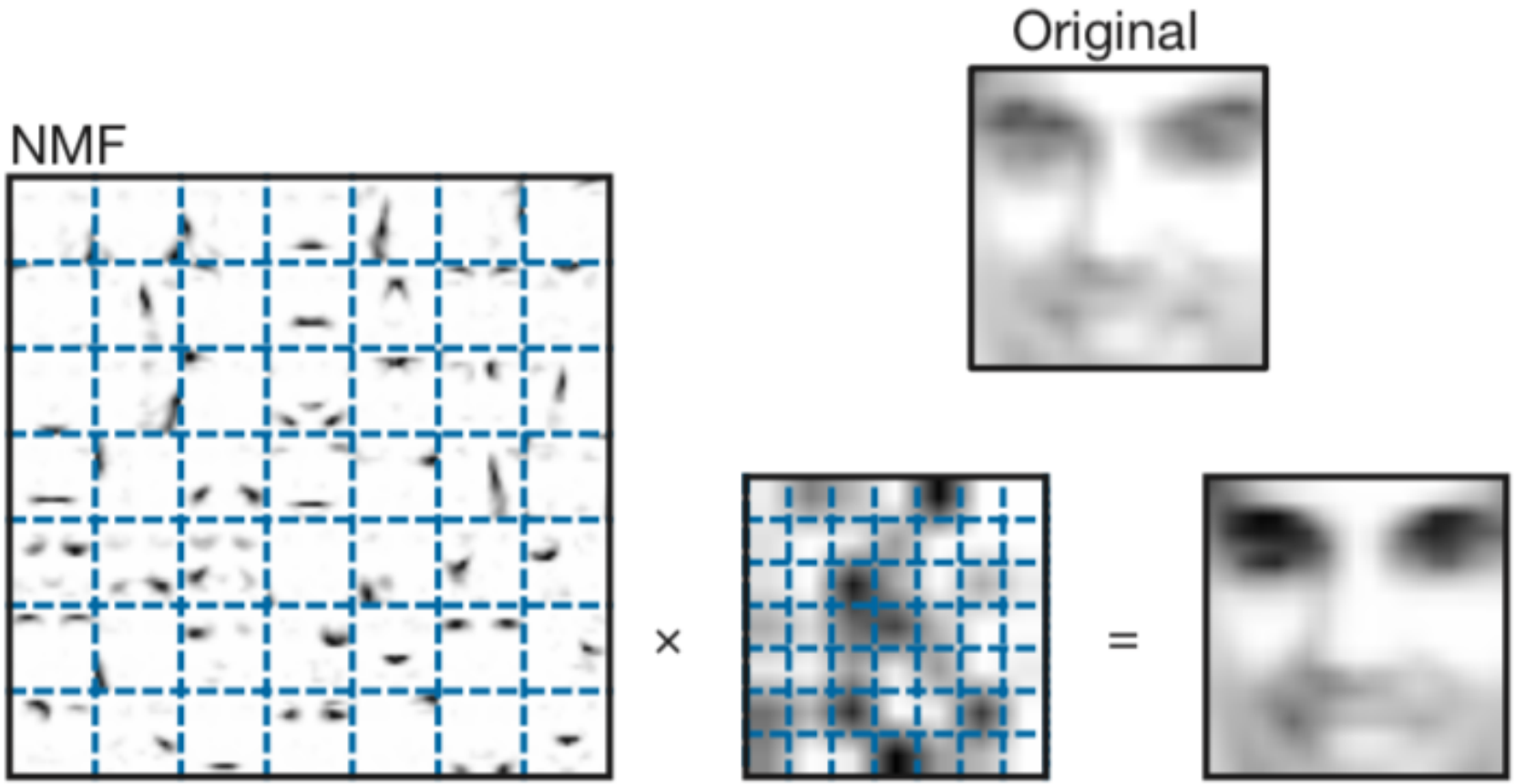
W has the *basis vectors*, showing how the words are clustered
H has the topic memberships for the documents.

Finding Features - NMF Algorithm

For this method, and for LSA, the size of the reduced dimensionality is chosen manually
Can chose based on the error from reconstruction
though like K means, and K nearest neighbours, this will be lower for higher values of d
Can run many times and build up a *consensus matrix*
Can also examine the *stability* of multiple random initialised runs for each value of d

Finding Features - NMF Examples

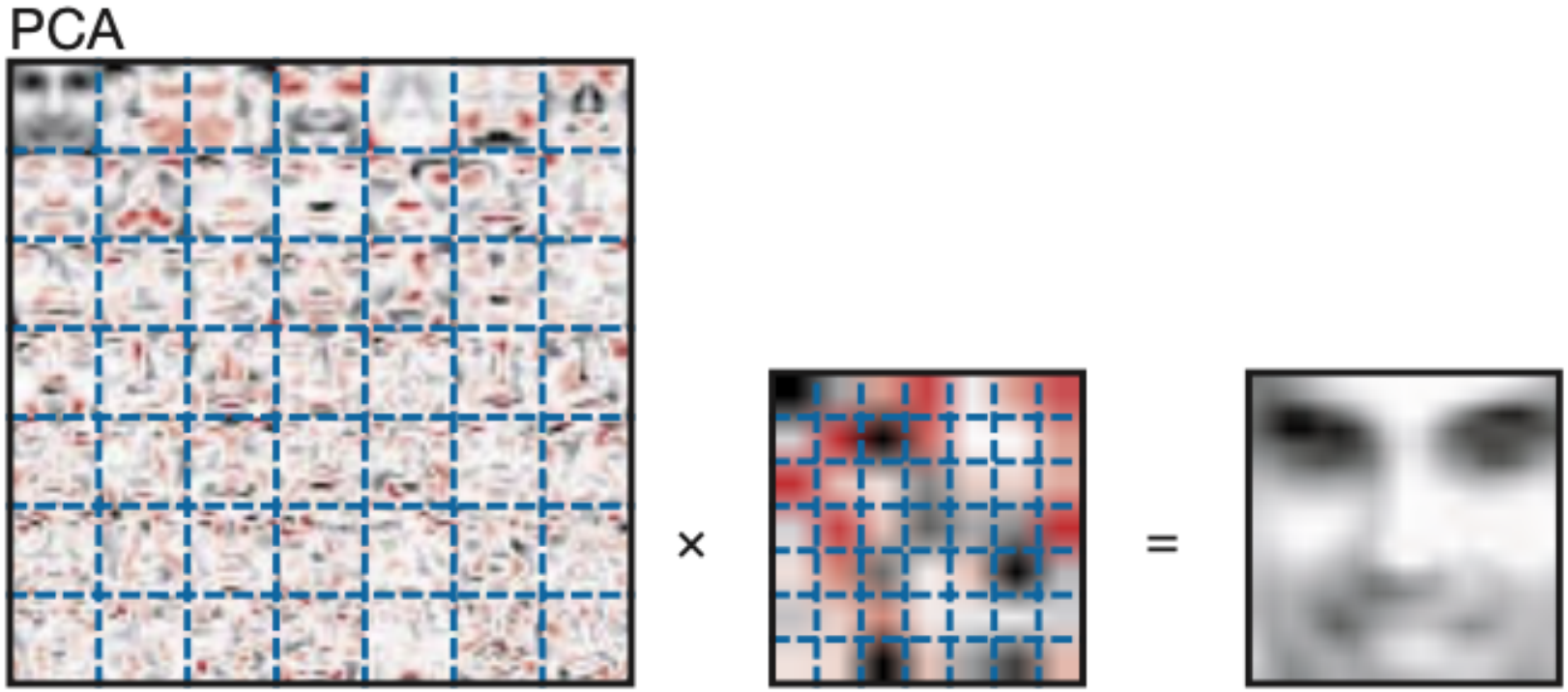
On a database of facial images, NMF constructed a decomposition of those faces in to parts, that H contained the weights to reconstruct every face from.



The face can be built up using a selection of the mouths, noses and eyes in the representation

Finding Features - NMF Examples

Constraining the values to be non negative forces the representation to be sparse, as they must all be additive.



if PCA is used then the representation does not decompose the data

Finding Features - NMF Examples

Using a collection of 2,225 BBC news articles with 5 manually annotated topics

So we know d

Top ranked terms for each topic:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
growth	mobile	england	film	labour
economy	phone	game	best	election
year	music	win	awards	blair
bank	technology	wales	award	brown
sales	people	cup	actor	party
economic	digital	ireland	oscar	government
oil	users	team	festival	howard
market	broadband	play	films	minister
prices	net	match	actress	tax
china	software	rugby	won	chancellor

from `http://derekgreene.com/slides/nmf_insight_workshop.pdf`

Finding Features - NMF Examples

21,000 news articles relating to the Irish economy.
Constructed matrix using *named entities*, $d = 8$
Top ranked terms for each topic:

Topic 1	Topic 2	Topic 3	Topic 4
nama	european_union	allied_irish_bank	hse
brian_lenihan	europe	bank_of_ireland	dublin
green_party	greece	anglo_irish_bank	mary_harney
ntma	lisbon_treaty	dublin	department_of_health
anglo_irish_bank	ecb	irish_life_permanent	brendan_drumm

Topic 5	Topic 6	Topic 7	Topic 8
usa	aer_lingus	uk	brian_cowen
asia	ryanair	dublin	fine_gael
new_york	dublin	northern_ireland	fianna_fail
federal_reserve	daa	bank_of_england	green_party
china	christoph_mueller	london	brian_lenihan

from `http://derekgreene.com/slides/nmf_insight_workshop.pdf`

Finding Features - NMF Examples

IMDB Keyword set for 21,000 films, $d = 10$
Top ranked terms for each topic:

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cowboy	bmovie	martialarts	police	superhero
shootout	atgunpoint	combat	detective	basedoncomic
cowboyhat	bwestern	hero	murder	superheroine
cowboyboots	stockfootage	actionhero	investigation	dccomics
horse	gangmember	brawl	policedetective	secretidentity
revolver	duplicity	fistfight	detectiveseries	amazon
sixshotter	gangleader	disarming	murderer	culttv
outlaw	deception	warrior	policeofficer	actionheroine
rifle	sheriff	kungfu	policeman	twowordtitle
winchester	povertyrow	onemanarmy	crime	bracelet

from `http://derekgreene.com/slides/nmf_insight_workshop.pdf`

Finding Features - NMF Examples

IMDB Keyword set for 21,000 films, $d = 10$
Top ranked terms for each topic:

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
worldwartwo	monster	love	newyorkcity	shotinthechest
soldier	alien	friend	manhattan	shottodeath
battle	cultfilm	kiss	nightclub	shotinthehead
army	supernatural	adultery	marriageproposal	punchedintheface
1940s	scientist	infidelity	jealousy	corpse
nazi	surpriseending	restaurant	engagement	shotintheback
military	demon	extramaritalaffair	party	shotgun
combat	occult	photograph	hotel	shotintheforehead
warviolence	possession	tears	deception	shotintheleg
explosion	slasher	pregnancy	romanticrivalry	shootout

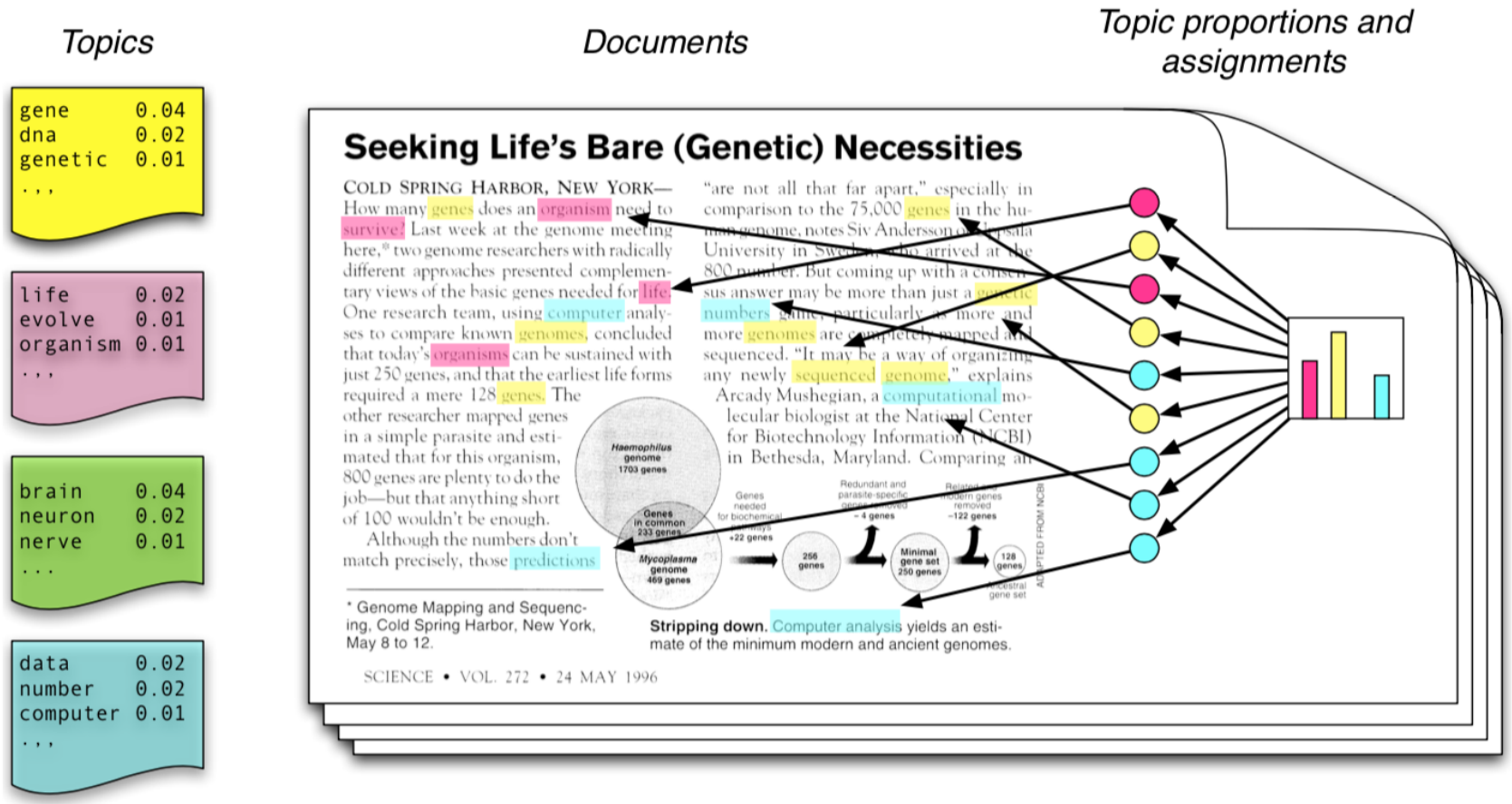
from `http://derekgreene.com/slides/nmf_insight_workshop.pdf`

Finding Features - Probabilistic Models

We model a **document** as a mixture of topics

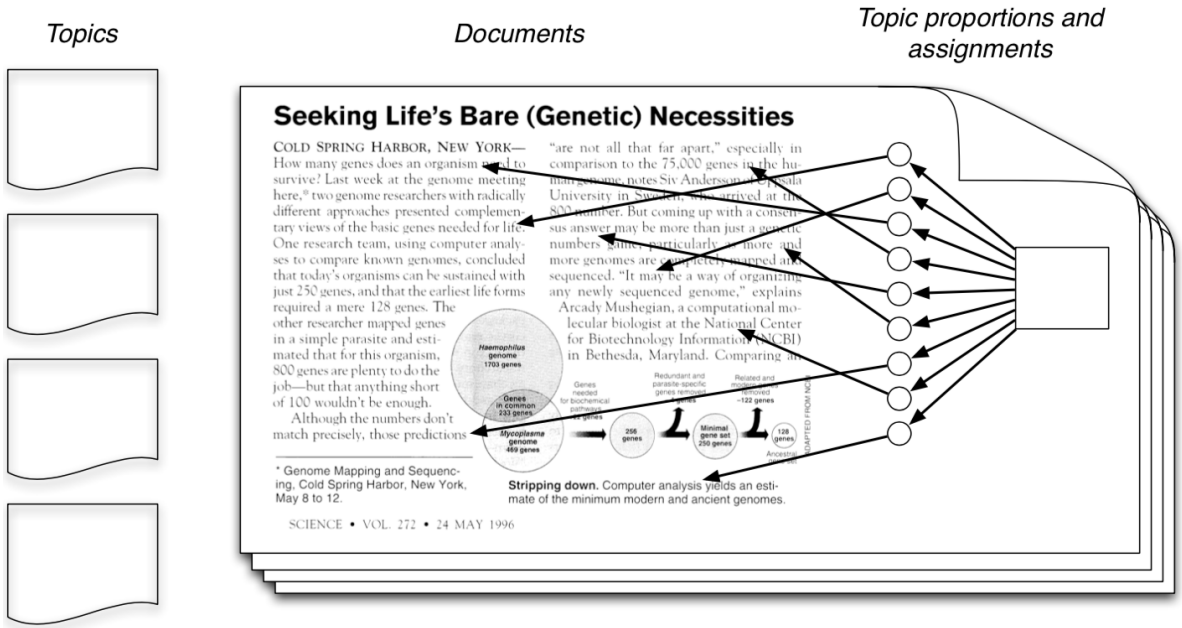
A **topic** is a distribution over words

Each **word** in the document is drawn from a topic



Finding Features - Probabilistic Models

In reality, only the document is visible
Topic distributions and assignments are *hidden*



We need to *infer* the hidden variables: i.e. compute the distribution conditioned on the documents
$p(\mathit{topics}, \mathit{props}, \mathit{assignments} | \mathit{documents})$

Finding Features - PLSA

Probabilistic Latent Semantic Analysis

- given a corpus
- observations are pairs of words and documents (w, d)
- each observation is associated with latent class variable c

Assumes probability of a co-occurrence of a word and document $P(w, d)$ is a mixture of conditionally independent multinomial distributions

$$P(w, d) = \sum_c P(c)P(d|c)P(w|c) = P(d) \sum_c P(c|d)P(w|c)$$

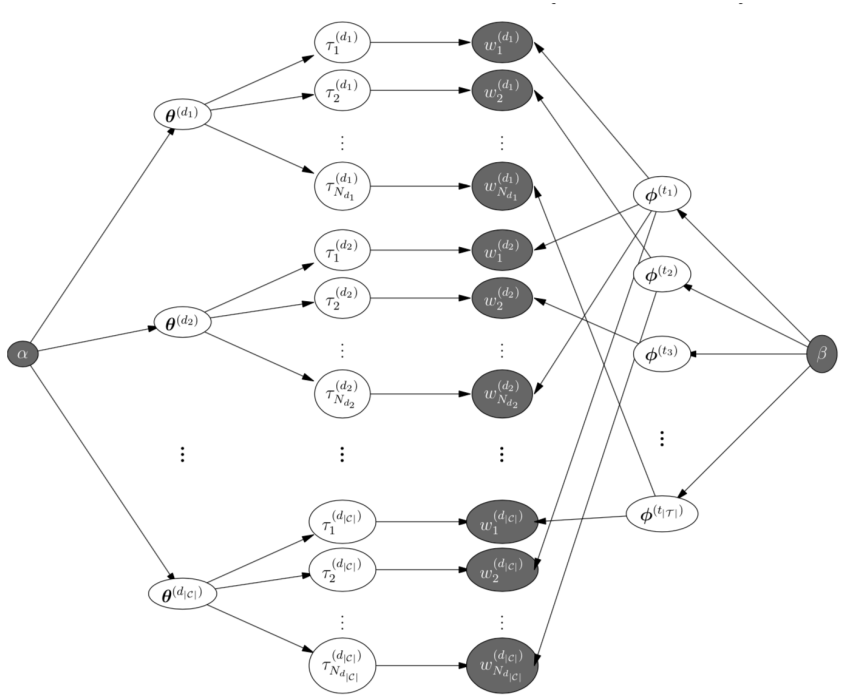
Mathematically equivalent to NMF with KL divergence, hybrid methods more successful (Ding, Li, Peng, 2008)

Finding Features - LDA

Latent Dirichlet Allocation (LDA) - also in AML
Bayesian Extension to PLSA

- Uses a Dirichlet prior on the topic distribution per document
- Fully Generative
- Baysian Inference to learn parameters
- Better than PLSA for small datasets, otherwise similar

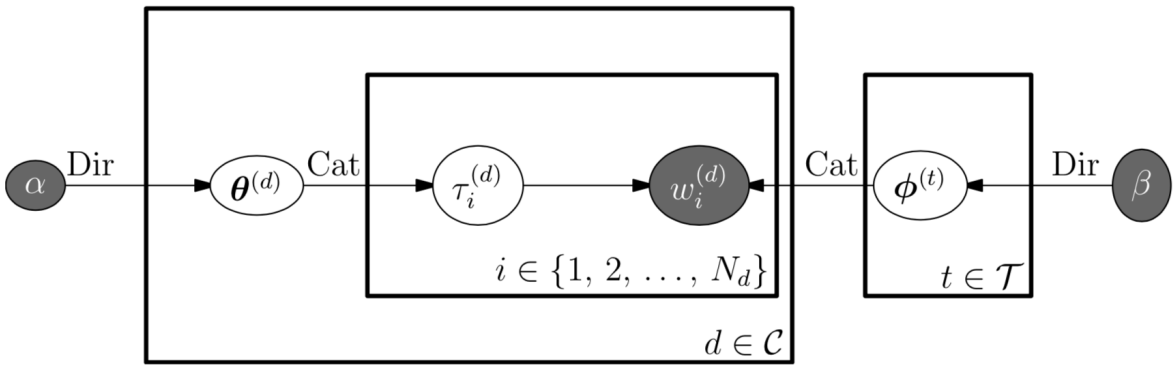
Finding Features - LDA



Where

- Dirichlet distribution parameters:
  - α for the document distributions
  - β for the topic distributions
- θ is the document topic proportions
- τ is the topic distribution over words

Finding Features - LDA



Condenses down to a plate diagram

When the parameters have been inferred, they can be used to generate new documents

Finding Features - LDA Examples

Science articles, 17,000 documents, stop words and rare words removed

100 topic LDA using variational inference

`http://www.cs.columbia.edu/~blei/talks/Blei_MLSS_2012.pdf`

Finding Features - Summary

Topic modelling is an important part of data mining unstructured data

Key Ideas:

- Items are made up from topics
- A small subset of topics for each item
- One key parameter to tune: number of topics