Вероятностное тематическое моделирование и нейросетевые модели языка для обучения векторного представления слов, контекстов и документов

Анна Потапенко

Научный руководитель: К.В. Воронцов

Специальность: 05.13.17, 2-ой год обучения

18 февраля 2016

Два направления исследований:

- Модели «мешка слов»
- Latent Semantic Analysis [Deerwester et al., 1990], Hyperspace Analogue to Language[Lund and Burgess, 1996]
- PLSA, LDA и десятки других тематических моделей [Hoffman, 1999; Blei et al., 2003]

- Модели «мешка слов»
- Нейросетевые модели языка [Bengio et al., 2003; Collobert and Weston, 2008; Mnih and Hinton, 2009]
- Word embeddings (векторные представления слов) [Mikolov et al., 2013, Pennington et al. 2014; Levy and Goldberg 2014]

Vector Space Models of Semantics [Pantel and Turney, 2010]

- обучаются по корпусу текстов (без учителя)
- улавливают близость слов, фраз и документов
- основаны на дистрибутивной гипотезе [Harris, 1954]

Приложения:

- IR, Q&A, классификация текстов, кластеризация и т.д.

Схема обработки: корпус сырых текстов $\rightarrow$ лингвистическая предобработка $\rightarrow$ матрица слова-документы или слова-слова $\rightarrow$ математическая предобработка (часто понижение размерности) $\rightarrow$ подсчет близости

Обучение скрытых векторных представлений:

- PISA, LDA, etc.: вероятности слов в темах ϕ_w и тем в документах θ_d. Ключевое свойство: интерпретируемость компонент.
- Word embeddings: вещественные вектора для слов ψ_w и контекстов ψ'_c (обычно также слов). Ключевое свойство: сохранение расстояний.

На основе предсказания статистик по корпусу:

- PLSA, LDA, etc.: $p(w|d)$
- CBOW: $p(w_j|w_{j-h}, \ldots, w_{j+h})$
- Skip-gram: $p(w_{j-h}, \ldots w_{j+h}|w_j)$
- Paragraph2vec: $p(w_j|w_{j-h}, \ldots, w_{j+h}, d)$

Вероятностные модели:

- PLSA (вероятность слова w в документе d):

$$p(w|d) = \sum_t p(w|t)p(t|d) = \langle \phi_w, \theta_d \rangle$$

- Skip-gram (вероятность слова w в контексте слова c):

$$p(w|c) = \frac{\exp \langle \psi_w, \psi'_c \rangle}{\sum_{w \in W} \exp \langle \psi_w, \psi'_c \rangle}$$

$$p(w_{j-h}, \dots w_{j+h}|w_j) = \prod_{-h \leqslant k \leqslant h, k \neq 0} p(w_{j+k}|w_j)$$

Два способа моделирования вероятностей: смесь распределений или softmax от вещественной оценки.

Логарифм правдоподобия:

- PLSA:

$$L = \sum_{d \in D} \sum_{w \in W} n_{wd} \log p(w|d)$$

- Skip-gram:

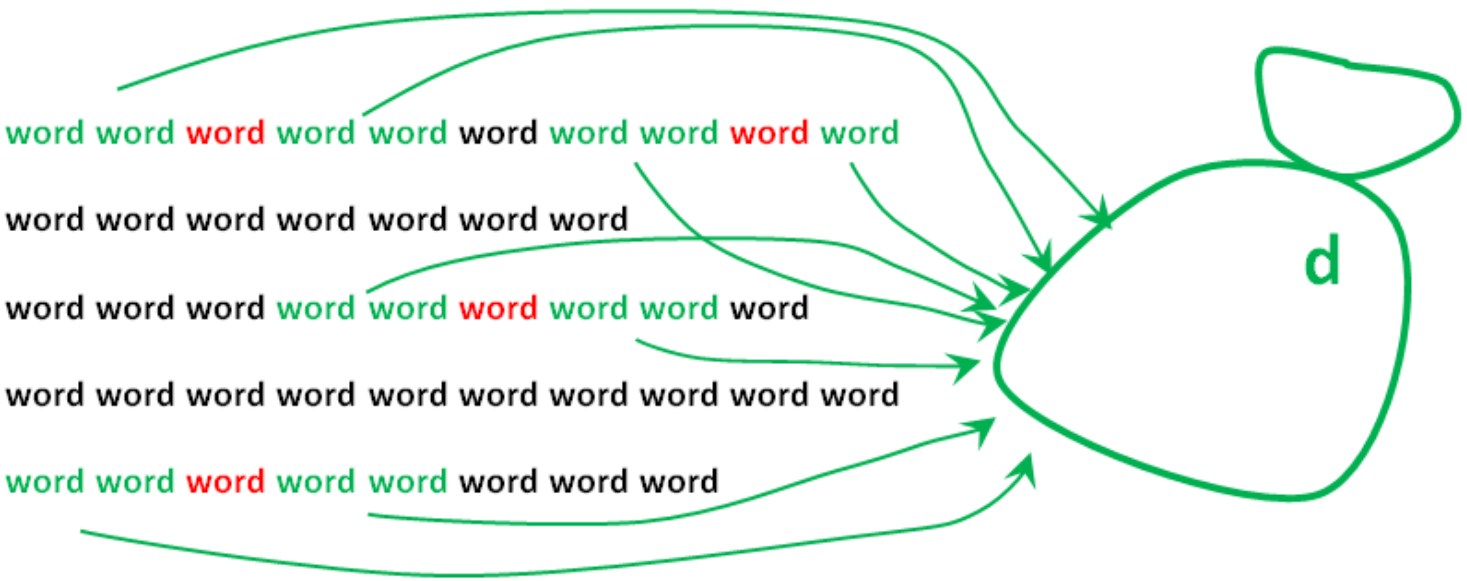
$$L = \sum_{j=1}^{N} \log p(w_{j-h}, \dots w_{j+h}|w_j) =$$

$$= \sum_{j=1}^{N} \sum_{-h \leqslant k \leqslant h, k \neq 0} \log p(w_{j+k}|w_j) = \sum_{c \in W} \sum_{w \in W} n_{wc} \log p(w|c)$$

Как определить «документы» d **так, чтобы** $n_{wd} = n_{wc}$**?**

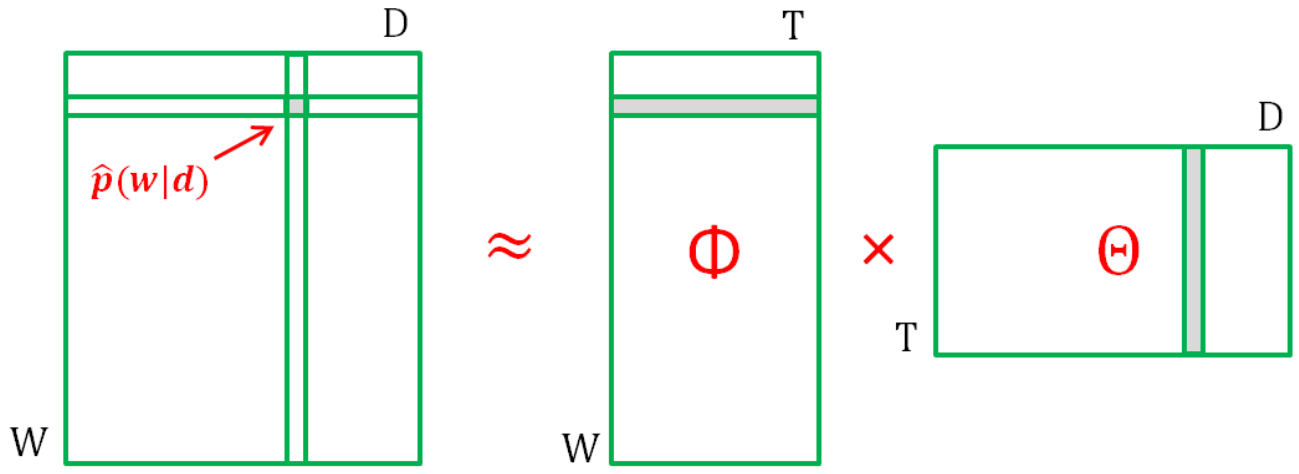
Тематическая модель для коллекции псевдо-документов

Определим псевдо-документ d, порожденный словом, как мешок всех слов, которые встречаются в локальном контексте любого вхождения этого слова:



Обучение тематической модели на W псевдо-документах – это то же самое, что предсказывать вероятности слов в контекстах.

Низкоранговые матричные разложения:



- PLSA:

$$\sum_{d\in D}\sum_{w\in W} n_{wd}\log\langle\phi_w,\theta_d\rangle \to \max_{\phi_w,\theta_d}$$

Метод оптимизации: EM-алгоритм.

Оптимум: $\hat{p}(w|d) = \frac{n_{wd}}{n_d}$.

Низкоранговые матричные разложения:

- GloVe (Global Vectors):

$$\sum_{w \in W} \sum_{c \in W} f(n_{wc}) \left(\langle \psi_w, \psi'_c \rangle + b_w + b'_c - \log n_{wc} \right)^2 \to \min_{\psi_w, \psi'_c, b_w, b'_c}$$

Метод оптимизации: AdaGrad.
Оптимум: $\log n_{wc}$.

- SGNS (Skip-gram Negative Sampling):

$$\sum_{w \in W} \sum_{c \in W} n_{wc} \log \sigma \left(\langle \psi_w, \psi'_c \rangle \right) + k \, \mathbb{E}_{\bar{c}} \log \sigma \left(- \langle \psi_w, \psi'_{\bar{c}} \rangle \right) \to \max_{\psi_w, \psi'_c}$$

Метод оптимизации: SGD (online).
Оптимум: $\text{sPMI} = \log \frac{n_{wc} n}{n_w n_c} - \log k$.

Рассматриваемые методы как векторные модели семантики

- По корпусу текстов строится матрица со-встречаемостей (иногда неявно)
- Производится матричное разложение
- Полученые представления полезны в ряде задач

	Вероятности	Вещественные оценки
Слова-документы	PLSA, LDA, ...	Semantic Word Vectors [Maas and Ng, 2010]
Слова-слова	?	word2vec, GloVe, ...

Слова-слова + Вероятности

1. способны решать задачи близости и аналогий
2. хорошо подходят для коротких документов

Тематические модели для задач близости и аналогий

Согласно нашим предварительным экспериментам, PLSA на матрице слова-слова:

- дает сравнимое качество с word2vec/GloVe на задаче близсти слов (датасет WordSim353)
- справляется с задачаей аналогий, например:

  Самые близкие к log('alexander') + log('girl') - log('boy'):
  tamara, anna, tatiana, natasha, nadia, olga,
  georgina, alexandra, katherine, anastasia.

  Самые близкие к log('edward') + log('girl') - log('boy'):
  katherine, georgina, susannah, louise, josephine,
  marjorie, amelia, emily, emma, jane.

Объединение преимуществ: интерпретируемость и арифметика

Темы для log('alexander') + log('girl') - log('boy'):

- **topic158:** daughter mary mother actress sister elizabeth lady marriage anne woman queen jane maria mrs miss husband anna margaret marie ann sarah princess daughters alice child whom louise girl helen catherine barbara charlotte couple female susan laura rose
- **topic133:** op orchestra minor symphony piano concerto sonata violin flat beethoven bach mozart suite composed string quartet movement pieces composer chamber cello theme variations opus allegro piece orchestral trio symphonies key opera philharmonic
- **topic69:** russian soviet russia moscow vladimir ukrainian ukraine alexander ivan petersburg ussr romanian kiev bulgarian mikhail boris bulgaria romania oblast sergei orthodox communist sofia saint georgia nikolai georgian armenian azerbaijan stalin
- **topic54:** championships olympic olympics gold medal event grand silver bronze nd cup race metres champion km athletics winter marathon indoor rd cross competed relay junior medals slalom ski individual jump skiing finished athlete prix sports athens

Объединение преимуществ: интерпретируемость и арифметика

Темы, в которых популярно имя 'anna':

- **topic158:** daughter mary mother actress sister elizabeth lady marriage anne woman queen jane maria mrs miss husband anna margaret marie ann sarah princess daughters alice child whom louise girl helen catherine barbara charlotte couple female susan laura rose
- **topic181:** santa monastery lady philippines maria saint our san philippine cathedral manila cruz fe parish abbey basilica mary santo chapel shrine barbara del virgin province convent rosa clara barangay monica sant municipality monks immaculate holy ana conception
- **topic69:** russian soviet russia moscow vladimir ukrainian ukraine alexander ivan petersburg ussr romanian kiev bulgarian mikhail boris bulgaria romania oblast sergei orthodox communist sofia saint georgia nikolai georgian armenian azerbaijan stalin
- **topic227:** di italian italy del il milan san rome giovanni della carlo maria francesco venice giuseppe antonio serie opera florence da monte naples province roma italia dei bologna dell paolo pietro luigi andrea marco turin palazzo palermo le villa comune mario

Динамическая тематическая модель британского искусства

Find similar objects

Artist

Richard Hamilton (...)

Type of object

artwork (...)
- painting (5,438)

Date

1964 (...)

Subject

emotions, concepts and ideas (14,339)
- emotions and human qualities (4,766)
  - anxiety (993)
- formal qualities (10,738)
  - photographic (3,989)

history (4,815)
- politics and society (2,183)
  - death: President John F. Kennedy, assassination, 22 Nov 1963 (2)

interiors (4,061)
- domestic (1,594)
  - living room (255)

literature and fiction (2,963)
- film, music and ballet (315)
  - film: Sirk, Douglas, 'Shockproof' (1)

objects (20,850)
- electrical appliances (334)
  - television (39)
- furnishings (2,762)
  - chair (858)

people (30,832)
- adults (22,449)
  - woman (8,810)
- named individuals (8,941)
  - Knight, Patricia (1)

work and occupations (10,758)
- arts and entertainment (4,544)
  - actor (209)

Context

Tate collection highlights (195)

Динамическая тематическая модель британского искусства

Примеры тем в 1800-х:

- 'river', 'river thames', 'england', 'bridge', 'boat and sailing', 'waterfront', 'boat and barge', 'figure'
- 'soldier', 'army', 'man', 'horse', 'sword', 'battle', 'woman'

Примеры тем в 2000-х:

- 'writing', 'lamp', 'reading', 'desk', 'telephone', 'photographic', 'office', 'man'
- 'dusk', 'ukraine', 'kharkov', 'politics ussr and dissolution and 1990-91', 'documentary', 'townscape', 'man'

По результатам ручного оценивания тем:

- **Слова-слова**: 22 хороших, 52 средних и 6 плохих
- **Слова-документы**: 18 хороших, 42 средних и 20 плохих

О прошедшем годе...

- Завершение исследования по отбору тем и доклад на международной конференции SLDS-2015 (Statistical Learning and Data Sciences) в Лондоне.
- Приглашенный доклад в Microsoft Research Cambridge по аддитивной регуляризации тематических моделей.
- Летняя школа по машинному обучению и анализу текстов на естественном языке LxMLS-2015 в Лиссабоне.
- Эксперименты с регуляризаторами локального контекста и постер на конференции Machine Learning: Prospects and Applications в Берлине в октябре 2015 г.

... и о годе следующем

- Подготовка статьи по комбинированию тематических моделей и нейросетевых моделей языка (word embeddings).
- Исследовательская стажировка в Google Zurich по анализу текстов (с апреля по июль 2016).