Natural Language Processing

Семинар 1

ВМК МГУ, 13 февраля 2017 г.

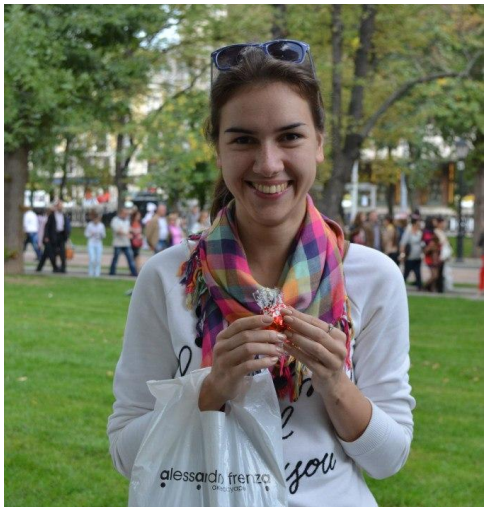
Сегодня:

- Знакомство и правила курса
- Векторное описание слов и документов:
  - Рецепты, как превратить сырой текст в матрицу
  - Дистрибутивная семантика (вектора слов)
  - Классификация (вектора документов)

Далее на семинарах:

- Рассказ про прикладные задачи в NLP
- Обзоры методов в дополнение к лекциям
- Тьюториалы в питоне и полезные инструменты
- Разбор свежих статей по разным темам

Ваши семинаристы



Анна Потапенко
(ВМК, ВШЭ, Яндекс)

Мурат Апишев (ВМК, Яндекс)

Правила игры

- Почта курса: nlp.msu@gmail.com
  - По всем вопросам туда!
  - Не в личку! :)
- Оценка:
  - накопленная (70%) + экзамен (30%)
- Задания:
  - 4 лабораторные работы в ipython notebook
  - конкурс на Kaggle in class
  - разбор научной статьи (выступление или реферат)
- Куда сдавать:
  - присылать на почту с темой вида “Имя Фамилия - Лабораторная 1”
  - дедлайн строгий, после него работа не засчитывается

Правила игры

- Страница на machinelearning.ru
  - В процессе наполнения
  - Вся информация (слайды, задания, новости...) будет там
- Обратная связь
  - Любой фидбек всячески приветствуется!
  - Почта группы для рассылок?
- Вопросы?

Темы некоторых семинаров

- Работа с последовательностями:
  - Методы: HMM, MaxEnt, CRF, LSTM, ….
  - Задачи: POS-tagging, Word Sense Disambiguation, Named Entity Recognition, Semantic Slot Filling, …

- Вероятностные модели языка
  - Методы: метод максимума правдоподобия, типы сглаживания, n-граммы, EM-алгоритм
  - Не приложения: векторные представления слов, тематическое моделирование
  - Приложения: исправление опечаток, статистический машинный перевод, распознавание речи, …

Темы некоторых семинаров

- Глубокие нейронные сети
  - Метод обратного распространения ошибки, модификации SGD
  - Рекуррентные нейронные сети (RNN, GRU, LSTM)
  - Sequence2sequence LSTM в машинном переводе
  - Рекурсивные нейронные сети (композиционная семантика)
  - Распознавание интента в диалоговых системах
  - Conversational Neural Networks

- Полезные инструменты и ресурсы
  - Python stack, NLTK, Gensim, Vowpal Wabbit, TensorFlow, ...
  - WordNet, BabelNet, НКРЯ (ruscorpora) …
  - Яндекс.Толока, Mechanical Turk

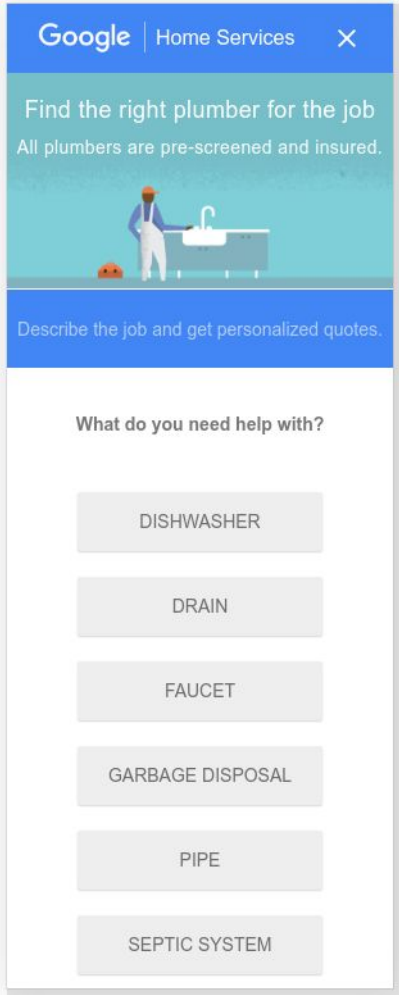
Знакомство

- Питон?
- Графические модели?
- Проекты по NLP?
- ШАД?
- Что хотелось бы узнать?
- Чему хотелось бы научиться?

Задача в Google Ads Quality

Хотим показывать специальные рекламные юниты для некоторых запросов. Что для этого нужно?

- понимать интент пользователя (аннотировать запрос)
- представлять, как устроена семантика конкретной предметной области
  - например, знать, что сантехник может помочь с “раковиной” и с “трубами”
- уметь на лету сопоставлять одно с другим



Автоматическое пополнение категорий

- Даны наборы ключевых слов для категорий
- Хотим автоматически пополнять близкими понятиями

Дистрибутивная гипотеза

“You shall know a word by the company it keeps.” -- Firth, 1957.

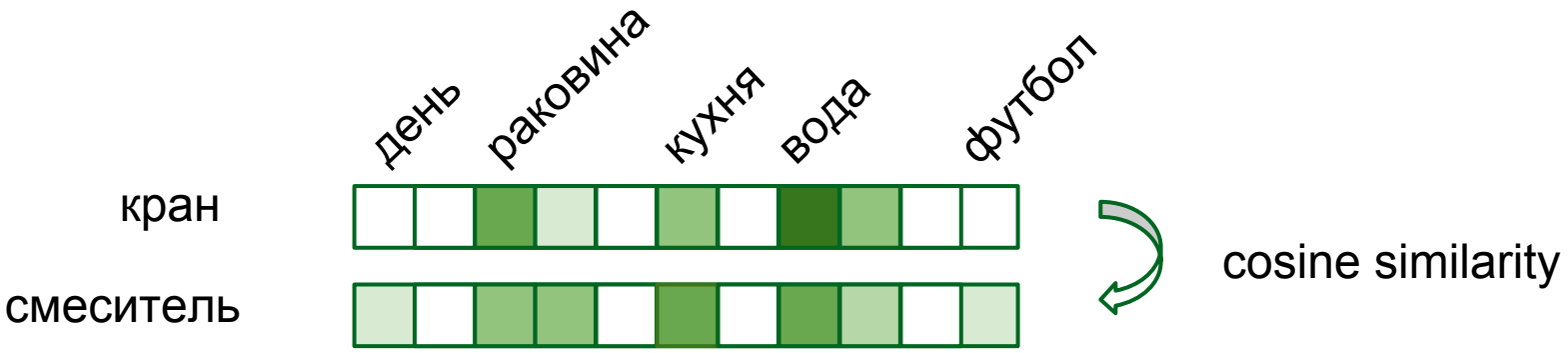
Счетчики совместной встречаемости слов:

- скользящее окно фиксированной ширины
- (positive) Pointwise Mutual Information

$$PMI = log\frac{p(u,v)}{p(u)p(v)} = log\frac{n_{uv}n}{n_u n_v}$$

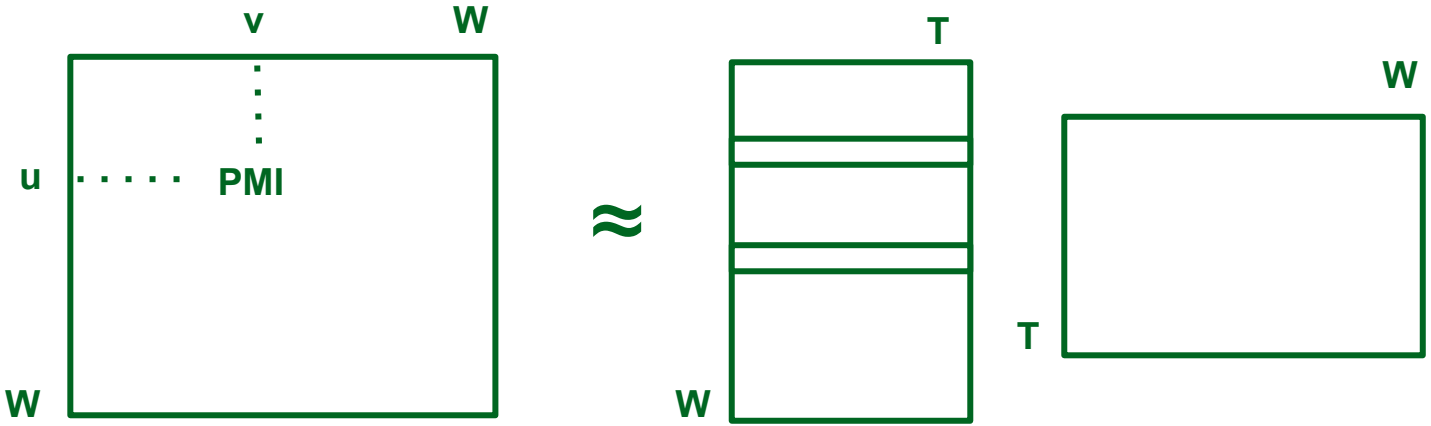
Turney, P.D., and Pantel, P. (2010), From frequency to meaning: ***Vector space models of semantics****, Journal of Artificial Intelligence Research (JAIR), 37, 141-188.*

- Со-встречаемости первого порядка:
  syntagmatic associates / relatedness (кран и вода)
- Со-встречаемости второго порядка:
  paradigmatic parallels / similarity (кран и смеситель)



Schutze, H., & Pedersen, J. (1993). A vector model for syntagmatic and paradigmatic relatedness. In Making Sense of Words: Proceedings of the Conference, pp. 104–113, Oxford, England.

Традиционные модели дистрибутивной семантики



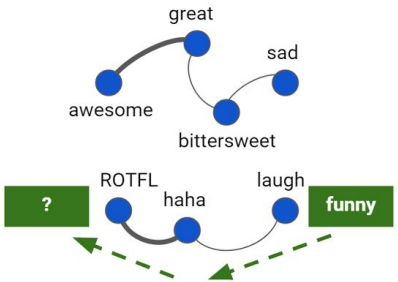
Архитектурные решения:

- Чем заполнить матрицу (слова-слова / слова-документы / слова-контексты)
- Различные способы понижения размерности (например, SVD-разложение)
- Различные меры сходства между итоговыми векторами

Что получилось:

1. Посчитали близости слов с помощью модели дистрибутивной семантики

2. Запустили распространение информации по графу



Learning Emotion Labels

PLUMBING_TARGET	ELECTRICAL_TARGET	LOCKSMITH_TARGET
basins	microwave_ovens	Safe
leak_detection	repair_refrigerators	Rekeying
Sump	Blender	Master key system
backflow_testing	Convection oven	Crash bar
gas_lines	Toaster	deadbolt_install
Grease trap	range_microwave	security_locks
pipe_lining	Self-cleaning oven	duplicate_key
Water well	blender	broken_key
water_softener	Griddle	key_duplication
Hydrant	fryer	key_cutting
trenchless_pipe	repair_appliance	combination_change
backflow_preventer	wine_cooler	mailbox_lock
Fire sprinkler	Deep fryer	key_extraction
hydro_jetting	DVD player	Break (locksmithing)
hydrojetting	fridge_freezer	locksets
water_filtration	Kitchen stove	Electromagnetic lock
Fire hydrant	Uline	Deadlock
Sewage treatment	Electric light	deadbolt_installation
Water supply	ice_makers	Key control
Relining	Cable television	Filing cabinet
water_purification	Compactor	locks_repaired
pipe_thawing	refrigerator_freezer	automotive_locksmith
ejector_pump	kettle	medeco

Спасибо за внимание!

Вопросы?