кафедра Алгоритмов
и Технологий
Программирования

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Автоматическая обработка текстов
Дистрибутивная семантика

Лекция 5

Емельянов А. А.
login-const@mail.ru

2018

https://github.com/king-menin/nlp-course

Коллокации

- **Коллокацией** называется словосочетание, имеющее признаки синтаксически и семантически целостной единицы, в котором выбор одного из компонентов осуществляется по смыслу, а выбор второго зависит от выбора первого (например, ставить условия — выбор глагола ставить определяется традицией и зависит от существительного условия, при слове предложение будет другой глагол — вносить).

https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BB%D0%BB%D0%BE%D0%BA%D0%B0%D1%86%D0%B8%D1%8F_(%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0)

Статистический подход: биграммы

- Топ биграмм обычно не то, что надо ☹

№	Словосочетание	Документы	Частота
1	и не	22732	201352
2	и в	27048	193983
3	потому что	14926	117401
4	я не	10675	113767
5	у меня	9734	97102
6	может быть	16086	96065
7	то что	17195	95251
8	что он	11786	92743
9	не было	13196	92729

НКРЯ: http://www.ruscorpora.ru/2grams.top.html

Биграммы с учетом частей речи

A
на расстоянии 1 от S

№	Вхождения	Документы	Фрагмент
1	19402	6104	крайней мере
2	18152	2791	российской федерации
3	12164	6528	настоящее время
4	11348	4160	должны были
5	11045	6067	последнее время
6	9720	2893	молодой человек

S
на расстоянии 1 от S

№	Вхождения	Документы	Фрагмент
1	45631	9556	а потом
2	21563	10492	том числе
3	17401	5932	друг друга
4	15362	6214	точки зрения
5	14925	5242	конце концов
6	12616	4597	т п

V
на расстоянии 1 от S

№	Вхождения	Документы	Фрагмент
1	8583	4415	три года
2	7404	2919	следующий день
3	7122	3044	три дня
4	6615	2851	было уже
5	5716	3262	данном случае
6	5345	2415	был уже

NUM
на расстоянии 1 от S

№	Вхождения	Документы	Фрагмент
1	14054	5071	несколько раз
2	12787	4376	несколько дней
3	10561	5359	два года
4	9248	5078	несколько лет
5	8583	4415	три года
6	8123	2852	несколько минут

НКРЯ: http://www.ruscorpora.ru/search-ngrams_2.html

Биграммы со словом большой

большой

№	Вхождения	Документы	Фрагмент
1	14299	4839	**больше** не
2	12725	6087	**больше** чем
3	10226	4365	и **больше**
4	9912	4733	с **большим**
5	8307	3992	**больше** всего
6	7978	3701	еще **больше**
7	7434	3772	все **больше**
8	5595	3039	в **большом**
9	5465	2852	**больше** и
10	5319	2862	не **больше**

большой
на расстоянии 1 от S

№	Вхождения	Документы	Фрагмент
1	4372	2678	**большая часть**
2	2872	1952	**большую часть**
3	1933	1417	**большое количество**
4	1692	1084	**большое значение**
5	1650	1058	**больше того**
6	1518	1066	**больше нет**
7	1190	763	**большую роль**
8	1164	863	**большим трудом**
9	1130	866	**большие деньги**
10	905	534	**большого театра**

НКРЯ: http://www.ruscorpora.ru/search-ngrams_2.html

Биграммы со словом огромный

огромный

№	Вхождения	Документы	Фрагмент
1	1809	1266	с **огромным**
2	1437	1105	**огромное** количество
3	1254	933	в **огромном**
4	924	722	с **огромными**
5	889	711	с **огромной**
6	723	579	**огромное** значение
7	717	585	в **огромной**
8	579	508	в **огромных**
9	436	382	на **огромном**
10	417	326	**огромную** роль

огромный
на расстоянии 1 от S

№	Вхождения	Документы	Фрагмент
1	1437	1105	**огромное количество**
2	723	579	**огромное значение**
3	417	326	**огромную роль**
4	321	221	**огромное большинство**
5	311	263	**огромные деньги**
6	266	237	**огромном количестве**
7	262	219	**огромное влияние**
8	245	227	**огромного количества**
9	222	199	**огромная толпа**
10	206	172	**огромное число**

НКРЯ: http://www.ruscorpora.ru/search-ngrams_2.html

Плюсы и минусы биграммного подхода

Плюсы и минусы биграммного подхода

- **Плюсы:**

  \+ простота

Плюсы и минусы биграммного подхода

- **Плюсы:**
  - \+ простота
  - \+ хорошо работает для фиксированных фраз

Плюсы и минусы биграммного подхода

- **Плюсы:**
  - \+ простота
  - \+ хорошо работает для фиксированных фраз
- **Минусы:**

Плюсы и минусы биграммного подхода

- **Плюсы:**

  \+ простота

  \+ хорошо работает для фиксированных фраз

- **Минусы:**

  – плохо работает для слов, не обязательно стоящих рядом:

  стучать во все **двери**
  стучать во все возможные **двери**
  в **дверь постучали**
  в **дверь** купе **постучали**
  постучал в новую **дверь**
  не ошибся **дверью** и **постучал**

Распределение расстояний между словами

- Посчитаем по выборке среднее расстояние (со знаком) между словами и его дисперсию:

$$\mu = \frac{1}{6}(3 + 4 - 1 - 2 + 3 + 2) = 1.5$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n - 1} \approx 2.42$$

- Чем меньше σ, тем больше слова похожи на коллокацию.

Распределение расстояний между словами

σ	μ	частота	w_1	w_2
0,43	0,97	11657	New	York
0,48	1,83	24	previous	games
0,15	2,98	46	minus	points
0,49	3,87	131	hundreds	dollars
4,03	0,44	36	editorial	Atlanta
4,03	0,00	78	ring	New
3,96	0,19	119	point	hundredth
3,96	0,29	106	subscribers	by
1,07	1,45	80	strong	support
1,13	2,57	7	powerful	organizations
1,01	2,00	112	Richard	Nixon
1,05	0,00	10	Garrison	said

1. Speech and Language Processing (3rd ed. draft) Dan Jurafsky and James H. Martin, URL: https://web.stanford.edu/~jurafsky/slp3/

Поточечная взаимная информация

$$PMI(w_1, w_2) = \log_2 \frac{P(w_1 w_2)}{P(w_1)P(w_2)}$$

- Хороший индикатор независимости, плохой показатель зависимости для редких слов.

$$PPMI(w_1, w_2) = \max(PMI(w_1, w_2), 0)$$

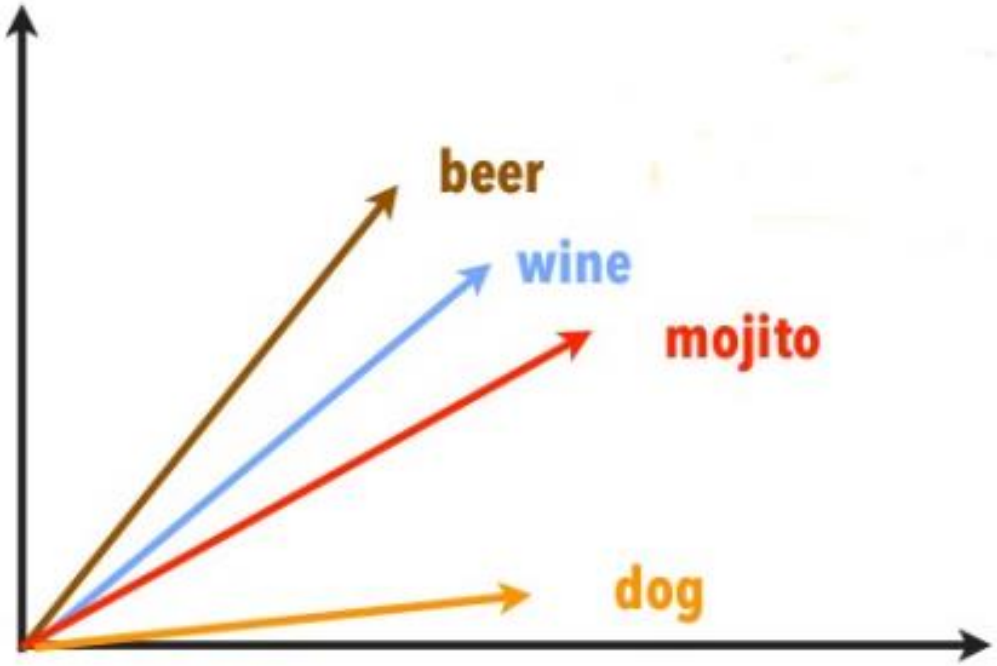
— положительная поточечная взаимная информация.

Близость между словами

- **Задача**: найти слова, синтаксически и/или семантически «ближайшие» к данному слову.

Векторное представление слов

- **Идея:** каждое слово представлять вектором в некотором пространстве R^n .



One hot encoding

- Представляем слова векторами с единственной единицей и остальными нулями в $R^{|V|}$ (где V — словарь).

- Суммируя такие векторы для всех слов документа, получаем представление **bag of words**.

One hot encoding

- Представляем слова векторами с единственной единицей и остальными нулями в $R^{|V_w|}$ (где V_w — словарь).

- Суммируя такие векторы для всех слов документа, получаем представление **bag of words**.

- Увы, такие векторы никак не связаны между собой. Нужно уменьшить размерность пространства. Например, чтобы каждая компонента вектора соответствовала некоторому «свойству».

Два подхода к векторным представлениям

- **count-based, или явный**
  - SVD матрицы совместной встречаемости
  - Eigenwords
  - Non-negative sparse embeddings
- **prediction-based, или неявный**
  - word2vec
  - fastText
  - StarSpace

Дистрибутивная гипотеза

- Лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения.

- A word is characterized by the company it keeps.
  - John Rupert Firth, 1957.

- Значит, векторы слов можно построить с помощью контекстов этих слов.

Явное представление слов контекстами

- Для словаря V_w и множества контекстов V_c построим разреженную матрицу $M_{[i,j]} = f(w_i, c_j)$ размера $|V_c| \times |V_w|$.

- Элемент $f(w_i, c_j)$ будет описывать связь слова w_i с контекстом c_j.

	c_1	c_2	…	$c_{\|V_C\|}$
w_1	f_{11}	f_{12}		$f_{1\|V_C\|}$
w_2	f_{21}	f_{22}		$f_{2\|V_C\|}$
…				
$w_{\|V_W\|}$	$f_{\|V_W\|1}$	$f_{\|V_W\|2}$		$f_{\|V_W\|\|V_C\|}$

Явное представление слов контекстами

- Для словаря V_w и множества контекстов V_c построим разреженную матрицу $M_{[i,j]} = f(w_i, c_j)$ размера $|V_c| \times |V_w|$.
- Элемент $f(w_i, c_j)$ будет описывать связь слова w_i с контекстом c_j.
- Как определить $f(w_i, c_j)$?

Явное представление слов контекстами

- Для словаря V_w и множества контекстов V_c построим разреженную матрицу $M_{[i,j]} = f(w_i, c_j)$ размера $|V_c| \times |V_w|$.
- Элемент $f(w_i, c_j)$ будет описывать связь слова w_i с контекстом c_j.
- Как определить $f(w_i, c_j)$?
  - $\#(w, c)$

Явное представление слов контекстами

- Для словаря V_w и множества контекстов V_c построим разреженную матрицу $M_{[i,j]} = f(w_i, c_j)$ размера $|V_c| \times |V_w|$.
- Элемент $f(w_i, c_j)$ будет описывать связь слова w_i с контекстом c_j.
- Как определить $f(w_i, c_j)$?
  - $\#(w, c)$
  - $P(w, c) = \#(w, c), (w, c) \in D$ – наблюдаемые пары (слово, контекст), всего пар $|D|$.

Явное представление слов контекстами

- Для словаря V_w и множества контекстов V_c построим разреженную матрицу $M_{[i,j]} = f(w_i, c_j)$ размера $|V_c| \times |V_w|$.
- Элемент $f(w_i, c_j)$ будет описывать связь слова w_i с контекстом c_j.
- Как определить $f(w_i, c_j)$?
  - $\#(w, c)$
  - $P(w, c) = \#(w, c)$, $(w, c) \in D$ – наблюдаемые пары (слово, контекст), всего пар $|D|$.
  - $PMI(w, c)$

Явное представление слов контекстами

- Для словаря V_w и множества контекстов V_c построим разреженную матрицу $M_{[i,j]} = f(w_i, c_j)$ размера $|V_c| \times |V_w|$.
- Элемент $f(w_i, c_j)$ будет описывать связь слова w_i с контекстом c_j.
- Как определить $f(w_i, c_j)$?
  - $\#(w, c)$
  - $P(w, c) = \#(w, c)$, $(w, c) \in D$ – наблюдаемые пары (слово, контекст), всего пар $|D|$.
  - $PMI(w, c)$
  - $PPMI(w, c)$

Оценка близости между векторами

- Косинусная мера близости:

$$\cos(u,v)=\frac{uv}{\|u\|_2\|v\|_2}=\frac{\sum_i u_i v_i}{\sqrt{\sum_i {u_i}^2}\sqrt{\sum_i {v_i}^2}}$$

- Мера Жаккара:

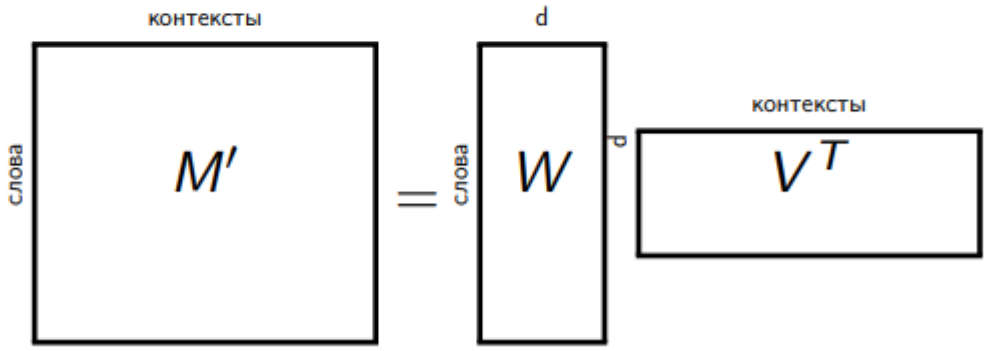
$$jc(u,i)=\frac{\sum_i \min(u_i, v_i)}{\sum_i \max(u_i, v_i)}$$

Уменьшение размерности

- С векторами такого размера работать неудобно.
- Будем строить векторы размерности $N \ll |V_c|$.
- **Факторизация** матрицы терм-контекст:

$$M' = W \times V^\top, W \in \mathbb{R}^{V_w \times d}, V \in \mathbb{R}^{V_c \times V_d}$$

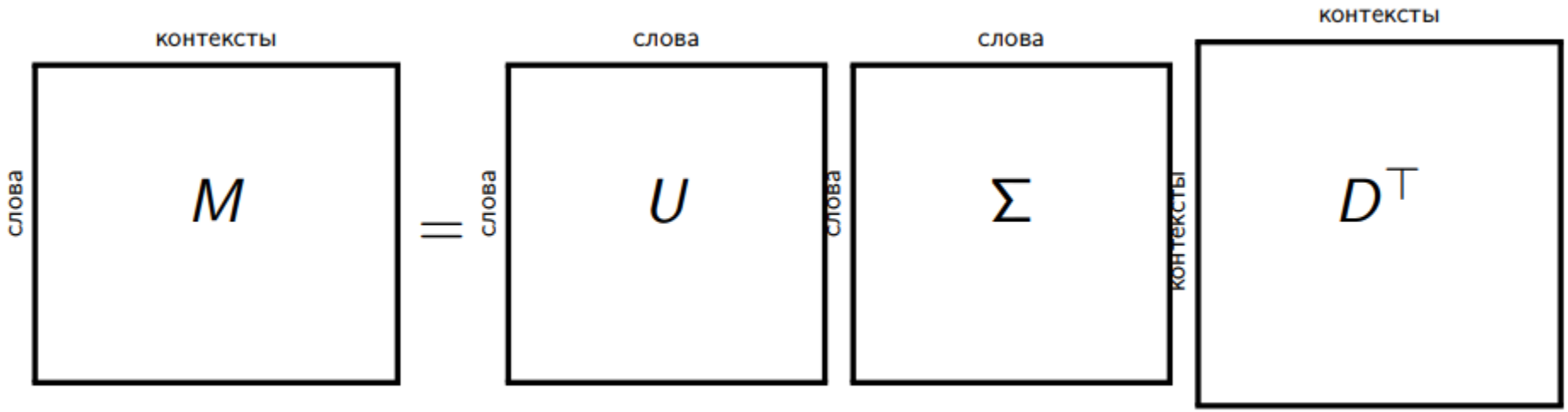
M' – лучшее приближение ранга d к M по L_2.



Уменьшение размерности

- **Сингулярное разложение** матрицы слово-контекст $M \in R^{V_W \times V_C}$:

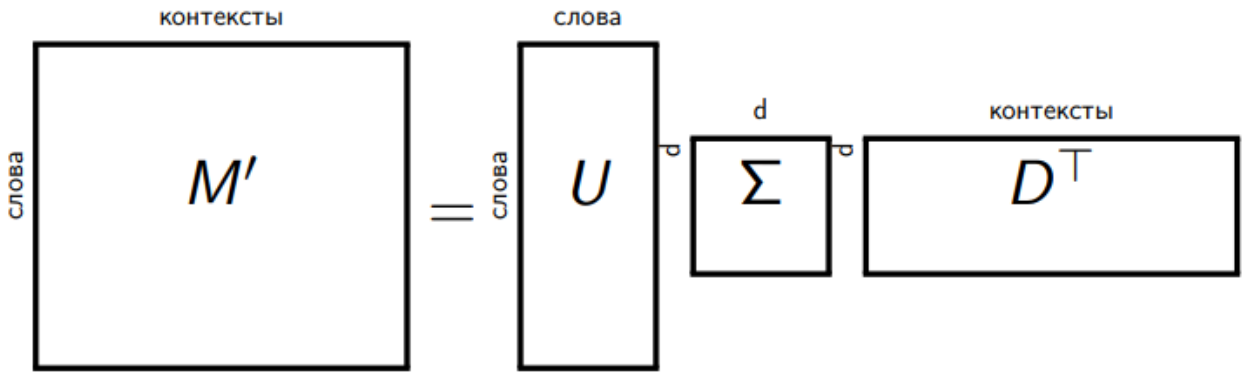
$$M = U\Sigma D^T$$



Уменьшение размерности

- **Аппроксимация** ранга d матрицы слово-контекст $M \in R^{V_W \times V_C}$:

$$M'_d = U_d \Sigma_d D_d^\top$$



- Искомое разложение:

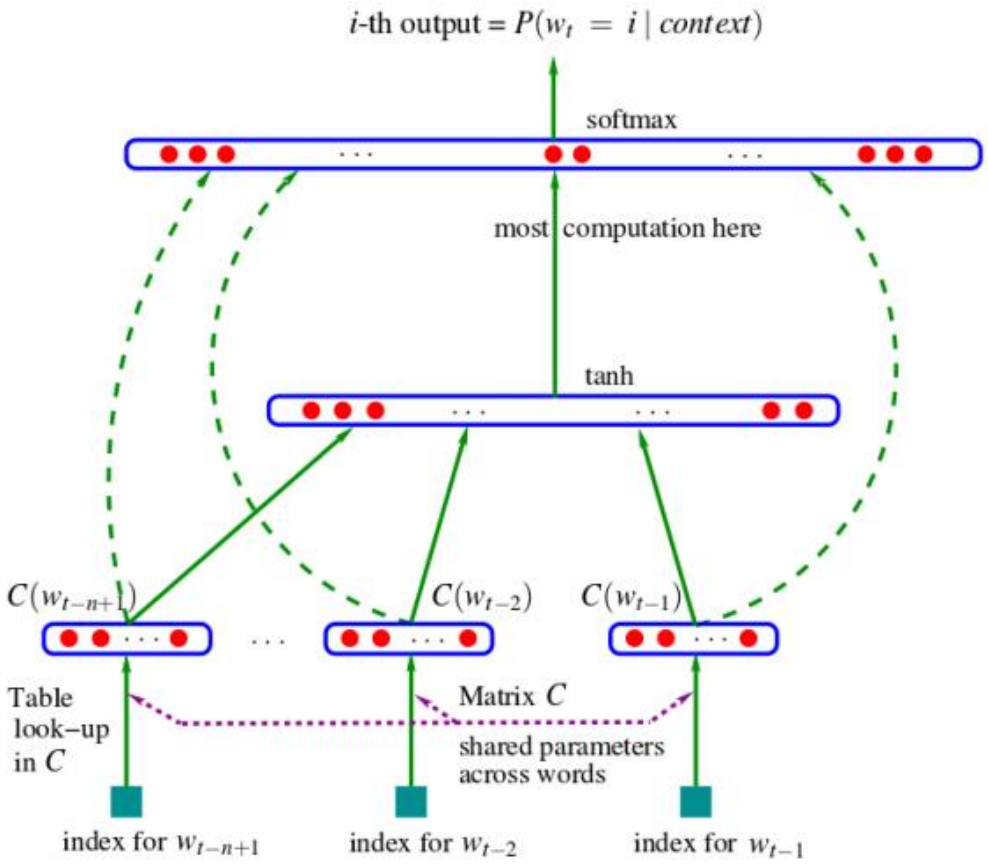
$$W = U_d \sqrt{\Sigma_d}, V^\top = \sqrt{\Sigma_d} D_d^\top$$

Латентно семантический анализ (1988)

- **Фактически** — применение SVD к матрице «терм–документ»
- **Возможности метода**:
  - оценка близости документов
  - оценка близости термов
  - кластеризация документов
  - оценка близости запроса и документа
- **Недостатки**:
  - низкая скорость
  - нет вероятностных предположений о распределении.

Неявное представление слов контекстами

- Другой способ получения векторного представления — нейронная сеть. Архитектуры нейронных сетей могут быть как последовательными, так и рекуррентными.



word2vec

- В 2013 г. Томас Миколов и его коллеги предложили word2vec — упрощенную нейронную сеть, которую можно быстро обучить на огромном объёме текстов для получения векторов слов.

- T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space (2013).

- T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed Representations of Words and Phrases and their Compositionality (2013).

word2vec

- **Две архитектуры:**
  - Continuous bag-of-words model (CBOW)
  - skip-gram
- **Два критерия оптимизации:**
  - Hierarchical softmax
  - Negative-sampling: для каждой пары $(w, c) \in D$ найти k слов, таких что $(w_k, c) \in D$

D – множество наблюдаемых пар слово-контекст.
D^- – множество ненаблюдаемых пар слово-контекст.

- **Вероятность** $(w, c) \in D$:

$$P(D = 1|w, c) = \frac{1}{1 + e^{-s(w,c)}}$$

- **Оптимизационная задача:**

$$L(\Theta, D, \bar{D}) = \sum_{(w,c)\in D} P(D = 1|w, c) + \sum_{(w,c)\in \bar{D}} P(D = 0|w, c)$$

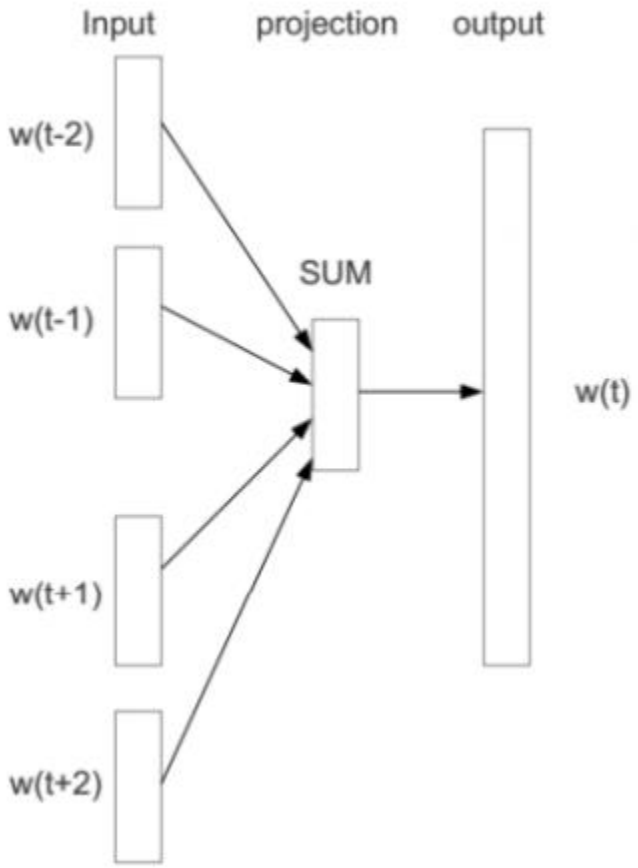
Continuous bag-of-words model (CBOW) [MCCD13]

- **Задача:** предсказание слова по заданному контексту.

- **Входной слой**:
  - контекст слова (+, $-\frac{k}{2}$ слова слева и справа)

- **Слой проекции**:
  - линейный

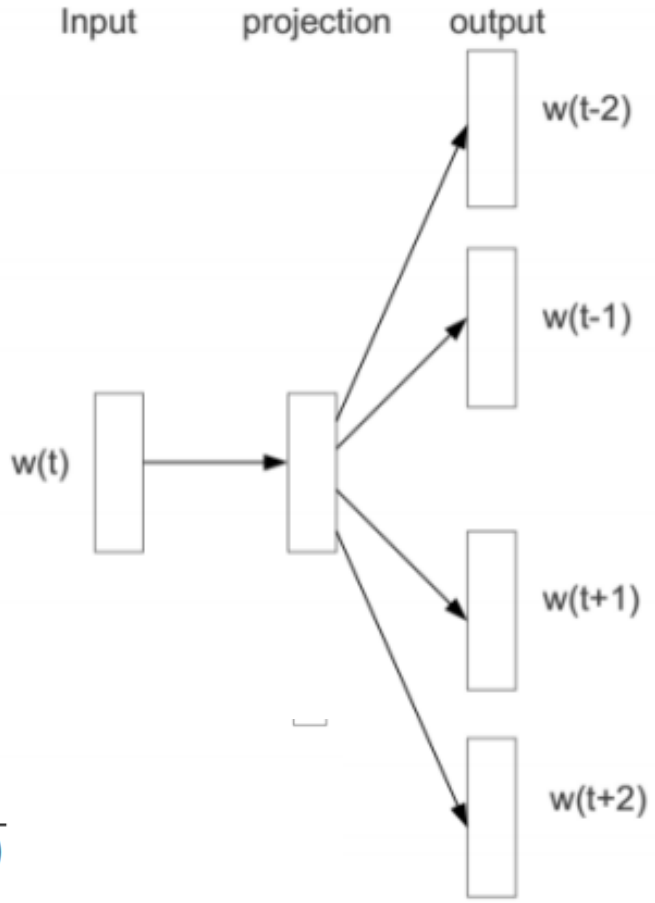
- **Выходной слой**:
  - вектор слова



$$P(D = 1|w, c_{1:k}) = \frac{1}{1 + e^{-(w \cdot c_1 + w \cdot c_2 + \ldots + w \cdot c_k)}}, c = \sum_{i=1}^{k} c_i$$

skip-gram [MCCD13]

- **Обратная задача**: предсказание векторов контекста по данному слову
- **Выходной слой**:
  - вектор слов
- **Все контексты независимы**:
  - $(w, c_1), \ldots, (w, c_k)$



$$P(D = 1|w, c_i) = \frac{1}{1 + e^{-(w \cdot c_i)}}$$

$$P(D = 1|w, c_{1:k}) = \prod_{i=1}^{k} P(D = 1|w, c_i) = \prod_{i=1}^{k} \frac{1}{1 + e^{-(w \cdot c_i)}}$$

Скрытый смысл подхода Negative Sampling

- O. Levy, Y. Goldberg. Neural Word Embedding as Implicit Matrix Factorization (2014).

- Minh Ngoc Le. https://minhlab.wordpress.com/2015/06/ (2015).

- SGNS неявно «факторизует» матрицу S «сдвинутого»

$$PMI: S_{ij} = \langle w_i , c_j \rangle = PMI(w, c) - \ln k.$$

Линейные свойства

- Оказывается, линейные операции над векторами v_w соотвествуют семантическим преобразованиям!

$$v_{king} - v_{man} + v_{woman} \approx v_{queen}.$$

$$v_{Paris} - v_{France} + v_{Italy} \approx v_{Rome}.$$

$$v_{big} - v_{small} + v_{smallest} \approx v_{biggest}.$$

Сравнение моделей эмбеддингов [SLMJ15]

- **Внутренние (intrinsic) задачи**
  - Определение похожих слов
  - Определение аналогий
  - Категоризация слов
  - Определение лишнего слова
  - Определение объектов глаголов
- **Внешние (extrinsic) задачи**
  - Классификация текстов
  - Извлечение именованных сущностей
  - Расширение запроса
- **Результаты** зависят от использованного корпуса для обучения, гиперпараметров обучения, корпуса для тестирования. Невозможно определить модель эмбеддингов, превосходящую остальные.

Другие модели

- **Word2Vec-f**
- **Doc2Vec**
- **GloVe**
- **FastText**
- **AdaGram**

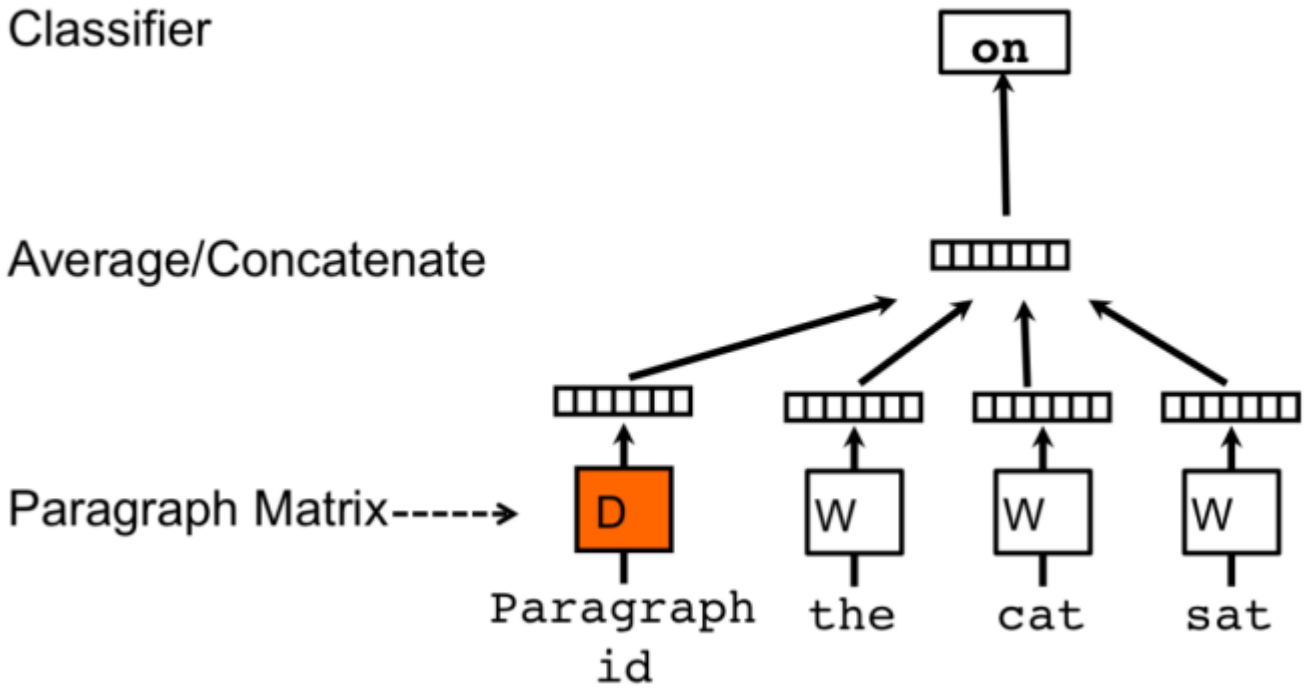
Word2Vec-f (dependency embeddings) [LG14a]

- Выбор контекста: синтаксически зависимые слова. Результат: функциональные зависимости.

Target Word	BoW5	BoW2	Deps
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl	superman superboy supergirl catwoman aquaman
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood	sunnydale collinwood calarts greendale millfield
turing	nondeterministic non-deterministic computability deterministic finite-state	non-deterministic finite-state nondeterministic buchi primality	pauling hotelling heting lessing hamming
florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas	texas louisiana georgia california carolina
object-oriented	aspect-oriented smalltalk event-driven prolog domain-specific	aspect-oriented event-driven objective-c dataflow 4gl	event-driven domain-specific rule-based data-driven human-centered
dancing	singing dance dances dancers tap-dancing	singing dance dances breakdancing clowning	singing rapping breakdancing miming busking

Насколько похожи два предложения (абзаца)? [LM14]

- Как найти вектор-предложения (абзаца)?
- Усреднить вектора слов, входящих в каждое предложение (с tf – idf весами)
- Doc2vec: что word2vec, только для предложений (абзацев).



FastText [BGJM16]

Слово w представляем символьными n-грамами:

$n = 3,\ G_{where} = \_wh, whe, her, re\_, \_where\_$

$sim_{w2v}(u, v) =< u, v >$

$sim_{ft}(u, v) = \sum_{e \in G_u} \sum_{g \in G_v} < e, v >$

AdaGram [BKOV16]

- Находит k смыслов слова.
- Демо

лето

Word ipm: 139.53, occurrences: 282349.

#2	0.46

Contexts: ...

Neighbours: *обо, ты, голубчик, похудеть, я*

Similar senses:

зима	0.70
осень	0.69
весна	0.65
выходной	0.58
сезон	0.53

#0	0.38

Contexts: ...

Neighbours: *год, 2012, 1919, 1941, 1940*

Similar senses:

весна	0.87
осень	0.75
осень	0.70
лето-осень	0.68
апрель-май	0.68

#1	0.16

Contexts: ...

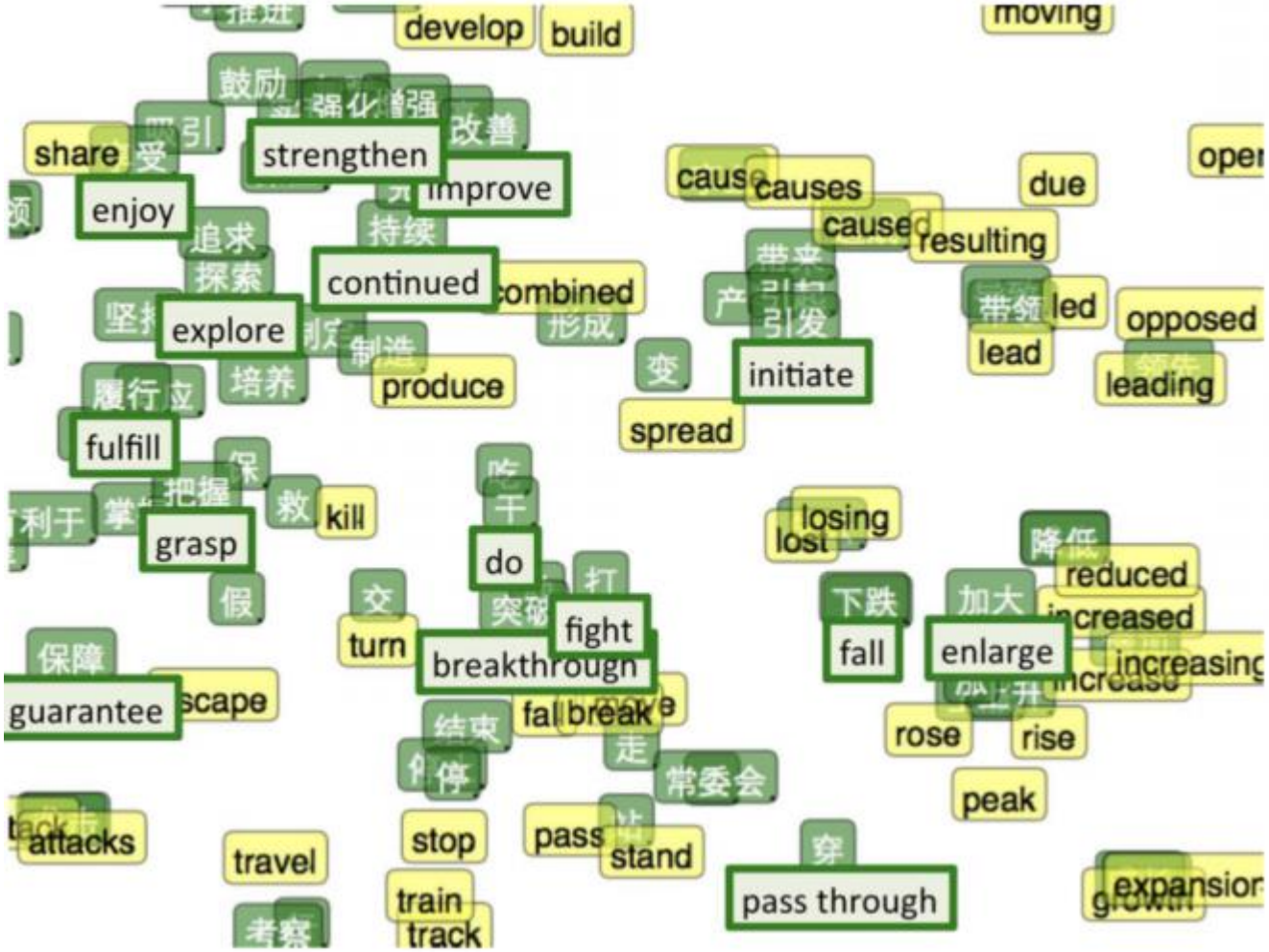
Neighbours: *жаркий, зима, влажный, густой, сухой*

Similar senses:

осень	0.80
весна	0.78
малоснежный	0.78
зима	0.77
холодный	0.73

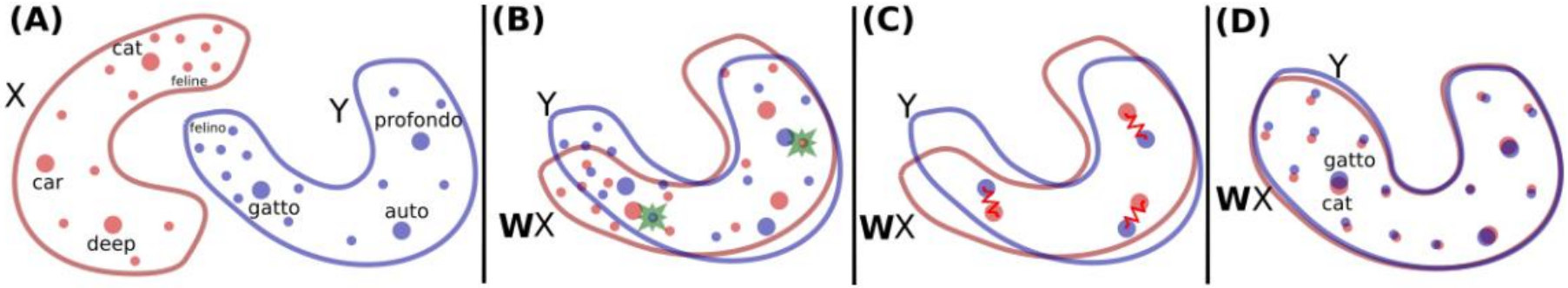
Двуязычные эмбедденги [ZSCM13]

- Дан (выровненный) параллельный корпус. Контекст слова: перевод этого слова на другой язык.



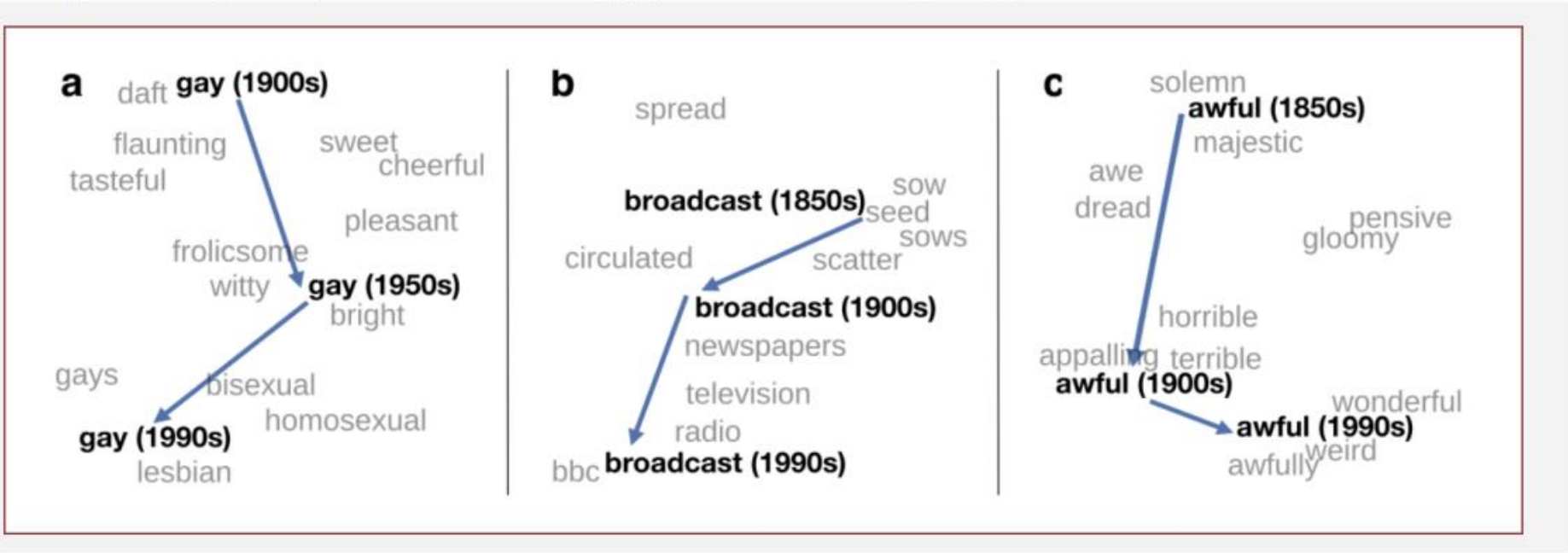
Двуязычные эмбедденги [CLR+17]

- Дано два невыровненных пространства слов
- Adversarial learning для определения матрицы поворота W
- Прокрустово преобразование для уточнения W
- k – NN-подобный метод для окончательного выравнивания



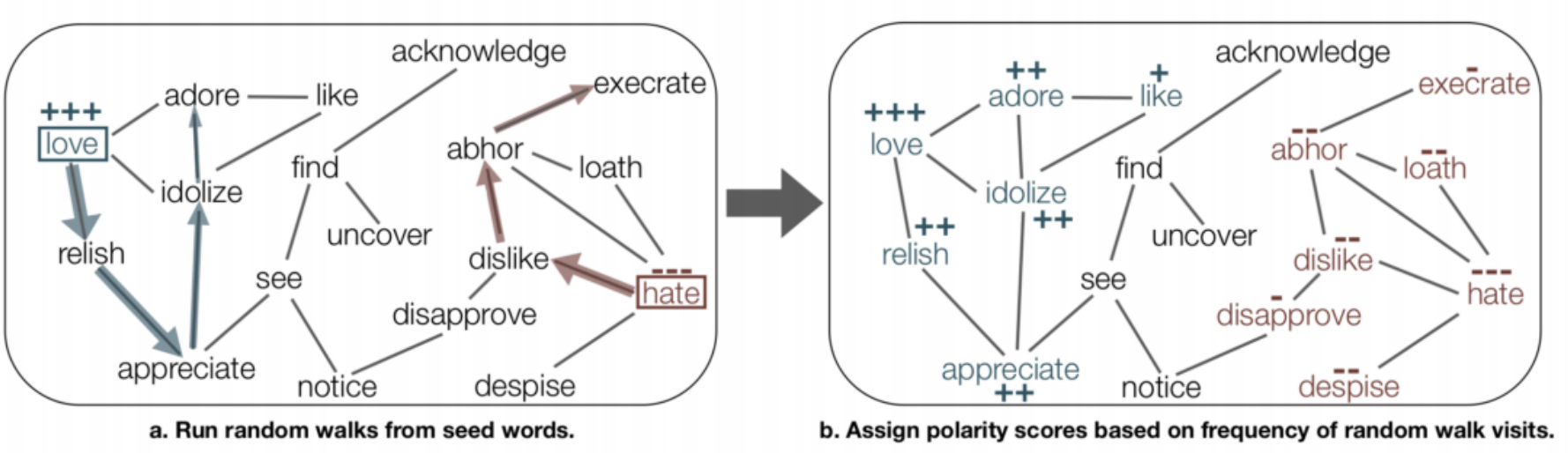
HistWords [HLJ16]

- Диахронические эмбеддинги: Прокрустово преобразование для поворота пространства эмбеддингов из периода t − 1 в t



Составления предметных словарей эмоционально-окрашенных слов [HCLJ16]

- Граф близости на словах
- Случайное блуждание для распространения метки



a. Run random walks from seed words.

b. Assign polarity scores based on frequency of random walk visits.

Домашнее задание 3

- **Целью этого задания является изучение языковых моделей и представления слов.**

- **Адрес:** login-const@mail.ru

- **Текс условия доступен по** ссылке.

Литература

- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, Enriching word vectors with subword information, arXiv preprint arXiv:1607.04606 (2016).
- Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov, Breaking sticks and ambiguities with adaptive skip-gram, Artificial Intelligence and Statistics, 2016, pp. 130–138.
- Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv´e J´egou, Word translation without parallel data, arXiv preprint arXiv:1710.04087 (2017).
- William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky, Inducing domain-specific sentiment lexicons from unlabeled corpora, Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2016, NIH Public Access, 2016, p. 595.
- William L Hamilton, Jure Leskovec, and Dan Jurafsky, Diachronic word embeddings reveal statistical laws of semantic change, arXiv preprint arXiv:1605.09096 (2016).

Литература

- Andrey Kutuzov and Elizaveta Kuzmenko, Webvectors: a toolkit for building web interfaces for vector semantic models, International Conference on Analysis of Images, Social Networks and Texts, Springer, 2016, pp. 155–16
- Omer Levy and Yoav Goldberg, Dependency-based word embeddings, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, 2014, pp. 302–308. 1.
- Quoc Le and Tomas Mikolov, Distributed representations of sentences and documents, International Conference on Machine Learning, 2014, pp. 1188–1196.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
- Jeffrey Pennington, Richard Socher, and Christopher Manning, Glove: Global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

Литература

- Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims, Evaluation methods for unsupervised word embeddings, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 298–307.
- Joseph Turian, Lev Ratinov, and Yoshua Bengio, Word representations: a simple and general method for semi-supervised learning, Proceedings of the 48th annual meeting of the association for computational linguistics, Association for Computational Linguistics, 2010, pp. 384–394.
- Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning, Bilingual word embeddings for phrase-based machine translation, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1393–1398.

СПАСИБО ЗА ВНИМАНИЕ