Vector-Space Modeling in R

Session # 3

Text Analytics for Batch 12

CBA @ ISB

Sudhir Voleti

1

Session Plan

- Factorizing data
  - factor-an PCA primer
  - Factor-an app and worked example
- Factorizing Text: Latent Topic Models
  - topicmodels simulation
  - LDAtuning fit assessment
- LDA in Py with genism
- In-Class exercise on Topic modeling
- Session Wrap-up

2

2

Factorizing Data

3

Quotable Quotes…

- "Firms are swamped with measures... it is commonplace for firms to have 50-60 top level measures, both financial and non-financial... includes 20 financial measures, 22 customer measures, 16 measures of internal process, 19 of renewal and internal development... many firms have struggled unsuccessfully to drive measures of shareholder value from the top to the bottom of the organization."
- (Fred Meyer 2002)
- Above familiar?
- What can be done about this? Let's head towards one solution …

4

The WHY of Factorizing Data

- Based on the example we just saw, it becomes clear that …
- (i) At least in some instances, businesses have more data in variables, metrics, measures etc than is optimal.*
- (ii) If there were some way to *reduce* the size (or dimensionality) of this data, without degrading its information content – such a method would be very useful indeed.
- (iii) If there exist *coherent* groups of variables, which are strongly inter-related, then identifying them and extracting their information content into a *single variable* would help a lot.
  - This single variable would ideally represent (and replace) the entire variable group.
- Enter Factorization of Data.

5

The HOW of Factorizing Data

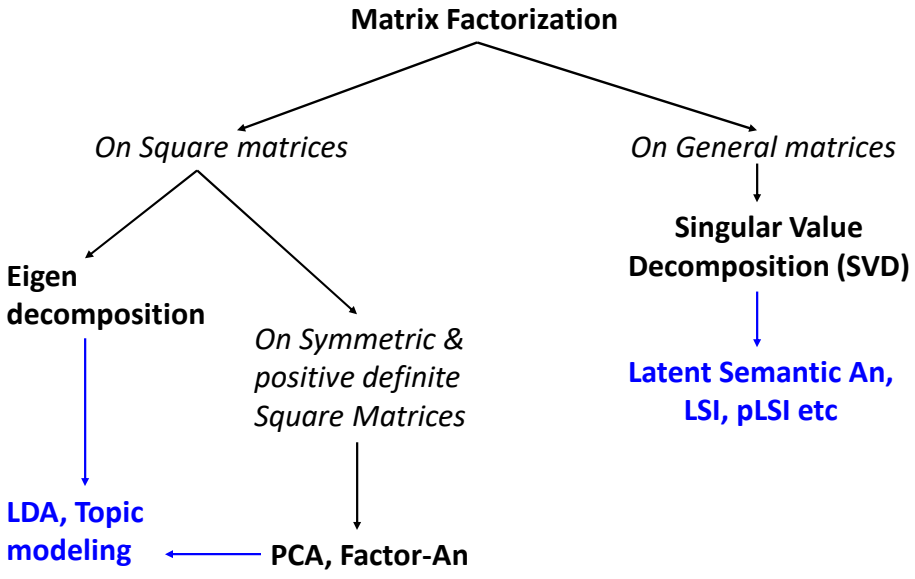
- A 'factor' is, quite literally, the inverse of 'multiple'. E.g. primes and composite numbers.
  - 3 x 4 = 12. 12 is a multiple of 3. Conversely, 3 is a *factor* of 12.
- Recall matrix multiplication: matrices A = i x j and B = j x k when multiplied yield AB = i x k
  - Conversely, AB can be 'factorized' into matrices A and B.
- And we are *very* interested in *factorizing* matrices, because datasets are often matrices.

Open file *'factor-an PCA primer'*

- Matrix Factors are a condensed form of the matrix's info content.
- Next, let's walk through a simple simulated example of matrix factorization & matrix reassembly (or recovery).

6

Matrix Factorization: Landscape Mapping



7

7

Factor-An on Objective data: mtcars

- Run '?mtcars' to overview data for 32 cars on 11 attributes such as:
- Are some of these variables inter-related?
- If so, how many factors or (coherent groups of variables) are there?
- Why not use the factor-An app and find out?

8

Factor-An on Objective data: mtcars

- The Factor-An app output shows us:
- [1] The correlation structure among the variables. (Do they seem sensible?)
- [2] The Factor loadings table with 2 factors.
- Can you interpret and label the factors?
- [3] A factor scores table for which cars score how much on each factor.
- [4] Uniqueness table: Which variables are most unique (i.e., independent of the factor solution?)

9

Factor-An Recap

- What is factor-An? Why is it used?
- When would we typically consider using factor-An?
- What are the inputs to and outputs from factor-An?
- What are the advantages and limitations of factor-an?
- We're finally ready to apply factor-an on text: topic modeling!

10

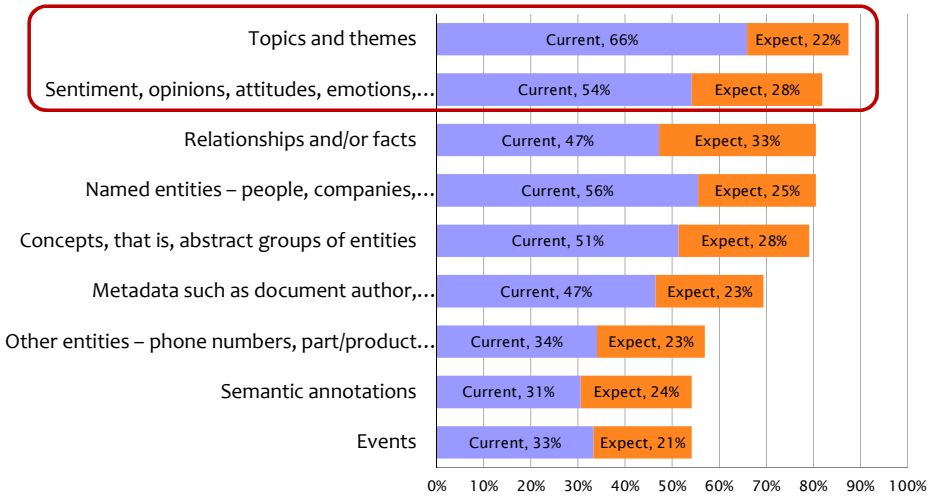
Factorizing Text:

Latent Topic Modeling

11

Recap: What's Industry looking for? …

Do you currently need (or expect to need) to extract or analyze…



**Source: JDAT (Journal of Data Analysis Techniques) 2014 conference, Paris.*

12

Why Topic-Mine Text?

- Recall from Text-An: A lot of data in Organizations is in text form.
- What people want to extract from text is *meaning*.
- One way to view meaning is in the form of coherent, condensed "*topics*" or "*themes*" that underlie a body of text.
- Currently, getting to meaning requires manual analysis of the text corpus.
  - Automated analysis relies on simple models that do not directly address themes or topics.
- This leads us to the central point of this module– how to mine directly for these latent topics or themes in text...

13

What happens when we Factorize a DTM?

- Recall 'regular' Factor-An. A dataset with R rows and C columns we factorized into two components – RxF scores matrix and a FxC loadings matrix.
- Further, we labeled & interpreted these factors in terms of the composite combinations of variables that loaded on them.
- What if our dataset was a DTM with D documents and T terms? And instead of conventional Factors, we had 'Topic Factors'?
- We'd get a DxF scores matrix of documents-on-Factors, and a FxT loadings matrix of terms-on-factors.
- What are these matrices? What do these scores and loadings mean in a text context? Let's find out …

14

Introducing Topic Models

- Consider a text corpus of 20 product reviews.
- Suppose there are **two** broad themes or topics in the structure of the corpus –"PRICE" and "BRAND".
- Further suppose that each document is a *mixture* of these two topics in different proportions.
- So a document that talks about Price & Brand 90% and 10% of the time respectively, should have 9 times more Price Terms than Brand Terms.
  - A topic model formalizes this intuition mathematically.
- The machine will tell you which tokens belong to which topics, and which documents load on which topics.
  - Then, we can sort, order, plot, analyze etc. the tokens and the documents…

15

15

A Topic Model Simulation

- Open file '*topicmodel simulation*', which presents a simple simulation to demonstrate the intuition behind how topic models work.
- Answer these Qs based on the simulation we just did.
- Q: How did we setup and go about the simulation?
- Q: What take-aways did you see from the simulation?
- Q: What are the challenges that would arise if we didn't know *a priori* how many and what the topics are?
- Q: How to know how many latent topics are there in a corpus?
- Open file *'LDAtuning for LDA fit metrics.Rmd'* and time permitting we'll cover the same.

16

16

How many topics are optimal?

- In the simulation example, we knew a priori that there *should* be 4 topics. But in most corpora, that won't happen.
- And if we choose a 5 topic solution when 3 (or say, 8) was optimal, we will get misleading results.
- So, how then to know how many latent topics are there in a corpus?
- A variety of metrics have been proposed in the literature. We will see an R library that neatly packages four of them into 1 function.
- Open file *'LDAtuning for LDA fit metrics.Rmd'*

17

17

Recap of LDA fit assessment

- What is the need for LDA fit assessment?
- What methods did we see reg LDA fit assessment?
- Which among the above is preferable. Why?
- What might you like to see in a *topic modeling shiny app*?
- How about an exercise on the same?

18

18

Latent Topic Modeling in Py:

Gensim explorations

19

LDA in Py - gensim

- Py implements LDA - mostly thru either *scikit-learn* or *genism*.
  - It's quite possible you may have to build a workflow around a py implementation of topic modeling.
- Gensim has soared in popularity owing to its speed, scalability and flexibility.
  - Its model output however needs some getting used to.
- Let's see what functionality genism in py has to offer.
- Open file *'genism explorations.ipynb'.*

20

LDA in genism - Recap

- What all ops did genism enable before we got to running LDA?
- Note: We'd done pre-processing and DTM building in py with NLTK previously, genism also provides the same functionality.
- What LDA fit assessment metrics did we see that genism provides?
- What advantages does genism provide over most other LTM modules?

21

Latent Topic Modeling:

A small, real world example.

22

Introducing Topic Models: A Real World Example

- What are Vision statements and Mission statements?
- Does *every* firm have a vision or mission statement?
- Open the 'Vision statements.csv' file and examine a few vision statements.
- Let's first do a basic Text-An on it using the basic-text-an-app of Session 5.
- Then run the Topic-mining-app (lines 42-43) and examine what we get.

23

Vision Statements: Basic Text-An Output

Weights Distribution of Wordcloud

	Weights
corporation	141
company	101
customers	75
solutions	65
world	55
energy	53
business	51
people	48
group	45
global	42
services	42
products-services	40
vision	40
employees	39
international	37
provide	37
work	37
holdings	36
industry	36
products	36
growth	35
innovative	34
financial	32
technology	32
customer	31

24

Vision Statements: Topic Mining App Output

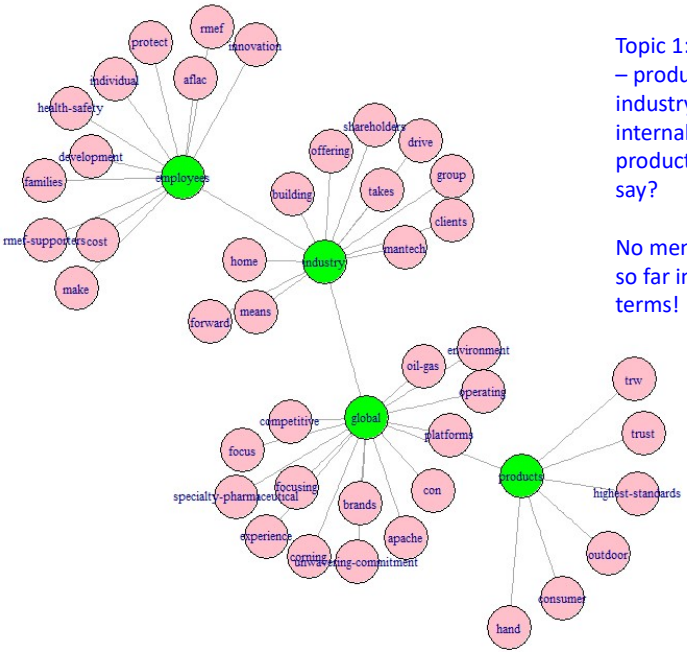
Topic 1

Topic 2

What might these topics mean? What initial guesses can you make just looking at the wordclouds? Let's also see the co-occurrences to sharpen our understanding …

25

Vision Statements: Topic Mining App Output

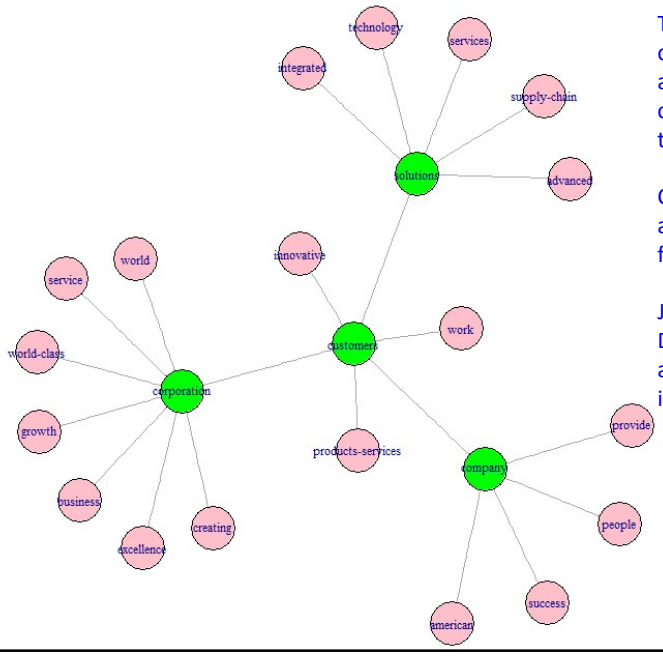


Topic 1: The central nodes – products, employees, industry, global - are all internal-company and product focused, could we say?

No mention of 'customer' so far in the top 50 odd terms!

26

Vision Statements: Topic Mining App Output



Topic 2: The *primary* central node is customer, and it connects the company and solutions together.

Could we say these firms are *customer centric* firms?

Just speculating only. Different folks can (& do) arrive at different interpretations

27

Vision Statements: Topic Mining App Output

Overview | Example dataset | Corpus Word Cloud | Topic Model - Summary

Show 5 entries

Doc.id	Topic 1	Topic 2	Document
6	2.994929	97.00507	FMC Corporation," As a global leader utiliz and cost-effective solutions to food and ag lubricants, structural pest control, turf & orı
7	70.764987	29.23501	The Carlyle Group L.P.," Carlyle professior visdom, knowledge and resourses requirec
8	15.786287	84.21371	"Office Depot, Inc.","Delivering Winning Sc doing what we say we're going to do ? effic heights.
9	3.039275	96.96072	And we don't settle for less than being the and innovative thinking that enable our cus to help people achieve their goals. Our mo passion for life . creating a fuller, more enri
10	28.540937	71.45906	FirstEnergy Corp.," FirstEnergy will be a le for long-term growth, investment value and
Doc.id	Topic 1	Topic 2	Document

Showing 6 to 10 of 884 entries

The last tab reveals the topic proportions in each document.

Thus, The Carlyle Group, a finance firm, is 70% product-focused and 30% customer-focused in its vision statement.

Whereas Office Depot, a stationery supplier, is 16% and 84% on those parameters.

28

Latent Topic Modeling: Recap

- Latent topics are powerful tools that dimension-reduce massive corpora into a handful of coherent, meaningful underlying themes or Topic factors.
  - Latent Topic models generalize across corpus size, #latent topics etc.
- Of course, it needs a good managerial mind with a good managerial question to deploy the model in the right contexts and divine topic interpretations for downstream impact.
- There's always some subjectivity in topic interpretation – analogous to what we have in conventional Factor-An interpretations.
  - But there's a lot more leeway in choosing what the optimal #topics should be. Model fit isn't the only criterion – meaningfulness is a bigger one.
- The R app automates workflows but its convenience comes at a cost – the flexibility to do free-flow analysis with actual data objects is perforce limited.

29

Q & A

30