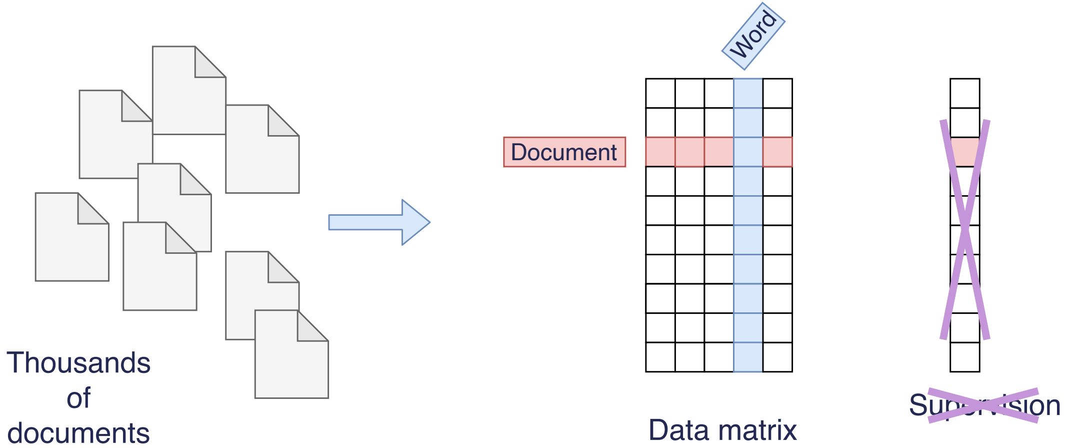
SEMANTIC MODELING & UNSUPERVISED APPROACHES

Agro–IODAA, semestre 1

Vincent Guigue

Introduction

What can we do... Without supervision?

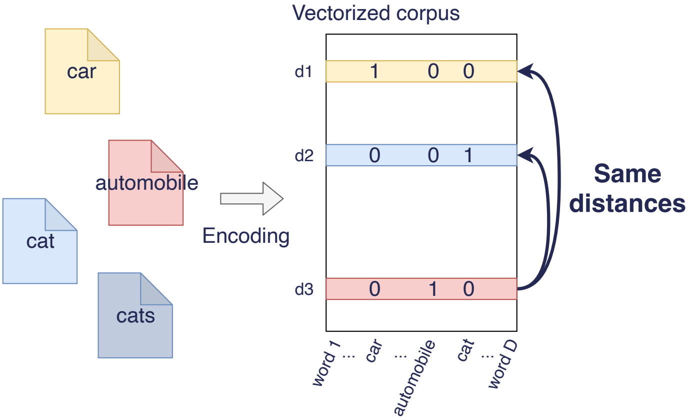


Clustering

Semantic analysis

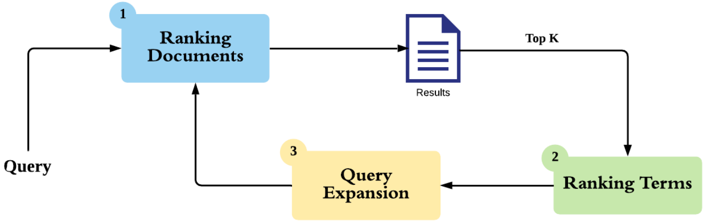
Bag of words limits

- No context modeling
  - Negative form
  - Disambiguation
- Semantic gap



Extensions

- N-gram encoding ⇒ group of words
  - *very good*
  - *not good*
  - Combinatorial dictionary ⇒ dimension issue !
- Lemmatization/stemming
  - 1 lexical stem = 1 column
- Rocchio's strategy
  - Pseudo Relevance Feedback
  - Query expansion



Semantic & ontologies

Rule based approaches

Adapted to several tasks... Especially the most complex: knowledge extraction.

Example:

- KEY = *event*
- Series of pattern to extract the location

$$KEY \begin{bmatrix} is\ located \\ is\ in \\ take(s)\ place \\ occurs \\ \dots \end{bmatrix} TARGET$$

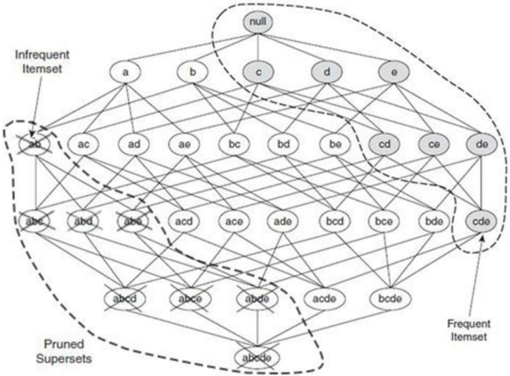
  - Kind of RDF triplets ⇔ Ontologies
- attendance, type of event, ...

+ Very good scaling
+ High precision
− Low recall

Learning Rules

Frequent item set

- Costly algorithm
- How taking into account synonyms?
  - Linguistic resources
- How taking into account language variations?
  - Handmade work



Word semantic

Objective

Understanding (automatically) word meaning

... And eliminating the semantic gap

$\Rightarrow$ Applications

- Information Retrieval
- Topic classification (& extraction)
- Information extraction
- Automated Summary
- Opinion classification
- ...

Linguistic resources

WordNet

- Description: Hierarchical description of words
  - Nouns
  - Verbs
  - Adjectives

Linguistic resources

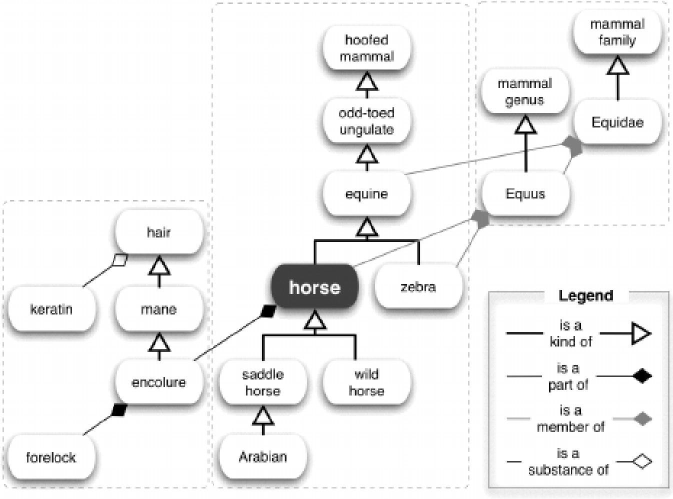
WordNet

- Description: Hierarchical description of words
  - Nouns
    - **hypernyms**: Y is a hypernym of X if every X is a (kind of) Y (canine is a hypernym of dog)
    - **hyponyms**: Y is a hyponym of X if every Y is a (kind of) X (dog is a hyponym of canine)
    - coordinate terms: Y is a coordinate term of X if X and Y share a hypernym (wolf is a coordinate term of dog, and dog is a coordinate term of wolf)
    - **meronym**: Y is a meronym of X if Y is a part of X (window is a meronym of building)
    - **holonym**: Y is a holonym of X if X is a part of Y (building is a holonym of window)
  - Verbs
  - Adjectives

Linguistic resources

WordNet

- Description: Hierarchical description of words



  - Nouns
  - Verbs
  - Adjectives

Linguistic resources

WordNet

- Description: Hierarchical description of words
  - Nouns
  - Verbs
    - **hypernym**: the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (to perceive is an hypernym of to listen)
    - **troponym**: the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (to lisp is a troponym of to talk)
    - **entailment**: the verb Y is entailed by X if by doing X you must be doing Y (to sleep is entailed by to snore)
    - **coordinate terms**: those verbs sharing a common hypernym (to lisp and to yell)
  - Adjectives

Linguistic resources

WordNet

- Description: Hierarchical description of words
  - Nouns
  - Verbs
  - Adjectives
    - **Antomyms** / **Synonyms**

WordNet: Metrics

- Metrics in WordNet
  - Length of the shortest path in the graph
  - Length of the shortest path in the *synonym* graph,
  - Distance of the first common ancestor,
  - cf: Leacock Chodorow (1998), Jiang Conrath (1997), Resnik (1995), Lin (1998), Wu Palmer (1993)
- WordNet & metrics are available in NLTK

WordNet: Limits & usage

- Fully depend on static resources
  - New expressions + technical/specialized vocabulary may lack
  - Social network mining, Hashtags ...

Existing extensions:

- Several translations
- More generally : **a powerful diffusion tool**
  - Characterizing one part of the vocabulary
    + using WordNet to spread characterization (synonyms...)
- Applications
  - IR: Information Retrieval
  - Word Desambiguation
  - Text Classification
  - Machine Translation
  - Summarization

[D. Jurafsky] Sentiment Lexicons

The General Inquirer

- Home page: `http://www.wjh.harvard.edu/~inquirer`
- List of Categories: `http://www.wjh.harvard.edu/~inquirer/homecat.htm`
- Spreadsheet: `http://www.wjh.harvard.edu/~inquirer/inquirerbasic.xls`
- Categories:
  - Positive (1915 words) and Negative (2291 words)
  - Strong vs Weak, Active vs Passive, Overstated versus Understated
  - Pleasure, Pain, Virtue, Vice, Motivation, Cognitive Orientation, etc
- Free for Research Use

Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. - MIT Press, 1966
The General Inquirer: A Computer Approach to Content Analysis

[D. Jurafsky] Sentiment Lexicons

LIWC (Linguistic Inquiry and Word Count)

- Home page: `http://www.liwc.net/`
- 2300 words, ¿70 classes
- Affective Processes
  - negative emotion (bad, weird, hate, problem, tough)
  - positive emotion (love, nice, sweet)
- Cognitive Processes
  - Tentative (maybe, perhaps, guess), Inhibition (block, constraint)
  - Pronouns, Negation (no, never), Quantifiers (few, many)
- $30 or $90 fee

Pennebaker, J.W., Booth, R.J., & Francis, M.E. 2007. Austin, TX
Linguistic Inquiry and Word Count: LIWC

[D. Jurafsky] Sentiment Lexicons

MPQA Subjectivity Cues Lexicon

- Home page: `http://www.cs.pitt.edu/mpqa/subj_lexicon.html`
- 6885 words from 8221 lemmas
  - 2718 positive
  - 4912 negative
- Each word annotated for intensity (strong, weak)
- GNU GPL

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, EMNLP 2005
Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis

[D. Jurafsky] Sentiment Lexicons

Bing Liu Opinion Lexicon

- Bing Liu's Page on Opinion Mining
  `http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`
- 6786 words
  - 2006 positive
  - 4783 negative

Minqing Hu and Bing Liu. ACM SIGKDD-2004.
Mining and Summarizing Customer Reviews

[D. Jurafsky] Sentiment Lexicons

SentiWordNet

- Home page: `http://sentiwordnet.isti.cnr.it/`
- All WordNet synsets automatically annotated for degrees of:
  - positivity, negativity, and neutrality/objectiveness
- Many contexts investigated

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. LREC-2010
SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining

[D. Jurafsky] Sentiment Lexicons

With an example: **short**

ADJECTIVE

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. LREC-2010
SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining

[C. Potts] Disagreements between polarity lexicons

	Opinion Lexicon	General Inquirer	SentiWordNet	LIWC
MPQA	33/5402 **(0.6%)**	49/2867 **(2%)**	1127/4214 **(27%)**	12/363 **(3%)**
Opinion Lexicon		32/2411 **(1%)**	1004/3994 **(25%)**	9/403 **(2%)**
General Inquirer			520/2306 **(23%)**	1/204 **(0.5%)**
SentiWordNet				174/694 **(25%)**
LIWC				

Building Lexicons or semantics (for sentiment analysis)

Overall Philosophy

Target:

- Extracting the meaning of words and patterns of words
- ... Namely, understanding the message and deducing the polarity

$\Rightarrow$ Building Universal Models

Important tasks and subtasks:

- Building/learning/using lexical resources
- Extracting complex sentiment patterns
- Dealing with different problems related to sentiment defintion ($(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$, entity, feature, polarity, holder, time)

Stanford NLP tools : `http://nlp.stanford.edu`
Named Entity Recognition, Dependency Tree Building, POS Tagging...

Opinionated Lexicons Building

Alternative 1:

1. Getting a lexicon with synonymous (e.g. WordNet)
2. Handmade opinion reference list:
   - *good, poor...*
3. Diffusion of the polarity according to the synonymous graph

Alternative 2:

1. Handmade opinion reference list:
   - *good, poor...*
2. Diffusion with external sources:
   - corpus (with labels or not)
   - search engines

Hatzivassiloglou and McKeown 1997

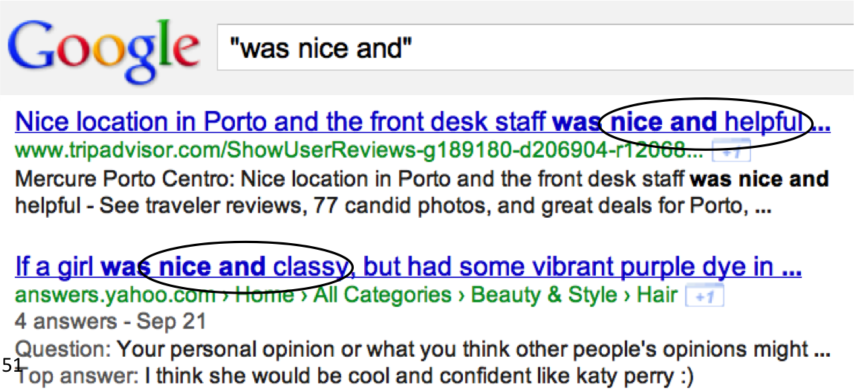
Hypothesis :

- Adjectives separated by **and** $\Rightarrow$ same polarity
  - Fair **and** legitimate, corrupt **and** brutal
  - fair **and** brutal, corrupt **and** legitimate
- Adjectives separated by **but** $\Rightarrow$ different polarity
  - fair **but** brutal
- Initialization: 1336 adjectives ($\approx$ 50/50 positive/negative)

Hatzivassiloglou McKeown 1997
Predicting the Semantic Orientation of Adjectives

Hatzivassiloglou and McKeown 1997

Expansion using external resources:

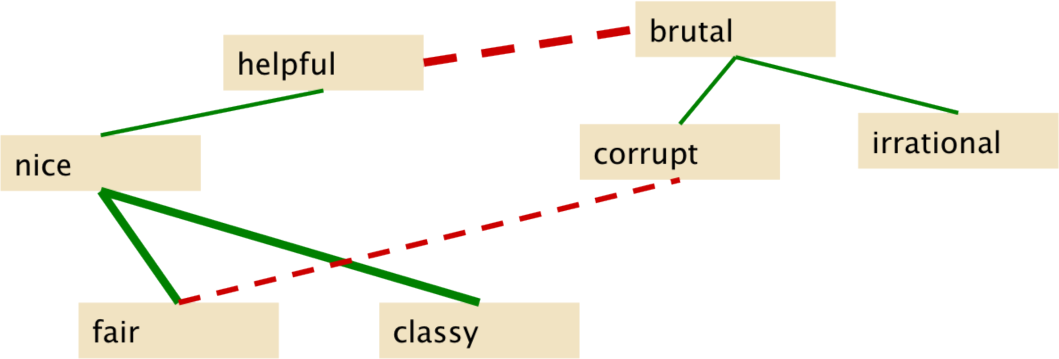


nice, helpful

nice, classy

Hatzivassiloglou McKeown 1997
Predicting the Semantic Orientation of Adjectives

Hatzivassiloglou and McKeown 1997



+ clustering

Hatzivassiloglou McKeown 1997
Predicting the Semantic Orientation of Adjectives

Hatzivassiloglou and McKeown 1997

Results :

- **Positive**
  bold decisive disturbing generous good honest important large mature patient peaceful positive proud sound stimulating straightforward strange talented vigorous witty...
- **Negative**
  ambiguous cautious cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor outspoken pleasant reckless risky selfish tedious unsupported vulnerable wasteful...

Hatzivassiloglou McKeown 1997
Predicting the Semantic Orientation of Adjectives

Hu & Liu 2004

Usage: one step of their summarization system:

- Initialization from an annotated corpus (user reviews)

The iPhone 4S: a smartphone and a whole lot more, September 30, 2012
By **SophieK** (Palo Alto, CA) - See all my reviews
This review is from: Apple iPhone 4S 16GB (White) - AT&T (Electronics)
I finally made the transition to the Apple iPhone 4S after over two years of a few highs and countless lows with an old Motorola Droid (model A855), which now serves as a paper weight. I'll make this short and sweet.

What I love:
1. The awesome camera, especially when paired with the Camera+ app, allows me to keep my bulky DSLR at home when I need a good, serviceable scenery shot for social

  - Part of Speech analysis
  - Adjectives annotated from document label
- frequential filtering

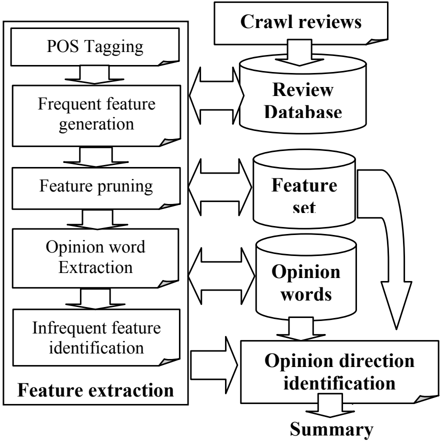


Figure 1: The opinion summarization system

Hu and Liu, AAAI NCAI 2004
Mining opinion features in customer reviews

Pointwise Mutual Information ,Turney, 2002

1. Documents ⇒ small patterns (=phrases)

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Nor NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

2. Phrases evaluation
   - Positive phrases co-occur more with *excellent*
   - Negative phrases co-occur more with *poor*
3. Score aggregation at the document level

Turney, ACL 2002

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

Pointwise Mutual Information ,Turney, 2002

1. Documents ⇒ small patterns (=phrases)

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Nor NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

2. Phrases evaluation
   - Positive phrases co-occur more with *excellent*
   - Negative phrases co-occur more with *poor*
3. Score aggregation at the document level

But how to measure co-occurrence?

Turney, ACL 2002
Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews

PMI ,Turney, 2002

Mutual Information:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right),$$

kind of similarity between X et Y.

Pointwise Mutual Information:

$$PMI(X,Y) = \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

How much more do events x and y co-occur than if they were independent? (i.e. $PMI = 0$ in case of independence)

PMI ,Turney, 2002

Probabilities estimation with Altavista:

- $P(word)$ is approximated by: $\text{hits}(word)/N$
- $P(word_1, word_2)$ by: $\text{hits}(word_1 \ NEAR \ word_2)/N^2$

Sentence Polarity

$$Pol(s) = PMI(s, "excellent") - PMI(s, "poor")$$

$$Pol(s) = \log\left(\frac{\text{hits}(s \ NEAR \ "excellent")\text{hits}("poor")}{\text{hits}(s \ NEAR \ "poor")\text{hits}("excellent")}\right)$$

PMI [Turney, 2002] : Results

Positive Reviews:

Phrase	POS tags	Polarity
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
...		
low fees	JJ NNS	0.33
true service	JJ NN	−0.73
other bank	JJ NN	−0.85
inconveniently located	JJ NN	−1.5
Average		0.32

Negative Reviews:

Phrase	POS tags	Polarity
direct deposits	JJ NNS	5.8
online web	JJ NN	1.9
very handy	RB JJ	1.4
...		
virtual monopoly	JJ NN	−2.0
lesser evil	RBR JJ	−2.3
other problems	JJ NNS	−2.8
low funds	JJ NNS	−6.8
unethical practices	JJ NNS	−8.5
Average		−1.2

⇒ External resources: finding some pattens that are topic-related and not universal

PMI [Turney, 2002] : Results

- 410 reviews from Epinions
  - 170 (41%) negative
  - 240 (59%) positive
  - 106,580 phrases
- Majority class baseline: 59%
- Turney algorithm: 74%
- Only 66% on movie reviews (average is not a good solution...)

Domain of Review	Accuracy
Automobiles	84.00 %
Honda Accord	83.78 %
Volkswagen Jetta	84.21 %
Banks	80.00 %
Bank of America	78.33 %
Washington Mutual	81.67 %
Movies	65.83 %
The Matrix	66.67 %
Pearl Harbor	65.00 %
Travel Destinations	70.53 %
Cancun	64.41 %
Puerto Vallarta	80.56 %
All	74.39 %

Key points:

- Phrases rather than words
- Learns domain-specific information
- Fast & require no labeled dataset

Extension of Kamps, 2004

Same methodology as Turney... But introducing other analysis axes :

$$\text{Evaluative factor: } EVA(m) = \frac{d(m, bad) - d(m, good)}{d(good, bad)} \tag{1}$$

$$\text{Potency factor: } POT(m) = \frac{d(m, weak) - d(m, strong)}{d(strong, weak)} \tag{2}$$

$$\text{Activity factor: } ACT(m) = \frac{d(m, passive) - d(m, active)}{d(active, passive)} \tag{3}$$

Quantitative results: 61% $\rightarrow$ 71%
Qualitative analysis: comparison with the General Inquirer

J. Kamps, MJ Marx, R.J Mokken et M. De Rijke, LREC 2004
Using wordnet to measure semantic orientations of adjectives

LSA: Latent Semantic Analysis

Statistical approch: vectorial semantics

- Modeling: Word count (and BoW storage)

$$X = \begin{array}{cc} & \begin{array}{ccc} \phantom{x_{1,1}} & \mathbf{t}_j & \phantom{x_{1,D}} \\ & \downarrow & \end{array} \\ \mathbf{d}_i \rightarrow & \begin{pmatrix} x_{1,1} & \cdots & x_{1,D} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,D} \end{pmatrix} \end{array}$$

- Basic proposal: semantics = metrics = similarity between columns in BoW

$$s(j,k) = \langle t_j, t_k \rangle, \qquad \text{Normalized: } s_n(j,k) = \cos(\theta) = \frac{\mathbf{t}_j \cdot \mathbf{t}_q}{\|\mathbf{t}_j\| \, \|\mathbf{t}_q\|}$$

  - If two terms appear in the same document, they are similar

SVD : Singular Value decomposition

- In NLP : SVD = LSA: Latent Semantic Analysis
- Idea : grouping similar documents / learning a representation of documents

$$
\begin{array}{ccccccc}
 & X^T & = & U & & \Sigma & V^T \\
 & \begin{array}{c}\mathbf{d}_i \\ \downarrow\end{array} & & & & & \begin{array}{c}\hat{\mathbf{d}}_i \\ \downarrow\end{array} \\
\mathbf{t}_j \rightarrow & \begin{pmatrix} x_{1,1} & \dots & x_{1,N} \\ \vdots & \ddots & \vdots \\ x_{D,1} & \dots & x_{D,N} \end{pmatrix} & = & \left(\begin{pmatrix} \\ \mathbf{u}_1 \\ \\ \end{pmatrix} \dots \begin{pmatrix} \\ \mathbf{u}_l \\ \\ \end{pmatrix} \right) & & \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{pmatrix} & \begin{pmatrix} (& \mathbf{v}_1 &) \\ & \vdots & \\ (& \mathbf{v}_l &) \end{pmatrix}
\end{array}
$$

- Good news: functions well on sparse matrices

Factorization = robustness & clustering ability

S. Deerwester, et al., JSIS 1990
Indexing by latent semantic analysis

Discussion : SVD, LSA

Selecting the k greatest singular values gives a rank-k approximation of the occurence matrix.

- Each $\mathbf{u} \in \mathbb{R}^D$ is a weight vector associated to the vocabulary
- The base $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is orthogonal
  - Each $\mathbf{u}$ corresponds to a different **lexical field**
- The new document representation $\mathbf{v}$ is a weight vector associated to the lexical fields
  - Clustering issue: the strongest weight gives the document class

Thomas K. Landauer, Peter W. Foltz et Darrell Laham, Discourse Processes, vol. 25, 1998
Introduction to Latent Semantic Analysis

Many applications

Usages:

- Clustering (each eigen vector describes a *topic*)
- Semantics: words have a representation over the topics
- IR Improvement:
  - Query expansion based on the topic definition
  - Detection of polysemic terms
- New representation $\Rightarrow$ new applications
  - opportunities in question answering
    - Finding the part of a document relating to a specific topic
  - Automated summarization
    - Document segmentation + sentence extraction
  - TDT : Topic detection & Tracking

LSA Limits

- Fully based on BOW: no word dependency modeling
  - issues regarding negative formulation
  - depends on document sizes
  - Not robust to stop words
    - associated to high singular values
    - + appear in many topics
- Topic modeling is link to a corpus
  - problem with rare words in small corpus
  - bias of the corpus

LSA variation : k-means

- Still a BOW modeling

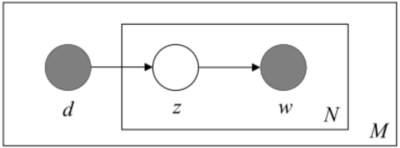
$$X = \begin{array}{cc} & \mathbf{t}_j \\ & \downarrow \\ \mathbf{d}_i \rightarrow & \begin{pmatrix} x_{1,1} & \cdots & x_{1,D} \\ \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,D} \end{pmatrix} \end{array}$$

- Algorithm that scale up well
  - Possible **on-line** version of the algorithm
  - Can be linked to chinese restaurant / indian buffet process
    - $\Rightarrow$ Discover k in an online process
- Orthogonality is not longer enforced

PLSA

Probabilistic Latent Semantic Analysis

- Idea: CEM $\Rightarrow$ EM (more complex / finer)
- All documents belongs to all clusters... With a weight $p(z|d)$
- Graphical model

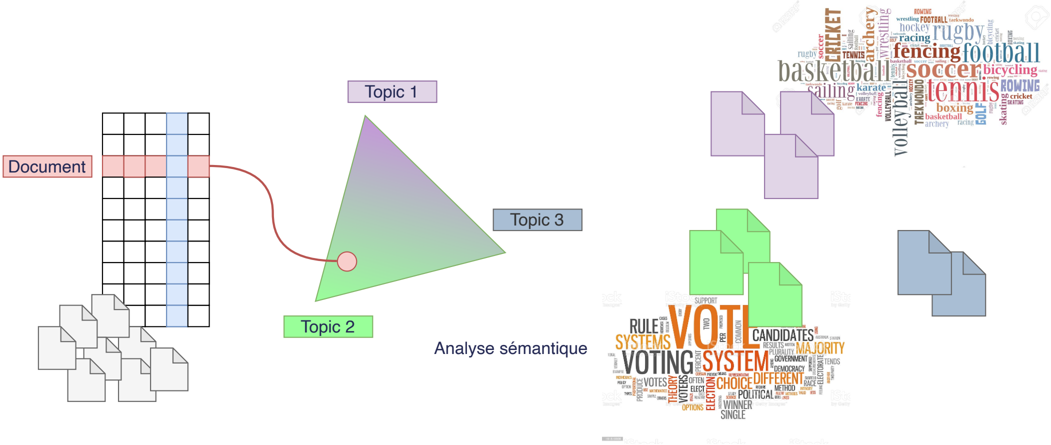


Given a number of topics K (and N documents in a vocabulary D)

We estimate the following parameters:

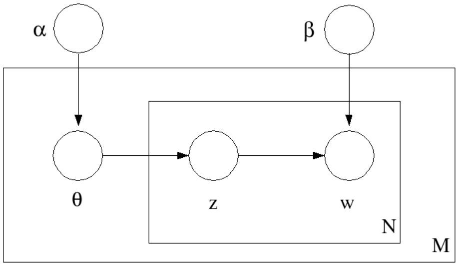
- Doc d is drawn from $P(d)$
- Topic z is drawn from $P(z|d)$
- Word w is drawn from $P(w|z)$
  - $p(d) \in \mathbb{R}^N$
  - $p(z|d) \in \mathbb{R}^{K \times N}$
  - $p(w|z) \in \mathbb{R}^{D \times K}$

PLSA: results



LDA

Latent Dirichlet Allocation:

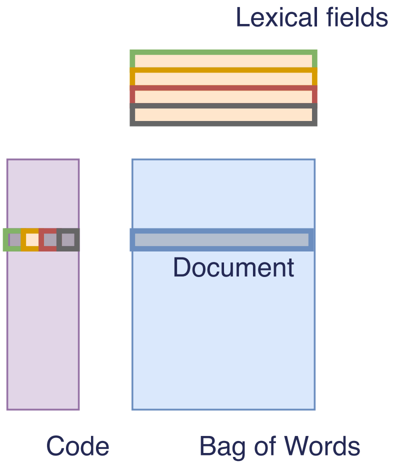


- Idea: adding a prior on the topic distribution
    - A document is supposed to belong to a topic **strongly or not**
- Learning through Gibbs sampling ($\sim$ MCMC)

not to be confused: LDA: Latent Dirichlet Allocation *vs* Linear Discriminant Analysis

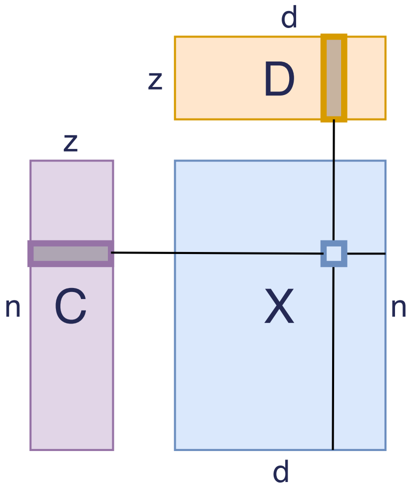
Conclusion on statistical semantic analysis

1. Quantitative results
   - Clustering
   - Major issue with frequent words
   - Human required in the loop (init., cluster selection, etc...)
   - Evaluation issue (purity, perplexity, ...)
2. Qualitative analysis
   - Word similarity
   - Lexical field extraction



Representation Learning & matrix factorization

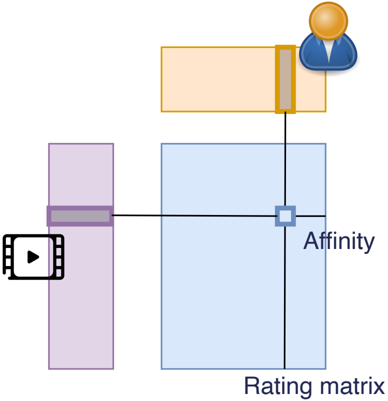
Matrix factorization = basic tool to understand the data



- Extract a compact representation
  - for words
  - for documents
- = focus on high-energy phenomenon
  - Eliminate noise in the data
- Optimal data compression [Mean Square criterion]

Representation Learning & matrix factorization

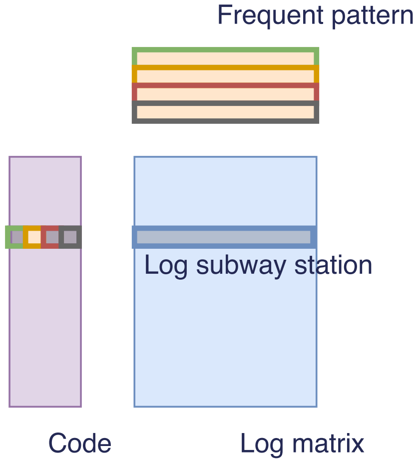
Matrix factorization = basic tool to understand the data



- Extract a compact representation
  - for words
  - for documents
- = focus on high-energy phenomenon
  - Eliminate noise in the data
- Optimal data compression [Mean Square criterion]

Representation Learning & matrix factorization

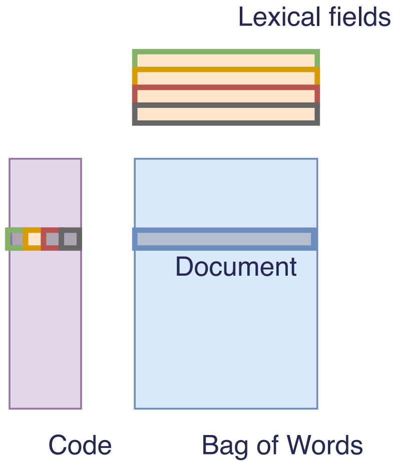
Matrix factorization = basic tool to understand the data



- Extract a compact representation
  - for words
  - for documents
- = focus on high-energy phenomenon
  - Eliminate noise in the data
- Optimal data compression [Mean Square criterion]

Representation Learning & matrix factorization

Matrix factorization = basic tool to understand the data



- Extract a compact representation
  - for words
  - for documents
- = focus on high-energy phenomenon
  - Eliminate noise in the data
- Optimal data compression [Mean Square criterion]

Conclusion

Conclusion

- Des algorithmes incontournables (k-means, PLSA et surtout LDA)
  - Rarement fonctionnel directement — Human-in-the-loop approaches
  - Réfléchir toujours en terme:
    1. Initialisation (probablement LDA)
    2. Stratégie **active-learning**: quels échantillons montrer à l'utilisateur
    3. Optimisation d'un classifieur (SVM, RegLog,...)
- Une évaluation problématique
  - Le qualitatif est rarement suffisant... Et ne s'optimise pas!
  - Crowd-sourcing = intéressant mais cher
- Une concurence accrue ces dernières années :
  - Pre-trained language model + few-shot finetuning