Лекция 3
Обработка текста. Классические алгоритмы, эмбеддинги слов

—

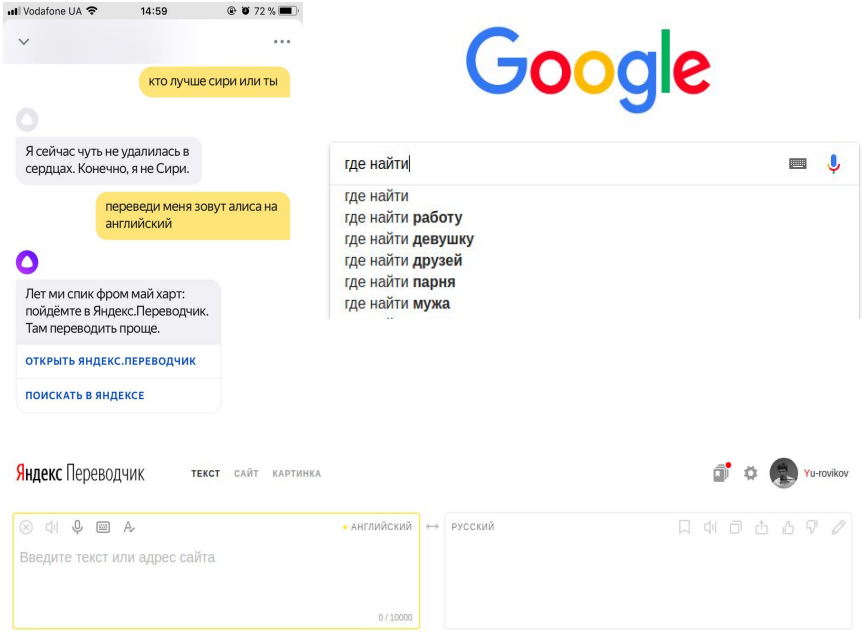
Юрий Яровиков

План лекции

- Основные задачи автоматической обработки текстов
- Простейшие текстовые признаки: Bag Of Words, TF-IDF
- Тематическое моделирование
- Эмбеддинги слов

Основные задачи обработки текста

- Классификация текстов
  - спам-фильтры
  - детекция токсичных комментариев
  - перенаправление запроса специалисту в службе поддержки
- Машинный перевод
- Тематическое моделирование
  - кластеризация большого корпуса текстов
- Ранжирование поисковых запросов



Простейшие признаки текстов

- Bag of words
  Пронумеруем все слова словаря. На k-ой позиции вектора признаков запишем количество вхождений k-ого слова в текст

Простейшие признаки текстов: TF-IDF

D — коллекция документов, t — слово (term), d — документ

Положим:

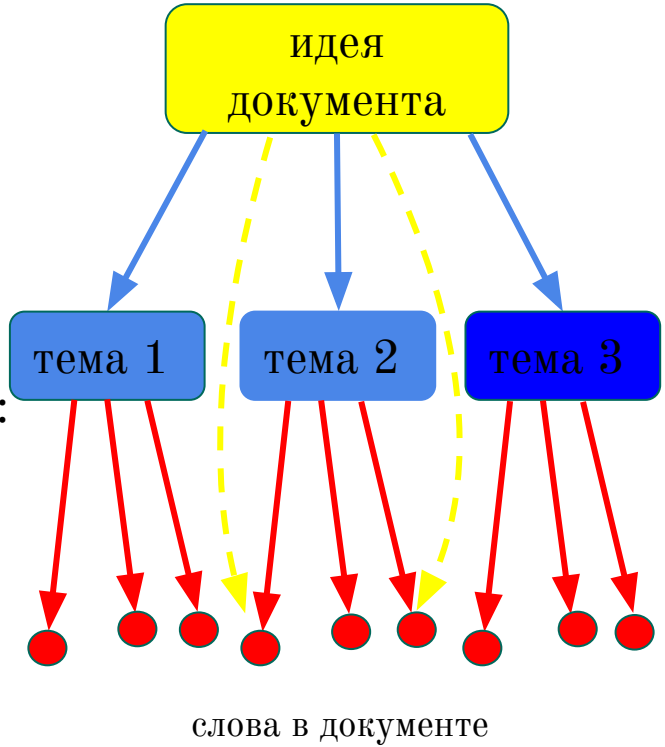
$$\mathrm{tf}(t, d) = \frac{n_t}{\sum_k n_k}$$

$$\mathrm{idf}(t, D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|}$$

$$\text{tf-idf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D)$$

Тематическое моделирование

- Тема — семантически однородный кластер текстов
- Для каждой темы типичны свои слова: для каждого t есть вероятностное распределение $\mathbf{P(w|t)}$ на множестве слов W
- У каждого документа есть некоторый список тем: для каждого d есть вероятностное распределение $\mathbf{P(t|d)}$
- Наблюдаем только $\mathbf{P(w|d)}$, необходимо восстановить $\mathbf{P(t|d)}$



Тематическое моделирование: связь с матрицей

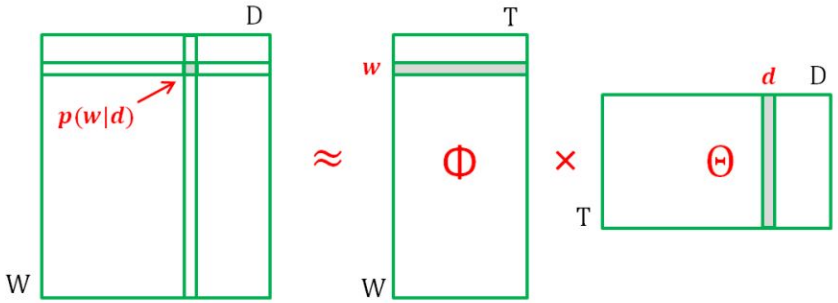
Дано: коллекция текстовых документов

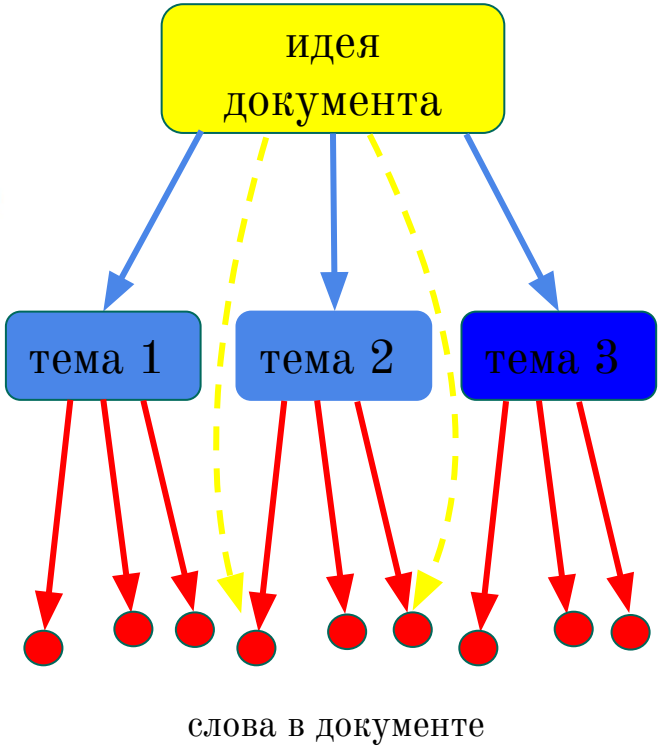
- n_{dw} — частоты термов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

Найти: параметры тематической модели $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

- $\phi_{wt} = p(w|t)$ — вероятности термов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Это задача стохастического матричного разложения:





Пример: тематическое моделирование

$$\begin{array}{l} \\ \text{Саша} \\ \text{Маша} \\ \text{Вася} \end{array} \begin{array}{c} \begin{array}{cc} \text{Р} & \text{Л} \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \\ 0.5 & 0.5 \end{pmatrix} \end{array} \times \begin{array}{c} \begin{array}{ccc} \text{банк} & \text{кино} & \text{встреча} \end{array} \\ \begin{array}{l} \text{Р} \\ \text{Л} \end{array} \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \end{pmatrix} \end{array} = \begin{array}{l} \\ \text{Саша} \\ \text{Маша} \\ \text{Вася} \end{array} \begin{array}{c} \begin{array}{ccc} \text{банк} & \text{кино} & \text{встреча} \end{array} \\ \begin{pmatrix} 0.68 & 0.2 & 0.12 \\ 0.44 & 0.4 & 0.16 \\ 0.5 & 0.35 & 0.15 \end{pmatrix} \end{array}$$

Пример: тематическое моделирование

$$\begin{array}{l} \\ \text{Саша} \\ \text{Маша} \\ \text{Вася} \end{array} \begin{array}{c} \begin{array}{cc} \text{Р} & \text{Л} \end{array} \\ \begin{pmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \\ 0.5 & 0.5 \end{pmatrix} \end{array} \times \begin{array}{l} \\ \text{Р} \\ \text{Л} \end{array} \begin{array}{c} \begin{array}{ccc} \text{банк} & \text{кино} & \text{встреча} \end{array} \\ \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \end{pmatrix} \end{array} = \begin{array}{l} \\ \text{Саша} \\ \text{Маша} \\ \text{Вася} \end{array} \begin{array}{c} \begin{array}{ccc} \text{банк} & \text{кино} & \text{встреча} \end{array} \\ \begin{pmatrix} 0.68 & 0.2 & 0.12 \\ 0.44 & 0.4 & 0.16 \\ 0.5 & 0.35 & 0.15 \end{pmatrix} \end{array}$$

Встреча кино кино кино встреча банк.

Пример: тематическое моделирование

$$\begin{array}{c} \\ \text{Саша} \\ \text{Маша} \\ \text{Вася} \end{array}\begin{array}{c}\begin{array}{cc} \text{Р} & \text{Л} \end{array}\\ \begin{pmatrix} ? & ? \\ ? & ? \\ ? & ? \end{pmatrix}\end{array} \times \begin{array}{c} \\ \text{Р} \\ \text{Л} \end{array}\begin{array}{c}\begin{array}{ccc} \text{банк} & \text{кино} & \text{встреча} \end{array}\\ \begin{pmatrix} ? & ? & ? \\ ? & ? & ? \end{pmatrix}\end{array} = \begin{array}{c} \\ \text{Саша} \\ \text{Маша} \\ \text{Вася} \end{array}\begin{array}{c}\begin{array}{ccc} \text{банк} & \text{кино} & \text{встреча} \end{array}\\ \begin{pmatrix} 0.68 & 0.2 & 0.12 \\ 0.44 & 0.4 & 0.16 \\ 0.5 & 0.35 & 0.15 \end{pmatrix}\end{array}$$

Встреча кино кино кино встреча банк.

Тематическое моделирование

- В результате восстановлено распределение тем в каждом документе
- Вероятности тем можно использовать как признаки документа

Проблемы классического подхода

- Признаки зависят от коллекции документов
- Ни один алгоритм не обрабатывает слов, ранее не встречавшихся в документах
- Признаков слишком много: столько же, сколько слов в словаре
- Признаки никак не учитывают связей между словами

Решение: Word Embeddings

- Кодируем каждое слово вектором из n элементов (например, $n = 100$)
- Хотим, чтобы близкие по смыслу слова имели близкие кодировки
- Для этого проходимся по большой коллекции текстов и наблюдаем, в каких контекстах встречается слово

Пример: Word Embeddings

What is bardiwac?

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwac.
- Nigel staggered to his feet, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.
- I dined off bread and cheese and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

Пример: Word Embeddings

What is bardiwac?

- He handed her a glass of bardiwac.
- Beef dishes are made to complement the bardiwac.
- Nigel staggered to his feet, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia's sunshine.
- I dined off bread and cheese and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

Bardiwac is a red alcoholic beverage made from grapes

Контекстуальная семантика

- A bottle of _____ is on the table. (1)
- Everybody likes _____. (2)
- Don’t have _____ before you drive. (3)
- We make _____ out of corn. (4)

Контекстуальная семантика

- A bottle of _____ is on the table. (1)
- Everybody likes _____. (2)
- Don’t have _____ before you drive. (3)
- We make _____ out of corn. (4)

What other words fit into these contexts?

Контекстуальная семантика

- A bottle of _____ is on the table. (1)
- Everybody likes _____. (2)
- Don't have _____ before you drive. (3)
- We make _____ out of corn. (4)

What other words fit into these contexts?

	(1)	(2)	(3)	(4)	…
bardiwac	1	1	1	1	
loud	0	0	0	0	
motor oil	1	0	0	1	
tortillas	0	1	0	1	
wine	1	1	1	0	
choices	0	1	0	0	

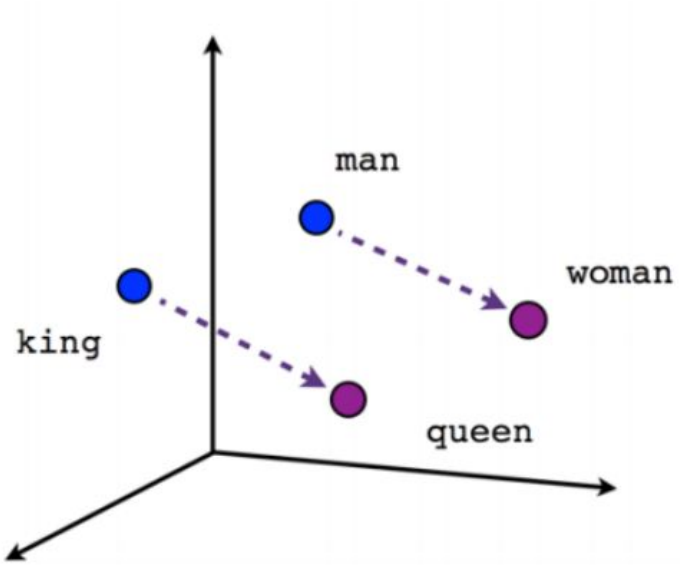
Контекстуальная семантика

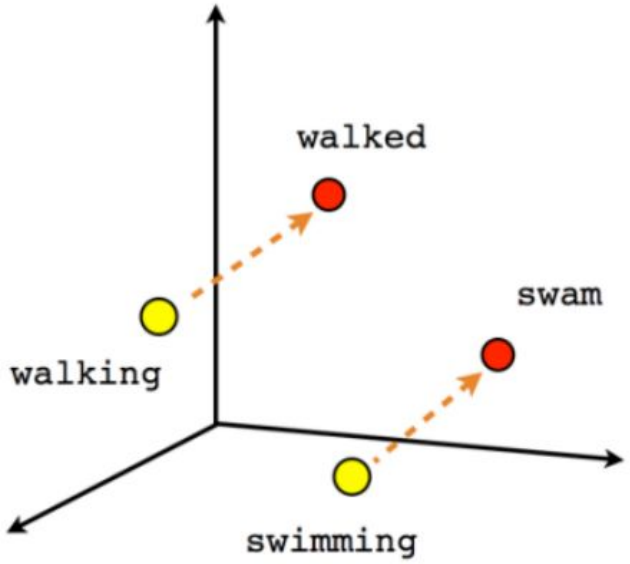
“You shall know the word by the company it keeps”

— Distributional hypothesis, J. Firth, 1957

Word2Vec

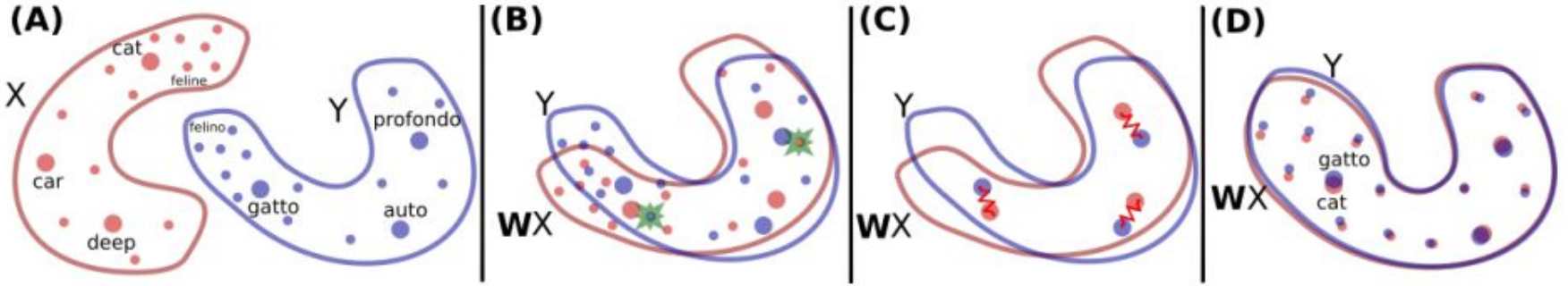
- Создан в 2012 году
- Поставляет обученные эмбеддинги слов





Машинный перевод на основе Word2Vec

- Предположение: тексты на разных языках похожи друг на друга
- Тогда и структура векторного пространства эмбеддингов должна совпадать
- Наложим пространство одного языка на пространство второго
- При поступлении нового слова находим его эмбеддинг в первом пространстве и восстанавливаем перевод из эмбеддинга на втором языке



Резюме: работа с текстами

- Исторически развитая область с множеством разработок
- В последние годы акцент области сместился на развитие нейросетевых методов, которые позволяют улучшить качество во многих задачах машинного обучения
- Эмбеддинги слов — основной способ представления слов в текстах