Анализ текстов

Лекция 11

Как научить компьютер читать?

Разрешение анафоры

“Мама вымыла раму, и теперь она блестит”

“Мама вымыла раму, и теперь она устала”

Задачи обработки текстов

1. Синтаксический анализ
2. Семантический анализ (классификация)
3. Порождение текста

Синтаксический анализ

1. Разметка частей речи
2. Морфологическая сегментация
3. Выделение границ предложений
4. Разрешение кореференций

Семантический анализ

1. Предсказание следующего слова по контексту
2. Анализ тональности текста
3. Ответы на вопросы (когда?, где?)
4. Выделение фактов
5. Поиск синонимов

Порождение текста

1. Автореферирование
2. Машинный перевод
3. Диалоговые модели

Классические модели обработки текста

Мешок слов

“Сантехник чинит кран”, “Сантехник из водоканала”, “Кран течёт”, “Электрик Анатолий”, “Анатолий чинит лампу”.

сантехник	чинить	кран	водоканал	течёт	электрик	лампа	Анатолий
1	1	1					
1			1				
		1		1			
					1		1
	1					1	1

TF-IDF

$$\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k}\ ,\ \mathrm{idf}(t,D) = \log \frac{|D|}{|\{\, d_i \in D \mid t \in d_i \,\}|}$$

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

https://ru.wikipedia.org/wiki/TF-IDF

Учёт порядка слов

“Мама мыла раму с мылом”

- n-граммы слов (“мама мыла”, “мыла раму”)
- k-skip-n-граммы (“мама раму”, “мыла мылом”)
- Буквенные n-граммы (“мыл”, “рам”, “мыл”)

Проблемы классической модели

- Вектора признаков очень разряженные
- Много 0 в признаках
- Количество признаков = размер словаря ~ миллионы
- В основе - гипотеза независимости слов

Итог: векторы слов не отражают семантику слов

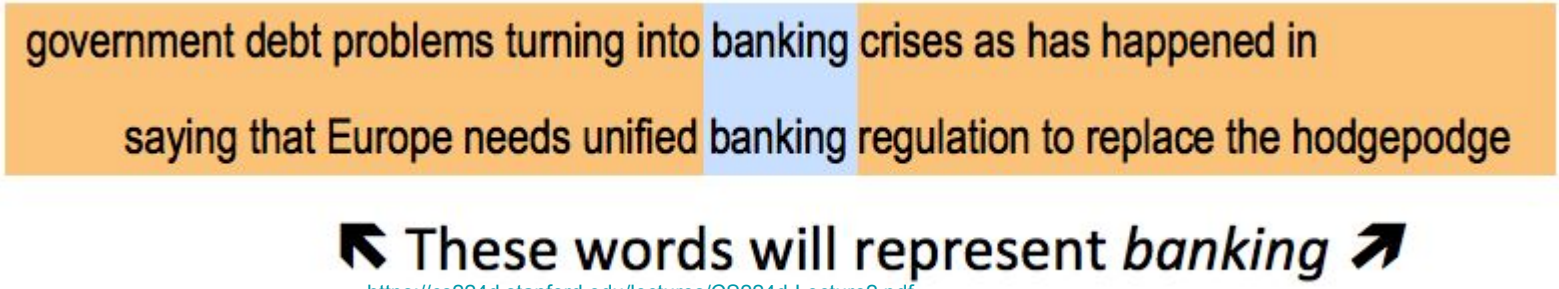
Векторные представления

Word Embedding

Дистрибутивная гипотеза

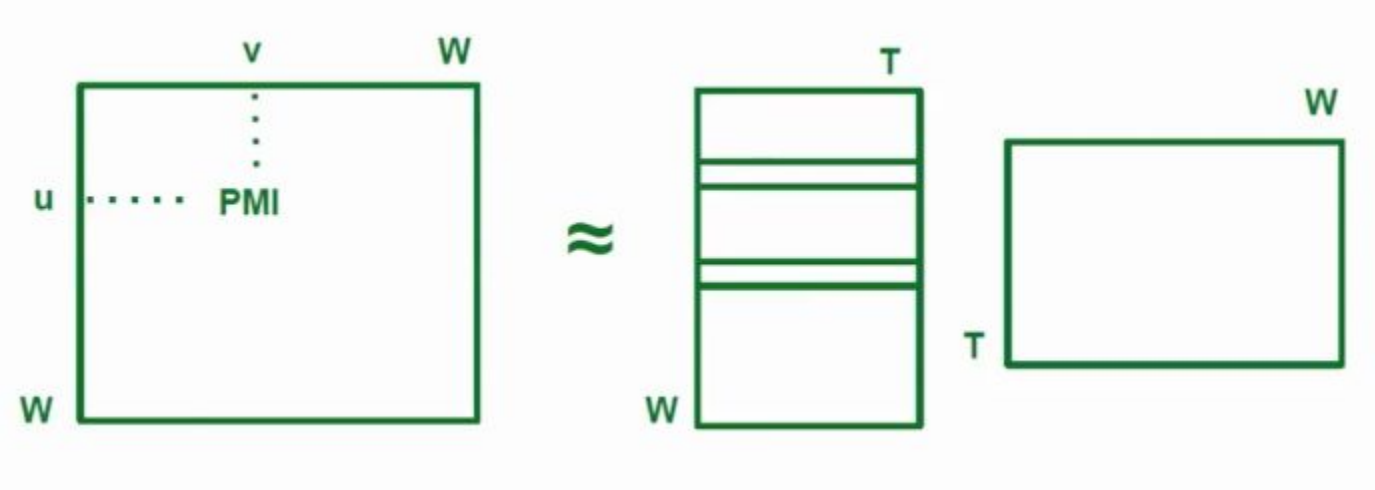
“You shall know a word by the company it keeps” – Firth, 1957.

Лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения. (1960-ые)

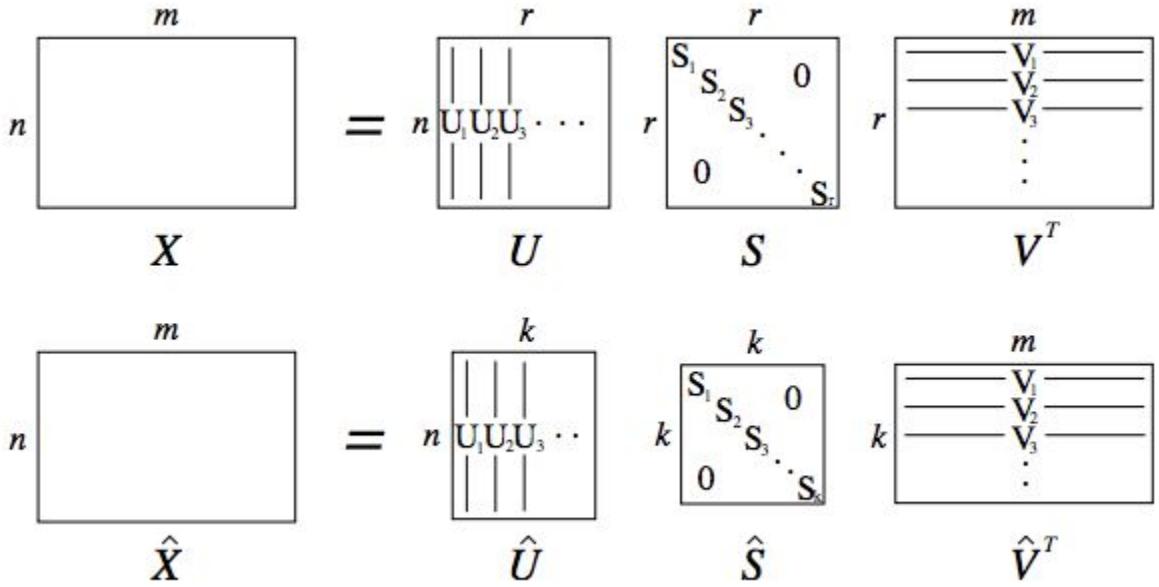


https://cs224d.stanford.edu/lectures/CS224d-Lecture2.pdf

Латентный семантический анализ LSA



Сингулярное разложение матрицы



https://cs224d.stanford.edu/lectures/CS224d-Lecture2.pdf

Распределенное представление

One-Hot encoding

Сантехник = [1,0,0,0,0……,1…,0]

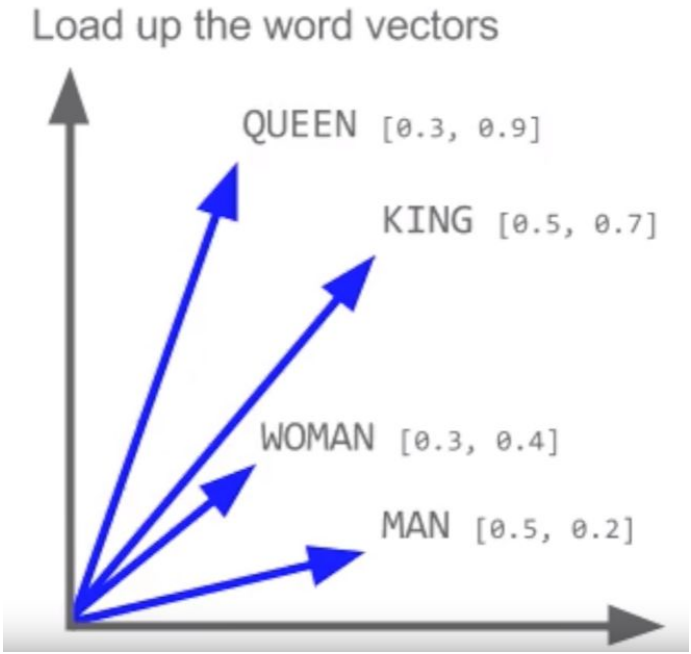
Электрик = [0,0,1,0,0,….,1,..,1]

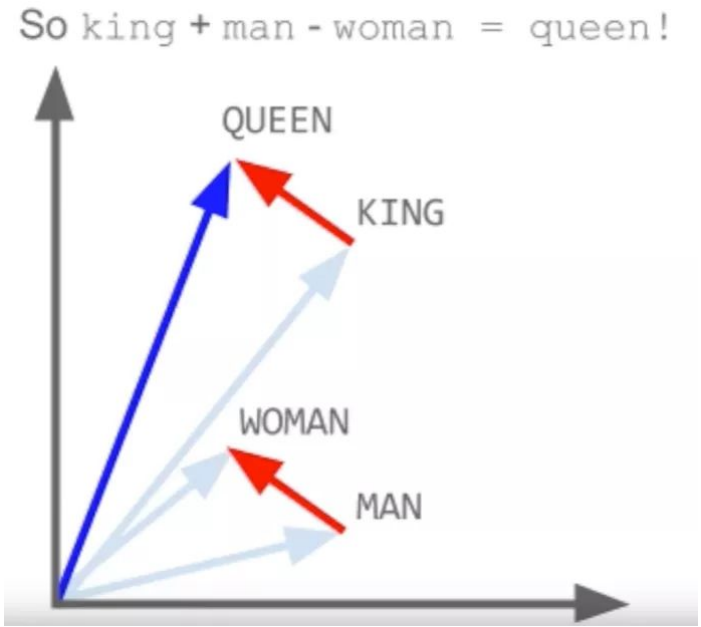
Word Embedding

Кран = [0.22, 0.42, 0.1, 0.5]

Труба = [0.23, 0.4, 0.15, 0.7]

word2vec


Load up the word vectors
QUEEN [0.3, 0.9]
KING [0.5, 0.7]
WOMAN [0.3, 0.4]
MAN [0.5, 0.2]


So king + man - woman = queen!
QUEEN
KING
WOMAN
MAN

word2vec алгоритм

- Каждому слову v сопоставляем вектор **w** размера d
- Считаем вероятности p(v|c1...cn), где ci - контекст
- p(v|c1..cn) = f(**w**,**o**), **o** - параметры функции
- Как найти **o** и **w** - нейросеть его знает...

Word2vec

https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

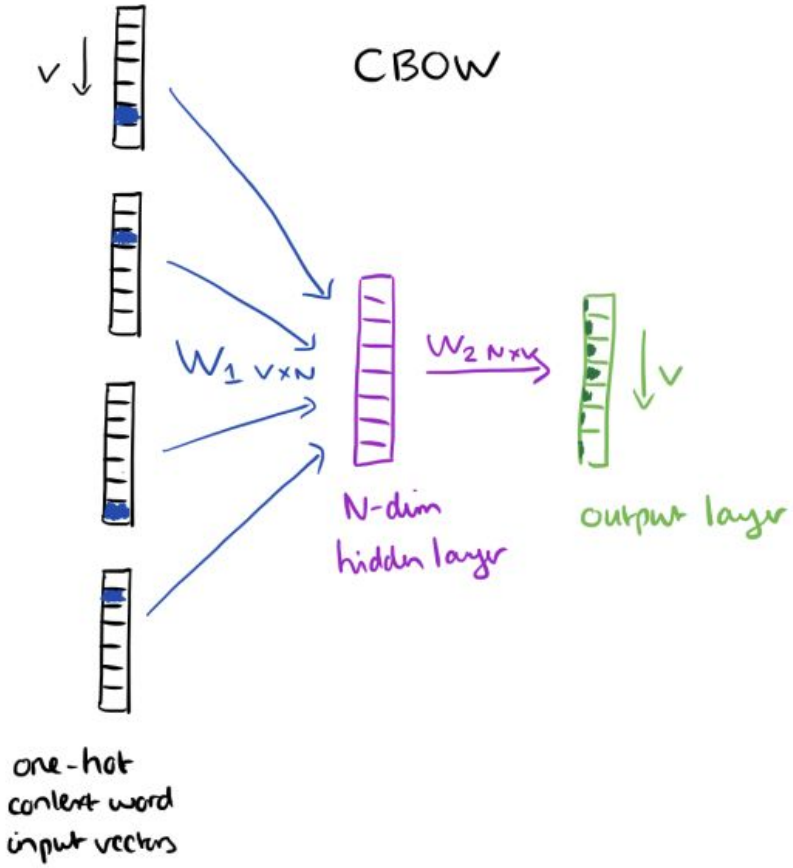
word2vec: CBOW

непрерывный мешок слов

Вход: контекст

Выход: целевое слово

Скрытый слой: векторное представление слов



https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

Модификации

- word2vec
  - CBOW
  - skip-gram
  - GLOVE - https://nlp.stanford.edu/projects/glove/
- doc2vec
- ...

Литература

1. Глубокое обучение. С. Николенко, А. Кадурин, Е. Архангельская.
2. Векторные представления слов в документах: https://www.youtube.com/watch?v=hiDBnEyoZS4
3. Мастер класс по Gensim: https://www.youtube.com/watch?v=oBb9aFmp0Hs
4. Лекции 4 семестра специализации ML&DS: https://www.coursera.org/specializations/machine-learning-data-analysis
5. https://cs224d.stanford.edu/lectures/CS224d-Lecture2.pdf
6. https://github.com/RaRe-Technologies/movie-plots-by-genre