TOMSK POLYTECHNIC UNIVERSITY

Обработка естественного языка

Сергей Аксёнов

Доцент отделения информационных технологий,

Инженерная школа информационных технологий и робототехники

Томский политехнический университет

Томск-2023

Задачи обработки естественного языка

Задача	Описание
Токенизация	Деление текстового корпуса на неделимые единицы
Устранение неоднозначности слов	Определение правильного значения слова
Распознавание именованных сущностей	Выделение сущностей (имен, компаний, лекарств, городов, и т.д.) из текстового корпуса
Морфологическая разметка	Определение частей речи в предложении и их аннотирование
Классификатор предложений	Отнесение текстов к определенным классам
Генерация языка	Генерация новых текстов с помощью примеров
Системы вопросов и ответов	Чат-боты, поиск информации и представление знаний
Машинный перевод	Преобразование текстов с одного языка в другой

Токены

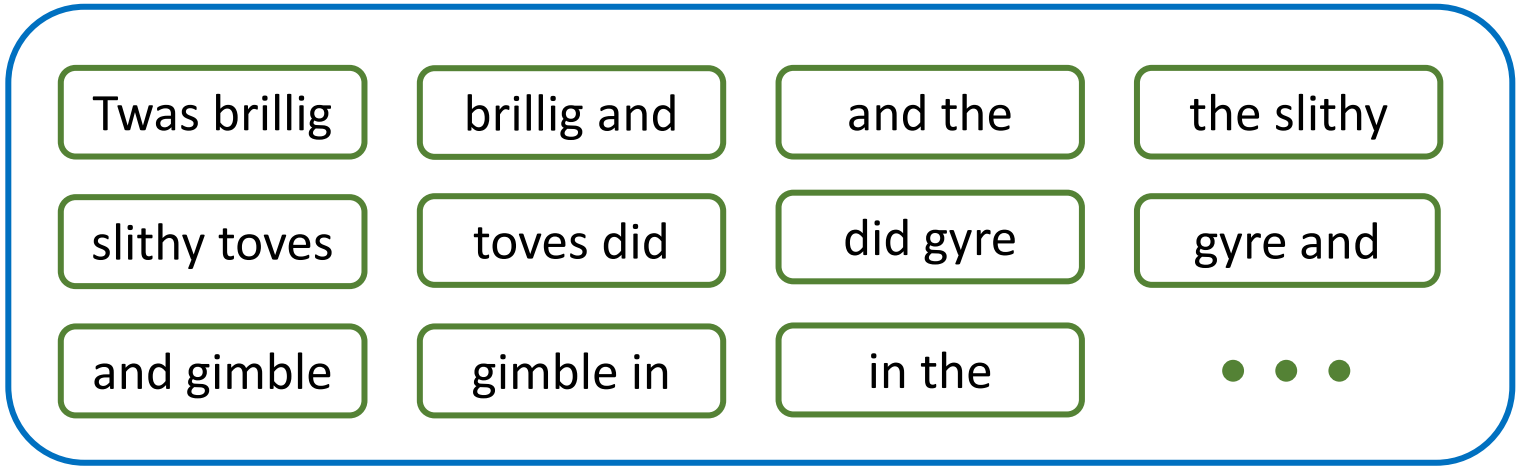
- Байты – символы ASCII
- Символы – многобайтные символы
- Части слов – слоги и общие группы символов
- Слова – слова словарей или их корни (стеммы и леммы)
- Наборы слов – используемые словосочетания, близкорасположенные слова

N-граммы

Группы из N близкостоящих слов, извлекаемых из текста.

«Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.»

Оригинальный текст



Набор биграмм

Токенизация

- Токенизатор – символы ASCII
- Словарь – многобайтные символы
- Парсер– слоги и общие группы символов
- Токен, терм, слово или n-грамма – слова словарей или их корни (стеммы и леммы)
- Выражение – используемые словосочетания, близкорасположенные слова

Унитарное кодирование (One-Hot Encoding)

Пусть имеется словарь размера *V*, то каждое *i*-е слово *wi* представляется в виде вектора длины *V* и вида [0, 0, 0, …, 0, 1, 0, …, 0, 0, 0].

Бактриан издавна являлся важным домашним животным в Азии.

Бактриан - [1, 0, 0, 0, 0, 0, 0]
Издавна - [0, 1, 0, 0, 0, 0, 0]
Является - [0, 0, 1, 0, 0, 0, 0]
Важное - [0, 0, 0, 1, 0, 0, 0]
Домашнее - [0, 0, 0, 0, 1, 0, 0]
Животное - [0, 0, 0, 0, 0, 1, 0]
Азия - [0, 0, 0, 0, 0, 0, 1]

Дополнительно может быть добавлено в словарь спецслово, соответствующее всем несловарным словам (OOV, out-of-vocabulary).

Матрица совместной встречаемости

Дыня – вид рода Огурец. Плод дыни – тыквина.

контекст

	вид	дыня	огурец	плод	род	тыквина
вид	0	1	0	0	1	0
дыня		0	0	1	0	1
огурец			0	0	1	0
плод				0	0	0
род					0	0
тыквина						0

Сложно использовать значение контекста, превышающее 1.

Вариант: вес слова в контексте уменьшается с расстоянием от слова интереса.

Мешок слов (Bag of Words - BoW): Пример

Словарь: blood, body, cell, forty, deep, give, make, percent, presence, red, shade.
Stop words: a, about, and, of, the.

Vector template:

blood	body	cell	forty	deep	give	make	percent	presence	red	shade

Red blood cells and white blood cells make about forty percent of the blood.

blood	body	cell	forty	deep	give	make	percent	presence	red	shade
3	0	2	1	0	0	1	1	0	1	0

The presence of the red blood cells gives the blood a deep-red shade.

blood	body	cell	forty	deep	give	make	percent	presence	red	shade
2	0	1	0	1	1	0	0	1	2	1

Частота термина-обратная частота документа (TF-IDF)

Term frequency (the number of times that term t occurs in document d):

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

Inverse document frequency (measure of how much information the word provides):

$$IDF(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}$$

Term frequency-Inverse document frequency:

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

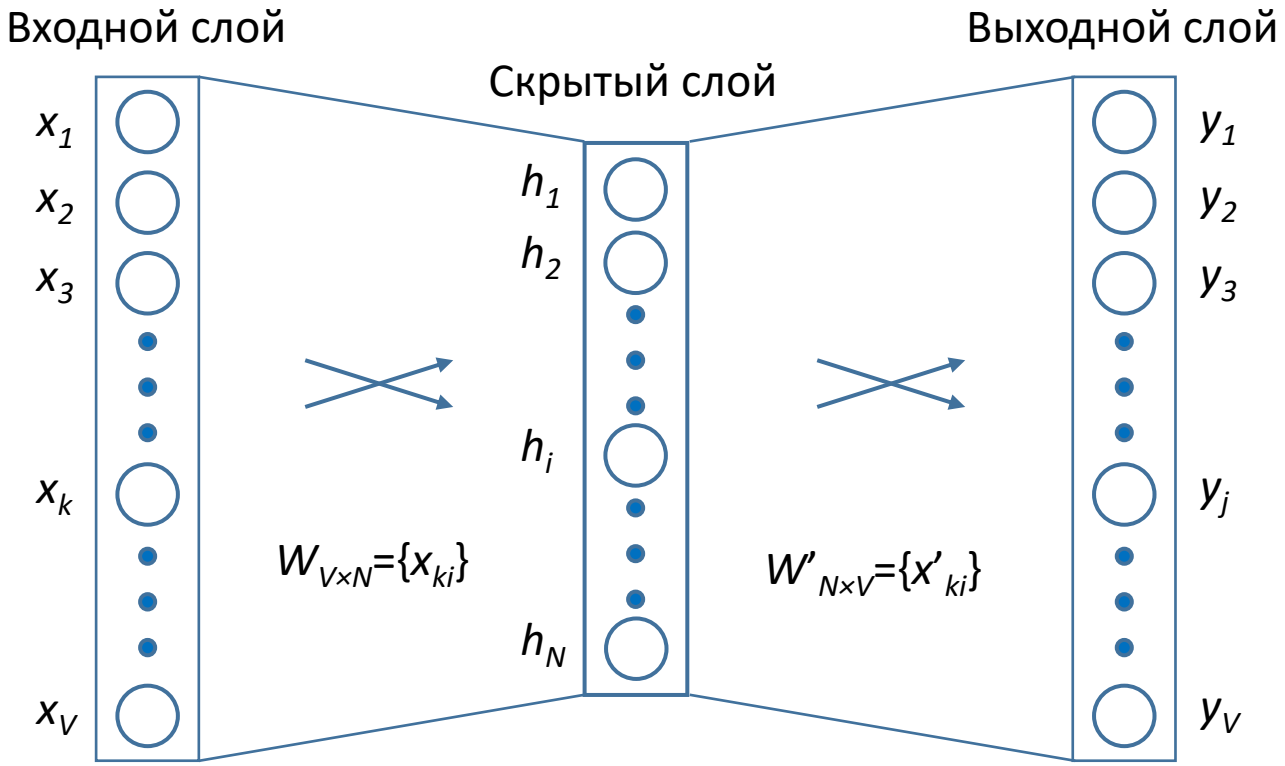
TF-IDF Example

Doc #1: *Eskimos leben im Norden*. (German: Eskimos live in the north)
Doc #2: *Eskimos wollen frische Fische essen*. (German: Eskimos want to eat fresh fish)
Doc #3: *Fische wollen das Futter fressen*. (German: The fish wants to eat)
Doc #4: *Frische Fische schmecken fein*. (German: Fresh fish tastes great)
Total number of words: 23.
The word *Fische* occurs in 3 sentences.
TF(word: *Fische*, Document: *Fische wollen das Futter fressen.*) = 1 /5 = 0.2.
IDF(word: *Fische*, Documents 1-5) = log10 (4/3) ≈ 0.125.
TF-idf(Word *Fische*, document *Fische wollen das Futter fressen.*, documents 1-5) ≈ 0.2 x 0.125 = 0.025.

Doc ID	das	Eskimos	essen	fein	fressen	Fische	Frische	Futter	im	leben	Norden	schmecken	wollen
1	0	0.075	0	0	0	0	0	0	0.15	0.15	0.15	0	0
2	0	0.06	0.12	0	0	0.025	0.12	0	0	0	0	0	0.06
3	0.12	0	0	0	0.12	0.025	0	0.12	0	0	0	0	0.06
4	0	0	0	0.15	0	0.03	0.15	0	0	0	0	0.15	0

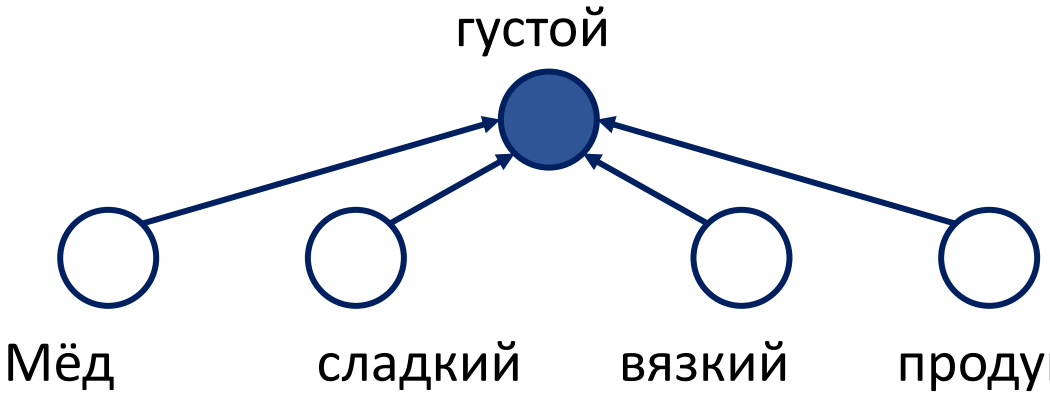
Предсказание следующего слова

- Задача предсказания слова по предыдущему слову
- Вход – one-hot вектор предыдущего слова
- Функция активации выходного слоя – softmax
- *W* или *W'*– матрицы эмбеддингов



Word2vec: The Continuous Bag of Words (CBOW) - 1

P(”густой” |”мёд”, ”сладкий”, ”вязкий”, ”продукт”)

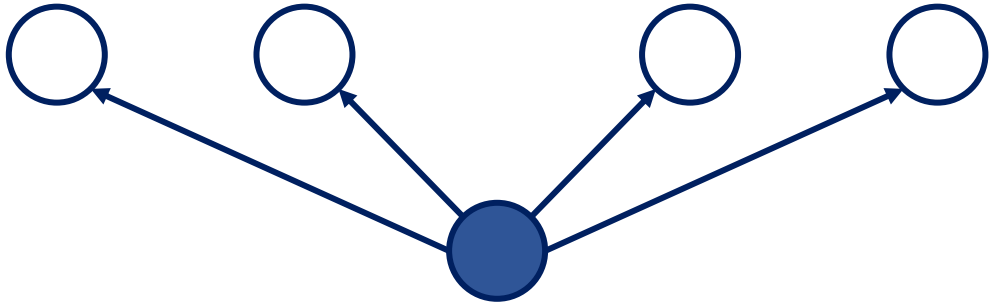


$$P(w_c|w_{o_1},\dots,w_{o_{2m}}) = \frac{\exp\left(\frac{1}{2m}u_c^\top\left(\mathrm{v}_{o_1}+\cdots+\mathrm{v}_{o_{2m}}\right)\right)}{\sum_{i\in v}\exp\left(\frac{1}{2m}u_i^\top\left(\mathrm{v}_{o_1}+\cdots+\mathrm{v}_{o_{2m}}\right)\right)} \qquad \mathcal{W}_o = \{w_{o_1},\dots,w_{o_{2m}}\}$$

$$\bar{\mathrm{v}}_o = \frac{\left(\mathrm{v}_{o_1}+\cdots+\mathrm{v}_{o_{2m}}\right)}{2m} \qquad P(w_c|\mathcal{W}_o) = \frac{\exp(u_c^\top\bar{\mathrm{v}}_o)}{\sum_{i\in v}\exp\left(u_i^\top\bar{\mathrm{v}}_o\right)}$$

Функция потерь: $J(\theta) = \prod_{t=1}^{T} P(w^{(t)}|w^{(t-m)},\dots,w^{(t-1)},w^{(t+1)},\dots,w^{(t+m)})$

Word2vec: Модель Skip-Gram - 1



$2m$ Слова контекста

Центральное слово

P(”мёд”, ”сладкий”, ”вязкий”, ”продукт” |”густой”)

P(”мёд”|”густой”) · *P*(”сладкий”|”густой”) · *P*(”вязкий”|”густой”) · *P*(”продукт”|”густой”)

$$P(w_o|w_c) = \frac{\exp(u_o^\top \mathrm{v}_c)}{\sum_{i\in\nu}\exp(u_i^\top \mathrm{v}_c)}$$

v_i - представление центрального слова

u_i - представление контекстного слова $\nu = \{0, 1, \ldots, |\nu| - 1\}$

Функция потерь: $J(\theta) = \prod_{t=1}^{T} \prod_{-m\le j\le m, j\neq 0} P(w^{(t+j)}|w^t)$

T - Длина текстовой последовательности

Skip-Gram Training

$$-\log\big(J(\theta)\big) = -\sum_{t=1}^{T} \prod_{-m \le j \le m, j \neq 0} \log\Big(P\big(w^{(t+j)} \big| w^{t}\big)\Big)$$

$$\log\big(P(w_o|w_c)\big) = u_o^{\top} \mathrm{v}_c - \log\Big(\sum_{i \in v} \exp(u_i^{\top} \mathrm{v}_c)\Big)$$

$$\frac{\partial \log(P(w_o|w_c))}{\partial \mathrm{v}_c} = u_o - \frac{\sum_{i \in v} \exp\big(u_j^{\top} \mathrm{v}_c\big)\, u_j}{\sum_{i \in v} \exp\big(u_i^{\top} \mathrm{v}_c\big)} = u_o - \sum_{j \in v} P(w_j|w_c) u_j$$

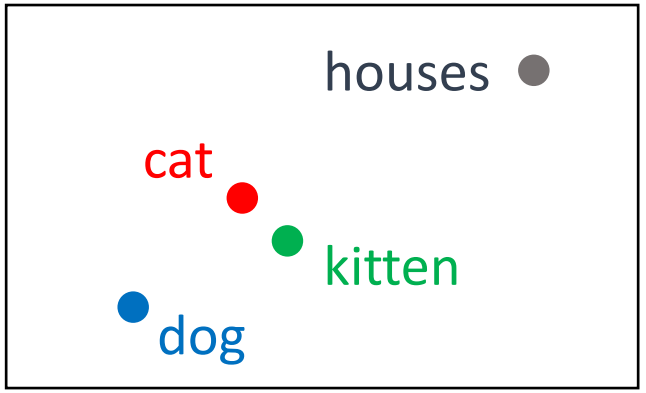
Center word vectors of the skip-gram model are typically used as the word representations.

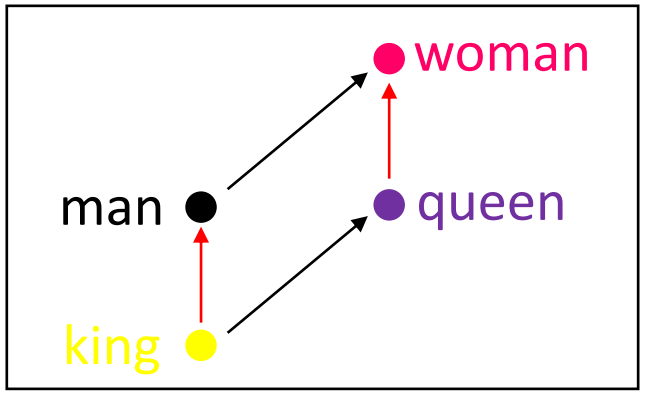
Эмбеддинги (вложения) слов (word2vec)

	living being	feline	human	gender	royalty	verb	plural
cat	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
kitten	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
dog	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
houses	-0.8	-0.4	-0.5	0.1	-0.9	-0.3	0.8
man	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
woman	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
king	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
queen	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

7D→2D

7D→2D





Слово

Эмбеддинги

Сокращение размерности

Визуализация векторов в 2D

Языковые модели (Language Model)

- Языковая модель позволяет оценить вероятность следующего слова в последовательности и оценить вероятность всей последовательности слов.
- Пример: Какое слово в последовательности вероятнее:

  Кошки ценятся человеком за умение забавлять …

  1) детей, 2) грызунов, 3) природу
- Какая последовательность вероятнее:

  Кошки ценятся человеком за умение забавлять детей.

  Забавлять человеком за кошки умение детей ценятся.