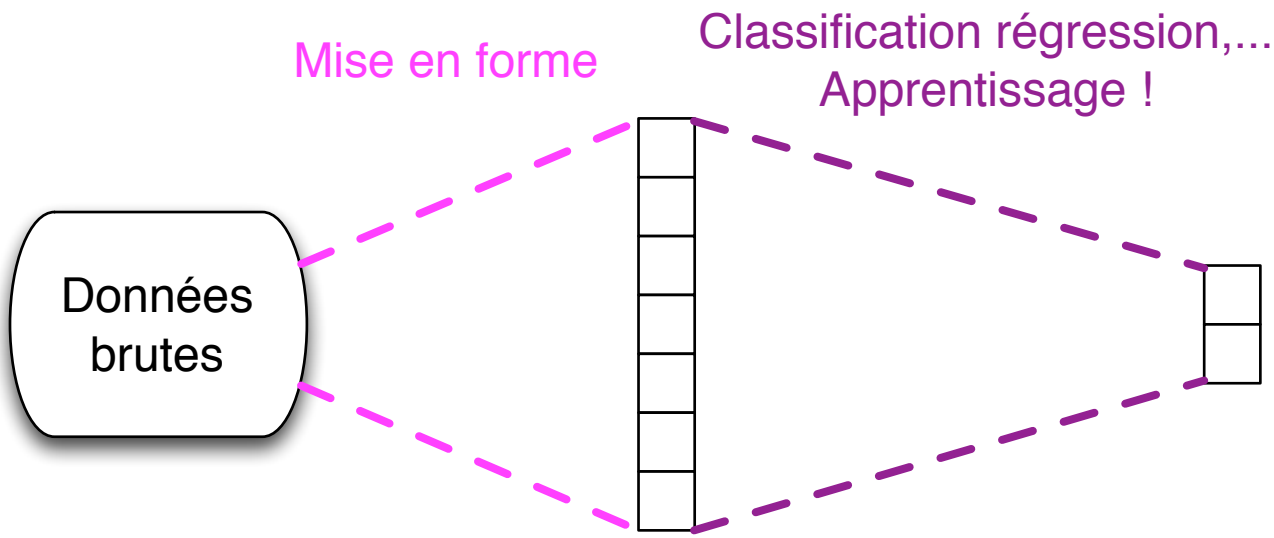
Textes, recommandation et données hétérogènes: apports des techniques d'apprentissage de représentations

Vincent Guigue

GDR ISIS - Journée thématique sur l'apprentissage de représentation

6 Octobre 2016

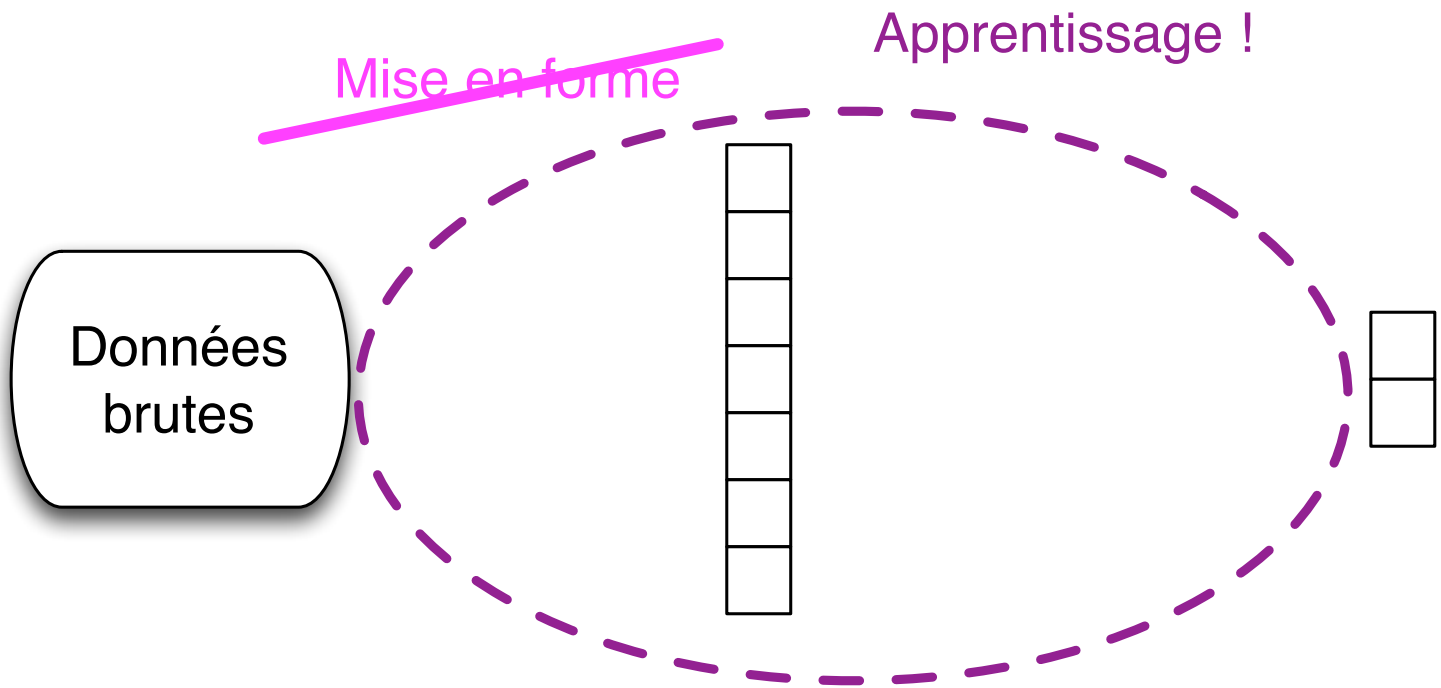
Apprentissage de représentations



Choix des mots, linguistique,
POS-tagging...
Filtrage des signaux...
Règles métiers, *a priori*

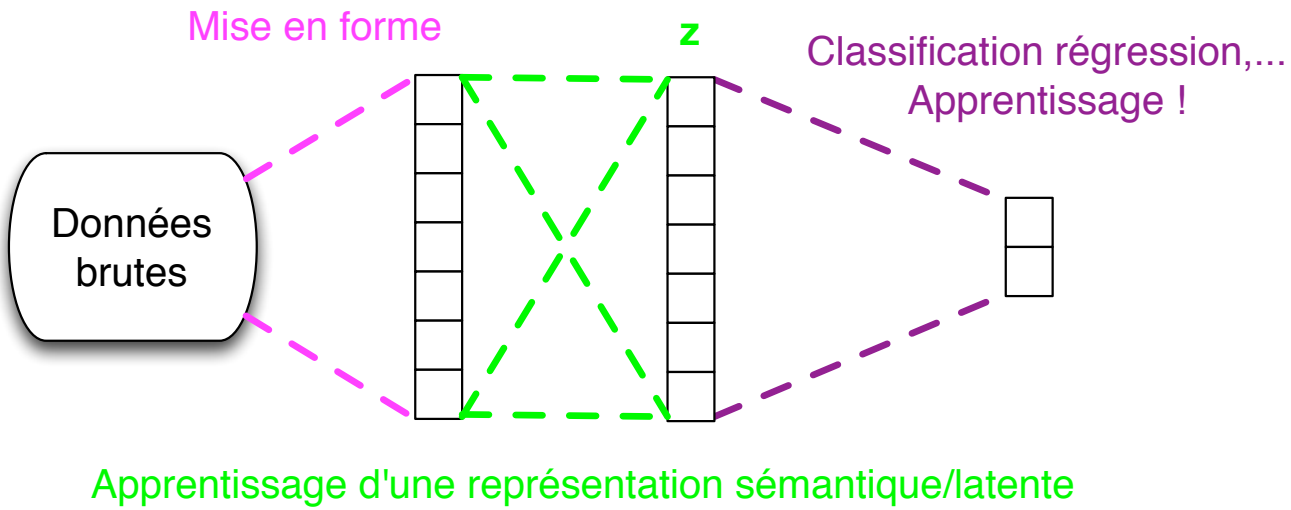
Arbre de décision, Ridge regression, SVM, ...

Apprentissage de représentations



Deep learning (par exemple)
Apprentissage des filtres, traitement du texte à partir des caractères...

Zone grise

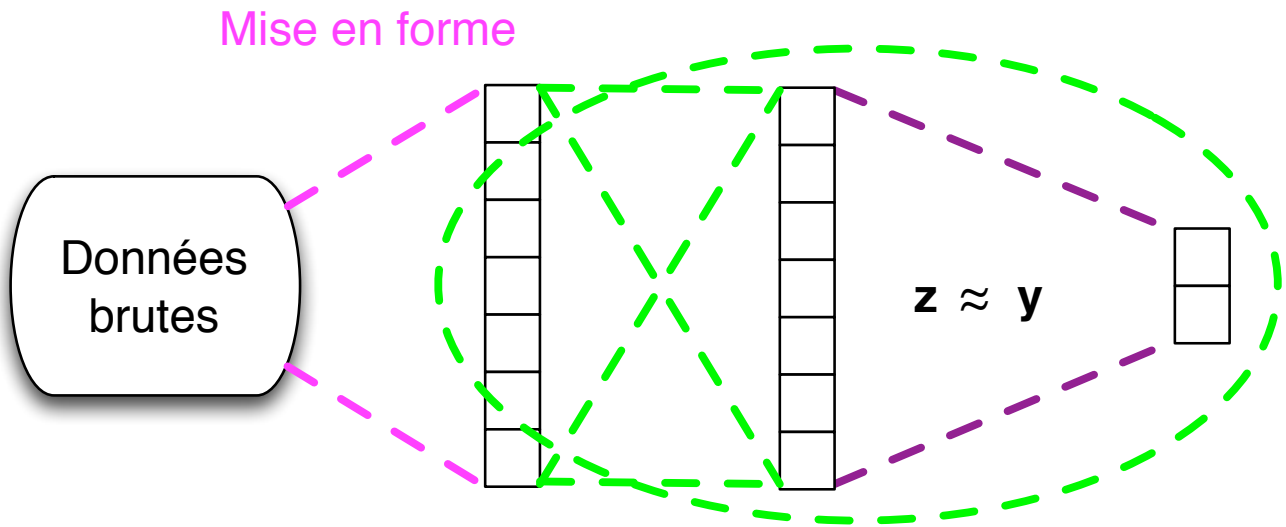


- Se passer de tous les pré-traitements...

  Pas très réaliste, pas toujours souhaitable

- Apprendre une représentation sémantique des éléments

Zone grise



- Se passer de tous les pré-traitements...

  Pas très réaliste, pas toujours souhaitable
- Apprendre une représentation sémantique des éléments
- Etirer progressivement la zone verte vers la droite...

  La représentation apprise répond-elle directement à la tâche?

Définition de la représentation

- **But:** obtenir une représentation vectorielle des éléments du problème
  - Textes, mots, utilisateurs... $\mathbf{x}_i \in \mathcal{X} \rightarrow \mathbf{z}_i \in \mathbb{R}^d$
  - Métrique(s) + opérateur(s) sur les $\mathbf{z}_i \Rightarrow$ réponse à la tâche

- **Critère:**
  - Capacité de reconstruction: $\mathbf{z}_i \in \mathbb{R}^d \rightarrow \mathbf{x}_i \in \mathcal{X}$
  - Proximité/Séparation: (pour certains couple) $\|\mathbf{z}_i - \mathbf{z}_j\|^2$
  - Lié à une tâche: $\|f(\mathbf{z}_i) - y_i\|^2$

- **Contraintes:** structuration de l'espace de représentation
  - Positions/formes des $\mathbf{z}_i$
  - Parcimonie des représentations

Positionnement de l'exposé

Domaine applicatif

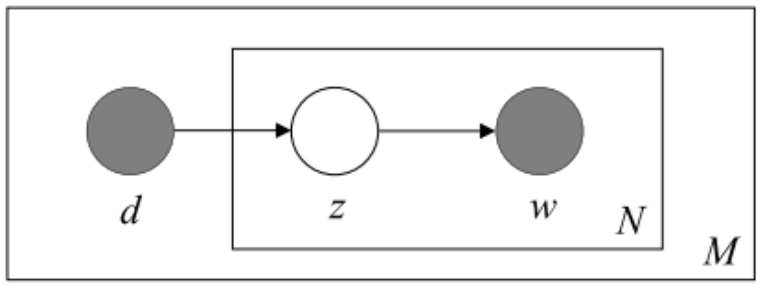
Le texte, la recommandation... Et les données hétérogènes.

1. [Texte] Pré-traitement des données / analyse sémantique latente
2. [Reco] Factorisation matricielle
3. [Hétérogénéité] Techniques de projections hétérogènes...
4. [Raisonnement] ... et opérateurs multiples dans les espaces latents

Modèles graphiques & clustering thématique

document = ensemble de mots

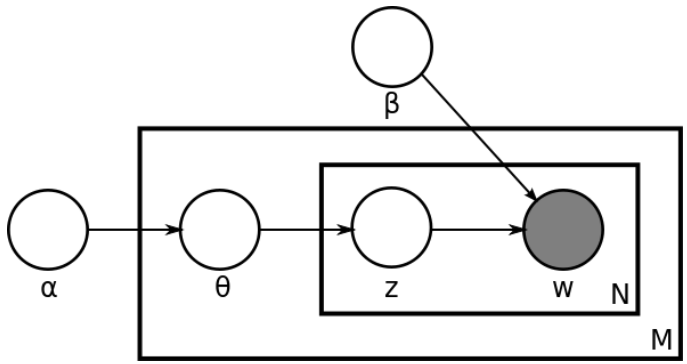
Probabilistic Latent Semantic Analysis



- Représentations des documents: $\mathbf{z}_i = p(z|d_i)$
- Représentations des mots : $\mathbf{w}_j = p(w_j|z)$

Hofmann, SIGIR 1999
Probabilistic latent semantic indexing

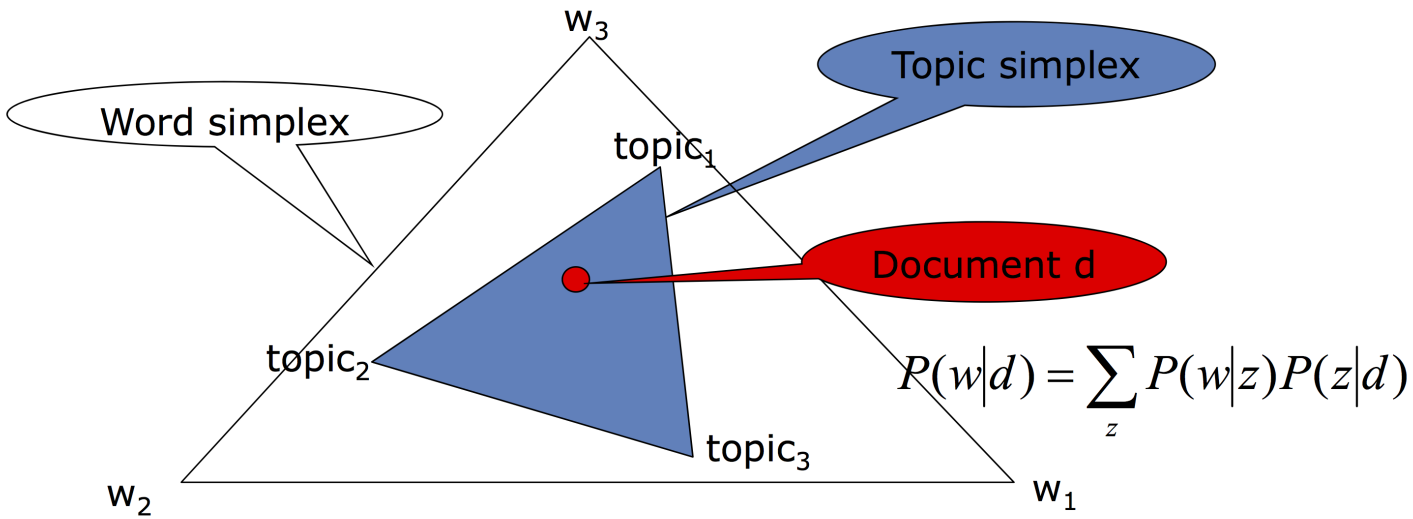
Latent Dirichlet Allocation



- Ajout d'une contrainte de parcimonie

Blei, Ng, Jordan, JMLR 2003
Latent dirichlet allocation

PLSA/LDA



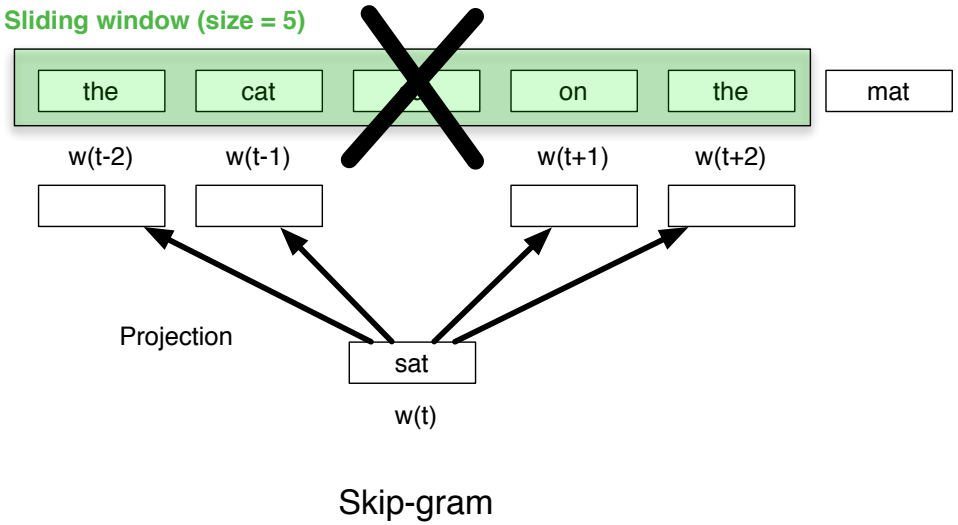
- Critère de reconstruction
  - Modèle génératif ⇒ maximum de vraisemblance
- Faible dimension / la représentation est une fin en soit

Est-ce de l'apprentissage de représentation?

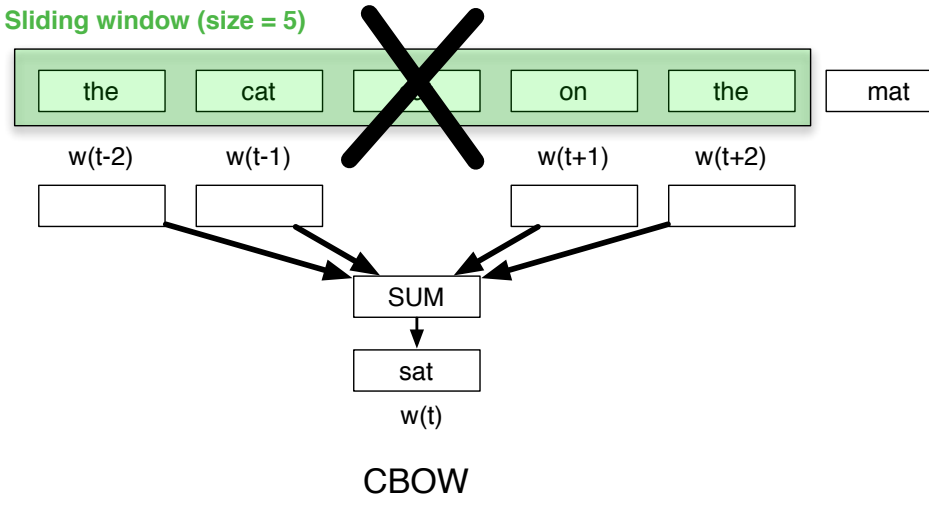
Apprentissage... OK

... de représentation - c'est moins sûr!

Sémantique locale: word2vec



Skip-gram



CBOW

Skip-Gram: apprentissage simple (par *negative sampling*)

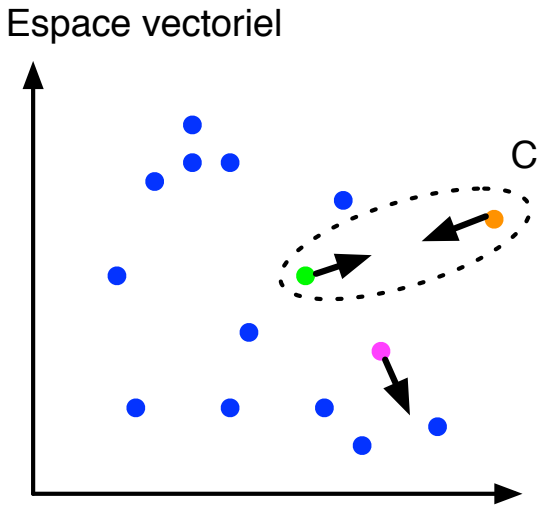
- Aspect local
- Critère *prédictif* : prédire l'information manquante

Mikolov, Sutskever, Chen, Corrado, Dean, NIPS 2013
Distributed representations of words and phrases and their compositionality

Explications détaillées

- Critères sur les mots w et les contextes C:

$$\text{Idée SG: } \arg\max_{\theta} \prod_{C} \prod_{w \in C} p(C|w;\theta)$$



- Soit $p(D = 1|w_i, w_j; \theta)$ la proba que w_i et w_j co-occurent dans le corpus:

$$\arg\max_{\theta} \prod_{i,j \in C} p(D = 1|w_i, w_j; \theta) + \underbrace{\prod_{i,j \in \bar{C}} p(D = 0|w_i, w_j; \theta)}_{\text{Negative Sampling}}$$

Goldberg, Levy, arXiv 2014
word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method

Hammer, NN 2002
Generalized Relevance Learning Vector Quantization

Skip-gram & negative sampling

$$\arg\max_{\theta} \prod_{i,j \in C} p(D=1|w_i, w_j; \theta) + \underbrace{\prod_{i,j \in \bar{C}} p(D=0|w_i, w_j; \theta)}_{\text{Negative Sampling}}$$

- Modélisation par fct logistique: $p(D=1|w_i, w_j) = \frac{1}{1+\exp(-\mathbf{z}_i \mathbf{z}_j)}$
- Passage au log:

$$Z^\star = \arg\max_{Z} \left(\sum_{i,j \in C} \log \sigma(\mathbf{z}_i \cdot \mathbf{z}_j) + \sum_{i,j \in \bar{C}} \log \sigma(-\mathbf{z}_i \cdot \mathbf{z}_j) \right)$$

σ : fct sigmoide, C : Set of Cooccurences, $\bar{C}$: Set of Non-Cooc

- Optimisation par descente de gradient stochastique efficace
- Trick sur la probabilité de tirage des mots fréquents:
  $p(w_i) = 1 - \sqrt{\frac{t}{freq(w_i)}}, \qquad t = 10^{-5}$

W2V: Résultats

- Dimension plus grande (*vs* PLSA/LDA & autres approches NN)
- Exploitation possible de la représentation
- Fonctionne sur les grands corpus... Et sur les petits!

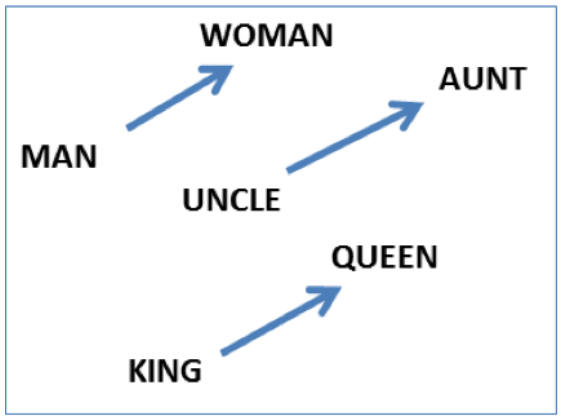
Model (training time)	Redmond	Havel	ninjutsu	graffiti	capitulate
Collobert (50d) (2 months)	conyers lubbock keene	plauen dzerzhinsky osterreich	reiki kohona karate	cheesecake gossip dioramas	abdicate accede rearm
Turian (200d) (few weeks)	McCarthy Alston Cousins	Jewell Arzu Ovitz	- - -	gunfire emotion impunity	- - -
Mnih (100d) (7 days)	Podhurst Harlang Agarwal	Pontiff Pinochet Rodionov	- - -	anaesthetics monkeys Jews	Mavericks planning hesitated
Skip-Phrase (1000d, 1 day)	Redmond Wash. Redmond Washington Microsoft	Vaclav Havel president Vaclav Havel Velvet Revolution	ninja martial arts swordsmanship	spray paint grafitti taggers	capitulation capitulated capitulating

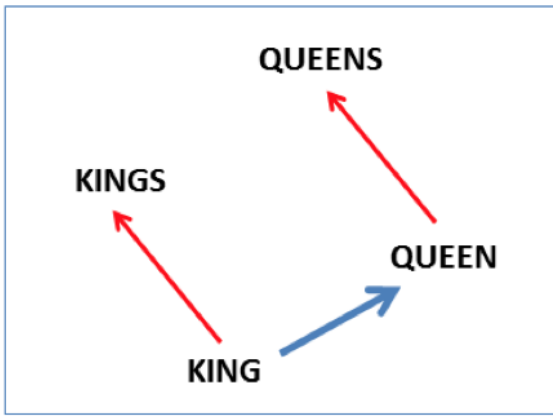
Table 6: Examples of the closest tokens given various well known models and the Skip-gram model trained on phrases using over 30 billion training words. An empty cell means that the word was not in the vocabulary.

W2V: Exploitation de l'espace latent

a est à *b* ce que *c* est à ??? $\Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Propriété syntaxique (1):





$\mathbf{z}_{woman} - \mathbf{z}_{man} \approx \mathbf{z}_{queen} - \mathbf{z}_{king}$

$\mathbf{z}_{kings} - \mathbf{z}_{king} \approx \mathbf{z}_{queens} - \mathbf{z}_{kings}$

Requête:

$\mathbf{z}_{woman} - \mathbf{z}_{man} + \mathbf{z}_{king} = \mathbf{z}_{req}$

Plus proche voisin:

$\arg\min_i \|\mathbf{z}_{req} - \mathbf{z}_i\| = \text{queen}$

$\Rightarrow$ Mise en place d'un test quantitatif de référence.

W2V: Exploitation de l'espace latent

a est à *b* ce que *c* est à ??? $\Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Propriété syntaxique (2):

Requête:

$\mathbf{z}_{easy} - \mathbf{z}_{easiest} + \mathbf{z}_{luckiest} = \mathbf{z}_{req}$

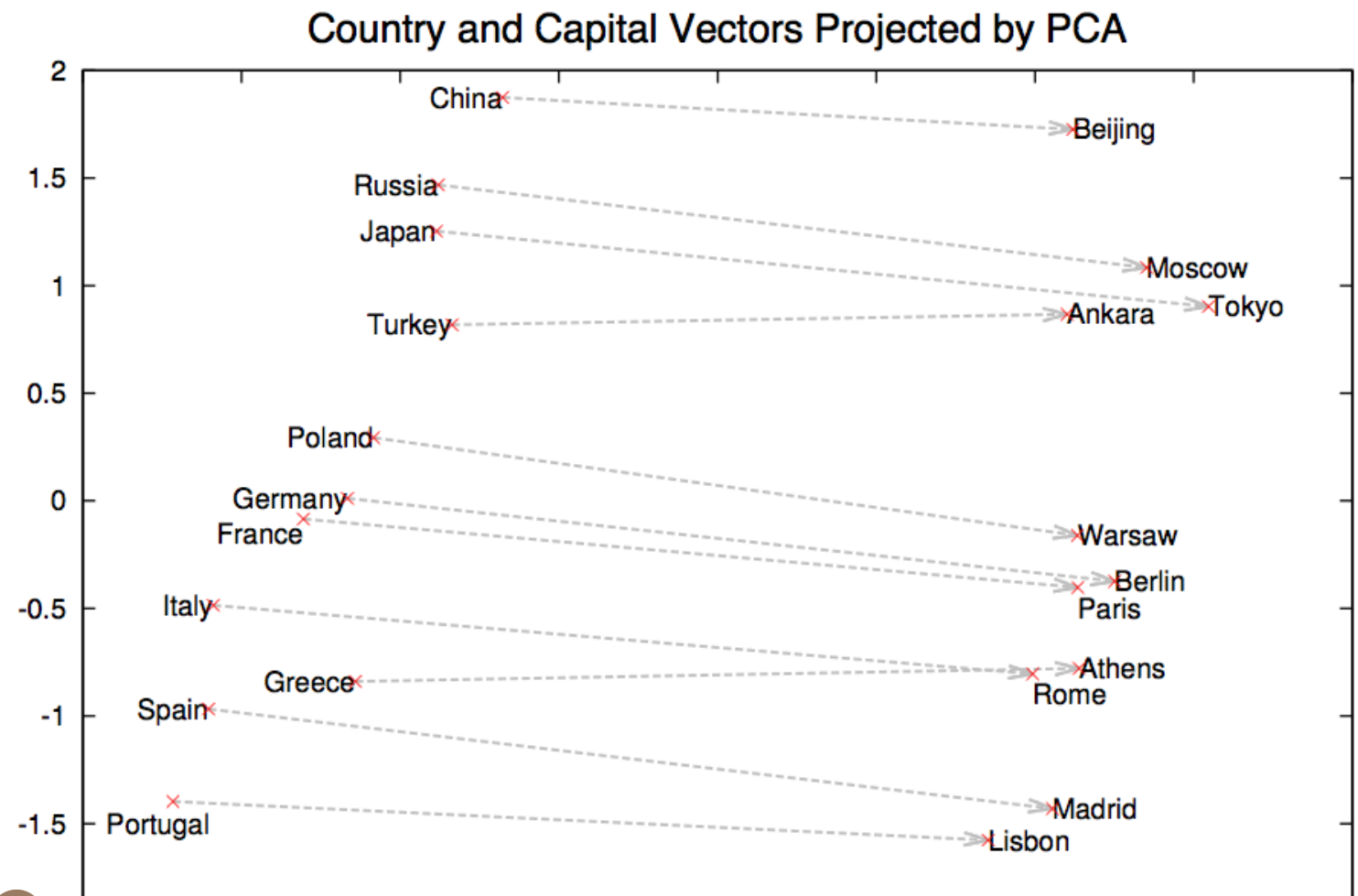
Plus proche voisin:

$\arg\min_i \|\mathbf{z}_{req} - \mathbf{z}_i\| = \text{lucky}$

W2V: Exploitation de l'espace latent

a est à *b* ce que *c* est à ??? $\Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Propriété sémantique (1)



W2V: Exploitation de l'espace latent

a est à *b* ce que *c* est à ??? $\Leftrightarrow \mathbf{z}_b - \mathbf{z}_a + \mathbf{z}_c$

Propriété sémantique (2)

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

Pourquoi ça marche?

- Prédire au lieu de compter? Pas sûr...
- Extraction locale de la sémantique ! $\Rightarrow$ Glove
  - $X \in \mathbb{R}^{V \times V}$ word co-occurrence matrix
  - X_{ij} frequency of word i co-occurring with word j
  - $X_i = \sum_k X_{ik}$ total number of occurrences of word i in corpus
  - $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ a.k.a. probability of word j occurring within the context of word i
  - $w \in \mathbb{R}^d$ a word embedding of dimension d
  - $\tilde{w} \in \mathbb{R}^d$ a context word embedding of dimension d

Baroni, Dinu & Kruszewski, ACL 2014
Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors

Pennington, Socher & Manning, EMNLP 2014
Glove: Global Vectors for Word Representation

Conclusion

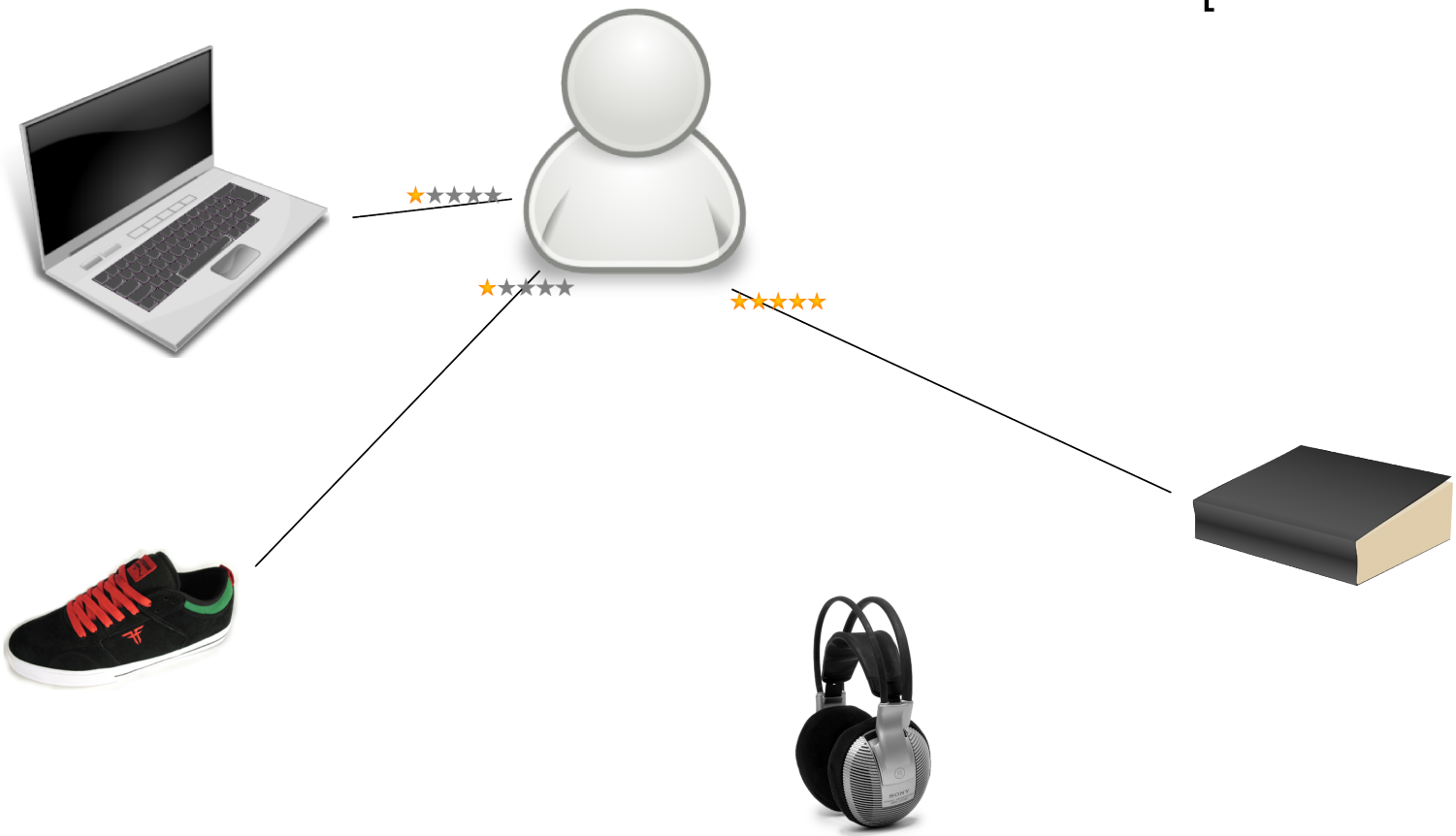
Les techniques récentes apportent:

- Légèreté
- Implémentation efficace [critère discutable]
- Représentation exploitable
  - un pas vers la représentation des connaissances
    - test quantitatif standard
  - la correction automatique
  - la traduction...

Collaborative filtering & factorisation matricielle

Collaborative filtering : réseau bipartie utilisateurs/produits

[cadre *netflix*]



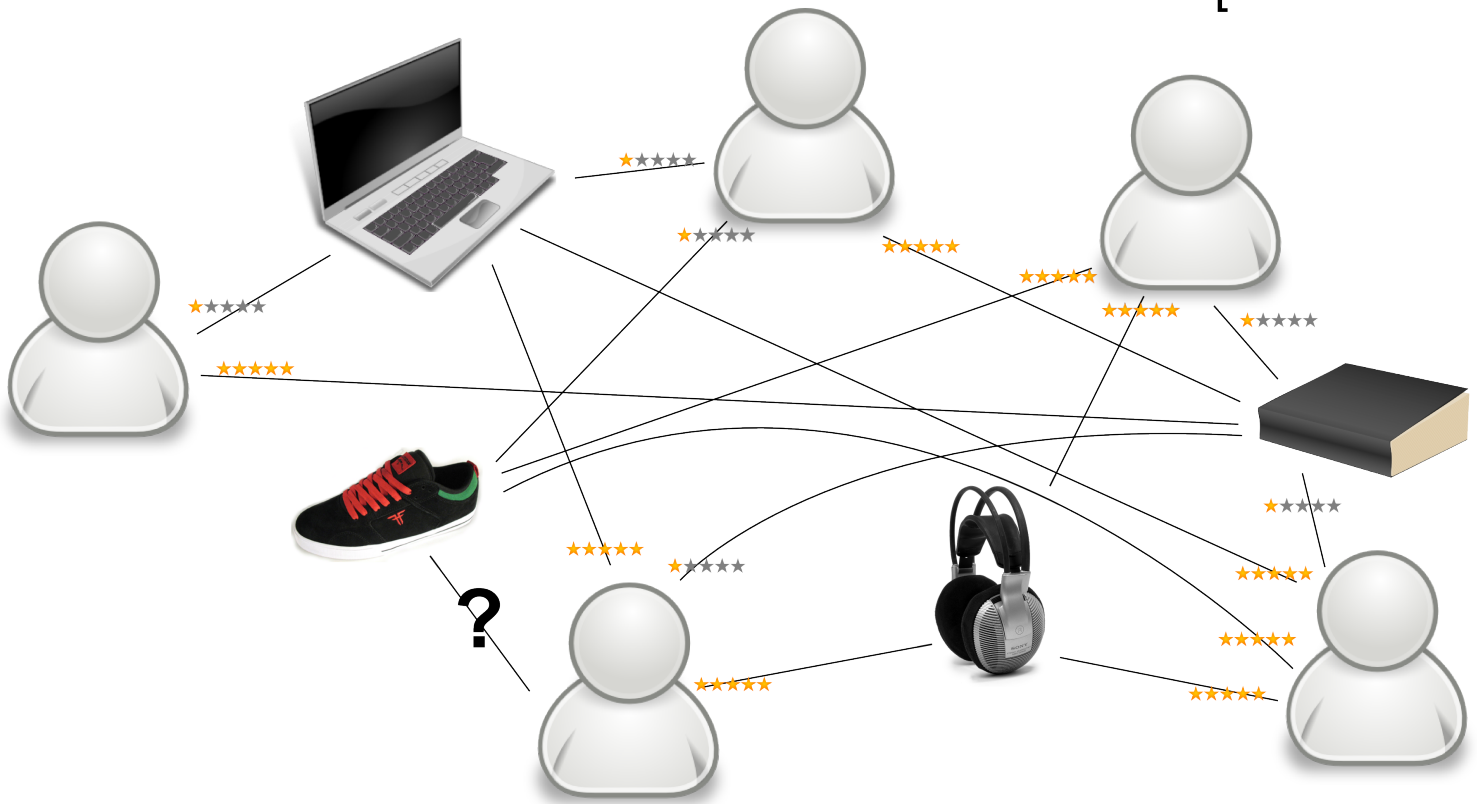
Koren, SIGKDD 2008

Factorization meets the neighborhood: a multifaceted collaborative filtering mode

Collaborative filtering & factorisation matricielle

Collaborative filtering : réseau bipartie utilisateurs/produits

[cadre *netflix*]



Koren, SIGKDD 2008

Factorization meets the neighborhood: a multifaceted collaborative filtering mode

Collaborative filtering & factorisation matricielle

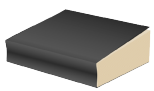
Collaborative filtering : réseau bipartie utilisateurs/produits

[cadre *netflix*]

★★★★★ ★★★★★ ★★★★★ ★★★★★

★★★★★ ? ★★★★★ ★★★★★

★★★★★ ★★★★★ ★★★★★



★★★★★ ★★★★★ ★★★★★ ★★★★★ ★★★★★

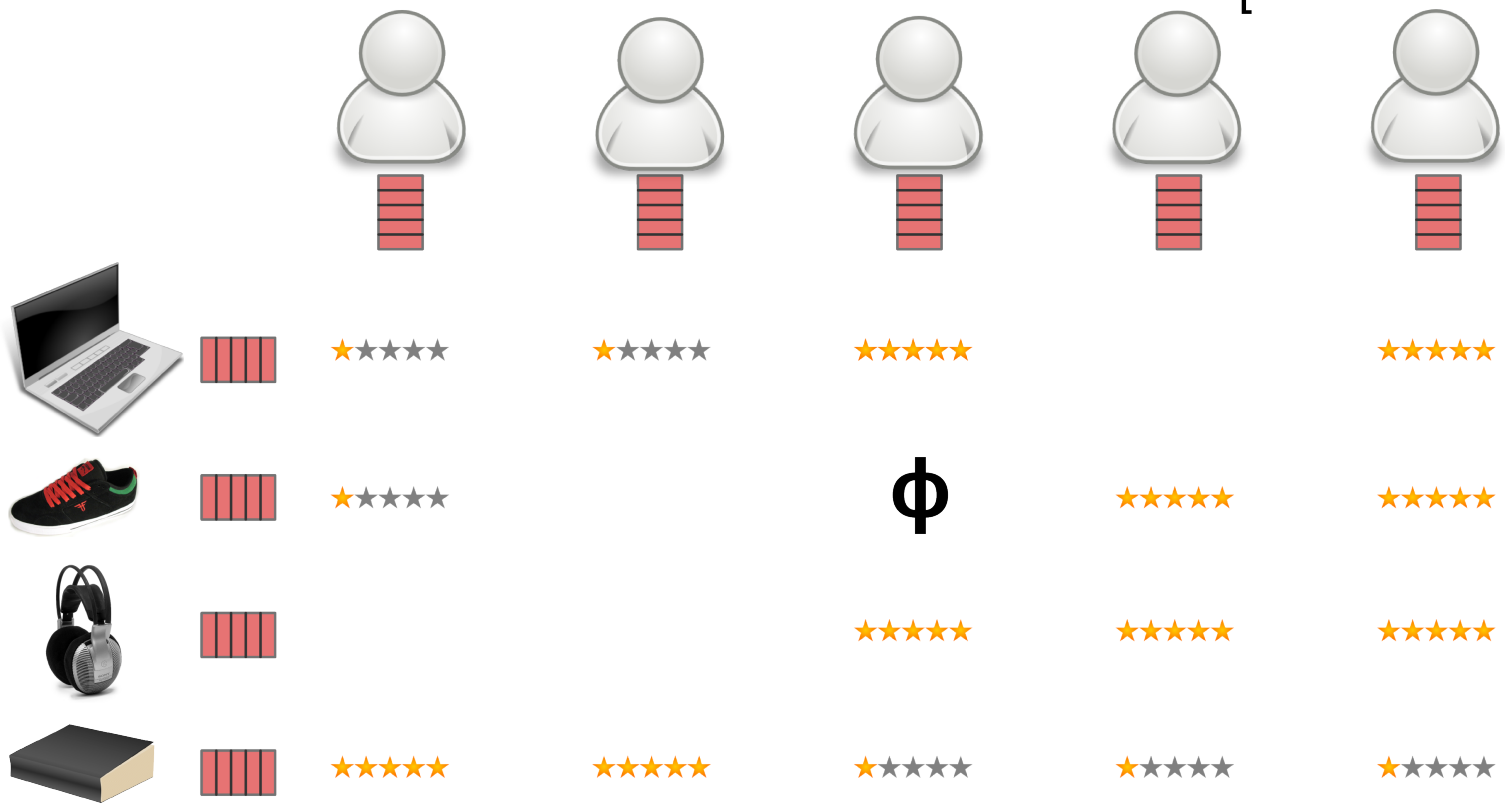
Koren, SIGKDD 2008

Factorization meets the neighborhood: a multifaceted collaborative filtering mode

Collaborative filtering & factorisation matricielle

Collaborative filtering : réseau bipartie utilisateurs/produits

[cadre *netflix*]



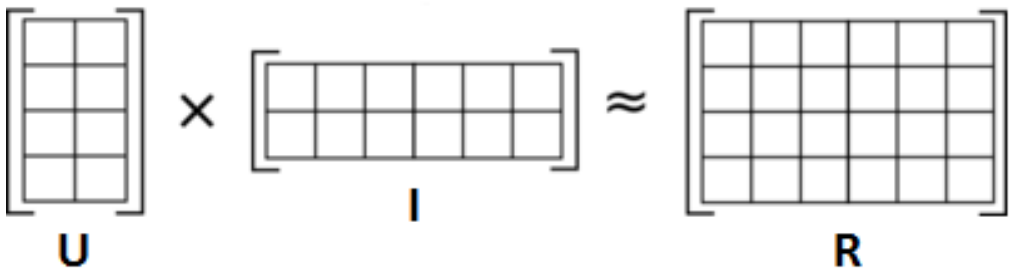
Koren, SIGKDD 2008

Factorization meets the neighborhood: a multifaceted collaborative filtering mode

Formalisation

- Apprendre des profils utilisateurs/items: $\mathbf{z}_u, \mathbf{z}_i$
- Critère de reconstruction des notes r:

$$Z_u^\star, Z_i^\star = \arg\min_{Z_u, Z_i} \sum_{(u,i,r)} \|r_{u,i} - \mathbf{z}_u \cdot \mathbf{z}_i\|^2$$



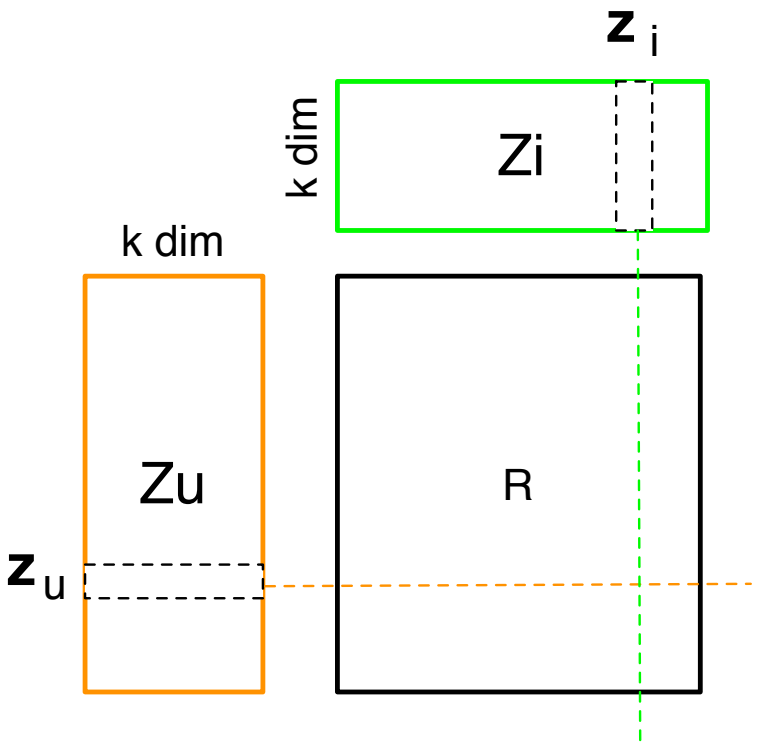
- Contraintes
  - Rang (= dimension de l'espace latent)
  - Non-négativité (reconstruction purement additive)
  - Parcimonie
- Implémentations nombreuses et efficaces

Hoyer, JMLR 2004
Non-negative matrix factorization with sparseness constraints

Régularité/rupture

- Régularité de la forme bilinéaire:

$$\hat{r}_{ui} = \mathbf{z}_i \mathbf{z}_j = \sum_k z_{ik} z_{jk}$$



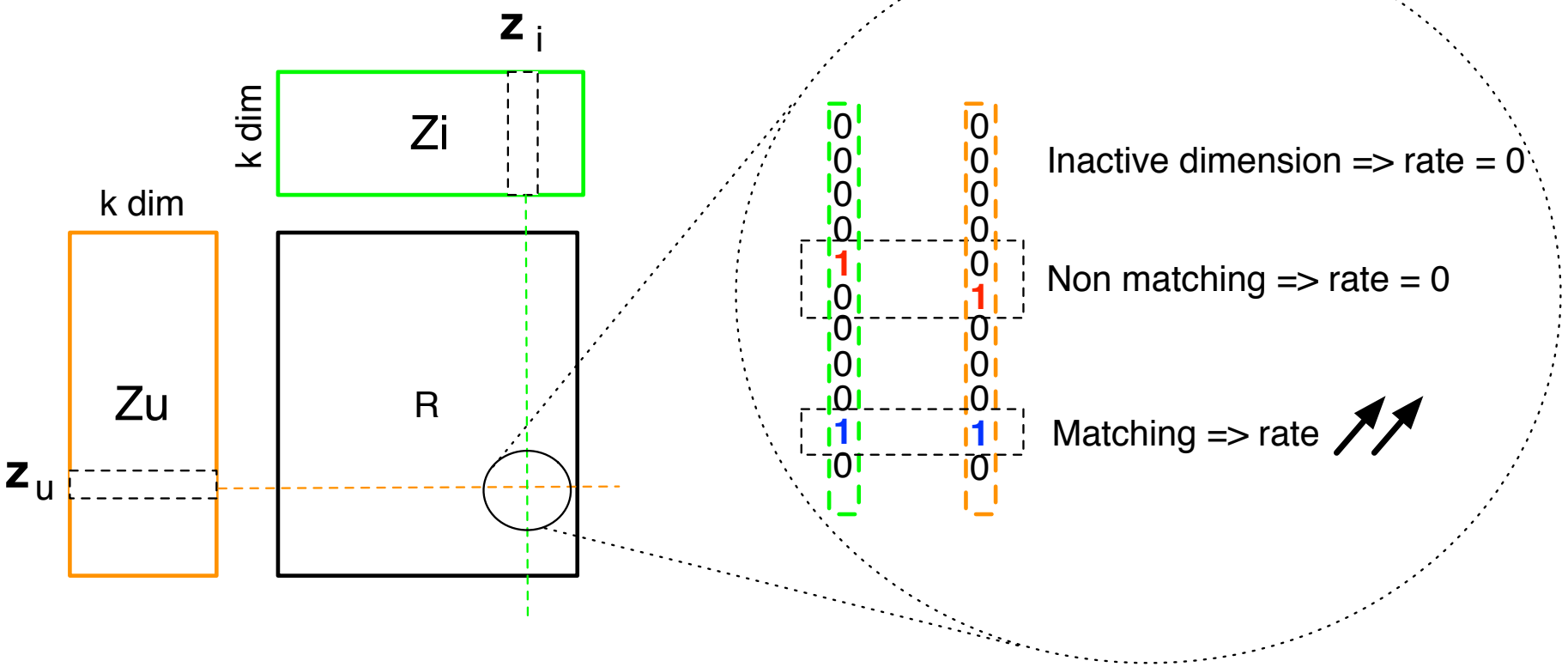
- Non-négativité + parcimonie

Régularité/rupture

- Régularité de la forme bilinéaire:

$$\hat{r}_{ui} = \mathbf{z}_i \mathbf{z}_j = \sum_k z_{ik} z_{jk}$$

- Non-négativité + parcimonie



Résultats: Amazon/Ratebeer

Prédiction optimale:

$$\tilde{r}_{ui} = \underbrace{\phi_0}_{\text{biais gén.}} + \underbrace{\phi_1(u)}_{\text{biais u}} + \underbrace{\phi_2(i)}_{\text{biais i}} + \underbrace{\phi_3(u,i)}_{\text{NMF}}$$

Evaluation: $\frac{1}{N} \sum_{(u,i,r)} \|r_{ui} - \tilde{r}_{ui}\|^2$

Perf **MSE**	ϕ_0		$\phi_2(i)$	$\phi_3(u,i)$
RB_U50_I200	0,67575			0,19776
RB_U500_I2k	0,56850			0,22377
RB_U5k_I20k	0,67744			0,28466
RB_U30k_I110k	0,70296			0,33157
A_U2k_I1k	1,53155			1,21357
A_U20k_I12k	1,47107			1,21267
A_U210k_I120k	1,50721			1,29709
A_U2M_I1M	1,60510			1,48153

Poussevin, Guigue, Gallinari, CORIA 2015

Extraction d'un vocabulaire de surprise par combinaison de recommandation et d'analyse de sentiments

Résultats: Amazon/Ratebeer

Prédiction optimale:

$$\tilde{r}_{ui} = \underbrace{\phi_0}_{\text{biais gén.}} + \underbrace{\phi_1(u)}_{\text{biais u}} + \underbrace{\phi_2(i)}_{\text{biais i}} + \underbrace{\phi_3(u,i)}_{\text{NMF}}$$

Evaluation: $\frac{1}{N} \sum_{(u,i,r)} \|r_{ui} - \tilde{r}_{ui}\|^2$

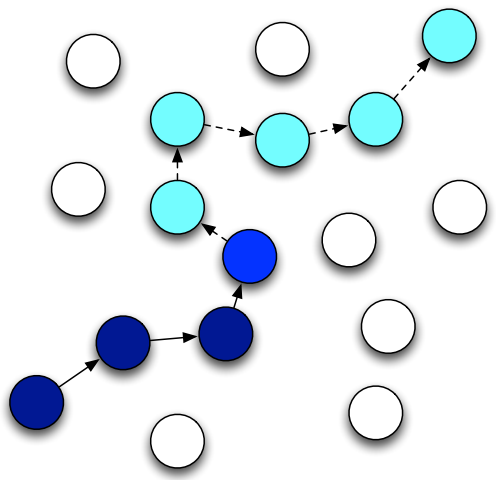
Perf **MSE**	ϕ_0	$\phi_1(u)$	$\phi_2(i)$	$\phi_3(u,i)$
RB_U50_I200	0,67575	0,65325	0,20913	0,19776
RB_U500_I2k	0,56850	0,52563	0,25089	0,22377
RB_U5k_I20k	0,67744	0,58782	0,30791	0,28466
RB_U30k_I110k	0,70296	0,60644	0,34876	0,33157
A_U2k_I1k	1,53155	1,30432	1,27850	1,21357
A_U20k_I12k	1,47107	1,28584	1,23608	1,21267
A_U210k_I120k	1,50721	1,44538	1,32229	1,29709
A_U2M_I1M	1,60510	1,63127	1,49281	1,48153

Poussevin, Guigue, Gallinari, CORIA 2015

Extraction d'un vocabulaire de surprise par combinaison de recommandation et d'analyse de sentiments

Exploitation temporelle

Est-on capable de coder la dynamique de l'utilisateur dans l'espace latent?



- Analyse par fenêtre
- Modélisation des tendances dans les biais
- Système de recommandation à niveau d'expérience

L. Baltrunas *et al.*, CARS, 2009
Towards time dependant recommendation based on implicit feedback.

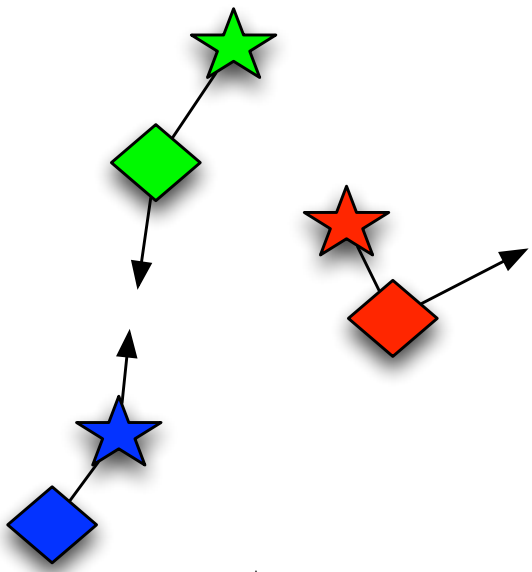
Y. Koren, SIGKDD, 2009
Collaborative filtering with temporal dynamics.

J. McAuley, Leskovec WWW, 2013,
From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews

Prédiction des déplacements dans l'espace latent

Trace utilisateur: user = séquence d'item

Plus de notion de temps, seulement d'ordre



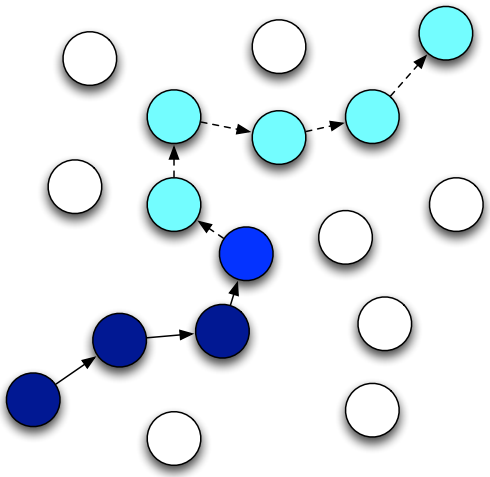
- $\theta_u = \{i_0, \ldots, i_t, \ldots, i_T\}$: **utilisateur = séquence d'items**
- $\mathbf{z}_i \in \mathbb{R}^d$: **profil items = position**
- Dédoublement des items: entrée $\mathbf{z}'$ / sortie $\mathbf{z}$
- Apprentissage $\approx$ *negative sampling* + gradient

$$Z^\star = \arg\min_{Z,Z'} \sum_u \sum_{i_t \in \theta_u, j \notin \theta_u} (||\mathbf{z}'_{i_t} - \mathbf{z}_j|| - ||\mathbf{z}'_{i_{t+1}} - \mathbf{z}_{i_t}||)$$

E. Guàrdia-Sebaoun, Guigue, Gallinari 2014, MARAMI, Recommandation Dynamique dans les Graphes Géographiques

S. Chen *et al.*, 2012, ACM SIGKDD, Playlist prediction via metric embedding.

Personnaliser la dynamique



1. Word2Vec sur les traces d'utilisateurs $\Rightarrow Z_i$
2. Regression linéaire pour chaque utilisateur

Deux tâches:

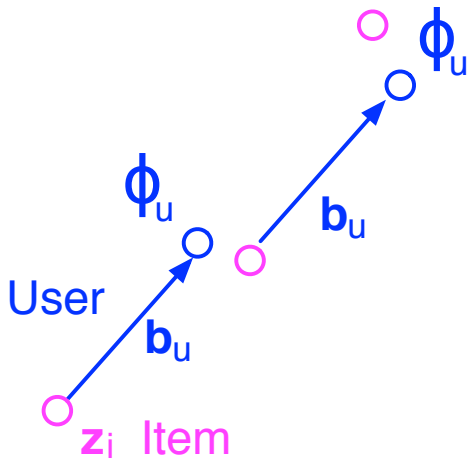
- Prédiction de l'item suivant
- Amélioration de la prédiction de note

Dire à un utilisateur:

Je pense que tu va aller voir ça... Mais c'est une mauvaise idée, tu ferais mieux d'aller de ce coté là.

Guàrdia-Sebaoun, Guigue, Gallinari, RecSys 2015,
Latent Trajectory Modeling: A Light and Efficient Way to Introduce Time in Recommender Systems

Personnalisation de la prédiction *next item*



- $\theta_u = \{i_0, \ldots, i_t, \ldots, i_T\}$: **utilisateur**
- $z_i \in \mathbb{R}^d$:**item**
- $b_u \in \mathbb{R}^d$: **profil utilisateur = direction** dans l'espace latent
- $A_c \in \mathbb{R}^{d \times d}$: **profil communauté = métrique** de l'espace latent
- $\phi_u(z_i)$: **prédicteur personnalisé** pour le prochain item de la trace de u.

(1) Modèle TRANS : $\phi_u = z_i + b_u$

(2) Modèle COMM : $\phi_u = A_c.z_i + b_u$

Guàrdia-Sebaoun, Guigue, Gallinari, RecSys 2015,
Latent Trajectory Modeling: A Light and Efficient Way to Introduce Time in Recommender Systems

Mélange des représentations latentes pour la prédiction de notes

- **Initialisation** Profils enrichis:

$$\gamma_u = \begin{bmatrix} \bar{\gamma}_u \\ b_u \end{bmatrix} \in \mathbb{R}^{2d}, \quad \gamma_i = \begin{bmatrix} A_{c_u} z_i \\ \bar{\gamma}_i \end{bmatrix} \in \mathbb{R}^{2d}$$

$$\begin{aligned} \textit{noir} &= \textit{rnd init.} + \textit{opti.} \\ \textit{rouge} &= \textit{constant} \end{aligned}$$

- **Critère** Moindres carrés:

$$\bar{\gamma}^\star = \arg\min_{\bar{\gamma}} \sum_{u \in U} \sum_{(i,r) \in \theta_u} \left[\mu + \mu_u + \mu_i + \langle \gamma_u, \gamma_i \rangle - r\right]^2 + \lambda \Omega\left(\bar{\gamma}\right)$$

- **Inférence** $\hat{r}_{u,i} = \mu + \mu_u + \mu_i + \langle \gamma_u, \gamma_i \rangle$

Résultats : des films et des bières

Cinq datasets, deux catégories:

- Bières : BeerAdvocate, Ratebeer.
- Cinéma : MovieLens-10m, Flixster, Amazon-Movies.

Dataset	# items	# users	# ratings
BeerAdvocate	66051	33387	1586259
RateBeer	110419	40213	2924127
MovieLens	10000	72000	10000000
Flixster	49000	1000000	8200000
Movies	253059	889176	7911684

Prédiction d'items :

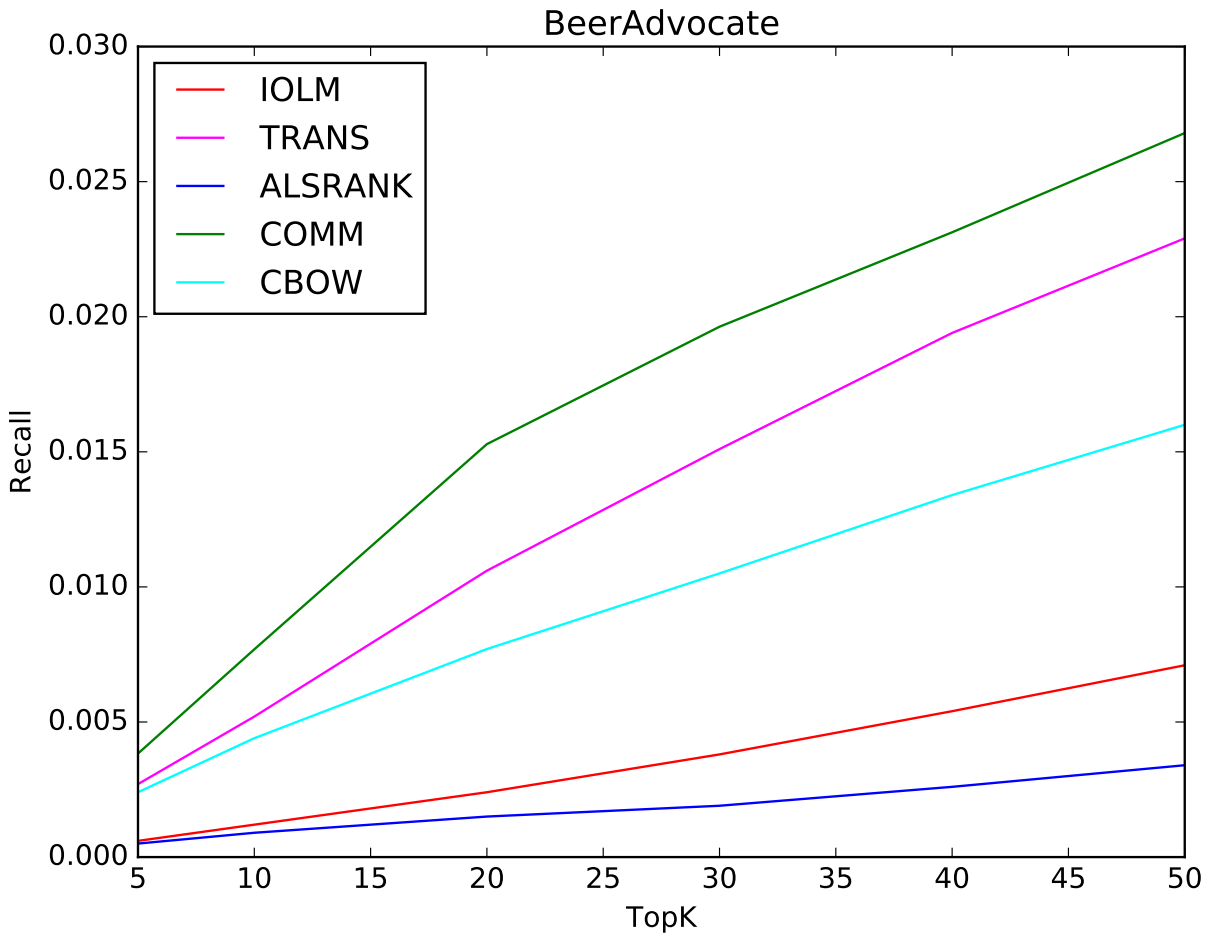
- Divers essais (Rang moyen, précision@K).
- Rappel@k = bon compromis.

Prédiction de notes :

- Erreur quadratique moyenne (MSE)

Résultats : des films et des bières

Prédiction du prochain item:



Résultats : des films et des bières

Table: Résultats en prédiction de notes, exprimés en MSE pour les modèles MF, TSVD, EXP, TRANS et COMM.

Dataset	MF	TSVD	EXP	TRANS	COMM
BeerAdvocate	0.4	0.381	0.367	**0.361**	**0.360**
RateBeer	0.331	0.301	0.297	**0.279**	**0.279**
MovieLens	0.691	0.681	0.684	0.663	**0.660**
Flixster	0.912	0.867	0.827	0.816	**0.811**
Movies	1.377	1.211	1.05	**0.913**	**0.913**

Conclusion

Recommandation =

- apprentissage de représentations...
- ... sur des données *hétérogènes* (mais que des utilisateurs et des items)
- ... dédiées à 1 tâche

NMF =

- Algo efficace, plein d'implémentations disponibles
- Algo flexible
- Nombreux domaines applicatifs

Perspective:

- Multiplier les tâches et les types de données

Toujours le même domaine applicatif...

Utiliser le **texte** dans les systèmes de recommandation...

[*Approches content based*]

... mais conserver le cadre du **filtrage collaboratif**

Et répondre au **démarrage à froid**!

Exploitation des revues d'utilisateurs :

- Segmentation thématique des textes
- Apprentissage de différents aspects

Ganu et al., WebDB 2009
Beyond the Stars: Improving Rating Predictions using Review Text Content

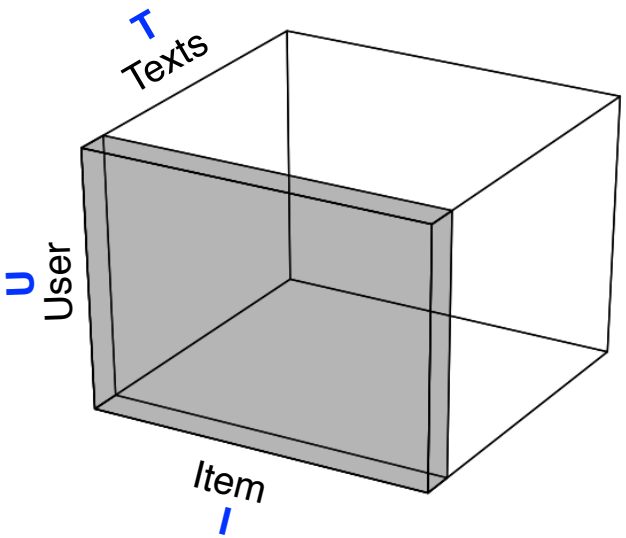
- LDA = profils textuels $\mathbf{z}_u, \mathbf{z}_i$
- Filtrage collaboratif = profils $\mathbf{z}'_u, \mathbf{z}'_i$
- Alignement $\mathbf{z}/\mathbf{z}'$

McAuley & Leskovec, RecSys 2013
Hidden factors and hidden topics: understanding rating dimensions with review text

Approche unifiée : factorisation tensorielle

Représentation *naturelle* d'une revue

- Utilisateur, Item, Texte ⇒ note



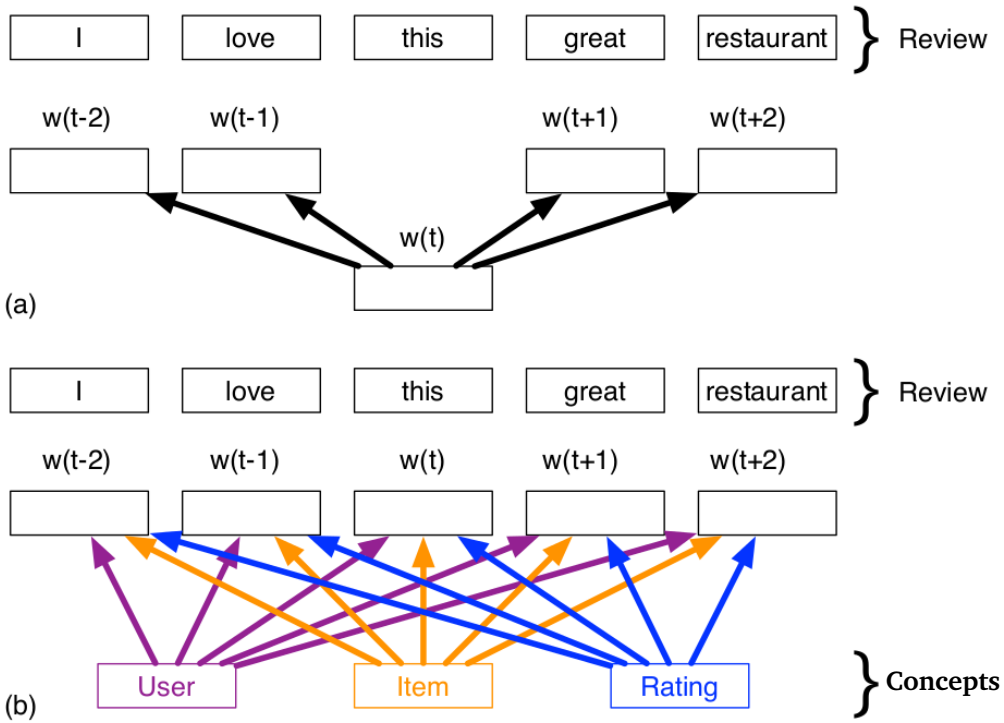
- Représentation simpliste
  - les cases codent des notes ou des présences
- Parcimonie
- Choix d'une technique de factorisation

⇒ Pas de résultats concluants

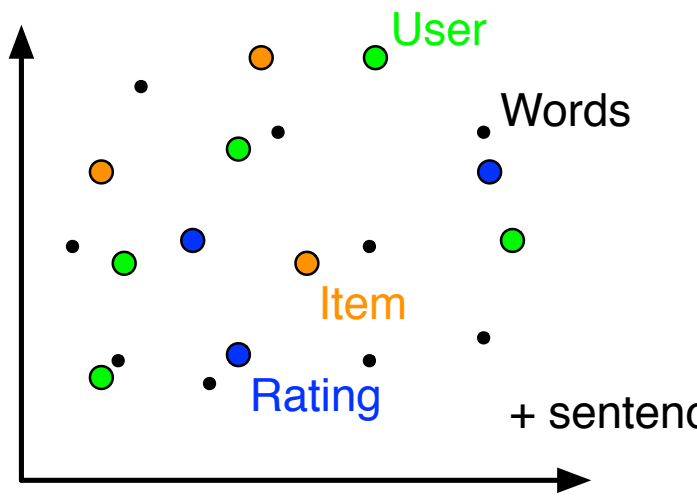
Usage de word2vec

Le texte comme socle pour un espace unifié

Contextual Skip-Gram



Unified vector space:



Dias, Guigue, Gallinari, CORIA 2016
Recommandation et analyse de sentiments dans un espace latent textuel

Exploitation des représentations

Navigation par plus proches voisins:

Autour d'une bière:

Bière n°12

Bière n°34

Bière n°227

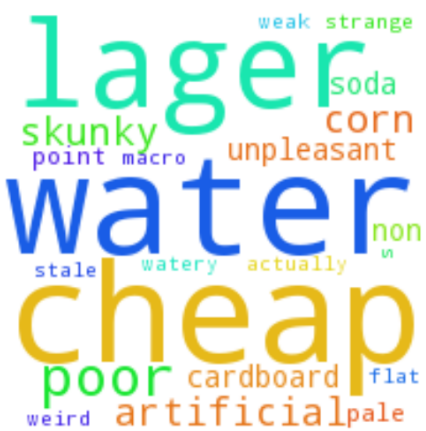
Bière n°254

Exploitation des représentations

Navigation par plus proches voisins:

Autour d'une note:

1 star



2 star

3 star

5 star

Exploitation des représentations

Navigation par plus proches voisins:

Estimer une note:

quelles notes u a-t-il données à des produits proches de i?
quelles notes les utilisateurs proches de u ont-ils données à i?

$$\hat{r_{ui}} = \underbrace{\mu_u + \frac{\sum_{v \in \mathcal{N}} \alpha_{uv}(r_{vi} - \mu_v)}{\sum_{v \in \mathcal{N}} \alpha_{uv}}}_{\text{user-user}}, \quad \hat{r_{ui}} = \underbrace{\mu_i + \frac{\sum_{j \in \mathcal{N}} \alpha_{ij}(r_{uj} - \mu_j)}{\sum_{j \in \mathcal{N}} \alpha_{ij}}}_{\text{item-item}}$$

Dataset	Moyenne	Fact. Mat	HFT*	CSG-kNN		k*
				util.	prod.	
Ratebeer	0.701	0.306	0.301	0.336	**0.286**	23
Beeradv.	0.521	0.371	0.366	0.382	**0.357**	29
Movies	1.678	**1.118**	1.119	1.39	1.304	33
Music	1.261	**0.957**	0.969	-	1.201	26
Yelp	1.890	1.49	-	1.591	**1.407**	27

Démarrage à froid

- Espace = appris comme précédemment
- Nouvel utilisateur = 50% des textes (seuls) pour la projection

Dataset	Moy.	Nouvel Utilisateur		Nouveau Produit	
		Moy. Prod.	CSG-kNN	Moy. Util.	CSG-kNN
Ratebeer	0.701	0.341	**0.333**	0.599	**0.371**
Beeradv.	0.518	0.397	**0.386**	0.490	**0.419**

Explication de la recommandation

Vérité terrain

Note: 5.0
Commentaire: "Old school. traditional mom n'pop quality and perfection. The best fish and chips you ll ever enjoy and equally superb fried shrimp. A great out of the way non corporate vestige of Americana. You will love it."

Prédiction:

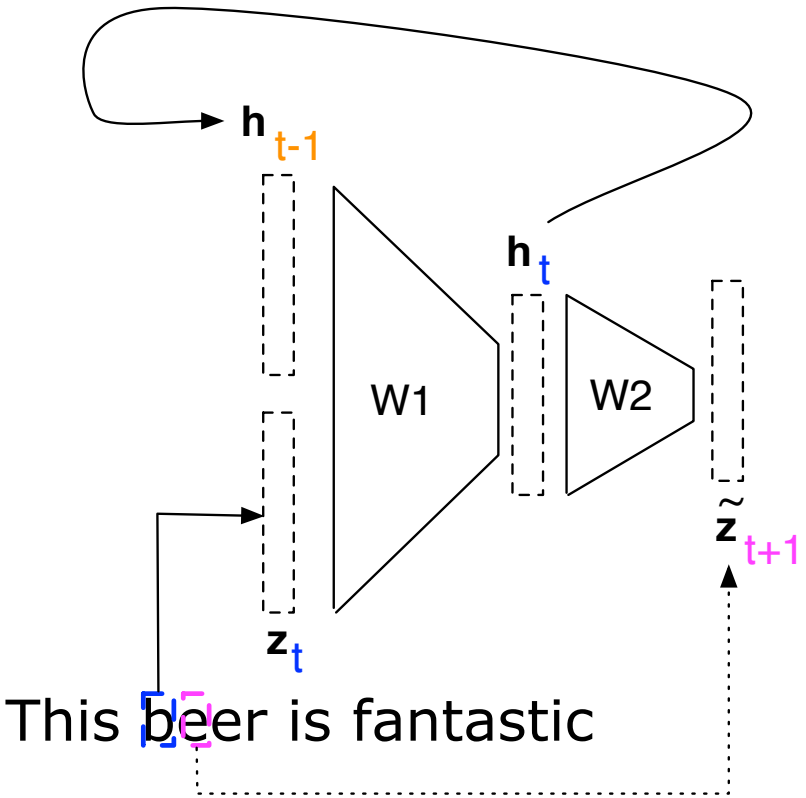
Note: 4.70
Avis complet: "Good fish sandwich"
Choix de n phrases: "The staff is extremely friendly. On top of being extremely large portions it was incredibly affordable. Most of girls are good one is very slow one is amazing. The fish was very good but the Reuben was to die for. Both dishes were massive and could very easily be shared between two people."

Figure: Exemple de prédiction de note et d'avis sur le corpus Yelp

Quid de l'évaluation?... ROUGE...

Approches par RNN

Karpathy ⇒ Générer du texte = Utiliser des RNN



Neurones exotiques / gestion de la mémoire:

- GRU
- LSTM

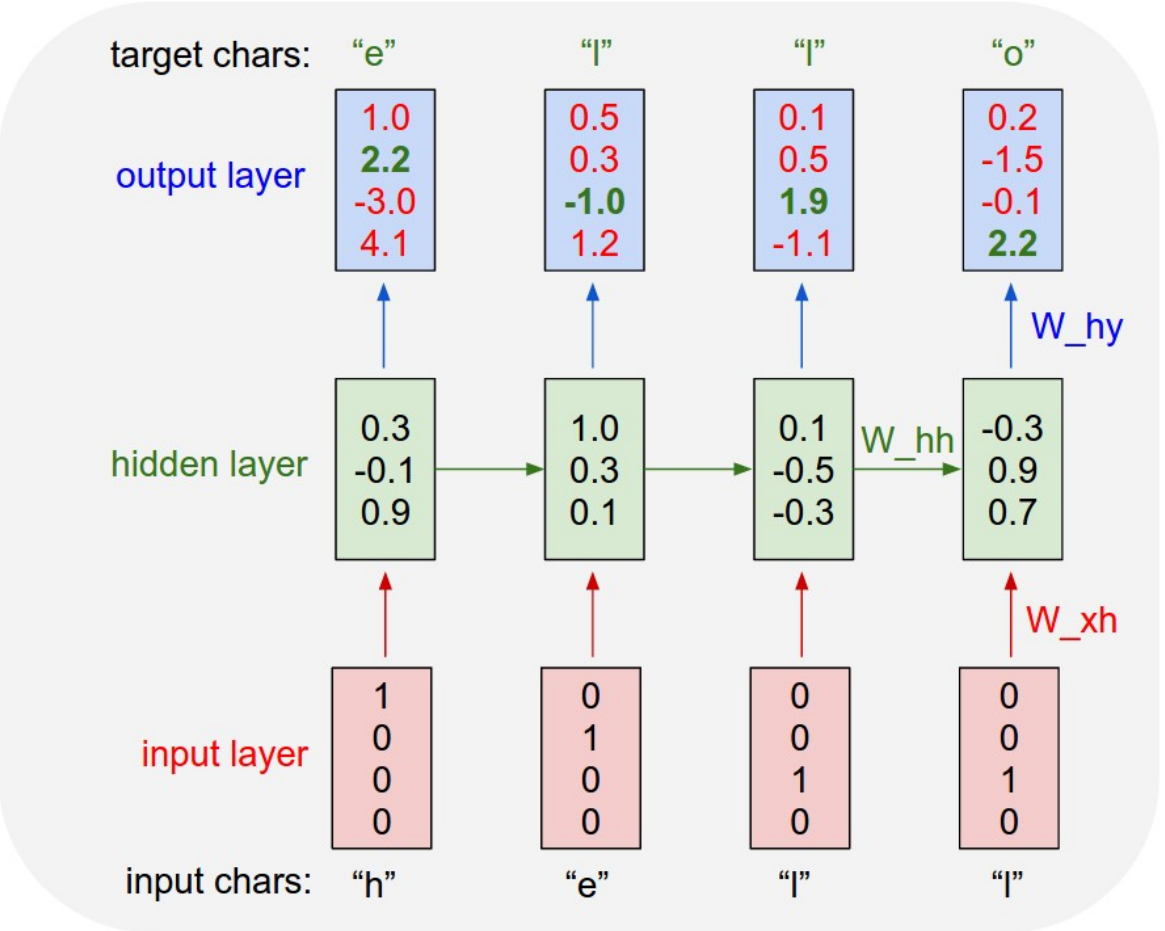
http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Inférence *beam search*: maximiser la probabilité de la séquence.

Andrej Karpathy blog, 2015
The Unreasonable Effectiveness of Recurrent Neural Networks

Lipton, Vikram, McAuley, arXiv, 2016
Generative Concatenative Nets Jointly Learn to Write and Classify Reviews

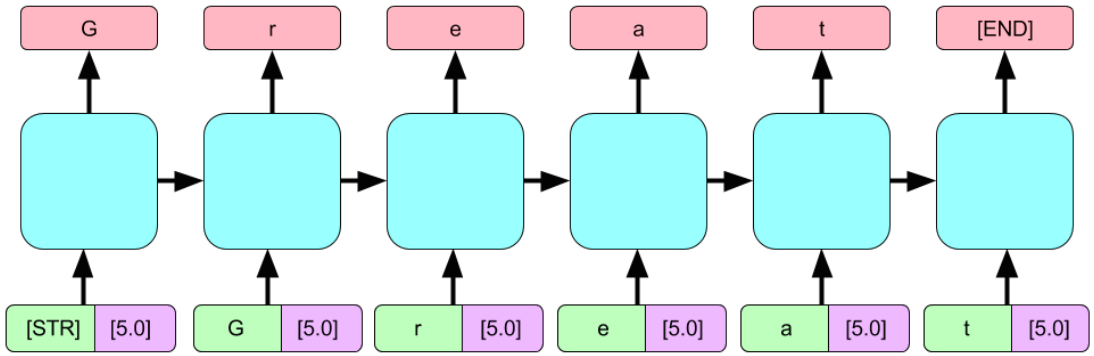
Approches par RNN



Andrej Karpathy blog, 2015
The Unreasonable Effectiveness of Recurrent Neural Networks

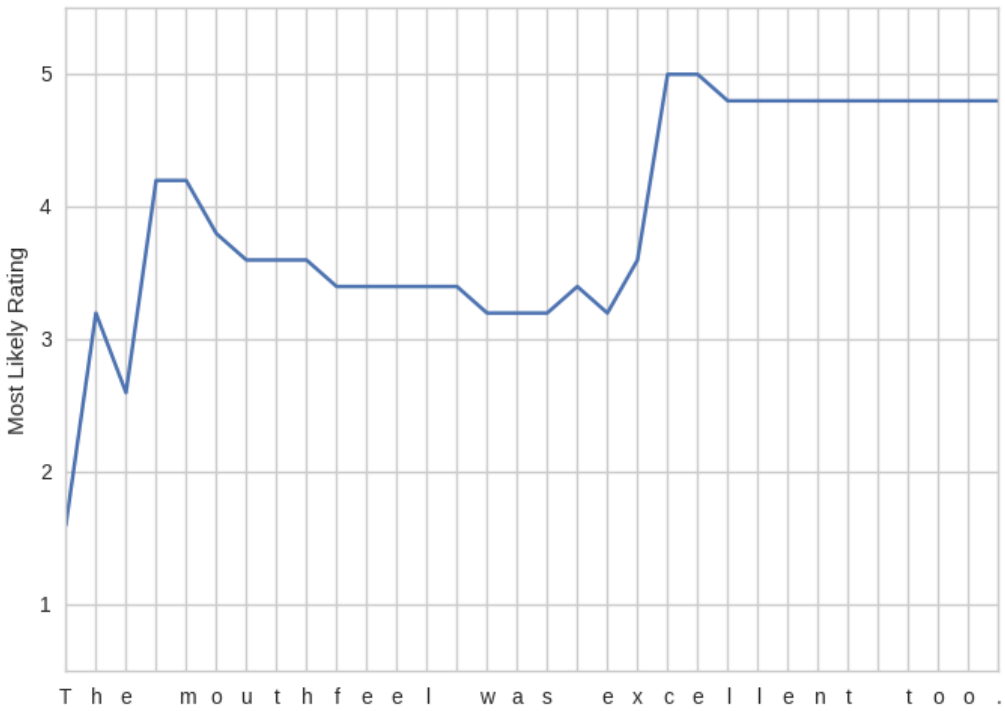
Lipton, Vikram, McAuley, arXiv, 2016

RNN & modélisation du contexte



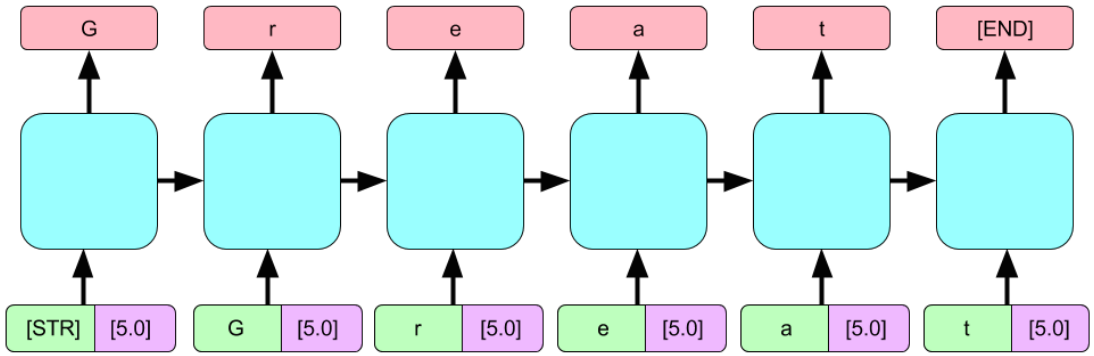
- Concatenation: input + contexte
- Inférence = test (combinatoire) des contextes + max de vraisemblance

Démo: `http://deepx.ucsd.edu/`



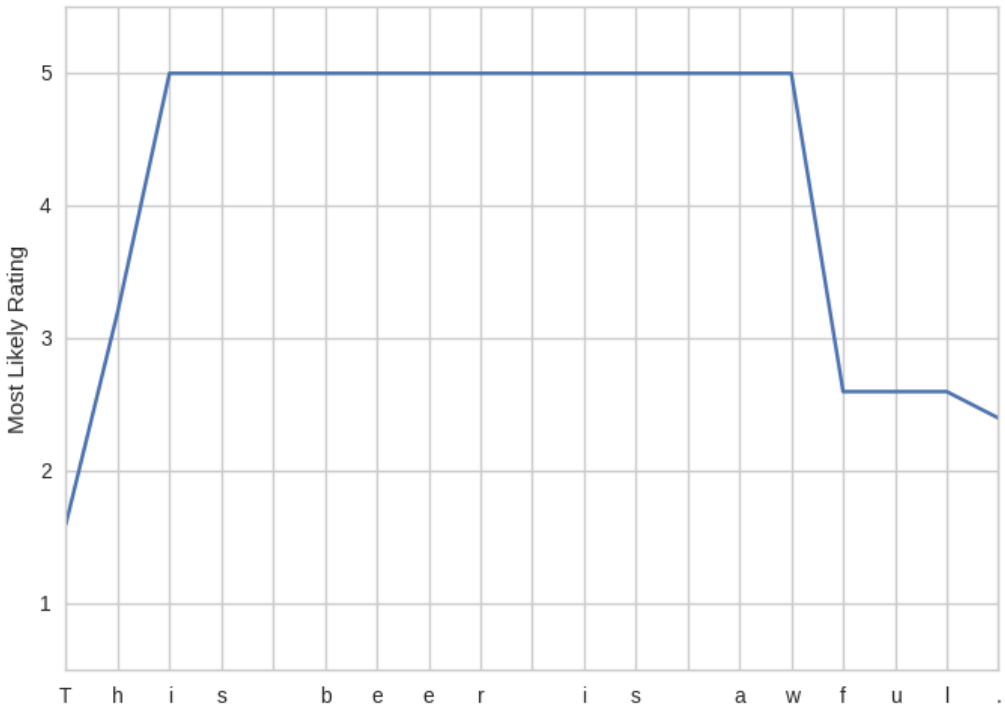
Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews

RNN & modélisation du contexte



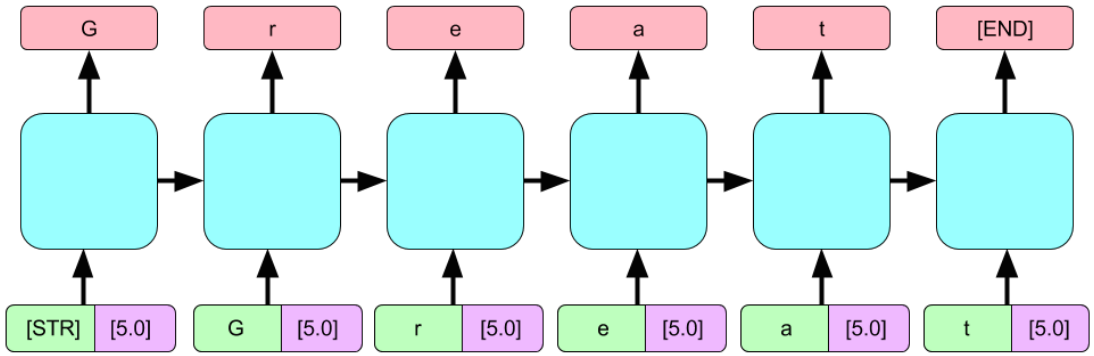
- Concatenation: input + contexte
- Inférence = test (combinatoire) des contextes + max de vraisemblance

Démo: `http://deepx.ucsd.edu/`



Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews

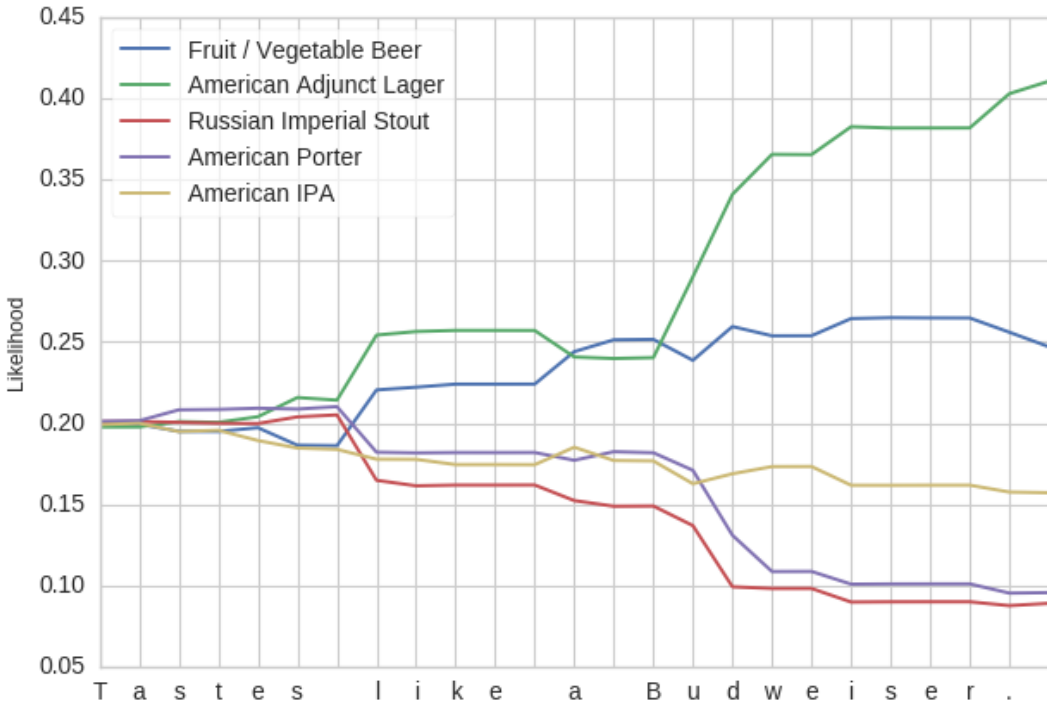
RNN & modélisation du contexte



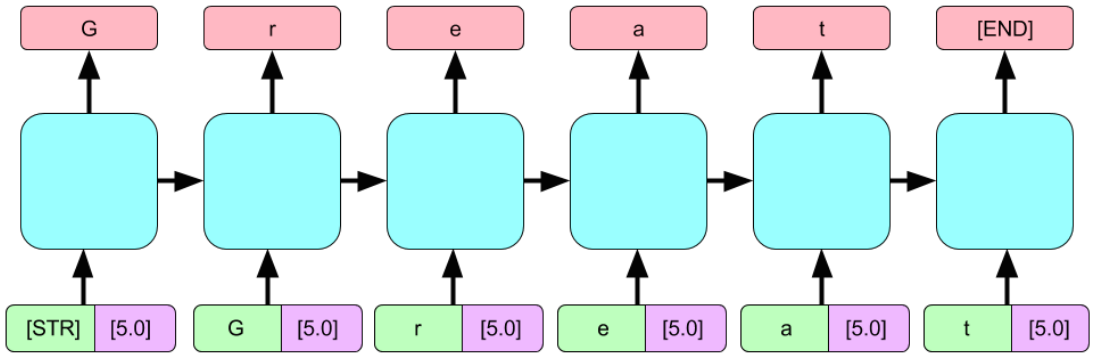
- Concatenation: input + contexte
- Inférence = test (combinatoire) des contextes + max de vraisemblance

Démo: `http://deepx.ucsd.edu/`

Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews



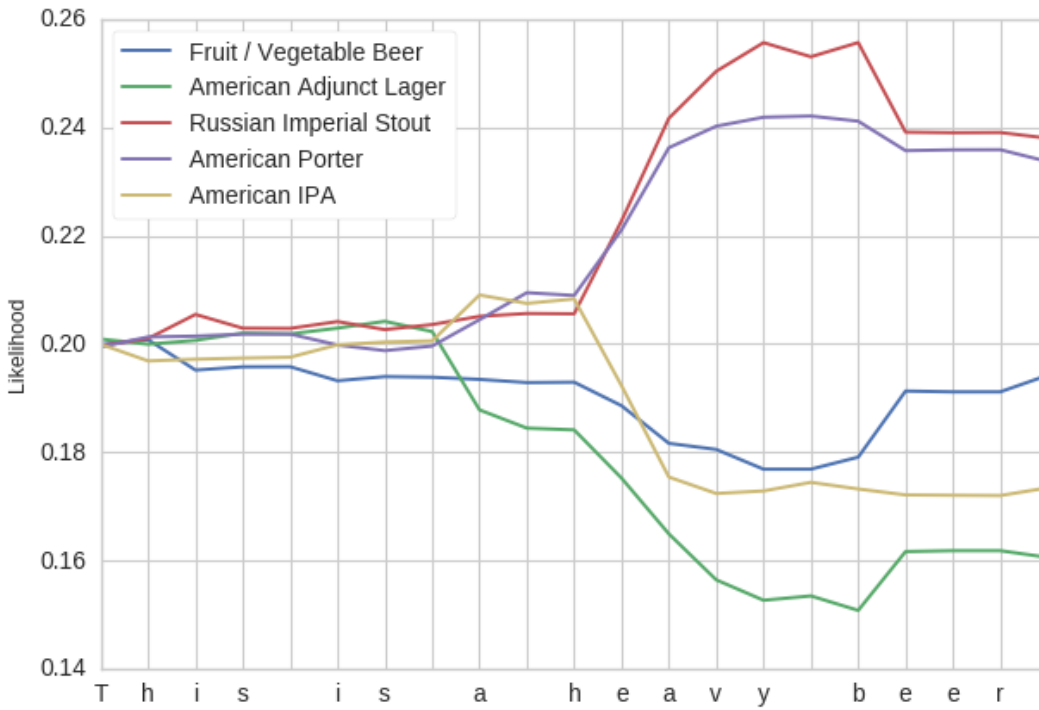
RNN & modélisation du contexte



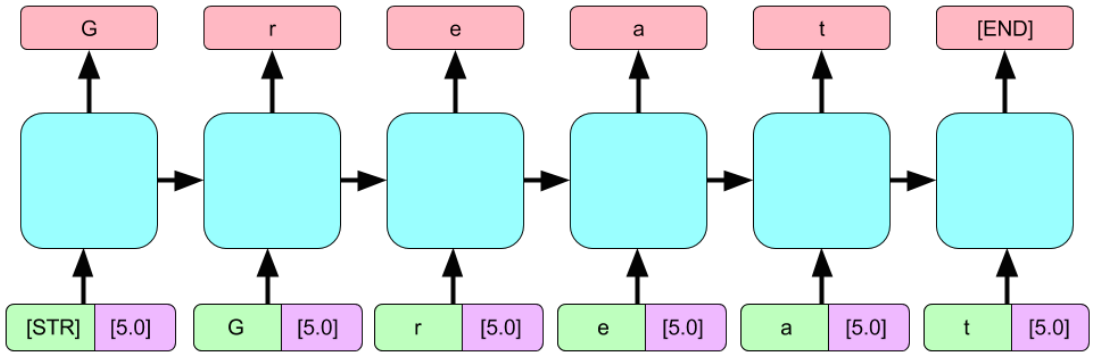
- Concatenation: input + contexte
- Inférence = test (combinatoire) des contextes + max de vraisemblance

Démo: `http://deepx.ucsd.edu/`

Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews



RNN & modélisation du contexte



- Concatenation: input + contexte
- Inférence = test (combinatoire) des contextes + max de vraisemblance

Démo: `http://deepx.ucsd.edu/`

Lipton, Vikram, McAuley, arXiv, 2016
Gen. Concatenative Nets Jointly Learn to Write and Classify Reviews

Perspective:

- Authorship
- $\Rightarrow$ Réduire la combinatoire...
- ... Ou traiter des contextes continus

Formulation Générale du problème

≈ GLVQ / semi-supervisé / données manquantes

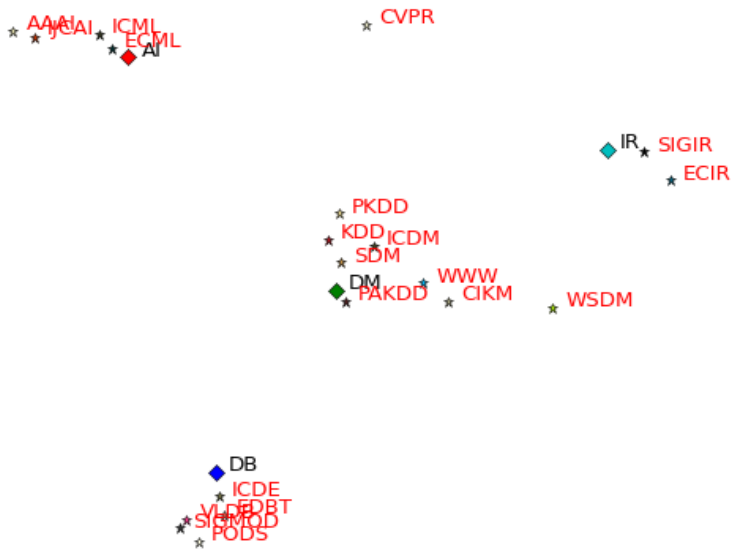
- Description des **données** = ensemble de **contraintes**
- Hypothèse : espace des représentations **homogène localement**
- Optimisation possible pour différentes tâches
- Classif. dans les graphes/ supervisée / non-supervisée / semi-supervisée

$$\mathcal{L}(f, Z) = \underbrace{\sum_{i \in S_\ell} \Delta(f(\mathbf{z}_i), y_i)}_{\text{Cost wrt labels}} + \lambda \underbrace{\sum_{i,j/w_{i,j} \neq 0} w_{i,j} \|\mathbf{z}_i - \mathbf{z}_j\|^2}_{\text{Link/cooc = constraints}}$$

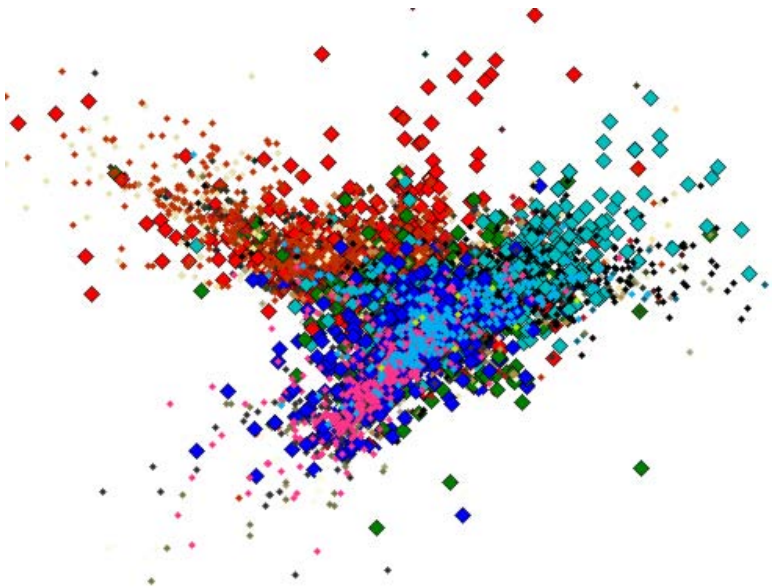
Jacob, Denoyer, Gallinari, WSDM 2014
Learning Latent Representations of Nodes for Classifying in Heterogeneous Social Networks

Résultats sur DBLP

- Contenu des articles,
- Co-auteurs
- Citations



(a) Centroid of different authors and papers labels (unlabeled nodes)



(b) All nodes

⇒ Apport : **liens hétérogènes** ≠ métrique unique

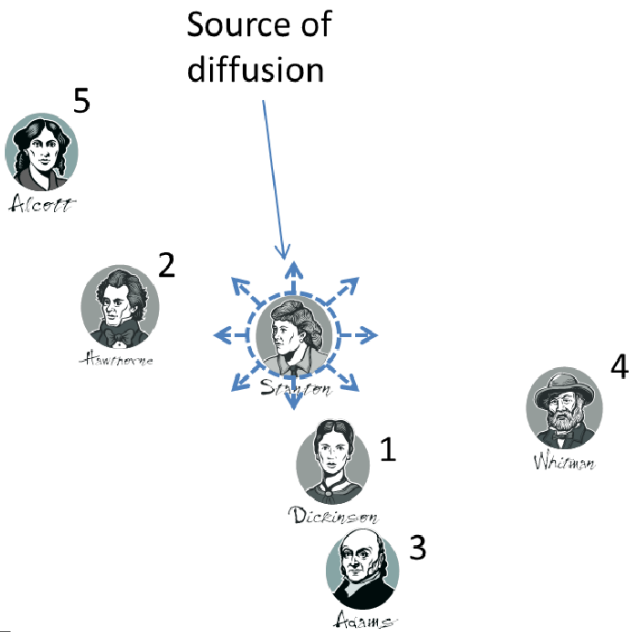
Applicable à d'autres domaines applicatifs

Diffusion de l'information dans les réseaux sociaux

Qui sont les **influenceurs**?

Ceux qui *émettent* l'information **en premier** dans une cascade

Apprentissage de représentation = **simplification** d'un problème difficile + **robustesse**



- Point chaud + diffusion de la chaleur

Bourigault, Lagnier, Lamprier, Denoyer, Gallinari, WSDM 2014
Learning social network embeddings for predicting information diffusion

Applicable à d'autres domaines applicatifs

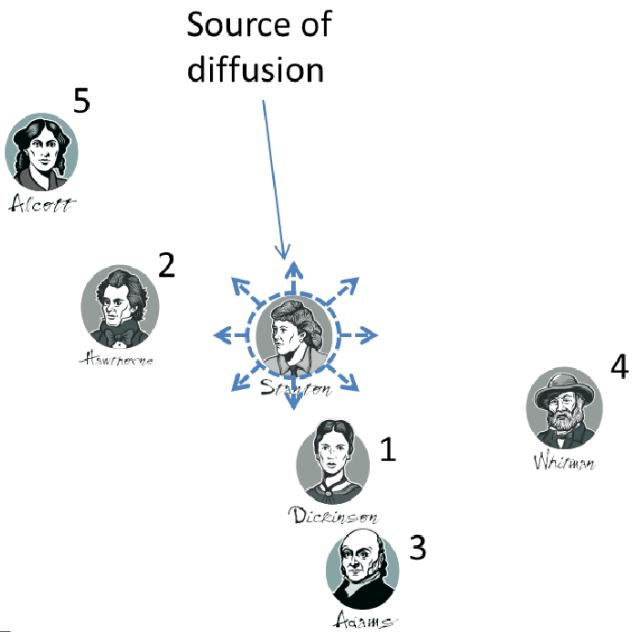
Diffusion de l'information dans les réseaux sociaux

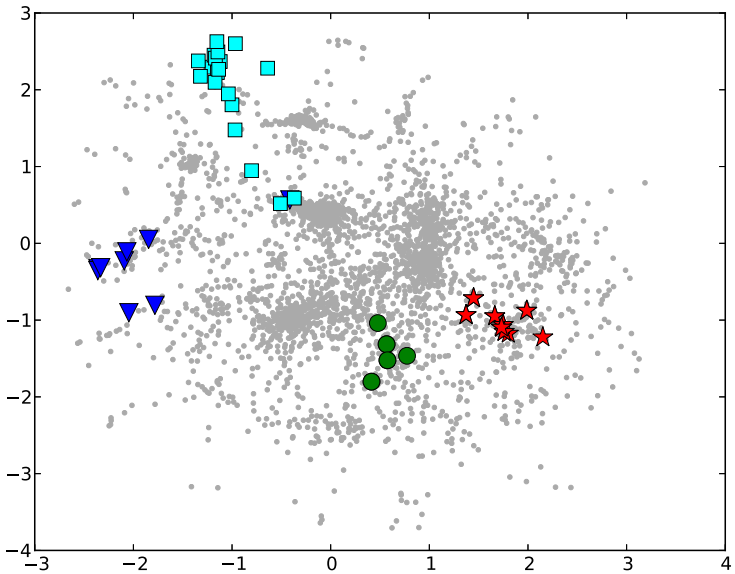
Qui sont les **influenceurs**?

Ceux qui *émettent* l'information **en premier** dans une cascade

Apprentissage de représentation = **simplification** d'un problème difficile

\+ **robustesse**





Bourigault, Lagnier, Lamprier, Denoyer, Gallinari, WSDM 2014

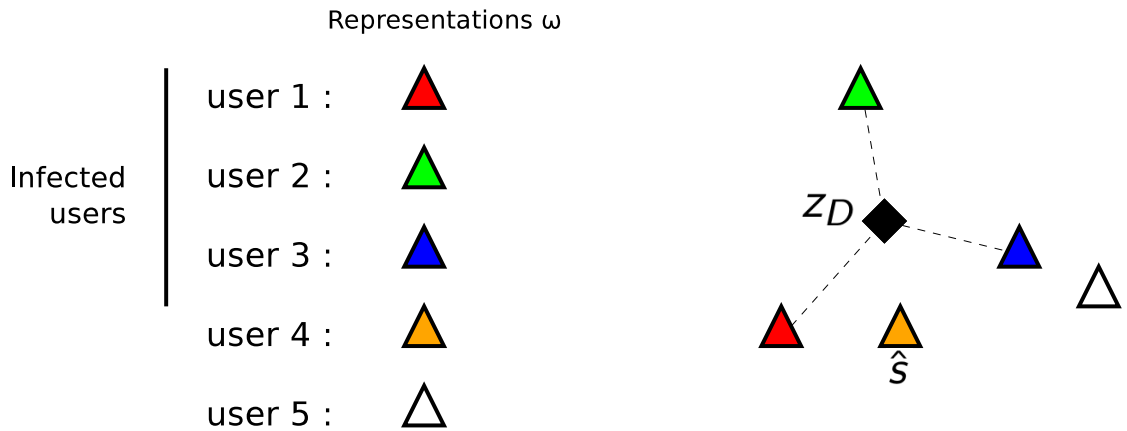
Learning social network embeddings for predicting information diffusion

Diffusion (suite)

Légèreté + robustesse...

⇒ **Complexification** du modèle + **nouvelles dimensions**

- Représentation latente des utilisateurs



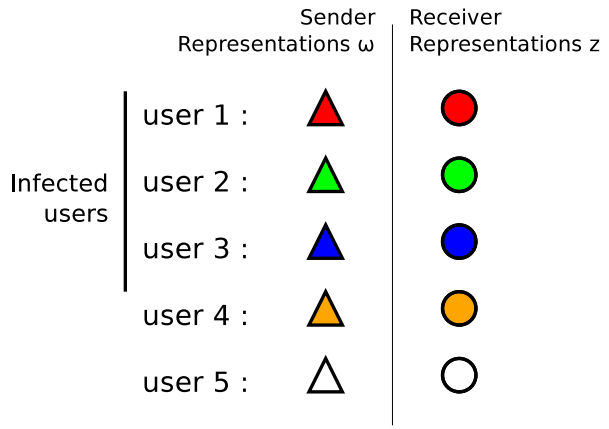
Bourigault, Lamprier, Gallinari, ECML 2016
Learning Distributed Representations of Users for Source Detection in Online Social Networks

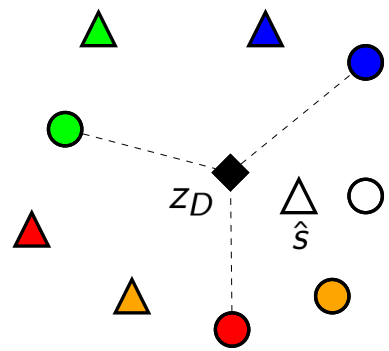
Diffusion (suite)

Légèreté + robustesse...

⇒ **Complexification** du modèle + **nouvelles dimensions**

- Représentation latente des utilisateurs
- Receveur *vs* Emetteur





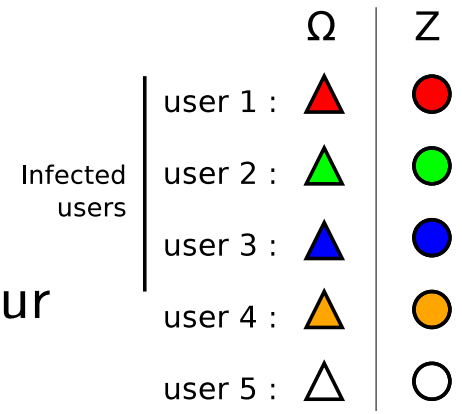
Bourigault, Lamprier, Gallinari, ECML 2016
Learning Distributed Representations of Users for Source Detection in Online Social Networks

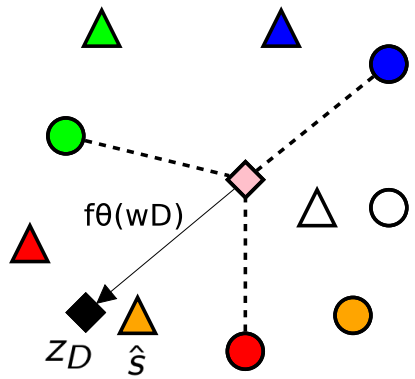
Diffusion (suite)

Légèreté + robustesse...

⇒ **Complexification** du modèle + **nouvelles dimensions**

- Représentation latente des utilisateurs
- Receveur *vs* Emetteur
- Aspect thématique





Bourigault, Lamprier, Gallinari, ECML 2016
Learning Distributed Representations of Users for Source Detection in Online Social Networks

Gestion de données hétérogènes

Espaces latents / apprentissage de représentation:

- adapté à la plupart des problématiques...
  - apprentissage supervisé classique,
  - semi-supervisé,
  - dans les graphes / réseaux sociaux ⇒ dépasser l'hypothèse *petit monde*
  - découverte de lien, influence...
- ... + de nouvelles perspectives applicatives
- un moyen élégant et léger

Perspectives:

Bonne gestion des données hétérogènes...

... Mais avec une métrique unique

⇒ vers le **multi-relationnel**

Raisonner = cadre multi-relationnel

Multi-tâches + pondération des types de liens :

$$\mathcal{L}(Z) = \underbrace{\sum_{i,k} \Delta_k(f_k(\mathbf{z}_i), y_i^k)}_{\text{Costs wrt labels}} + \lambda \underbrace{\sum_{i,j/w_{i,j}\neq 0} w_{i,j}\|\mathbf{z}_i - \mathbf{z}_j\|^2}_{\text{Link/cooc = constraints}}$$

Structuration des tâches / contraintes

$$\mathcal{L}(Z) = \sum_{i,j,k} \Delta_k(f_k(\mathbf{z}_i, \mathbf{z}_j), y_i)$$

Jacob, Denoyer, Gallinari, WSDM 2014
Learning Latent Representations of Nodes for Classifying in Heterogeneous Social Networks

Dos Santos, Piwowarsky, Gallinari, ECML 2016
Multilabel classification on heterogeneous graphs with gaussian embeddings

Raisonner, pour nous c'est...

Transformer la recommandation en un problème d'opération simple dans les espaces latents

- Quel opérateur?
- Comment apprendre ensemble les **opérateurs** et les **représentations**

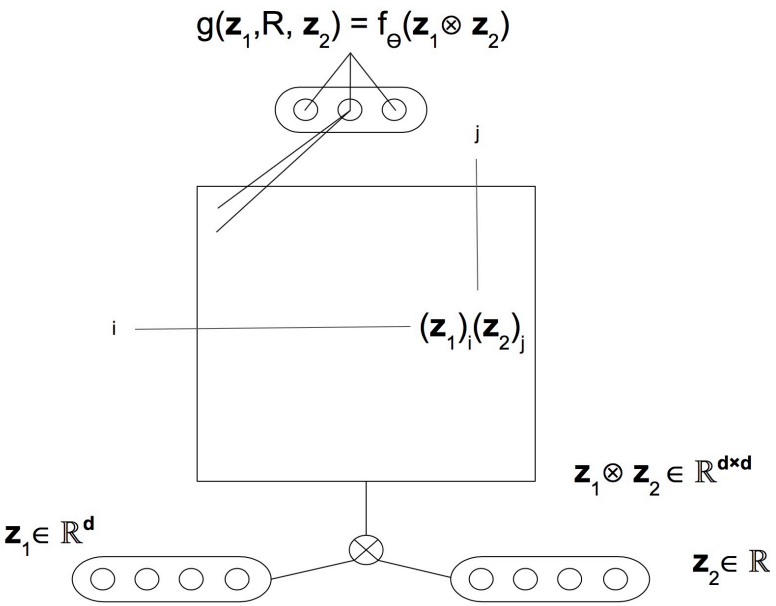
Cadre MovieLens

- $\mathbf{z} \in \mathbb{R}^d$
- $f_{age}(\mathbf{z}_u) \approx \mathbf{z}_{age}$, $f_{rate}(\mathbf{z}_u, \mathbf{z}_i) \approx \mathbf{z}_{rate} \dots$
- Apprendre $\mathbf{z}$ et f_k

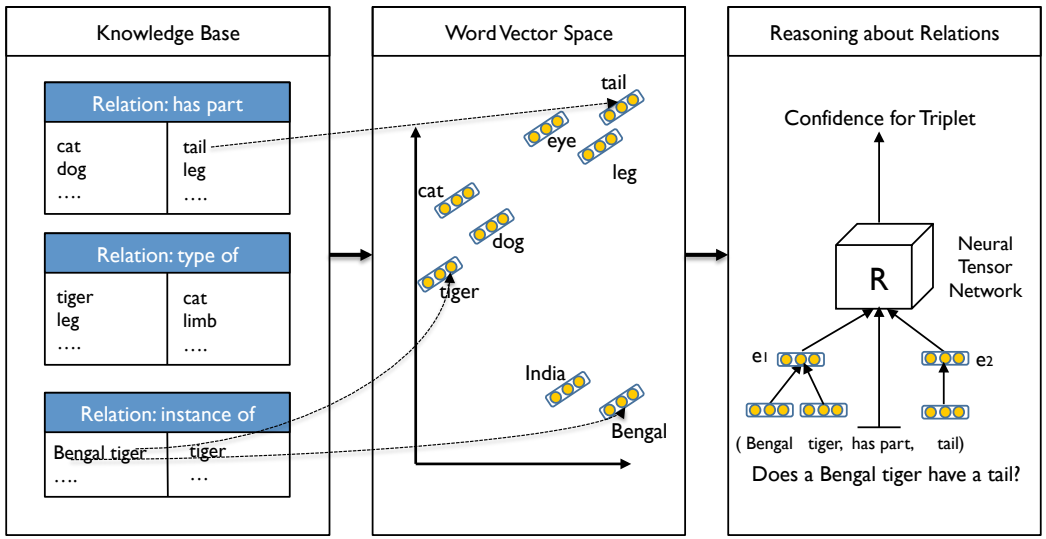
Sileo, Guigue, CAp 2016
Apprentissage relationnel pour la recommandation et la prédiction de données manquantes

Quelques conclusions préliminaires

- Echec des modèles additifs
  - La parcimonie n'a plus de signification...
- Bonne performance des modèles multiplicatifs



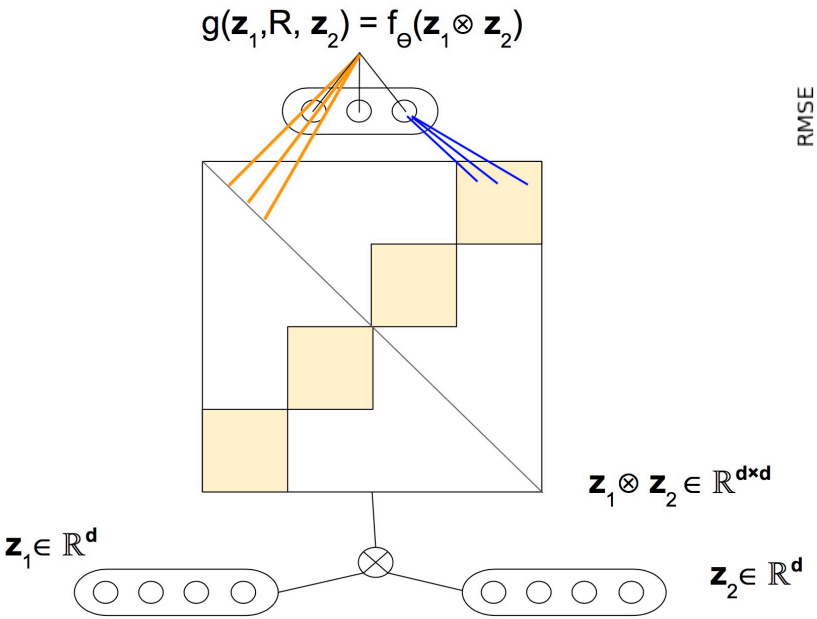
- Réduire d d'un facteur 10
- Trop de paramètre dans f

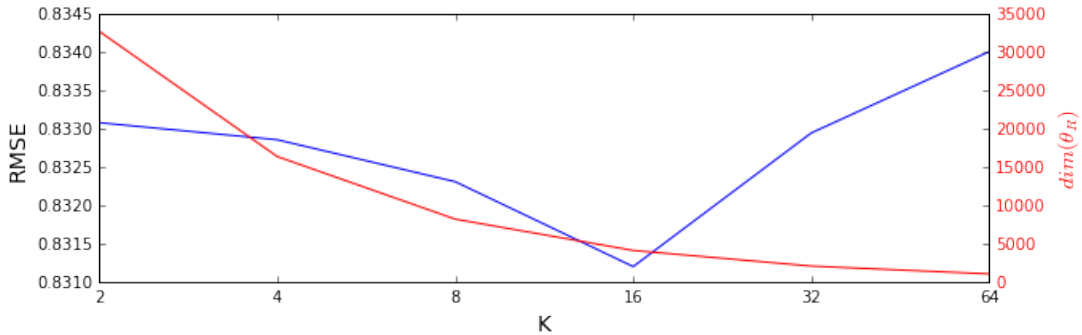


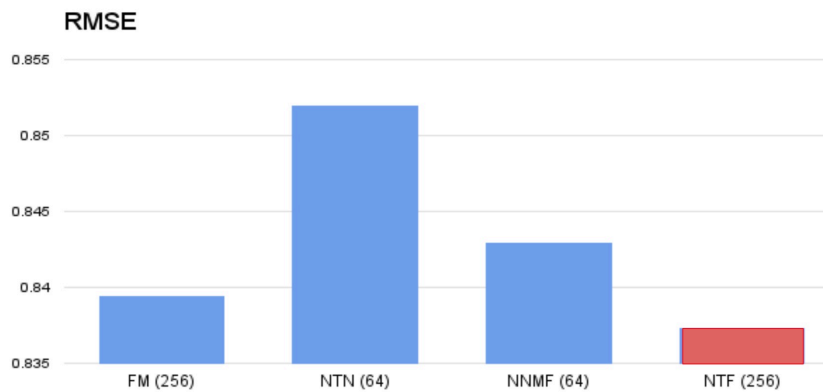
Socher, Chen, Manning, Ng, NIPS 2013
Reasoning With Neural Tensor Networks for Knowledge Base Completion

Résultats en recommandation

$$g(\mathbf{z}_1, R, \mathbf{z}_2) = f_\theta(\underbrace{s(\mathbf{z}_1 \otimes \mathbf{z}_2)}_{\text{Echantillonnage}}) + W \underbrace{(\mathbf{z}_1 \cdot \mathbf{z}_2)}_{\text{Reco classique}}$$





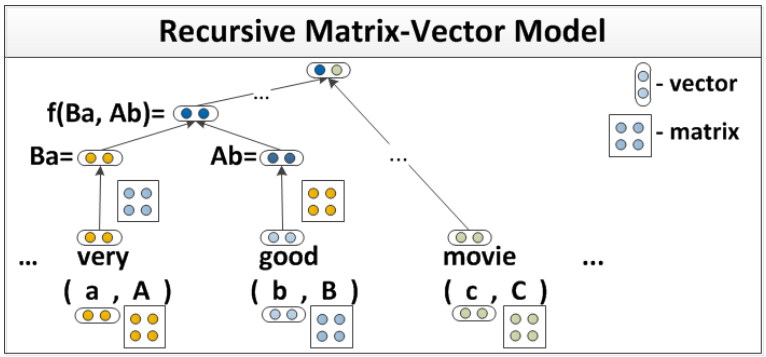


param. K = nb de blocs

Sileo, Guigue, CAp 2016
Apprentissage relationnel pour la recommandation et la prédiction de données manquantes

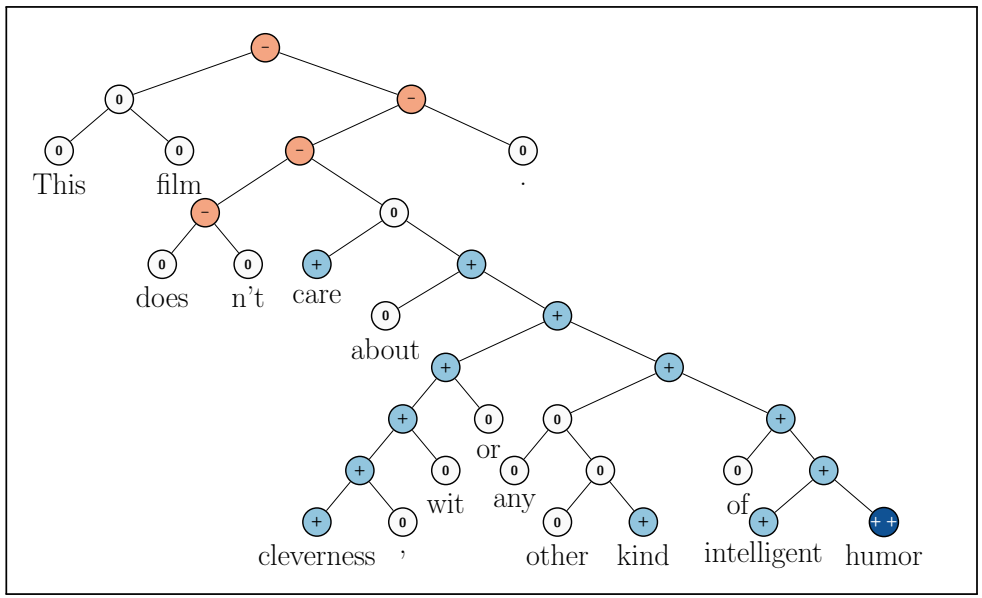
Résultats en classification de sentiments

Modèle:



- Critère de reconstruction
- Mot =
  - Représentation **z**
  - Matrice de transf. M

Résultats:



Socher et al., EMNLP 2013
Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Conclusion

Le *chatbot* de demain:

- Stocker des connaissances de manière légère
- Robuste au bruit
- Inférer les données manquante
- Répondre au question...
- ... tout en prenant en compte le profil de l'utilisateur
  - sur le plan thématique, de l'opinion...

Construire un profil utilisateur universel

- Transparent, réversible
- Manipulable par l'utilisateur