

The Supervised Word Mover's Distance

Поляков Павел

НИУ ВШЭ

25 апреля 2017 г.

Постановка задачи

Задача:

- Определение схожести документов

Применение:

- Поисковое ранжирование
- Классификация текстов
- Рекомендательные системы

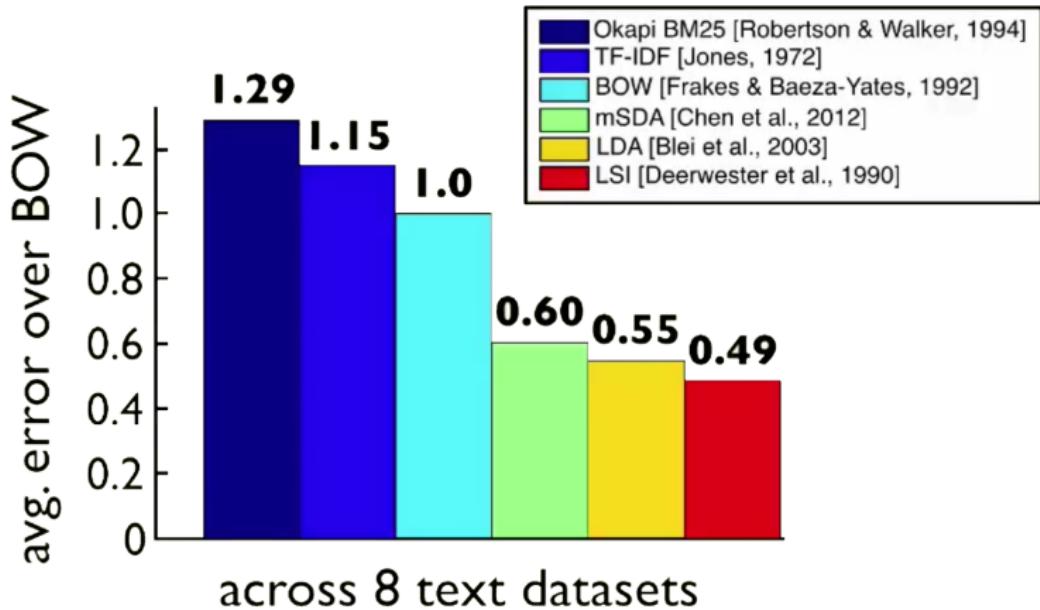
Стандартный подход

Документы любым способом переводятся в конечномерные вектора, после чего между ними вычисляется евклидово расстояние (или другая р-метрика)



Какого качества это позволяет достичь?

kNN error compared to BOW



Причина плохого качества

Может быть причина в синонимии/полинимии/потери информации о порядке слов? На самом деле проблема в том, что мы забыли определить, что такое "похожесть".

Сравним следующие 2 текста:

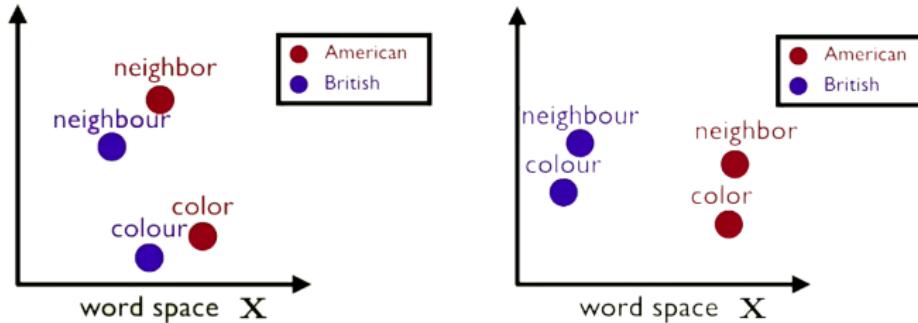
- (i) I just cooked this awesome recipe, it is my favourite dish.
The food is yummy.
- (ii) My favourite politician gave an awesome speech about an important topic.

Они похожи, как два позитивных текста, или различны, как тексты о еде и политике?

Обучение метрики

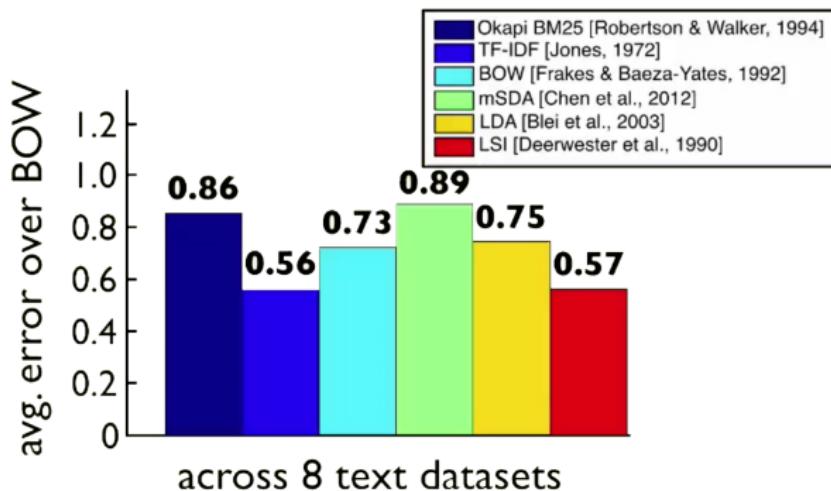
Предлагается обучать метрику расстояния относительно матрицы A , где $A \in \mathbb{R}^{r \times d}$, $r \leq d$, так, чтобы минимизировать некоторую аппроксимацию LOO kNN classification error

$$D(x_i, x_j) = \|A(x_i - x_j)\|^2$$



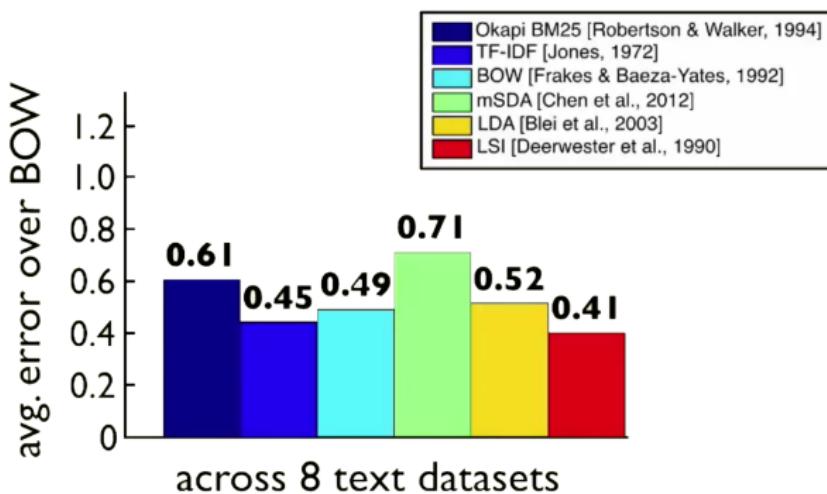
Способы обучения и получаемое качество

Information Theoretic Metric Learning (ITML)



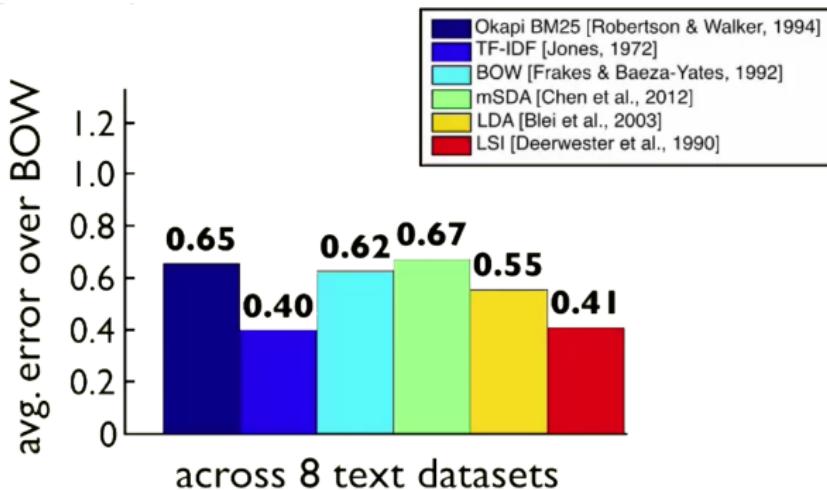
Способы обучения и получаемое качество

Large margin nearest neighbor (LMNN)



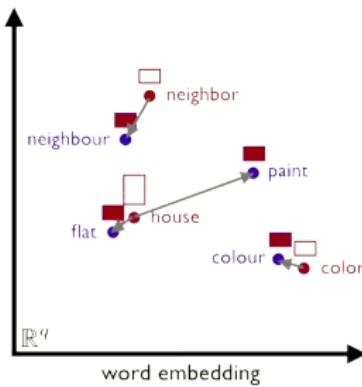
Способы обучения и получаемое качество

Neighborhood Components Analysis (NCA)



Определение

WMD - расстояние между двумя документами как опт. стоимость перемещения слов из одного документа в другой с помощью векторного представления слов (word2vec).

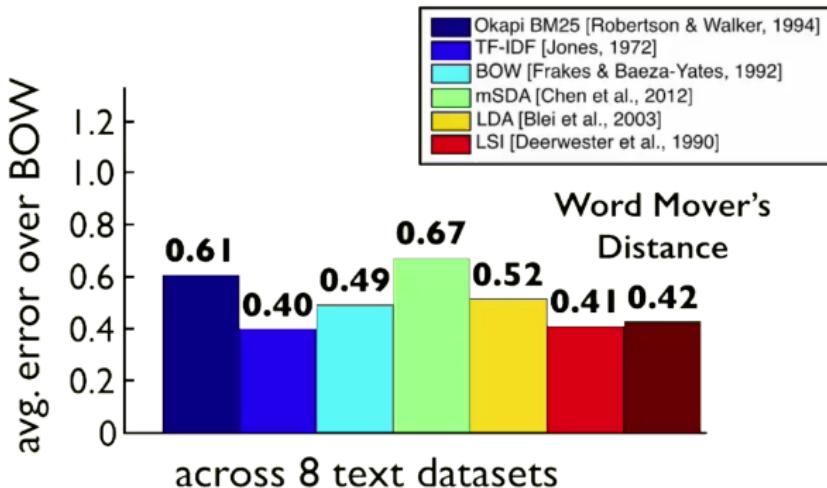


Оптимизационная задача

$X \in \mathbb{R}^{d \times n}$ - матрица представлений слов
 $d^a, d^b \in \mathbb{R}^n$ - нормированные BOW для документов

$$D(d^a, d^b) = \min_{T \geq 0} \sum_{i,j=1}^n T_{ij} \|x_i - x_j\|_2^2, \quad \text{subject to,}$$
$$\sum_{j=1}^n T_{ij} = d_i^a, \quad \sum_{i=1}^n T_{ij} = d_j^b \quad \forall i, j$$

Качество по сравнению с предыдущими методами



Необходимо построить supervised версию для WMD

Leave-one-out (LOO) classification

Метрика: $D(x_i, x_j) = \|A(x_i - x_j)\|^2$

Чтобы найти A , мы определяем целевую функцию как точность (accuracy) классификации в новом пространстве и пытаемся найти A , как $A^* = \text{argmax}_A f(A)$

Пусть мы предсказываем метку для одной точки по kNN в заданной метрике (LOO classification). Множество соседей будет изменяться дискретно в ответ на непрерывные изменения A , т.е. $f(A)$ не будет дифференцируемой.

Stochastic nearest neighbours

Вместо kNN будем рассматривать всю трансформированную выборку как стохастических ближайших соседей. Пусть

$$p_{ij} = \frac{\exp(-\|A(x_i - x_j)\|_2^2)}{\sum_{k \neq i} \exp(-\|A(x_i - x_k)\|_2^2)}, \quad i \neq j, \quad p_{ii} = 0, \quad i = j$$

Вероятность корректно классифицировать объект i и целевая функция равны соответственно

$$p_i = \sum_{j:y_i=y_j} p_{ij}, \quad f(A) = \sum_i p_i = \sum_i \sum_{j:y_i=y_j} p_{ij}$$

NCA and KL-divergence

$$A^* = \operatorname{argmax}_A f(A) = \operatorname{argmin}_A (-\ln(f(A))) = \\ \operatorname{argmin}_A (-\sum_i \ln(p_i)) = \operatorname{argmin}_A \sum_i q_i \ln(\frac{q_i}{p_i}), \quad q_i = 1 \quad \forall i$$

Т.е. NCA минимизирует KL-дивергенцию между распределением предсказанных меток и истинным, где $p_i = 1$

Обучение метрики в WMD

$X \in \mathbb{R}^{d \times n}$ - матрица представлений слов

$w \in \mathbb{R}^n$ - вектор важности слов для отличия классов

$d^a, d^b \in \mathbb{R}^n$ - нормированные BOW для документов

$\tilde{d}^a = \frac{w \circ d^a}{w^T d^a}, \tilde{d}^b = \frac{w \circ d^b}{w^T d^b}$ - перевзвешенные BOW

$$D(d^a, d^b) = \min_{T \geq 0} \sum_{i,j=1}^n T_{ij} \|A(x_i - x_j)\|_2^2, \quad \text{subject to,}$$
$$\sum_{j=1}^n T_{ij} = \tilde{d}_i^a, \quad \sum_{i=1}^n T_{ij} = \tilde{d}_j^b \quad \forall i, j$$

Добавление регуляризатора

Добавим к функционалу энтропийный регуляризатор:

$$D(d^a, d^b) = \min_{T \geq 0} \sum_{i,j=1}^n T_{ij} \|A(x_i - x_j)\|_2^2 + \frac{1}{\lambda} \sum_{i,j=1}^n T_{ij} \ln T_{ij},$$

subject to, $\sum_{j=1}^n T_{ij} = \tilde{d}_i^a$, $\sum_{i=1}^n T_{ij} = \tilde{d}_j^b \forall i, j$

Это сделало задачу строго выпуклой. Теперь T^* единственно и, как следствие, существует градиент по A :

$$\nabla_A D(d^a, d^b) = 2A \sum_{i,j=1}^n T_{ij}^* (x_i - x_j)(x_i - x_j)^T$$

Добавление регуляризатора

Добавление регуляризатора также позволяет аппроксимировать T_λ^* за $O(q^2)$, где q - число уникальных слов в d^a и d^b , что намного быстрее точного решения оригинальной задачи за $O(q^3 \ln(q))$:

$$T_\lambda^* \approx \text{diag}(u) K \text{diag}(v), \quad K_{ij} = \exp(-\lambda \|A(x_i - x_j)\|_2^2),$$

Где u, v находятся итеративно:

$$(u, v) \leftarrow (\tilde{d}^a / (Kv), \tilde{d}^b / (K^T u))$$

SWMD in NCA

Итого, задача оптимизации:

$$f(A, w) = - \sum_{l=1}^n \ln \left(\sum_{k: y_k = y_l} \frac{\exp(-D_{A,w}(d^l, d^k))}{\sum_{o \neq l} \exp(-D_{A,w}(d^l, d^o))} \right)$$
$$D(d^l, d^k) = \min_{T \geq 0} \sum_{i,j=1}^n T_{ij} \|A(x_i - x_j)\|_2^2 + \frac{1}{\lambda} \sum_{i,j=1}^n T_{ij} \ln T_{ij},$$

subject to, $T \mathbf{1}_n = \frac{w \circ d^l}{w^T d^l}$, $T^T \mathbf{1}_n = \frac{w \circ d^k}{w^T d^k} \quad \forall i, j$

$\nabla_A D(d^a, d^b)$ мы уже знаем. Чтобы найти $\nabla_w D(d^a, d^b)$, необходимо перейти к двойственной задаче

Двойственная задача

$$D_{A,w}^*(d^a, d^b) = \max_{\alpha, \beta} (\alpha^T \tilde{d}^a + \beta^T \tilde{d}^b - \frac{1}{\lambda} \sum_{i,j=1}^n e^{\lambda(\alpha_i + \beta_j - \|A(x_i - x_j)\|_2^2) - 1}, \alpha_i + \beta_j \leq \|A(x_i - x_j)\|_2^2)$$

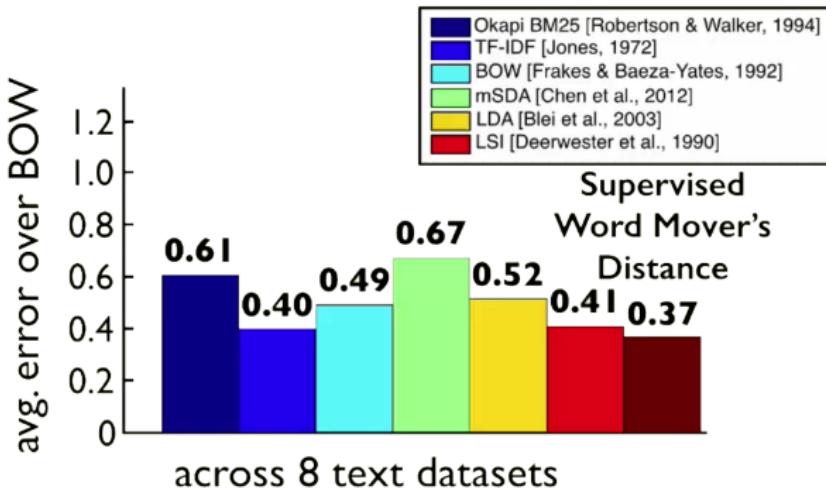
Её аппроксимированное решение:

$$\alpha_\lambda^* = \frac{\ln(u)}{\lambda} - \frac{\ln(u)^T \mathbf{1}}{p} \mathbf{1}, \quad \beta_\lambda^* = \frac{\ln(v)}{\lambda} - \frac{\ln(v)^T \mathbf{1}}{p} \mathbf{1}$$

Градиент по w :

$$\nabla_w D(d^a, d^b) = \frac{\alpha^* \circ d^a - (\alpha^* \tilde{d}^a) d^a}{w^T d^a} + \frac{\beta^* \circ d^b - (\beta^* \tilde{d}^b) d^b}{w^T d^b}$$

Сравнение результатов



Выделенные слова

ohsumed

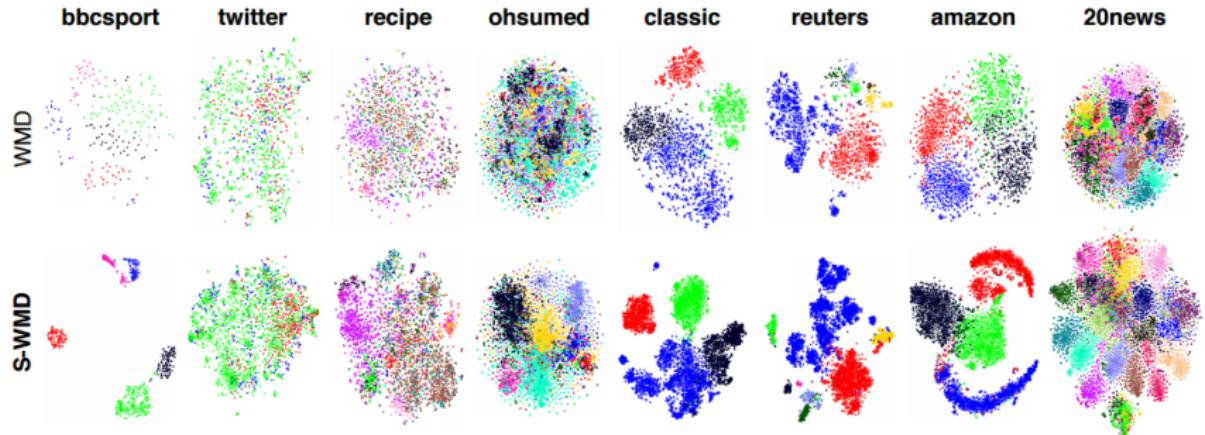
reuters

BBL, Tari, CSR
declared, England, OTR
Gulf, oilfield, buy
distribution, department
sets, DLRS, exchange
loss, shareholders
currencies, stock discount
freight, sale
payout, gain acquisition, barrels
uranium, petroleum, crop
system, fleet, profit, earnings
marks, ship, oil rate, interest
drilling, growth

20news

Mazda
Boone
motherboard
happening
homosexuals
playoff
motorcycles
dolphins
computer
room
Keith
motif
graphics
pro
orbit
driver
gun
bike
apple
Rutgers
clipper
controller
security
hell
virtual
self
leihende
Islamic
Windows
DOD
sale
card
bikes
mac
space
encryption
copy
warning
Amerson
forsaler
books
Computer
powerbook
SCSI
government
chip
SUN
AT&T
Nasa
DOS

t-SNE



Спасибо за внимание!