Обучение векторных моделей без учителя в задачах информационного поиска

Олег Найдин

2016

План

- Задача информационного поиска
- Векторные модели языка
  - обученные с учителем
  - обученные без учителя
- Некоторые модели под микроскопом
  - задача аналогии
  - задача релевантности

Информационный поиск

Text REtrieval Conference (TREC) http://trec.nist.gov

информационный поиск

Яндекс × информационный поиск

Яндекс

информационный поиск — 89 млн ответов

Найти

Войти

Поиск

Картинки

Видео

Карты

Маркет

Ещё

Информационный поиск — Википедия

ru.wikipedia.org › **Информационный поиск**

Информацио́нный по́иск (англ. information retrieval) — процесс **поиска** неструктурированной документальной **информации**, удовлетворяющей **информационные** потребности, и наука об этом **поиске**.

Основные понятия **информационного поиска**

koriolan404.narod.ru › tipis/25.htm

Информационный поиск (ИП) (англ. Information retrieval) — процесс **поиска** неструктурированной документальной **информации** и наука об этом **поиске**.

Информационный поиск - ...Что такое Информационный...

dic.academic.ru › dic.nsf/bse/90912/**Информационный**

Информационный поиск это: Толкование Перевод. ... Смотреть **что такое** "**Информационный поиск**" в других словарях

Информационный поиск

refleader.ru › otrbewpolbew.html

Информационный поиск. Вопросы: 1. Понятие **информационного поиска**. ... 4. **Информационный поиск** в Интернете.

Информационный поиск - Наука - Wikia

ru.science.wikia.com › Викинаука › **Информационный_поиск**

Информационный поиск (ИП) (английский термин Information retrieval) — наука о **поиске** неструктурированной документальной информации. В частности это относится к **поиску информации** в документах, **поиск** самих документов...

информационный поиск

Яндекс × информационный поиск

Яндекс

информационный поиск — 89 млн ответов

Найти

Войти

Поиск

Картинки

Видео

Карты

Маркет

Ещё

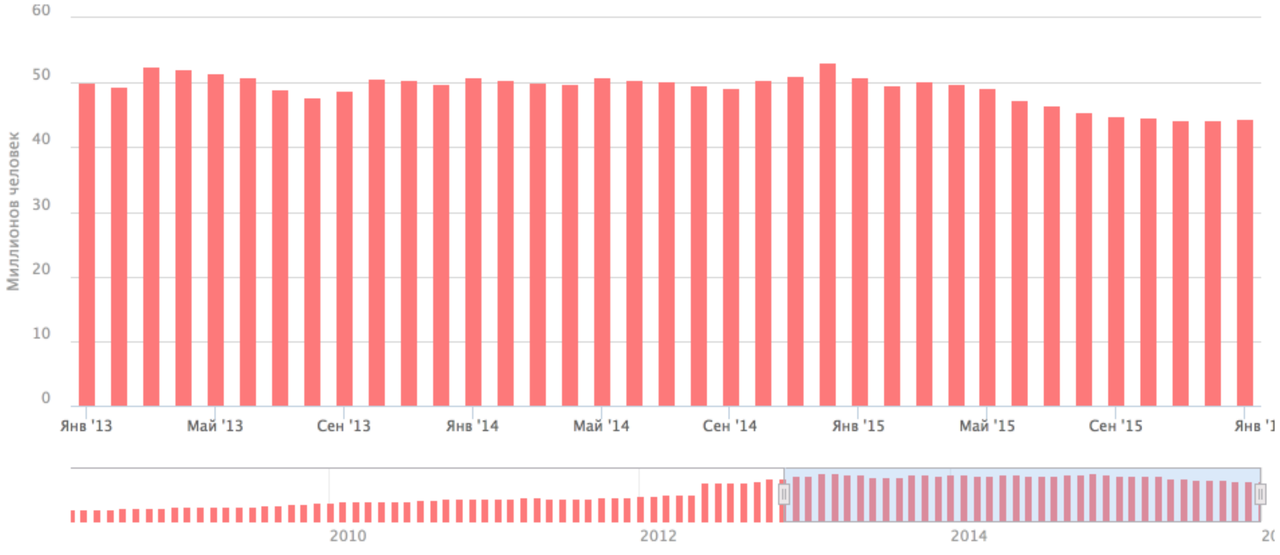
Информационный поиск — Википедия
ru.wikipedia.org › **Информационный поиск**
Информацио́нный по́иск (англ. information retrieval) — процесс **поиска** неструктурированной документальной **информации**, удовлетворяющей **информационные** потребности, и наука об этом **поиске**.

Основные понятия **информационного поиска**
koriolan404.narod.ru › tipis/25.htm
Информационный поиск (ИП) (англ. Information retrieval) — процесс **поиска** неструктурированной документальной **информации** и наука об этом **поиске**

Информационный поиск - ...Что такое Информацио
dic.academic.ru › dic.nsf/bse/90912/**Информационный**
Информационный поиск это: Толкование Перевод. ... Смотреть **что такое** "**Информационный поиск**" в других словарях

Информационный поиск
refleader.ru › otrbewpolbew.html
Информационный поиск. Вопросы: 1. Понятие **информационного поиска** **Информационный поиск** в Интернете.

Информационный поиск - Наука - Wikia
ru.science.wikia.com › Викинаука › **Информационный_поиск**
Информационный поиск (ИП) (английский термин Information retrieval) — наука о **поиске** неструктурированной документальной информации. В частности это относится к **поиску информации** в документах, **поиск** самих документов...



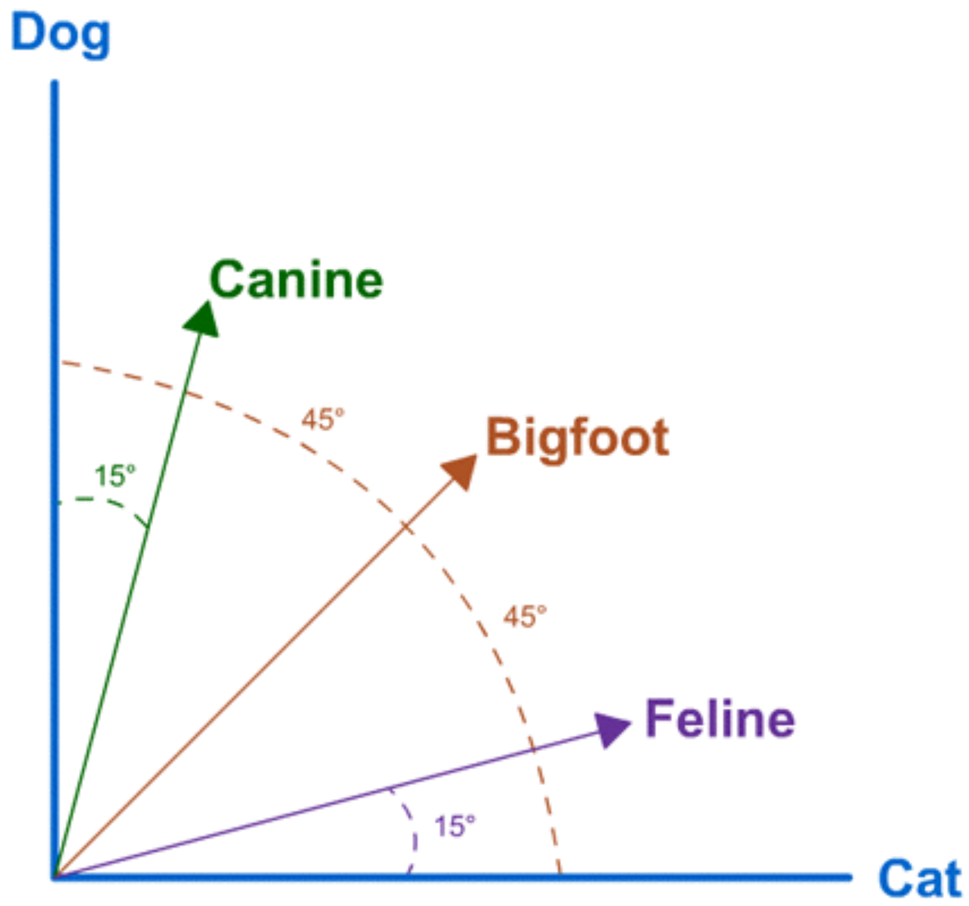
Информационный поиск

- Теоретико-множественные модели
  - различные модели булева поиска
- Алгебраические модели
  - латентно-семантический анализ
  - векторные модели
- Вероятностные модели
  - BM-25
  - латентное размещение Дирихле
- Ансамбли моделей

Информационный поиск

- Теоретико-множественные модели
  - различные модели булева поиска
- Алгебраические модели
  - латентно-семантический анализ
  - векторные модели
- Вероятностные модели
  - BM-25
  - латентное размещение Дирихле
- Ансамбли моделей

Как учить?



Векторные модели языка

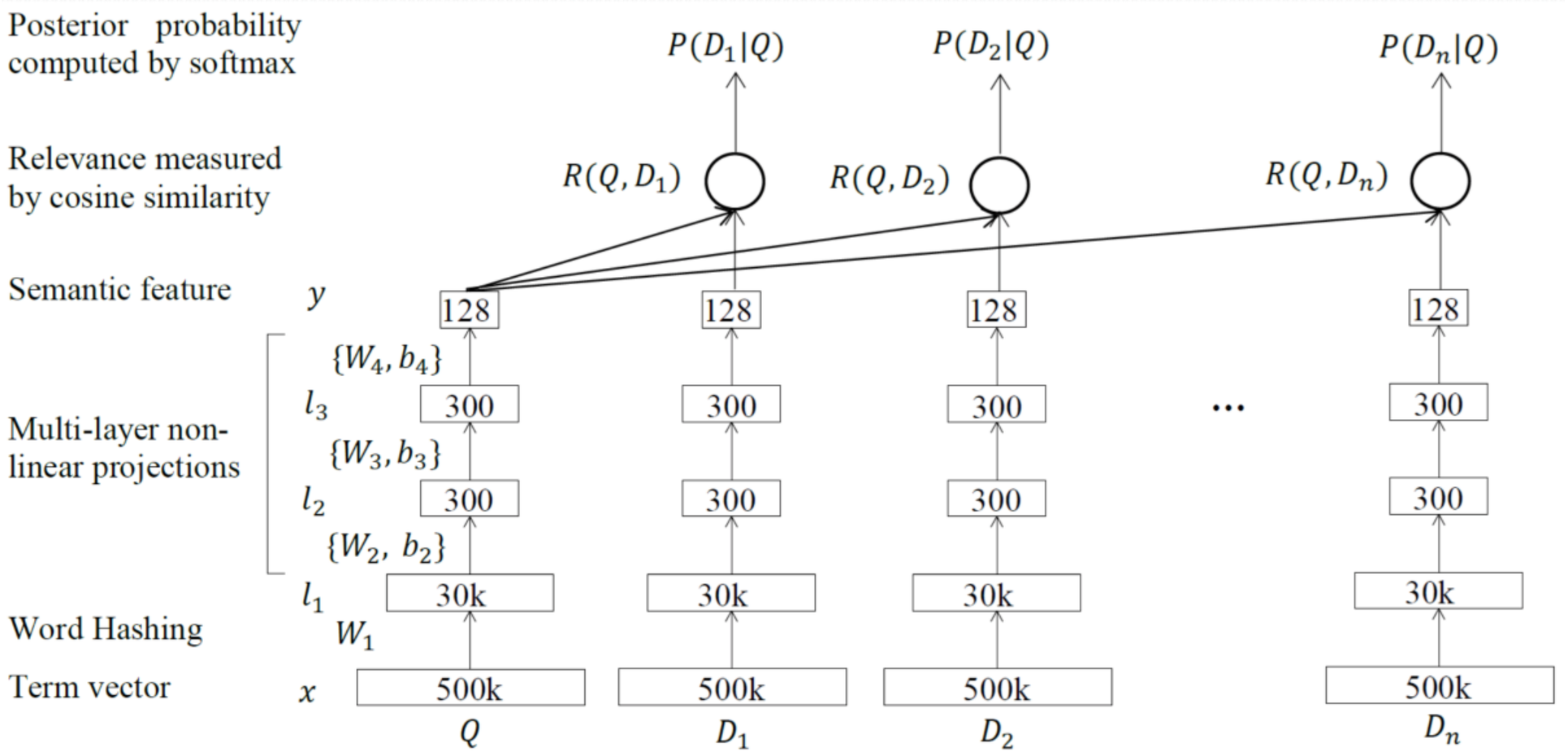
cos(**Dog**, **Canine**) > cos(**Dog**, **Feline**)

Векторные модели языка

- Локальные модели
  - унитарные коды
  - TF-IDF
- Непрерывные модели
  - латентно-семантический анализ
  - латентное размещение Дирихле
  - распределенные модели

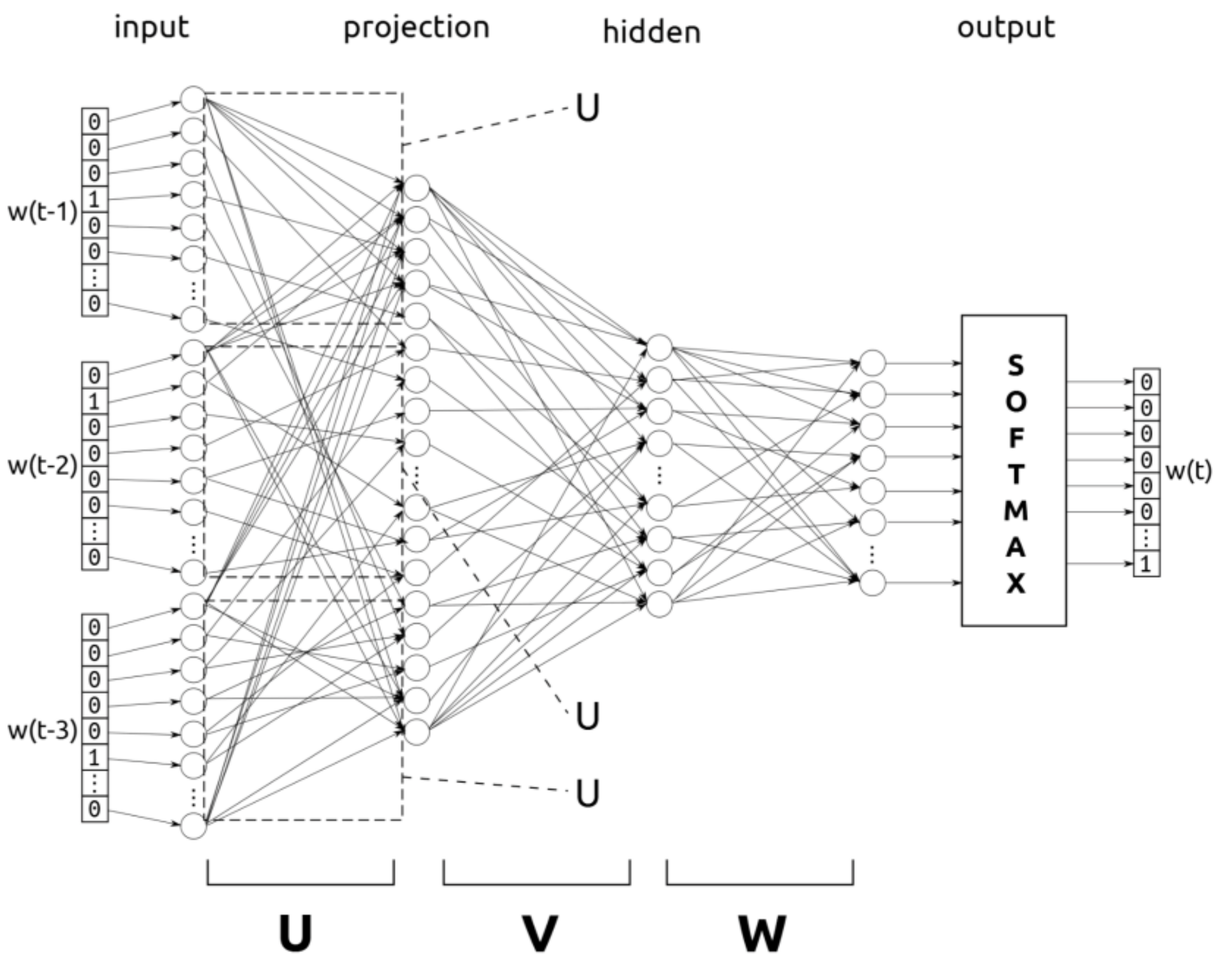
Векторные модели языка

- Локальные модели
  - унитарные коды
  - TF-IDF
- Непрерывные модели
  - латентно-семантический анализ
  - латентное размещение Дирихле
  - распределенные модели
- Обученные с учителем
  - обучение классификаторов
  - обучение метрик
- Обученные без учителя



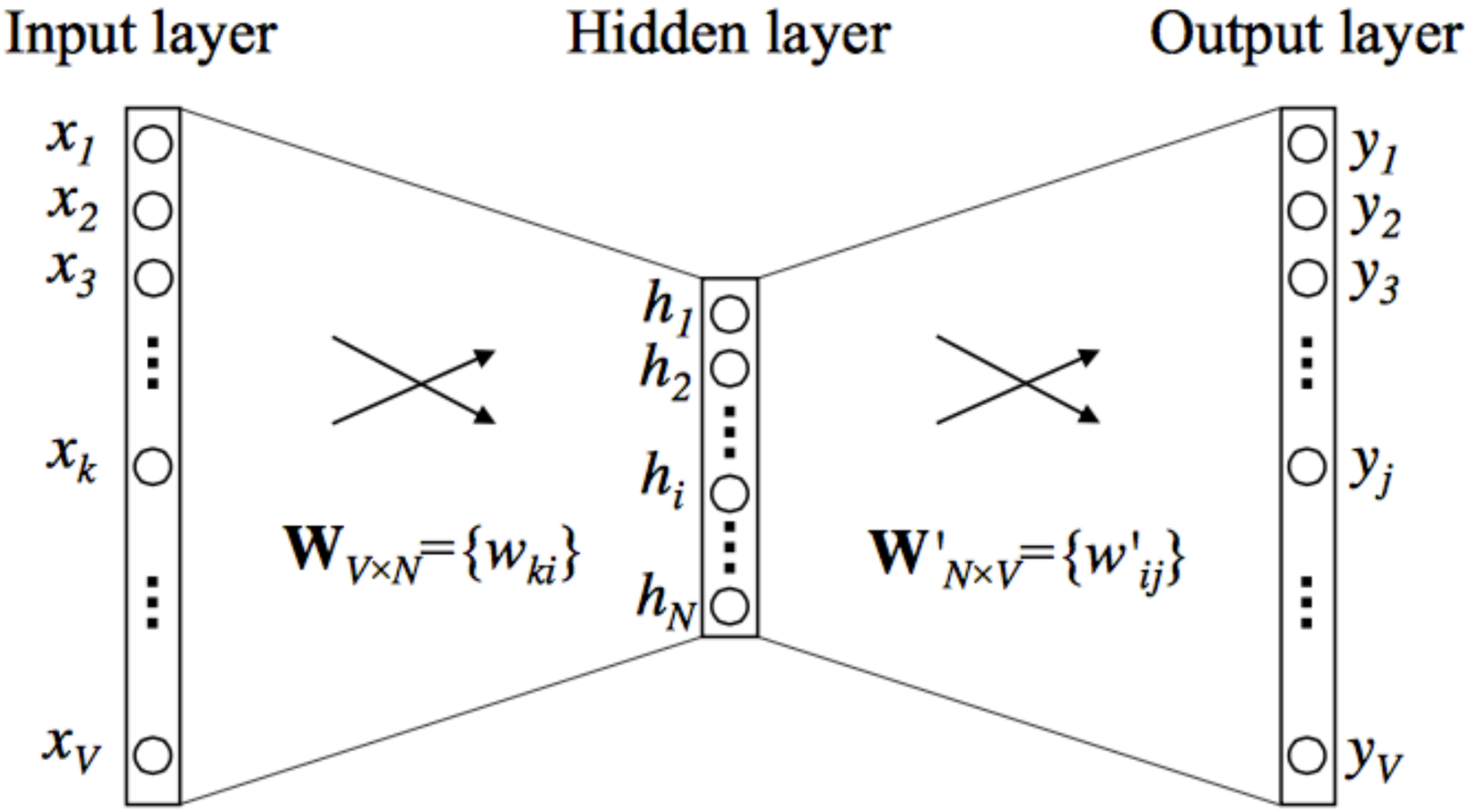
DSSM

P.-S. Huang et al. [Learning Deep Structured Semantic Models for Web Search using Clickthrough Data](), 2013



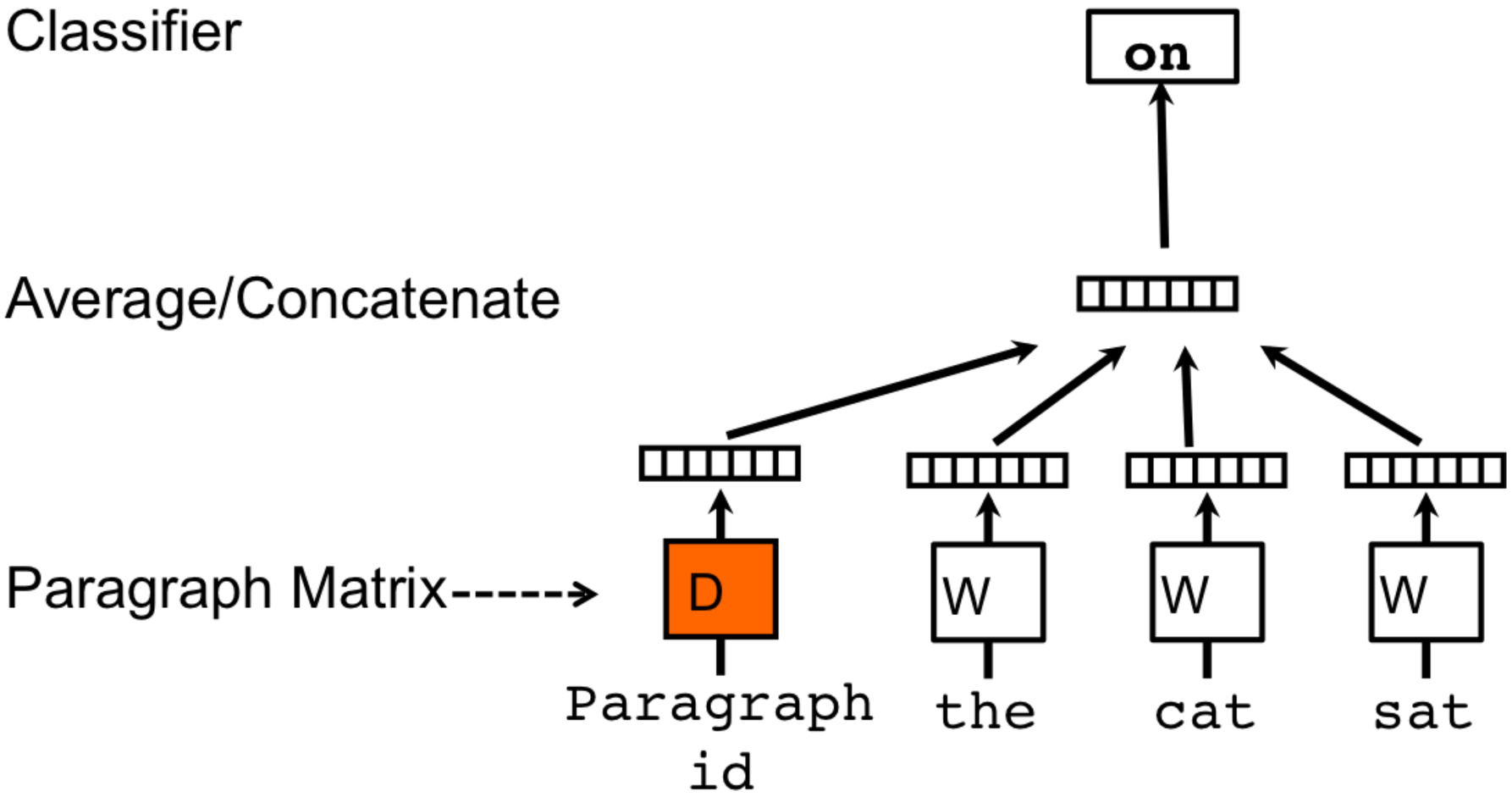
N-gram language model

Y. Bengio et al. *A Neural Probabilistic Language Model*, 2001



Word2vec

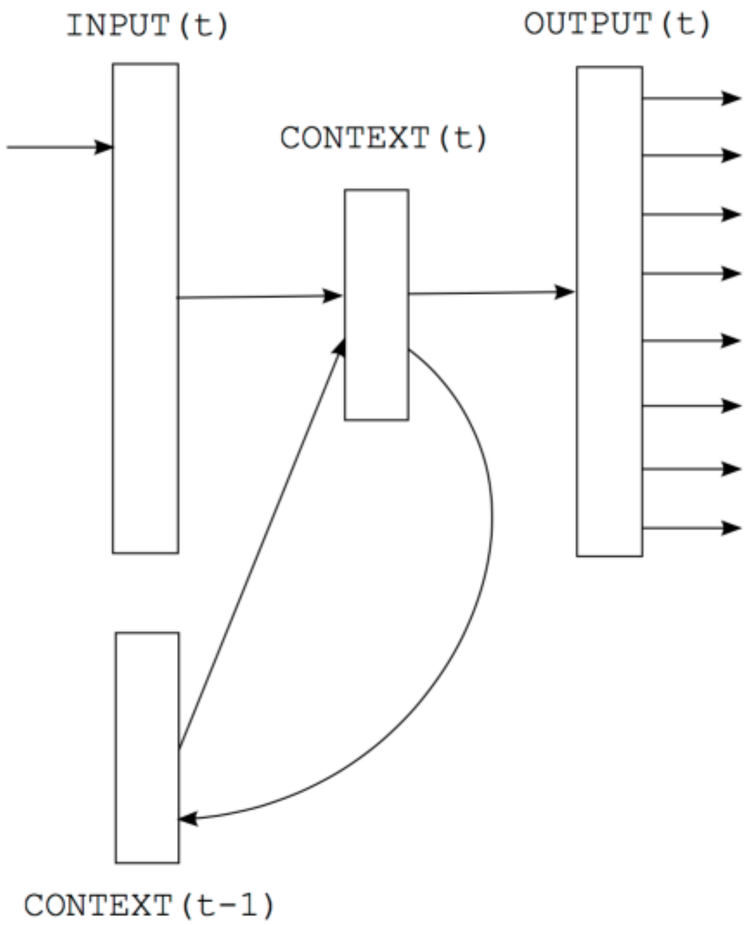
T. Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*, 2013
T. Mikolov et al. *Distributed Representations of Words and Phrases and their Compositionality*, 2013



Paragraph vector

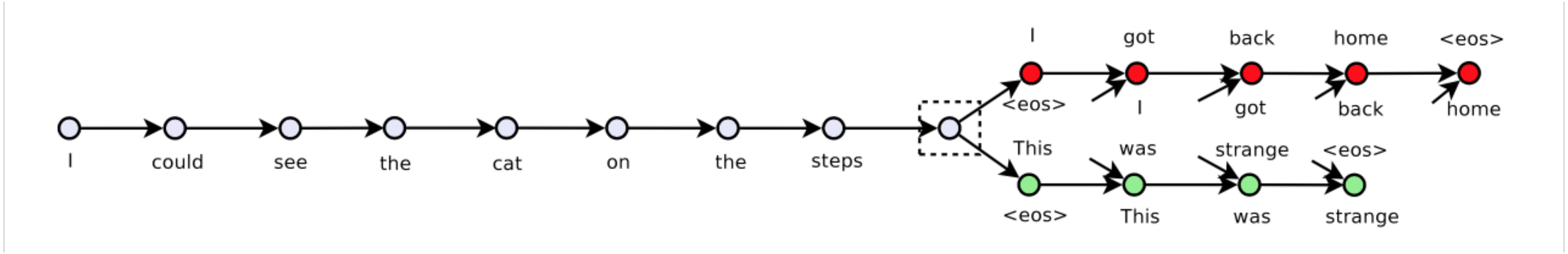
~~Q. Le & T. Mikolov *Distributed Representations of Sentences and Documents*, 2014~~

G. Mesnil et al. *Ensemble of Generative and Discriminative Techniques forSentiment Analysis of Movie Reviews*, 2014



RNNLM

T. Mikolov et al. *Recurrent neural network based language model*, 2010

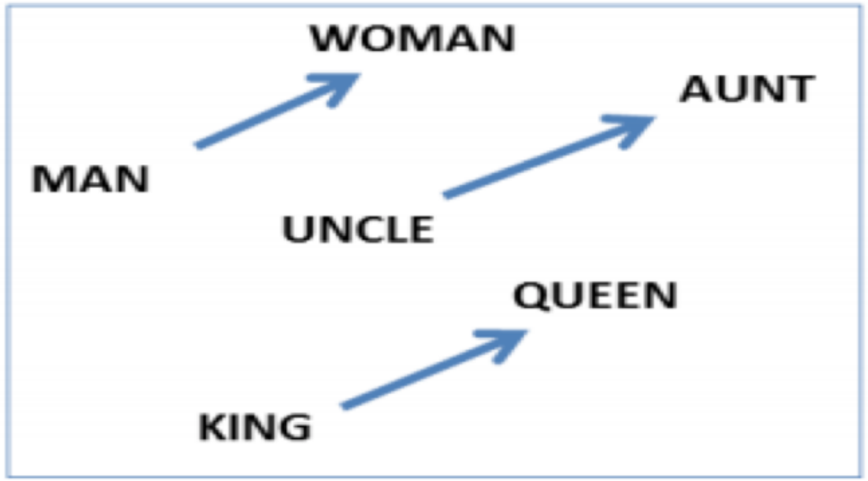


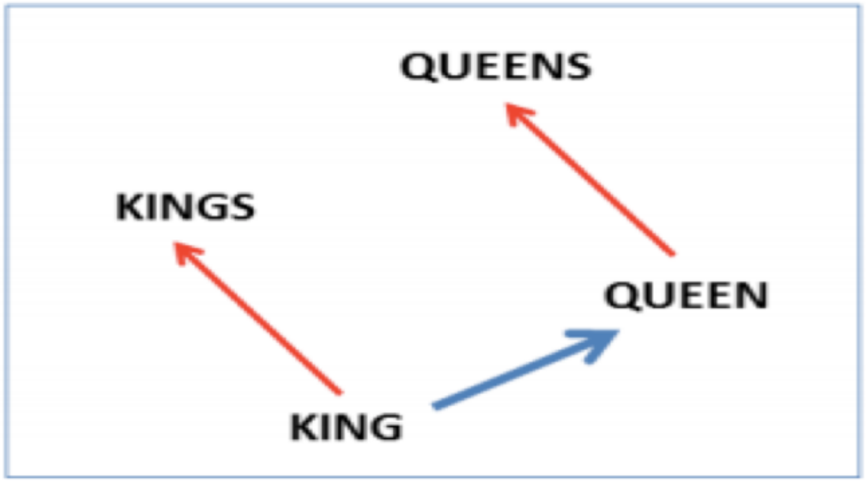
Skip-thought vectors

R. Kiros et al. *Skip-Thought Vectors*, 2015

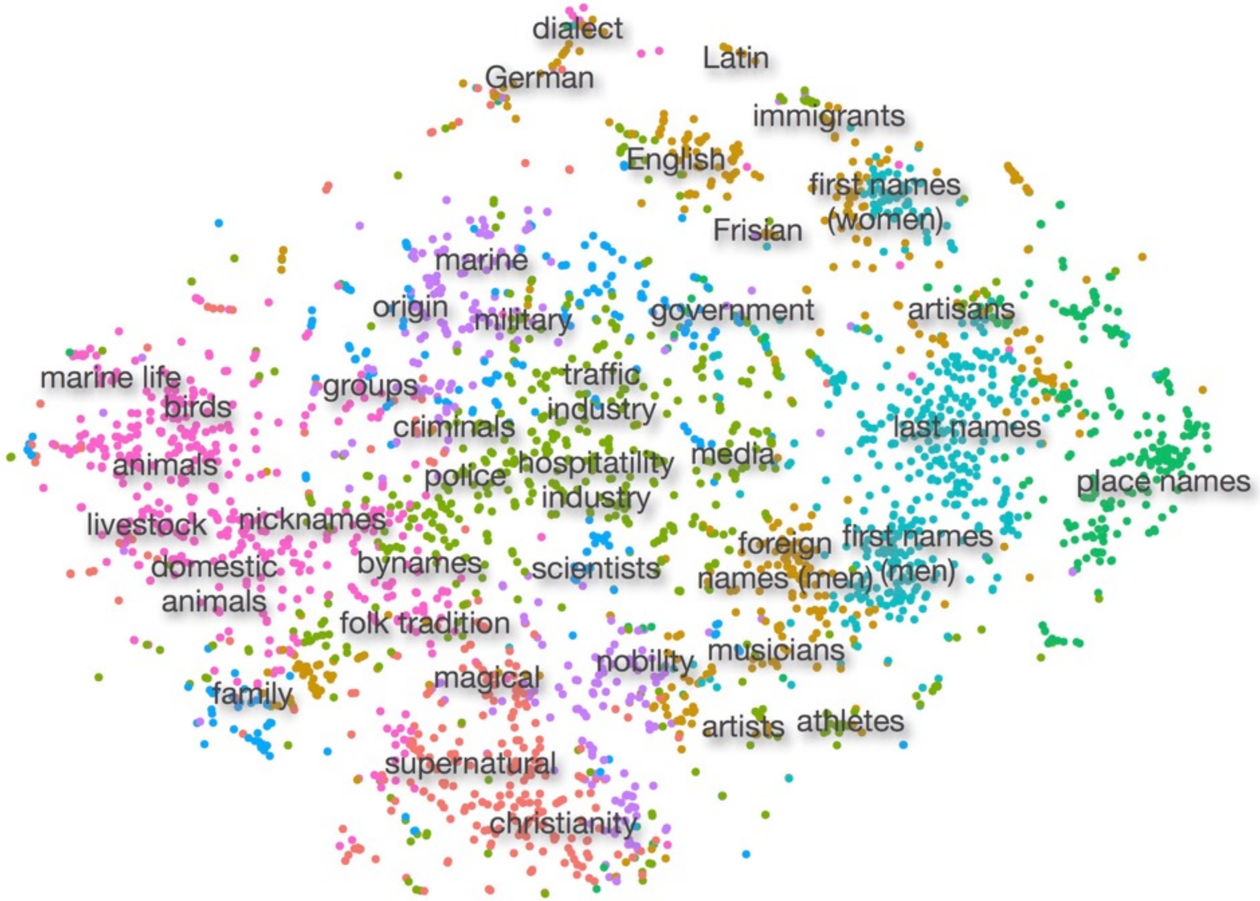
Векторные модели языка

- NBSVM (S. Wang & C. Manning *Baselines and Bigrams: Simple, Good Sentiment and Topic Classification*, 2012)
- GloVe (J. Pennington at al. *GloVe: Global Vectors for Word Representation*, 2014)
- DictRep (F. Hill et al. *Learning to Understand Phrases by Embedding the Dictionary*, 2015)
- LSTM-RNN (K. Greff at al. *LSTM: A Search Space Odyssey*, 2015)
- C-PHRASE (N. T. Pham et al. *Jointly optimizing word representations for lexical and sentential tasks with the C-PHRASE model*, 2015)
- Char-RNN (A. Karpathy *The Unreasonable Effectiveness of Recurrent Neural Networks*, 2015)





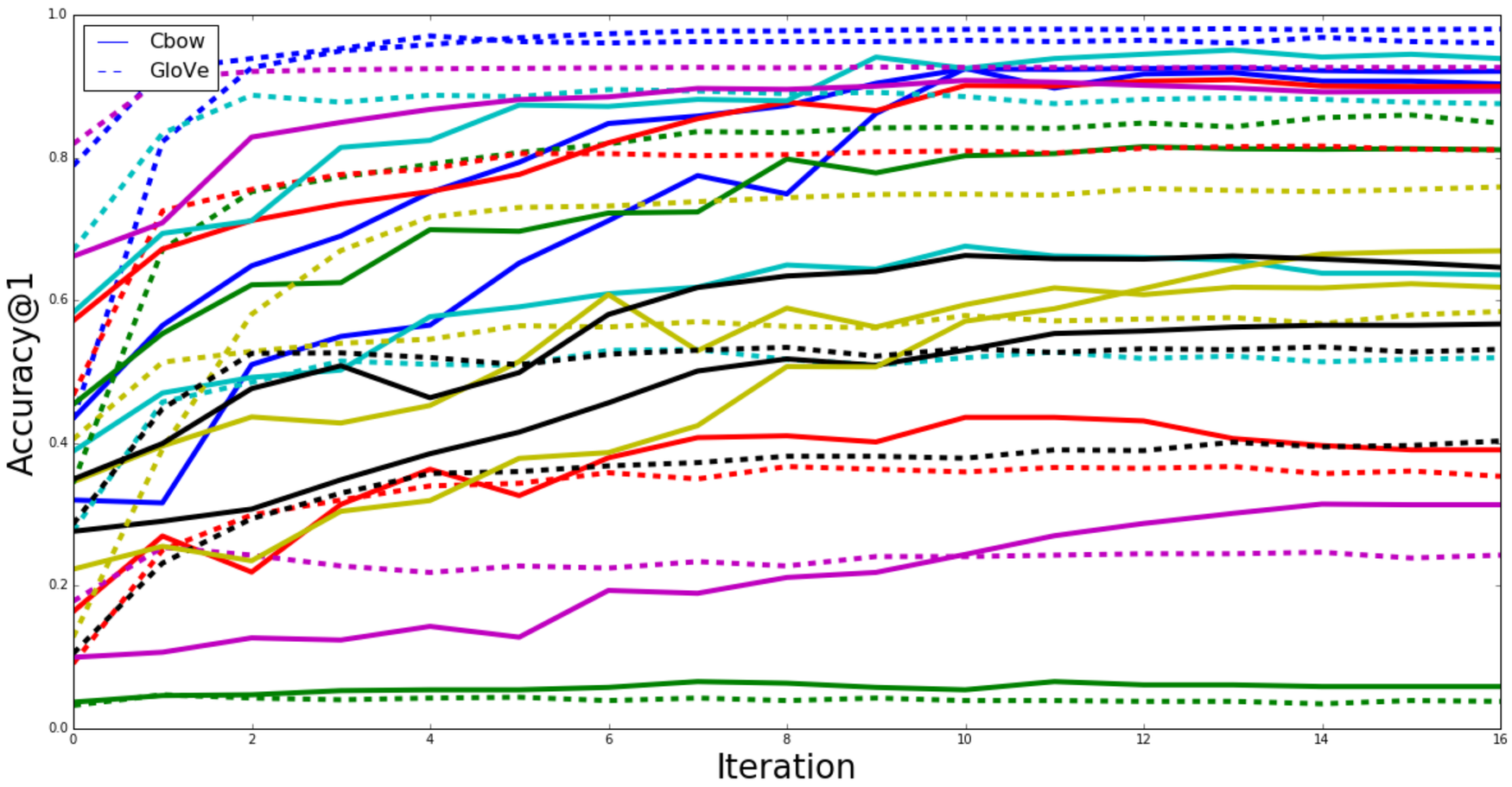
Лингвистические аналогии



Семантическая близость

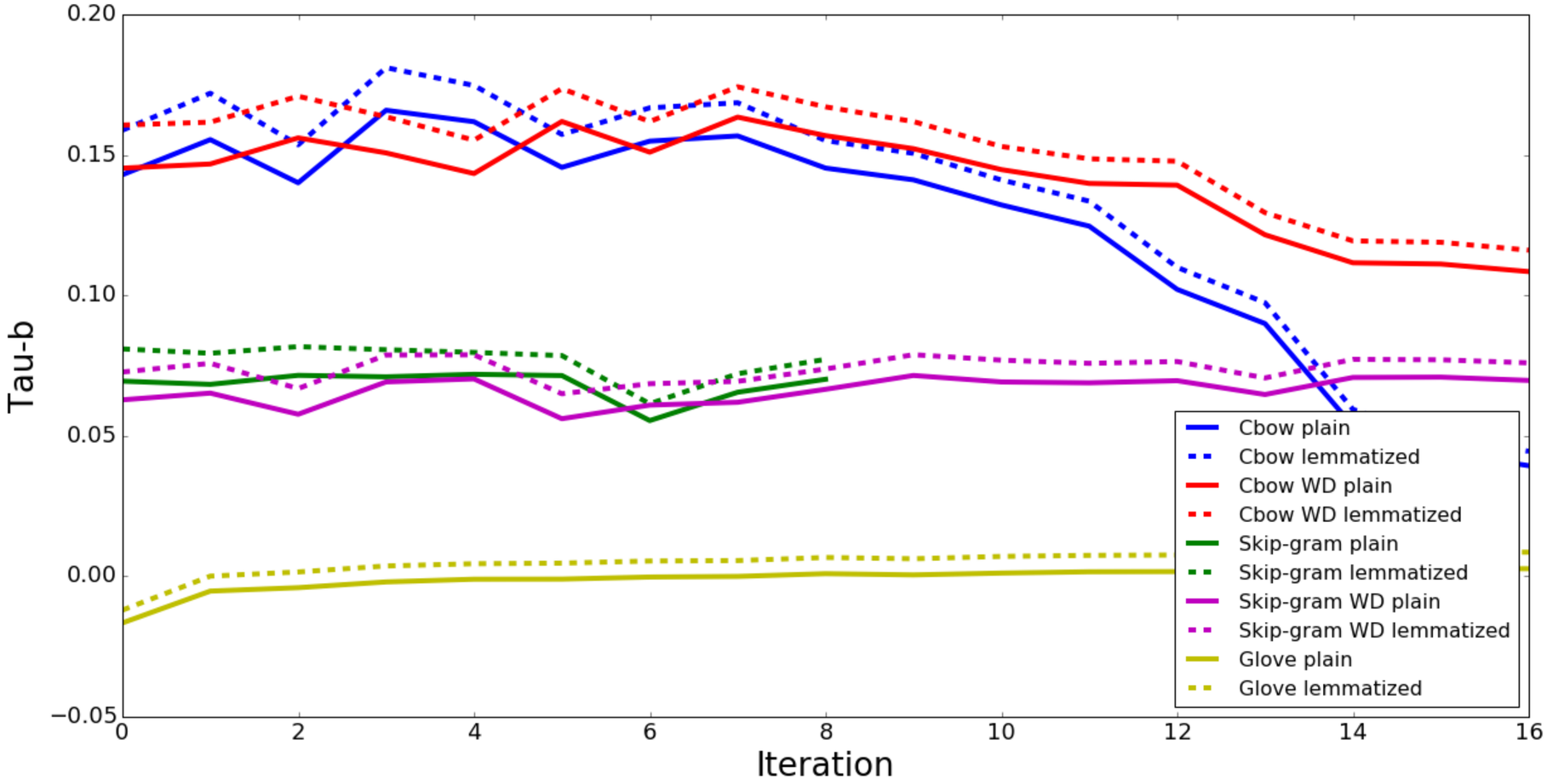
Векторные модели языка

- Задача аналогии
  - T. Mikolov et al. *Linguistic Regularities in Continuous Space Word Representations*, 2013
  - T. Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*, 2013
- Задача релевантности
  - Kaggle competition *Crowdflower Search Results Relevance* (8 months ago)
  - Kaggle competition *Home Depot Product Search Relevance* (53 days to go)



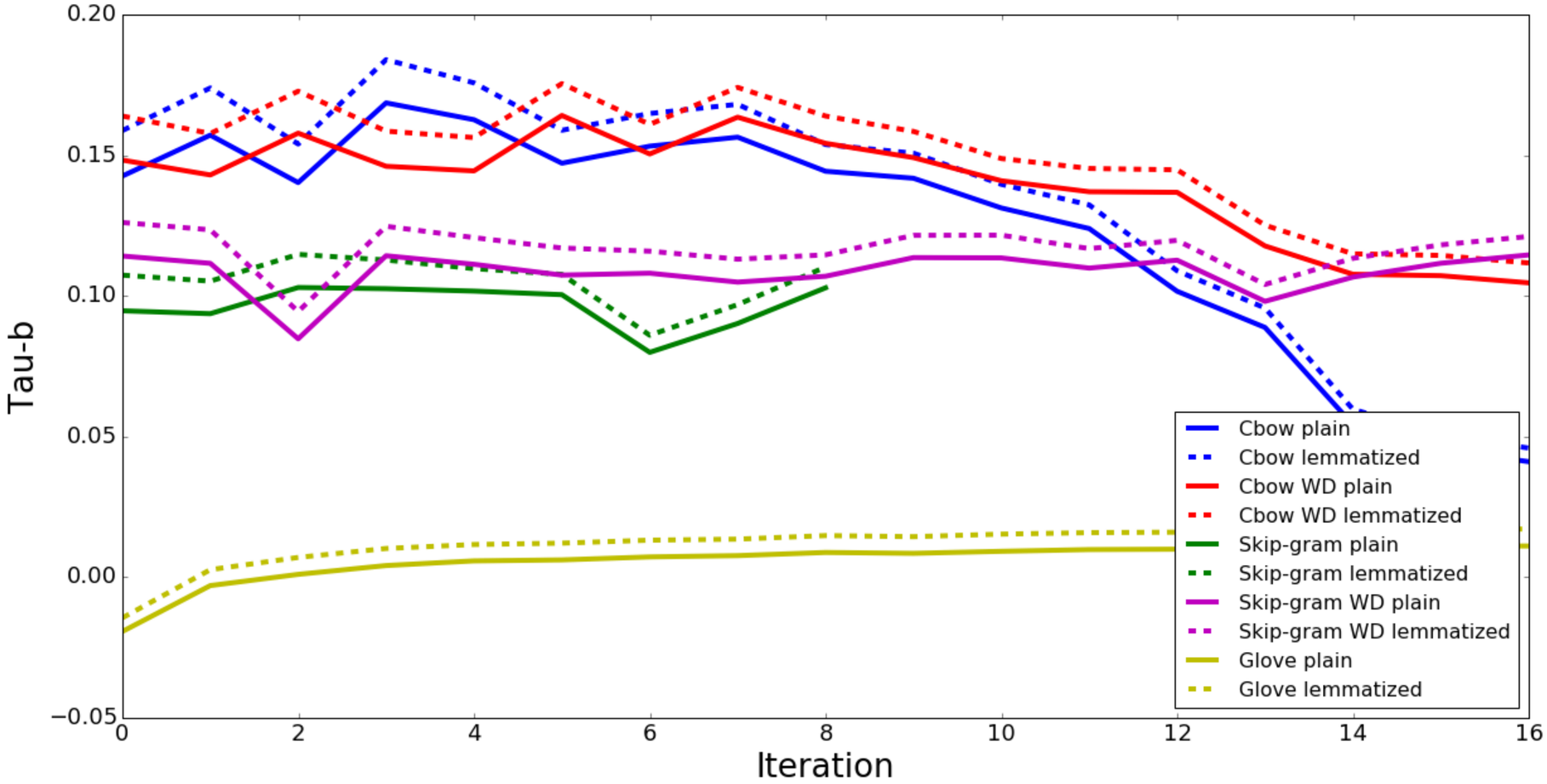
Задача аналогии

Размерность векторных представлений: 512
Предобработка векторов слов: нормирование



Задача релевантности

Размерность векторных представлений: 512
Способ агрегирования векторов слов: усреднение



Задача релевантности

Размерность векторных представлений: 512
Предобработка векторов слов: центрирование
Способ агрегирования векторов слов: усреднение

Векторные модели языка

- A. M. Dai et al. *Document Embedding with Paragraph Vectors*, 2015
- F. Hill et al. *Learning Distributed Representations of Sentences from Unlabelled Data*, 2016
- R. Jozefowicz *Exploring the Limits of Language Modeling*, 2016

Спасибо!