

LLM-инженер

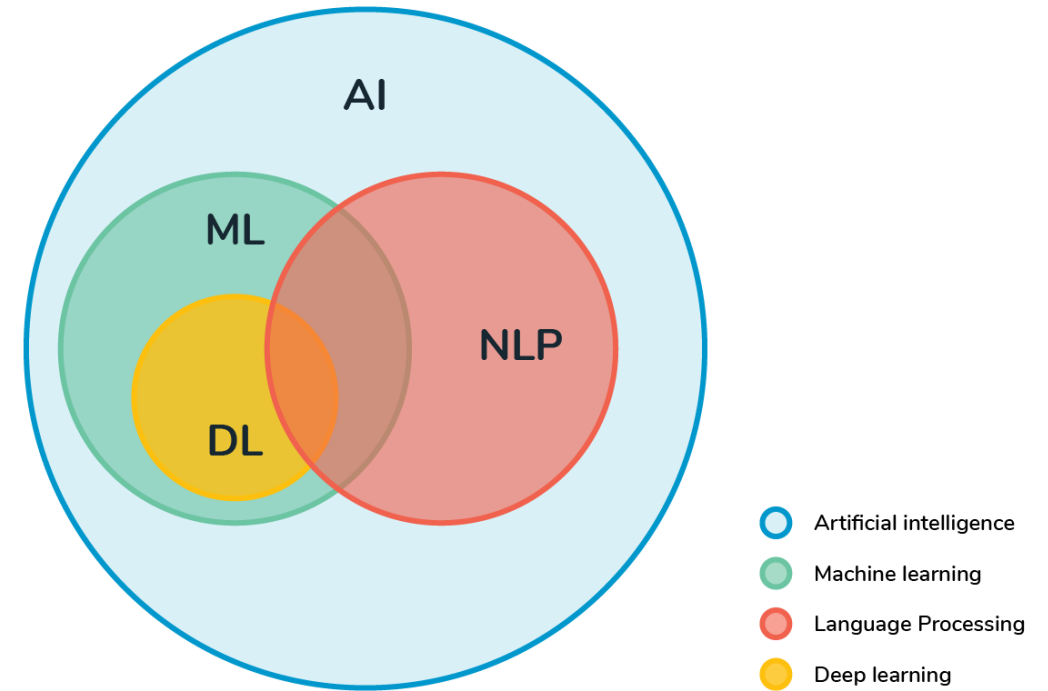
online course ✨

модуль #1

лекция: NLP - от базы до DL

NLP - Обработка естественного языка

NLP – это технология машинного обучения, которая дает компьютерам возможность интерпретировать, манипулировать и понимать человеческий язык.





Почему решать задачи NLP сложно?

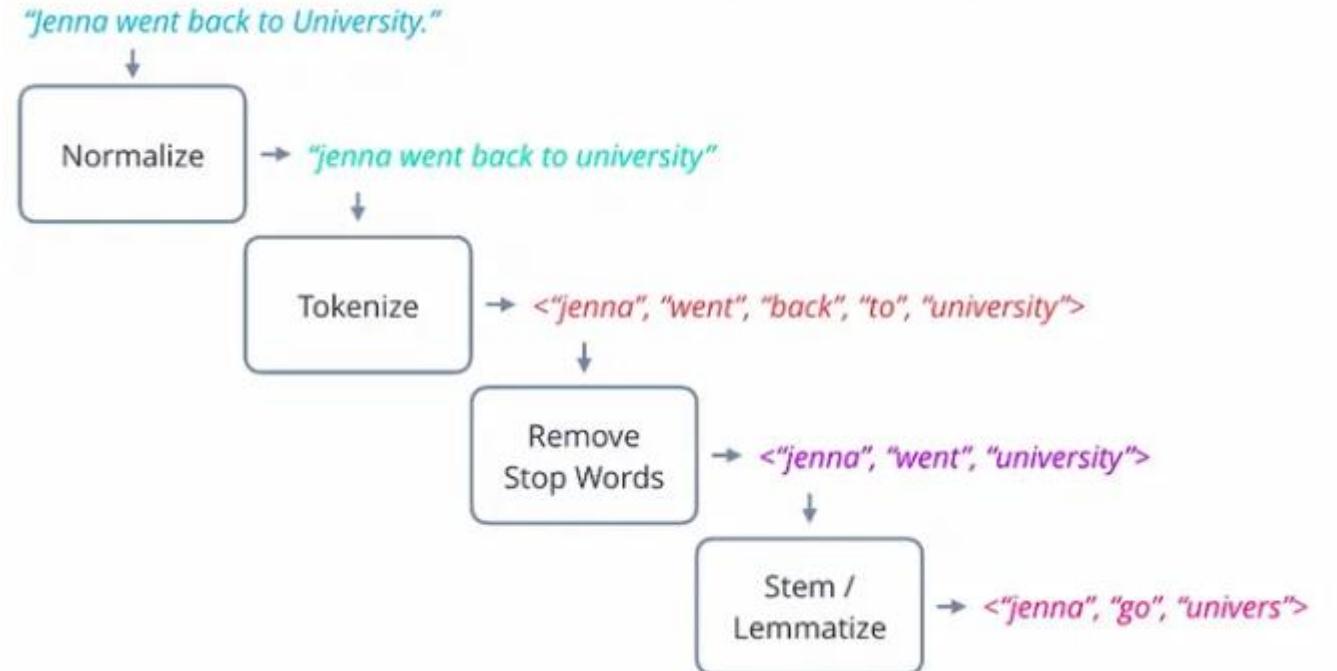
- **Полисемия** (многозначность): остановка (процесс или здание), стол (организация или объект), дятел (птица или человек).
- **Омонимия**: ключ, лук, замок, печь.
- **Местоименная анафора**: пусть нам дан текст «Дворник два часа мел снег, он был недоволен». Местоимение «он» может относиться как к дворнику, так и к снегу. По контексту мы легко понимаем, что он – это дворник, а не снег. Но добиться, чтобы компьютер это тоже легко понимал, непросто.

Уровни абстракции текстовых данных

- Буквы
- Слова
- n-граммы
- Предложения
- Документы

Предобработка текста

- Сегментация предложений
- Нормализация
- Токенизация
- Морфологический анализ
 - Стемминг
 - Лемматизация



Сегментация предложений

Sentence Segmentation

Hello world. This blog post is about sentence segmentation. It is not always easy to determine the end of a sentence. One difficulty of segmentation is periods that do not mark the end of a sentence. An ex. is abbreviations.



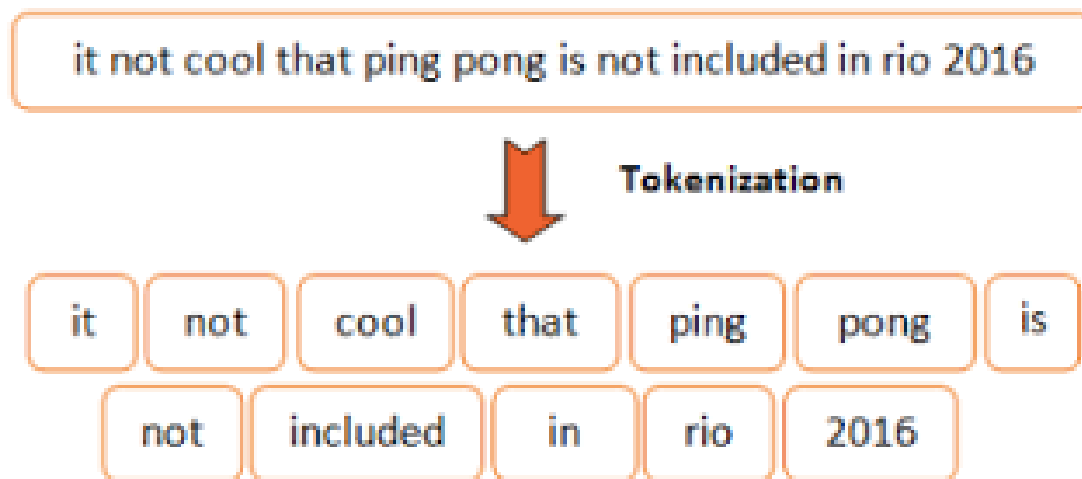
- Hello world.
- This blog post is about sentence segmentation.
- It is not always easy to determine the end of a sentence.
- One difficulty of segmentation is periods that do not mark the end of a sentence.
- An ex. is abbreviations.

Нормализация

- Обработка чисел
- Очистка от html-разметки
- Обработка символов на иностранном языке
- Удаление ненужной пунктуации
- Обработка аббревиатур и сокращений
- Проверка правописания

Токенизация

Токенизация — процесс разделения текста на составляющие (их называют «токенами»).



Удаление стоп-слов

В языке есть слова, которые предсказуемо удаляются из текста без потери смысла. Если их удалить, то, скорее всего, текст станет яснее, информативнее и проще для чтения. Это — **стоп-слова**.

Стоп-слова

Междометия + вводное

Ах, значит, в общем, как

Определения и наречия

Усилители: абсолютный, безусловный, весьма,

Обобщающие и неопределённые: в целом, всякий, общий, около, разнообразный

И все остальные заодно: активный, актуальный, взыскательный, длительный, знаковый, инновационный

Местоимения и очевидные сущности

Мы, вы, посетитель, документ, сайт, страница, меню, ссылка, информация, здесь, тут

Штампы

а также, в лучших традициях, как говорится, шаг за шагом

Паразиты времени

В настоящий момент, в наши дни, сейчас, нынче, на сегодняшний день

Отглагольное

осуществлять деятельность, реализовывать план, производить работы

Модальность

может, должен, нужно

Неопределённое

какой-то, что-то, как-то, зачем-то, сколько-то, несколько, неизвестно как, примерно, приблизительно

Лишние «бы»

я бы хотел, мне бы не хотелось

Стемминг

Стемминг — это процесс нахождения основы слова для заданного исходного слова.

- сокращает размерность
- повышает полноту поиска
- снижает точность (одинаковые основы для разных слов)

chang^{ing}
chang^{ed}
chang^e *stemming* → chang
chang
chang

stud^ying
stud^{ies}
stud^y *stemming* → studi
studi
studi

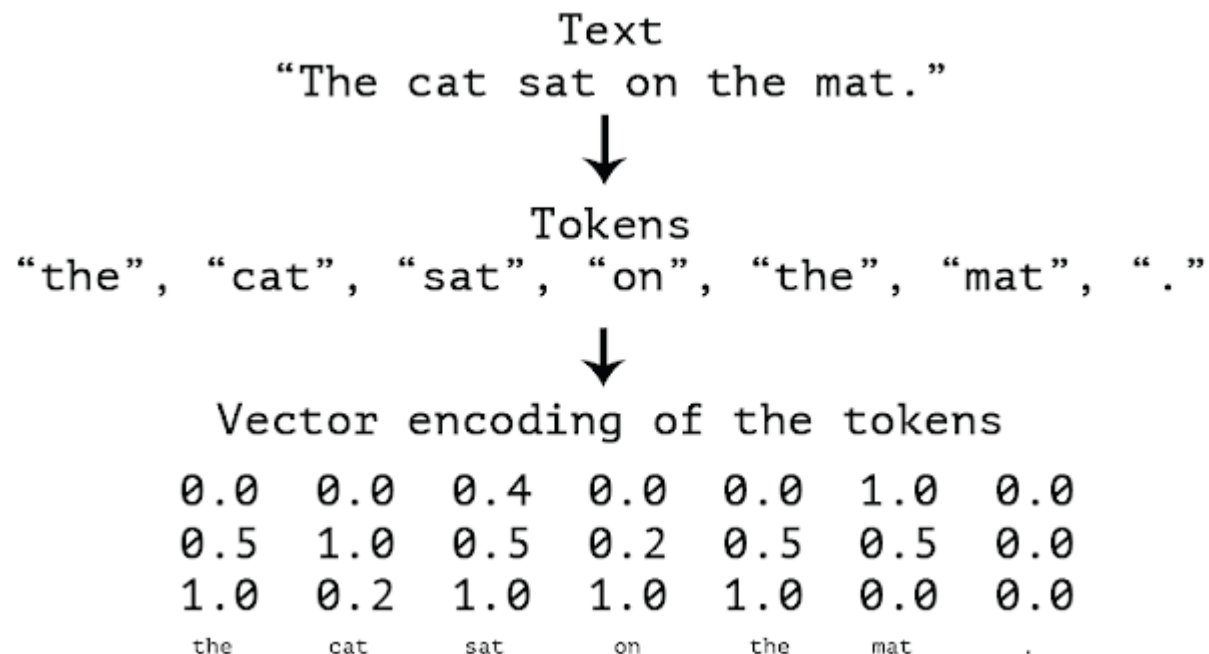
Лемматизация

Лемматизация — процесс приведения словоформы к лемме — её нормальной (словарной) форме

Original	Stemming	Lemmatization
New	New	New
York	York	York
is	is	be
the	the	the
most	most	most
densely	dens	densely
populated	popul	populated
city	citi	city
in	in	in
the	the	the
United	Unite	United
States	State	States

Векторизация текстов как часть предобработки

- Bag of words / n-grams
- TF-IDF
- Word Embeddings



Bag of words (мешок слов)

good movie		good	movie	not	a	did	like
not a good movie	→	1	1	0	0	0	0
did not like		1	1	1	1	0	0
		0	0	1	0	1	1

Bag of words (мешок слов)

- Признаковое пространство имеет размерность, равную мощности словаря коллекции текстов.
- Количество строк определяется количеством документов.
- Минусы:
 - Словарь часто огромный, и на деле получаются разреженные векторы большой размерности, что неудобно на практике.
 - Не учитывается семантическая близость слов, все векторы одинаково далеки друг от друга в признаковом пространстве.
 - Не учитывается порядок слов.

Статистическая мера

- Самые популярные слова будут встречаться в большинстве документов. В результате такие слова усложняют подбор текстов, представленных с помощью модели мешка слов. Кроме того, самые популярные слова часто являются функциональными словами без смыслового значения. Они не несут в себе смысл текста.
- Мы можем применить статистическую меру TF-IDF (частота слова - обратная частота документа), чтобы уменьшить вес слов, которые часто используются в тексте и не несут в себе смысловой нагрузки.

TF-IDF

- TF-IDF — статистический показатель, применяемый для оценки важности слова в контексте категории, документа или коллекции документов.
- Как правило, TF-IDF определяется для каждого слова. Чем выше значение данного показателя, тем значимее слово в контексте категории, документа, коллекции.
- Мера TF-IDF часто используется для представления документов коллекции в виде числовых векторов, отражающих важность использования каждого слова из некоторого набора слов.

TF-IDF

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

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

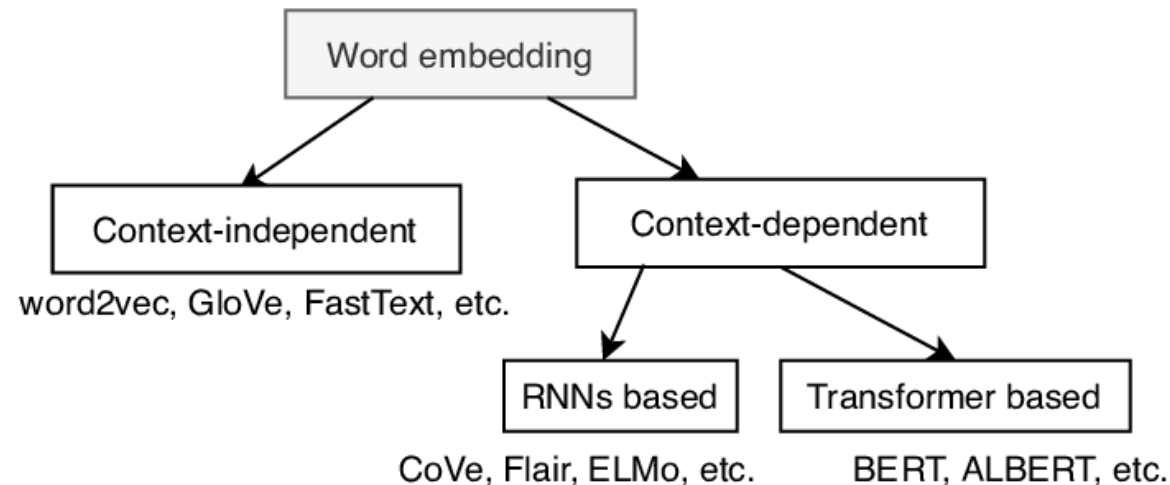
- TF (частота слов) характеризует отношение числа вхождений конкретного слова к общему набору слов в документе. Чем выше TF, тем весомее конкретное слово в рамках документа.
- IDF (обратная частота документа) характеризует инверсию частотности, с которой конкретное слово используется в тексте. С помощью этой метрики можно снизить важность слов — например, союзов или предлогов.

TF-IDF

- Мера TF-IDF часто используется для представления документов коллекции в виде числовых векторов, отражающих важность использования каждого слова из некоторого набора слов.
- Быстро вычисляется. Для формирования оценки достаточно просканировать все документы в пределах одной коллекции.
- Минусы:
 - Оценка является статической. Может измениться только при изменении одного из документов коллекции.
 - Частота встречаемости слова далеко не самый надёжный показатель релевантности, особенно для русского языка. Можно составить документ, в котором релевантное слово не будет повторяться (с использованием синонимов), или же, наоборот, текст будет перегружен омонимами нерелевантного слова.

Word embeddings

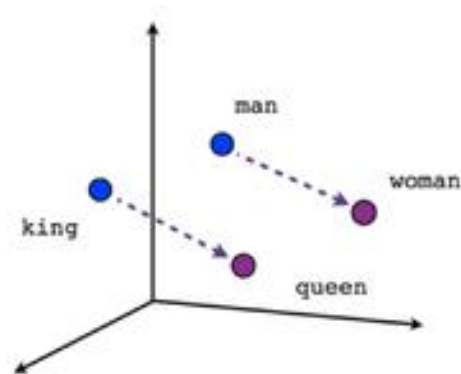
Word embeddings - общее название для различных подходов к моделированию языка и усвоению представлений в обработке естественного языка, направленных на сопоставление словам (и, возможно, фразам) из некоторого словаря векторов из R^n для n , значительно меньшего количества слов в словаре.



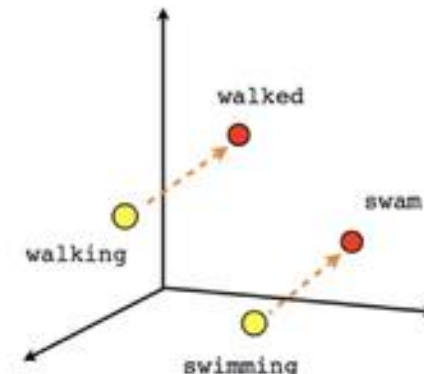
Word2Vec

Word2vec — способ построения сжатого пространства векторов слов, использующий нейронные сети.

Принимает на вход большой текстовый корпус и сопоставляет каждому слову вектор. Сначала он создает словарь, а затем вычисляет векторное представление слов. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а следовательно, имеющие схожий смысл), в векторном представлении имеют высокое косинусное сходство



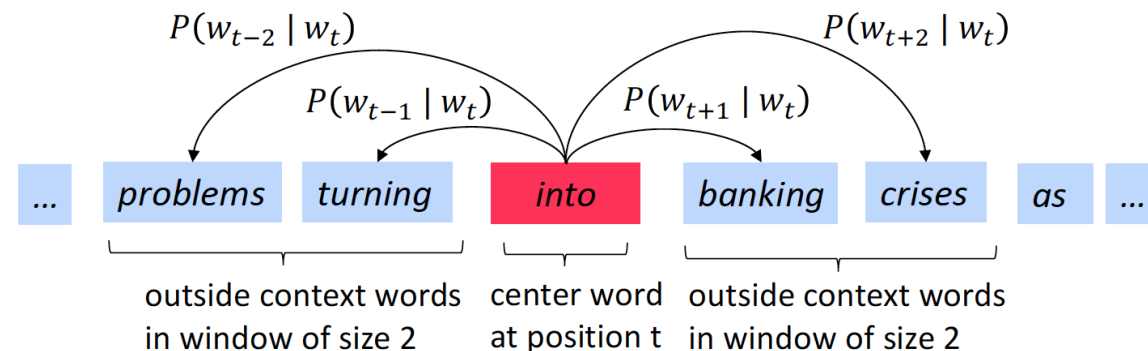
Male-Female



Verb tense

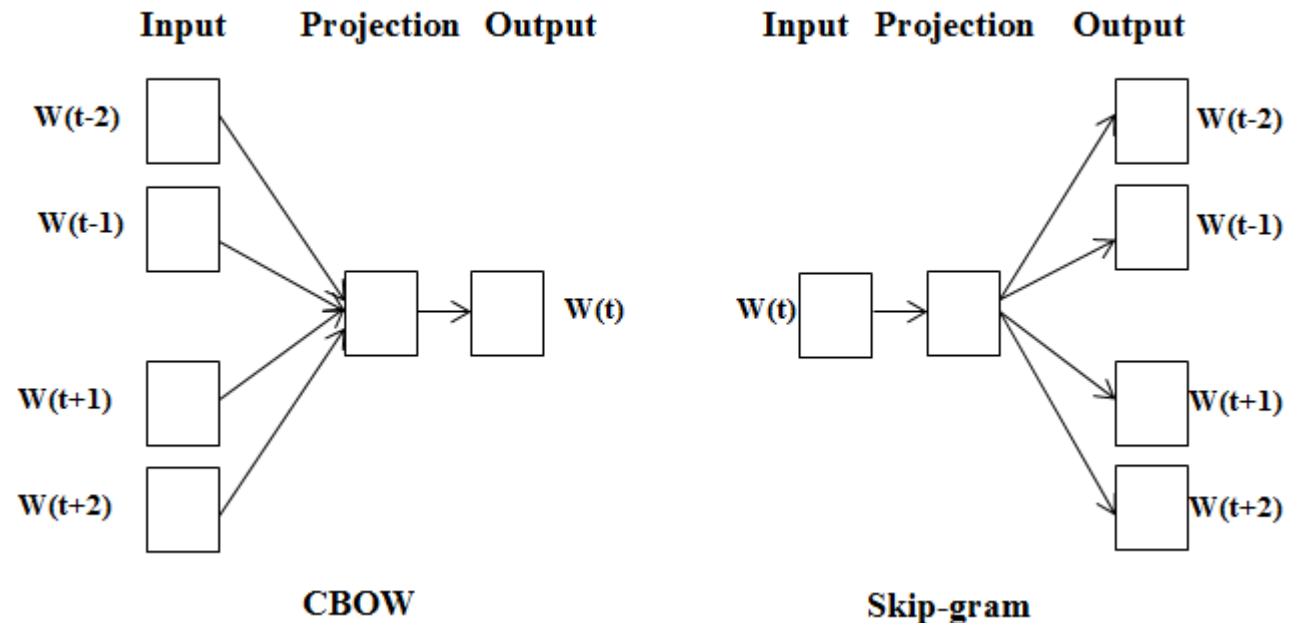
Word2Vec: идея

- Берем большой набор текстов. Каждое слово представляется вектором.
- Проходим по текстам скользящим «окном», перемещаясь на одно слово за раз
- Таким образом, на каждом шаге есть центральное слово и контекстные
- Вычисляем вероятность центрального слова при условии контекстных
- Корректируем вектора для максимизации вероятности



Word2Vec: Skip-Gram & CBOW

Основные архитектуры Word2Vec — Continuous Bag of Words (**CBOW**) и **Skip-gram**. Принцип работы CBOW — предсказывание слова при данном контексте, а Skip-Gram наоборот — предсказывается контекст при данном слове.



Word2Vec: обучение

- Чтобы обучить модель, мы подбираем параметры θ , минимизирующие функцию потерь. Фактически параметры – вектора v_w и u_w
- Для оптимизации используется градиентный спуск
- За раз обновляется одна пара из центрального и контекстного слова

$$\text{Loss} = J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{-m \leq j \leq m, j \neq 0} \log P(w_{t+j} | w_t, \theta) = \frac{1}{T} \sum_{t=1}^T \sum_{-m \leq j \leq m, j \neq 0} J_{t,j}(\theta).$$

GloVe

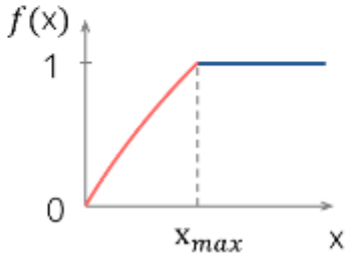
- Модель **GloVe** (Global Vectors) представляет собой комбинацию count-based методов и методов прогнозирования (например, Word2Vec).

context vector word vector bias terms (also learned)

$$J(\theta) = \sum_{w,c \in V} \underbrace{f(N(w, c))}_{\text{weighting function}} \cdot (u_c^T v_w + b_c + \bar{b}_w - \log N(w, c))^2$$

Weighting function to:

- penalize rare events
- not to over-weight frequent events

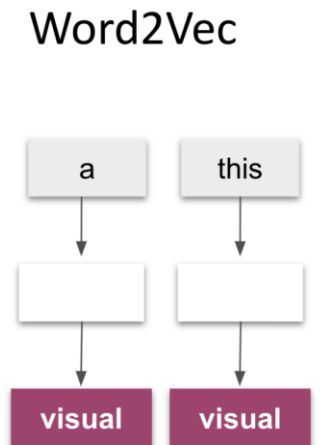
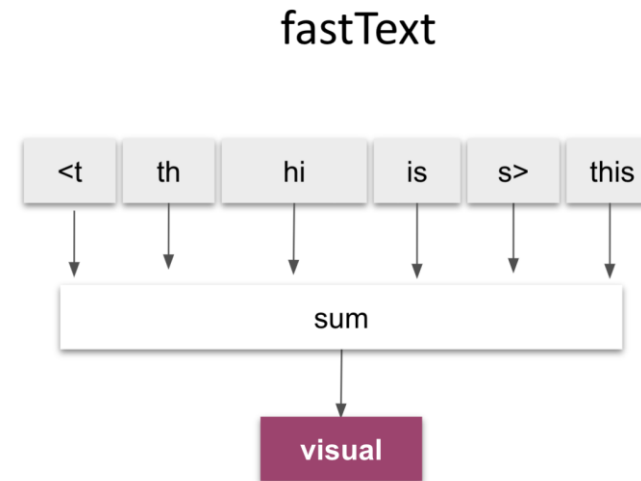

$$f(x) = \begin{cases} (x/x_{max})^\alpha & \text{if } x < x_{max}, \\ 1 & \text{otherwise.} \end{cases}$$

$\alpha = 0.75, x_{max} = 100$

- Как и в Word2Vec, здесь разные векторы для центральных и контекстных слов — параметры. Кроме того, метод имеет скалярный член смещения для каждого вектора слов.

FastText

- Для модели векторных представлений слов используется skip-gram с негативным сэмплированием.
- Skip-gram игнорирует структуру слова, но в некоторых языках есть составные слова, как, например, в немецком. Поэтому к основной модели была добавлена subword-модель. Subword-модель — это представление слова через n-граммы с n от 3 до 6 символов от начала до конца слова плюс само слово целиком.



Меры близости

- Меры близости двух векторов u, v в многомерном пространстве можно выбрать различным образом:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

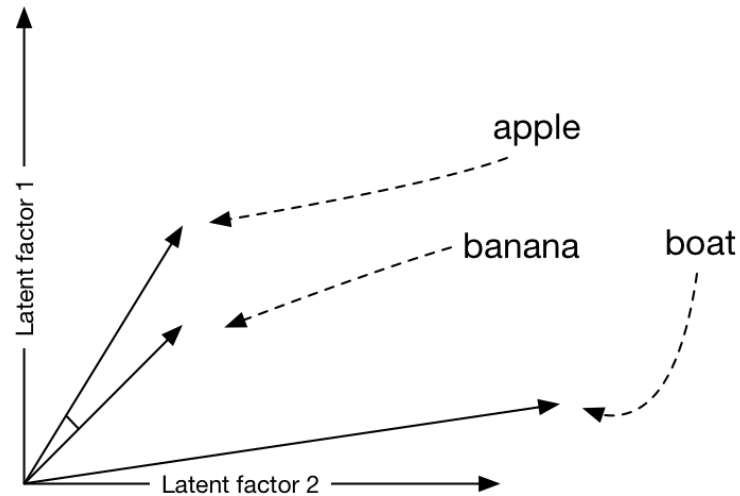
$$\text{soft_cosine}_1(a, b) = \frac{\sum_{i,j}^N s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^N s_{ij} a_i a_j} \sqrt{\sum_{i,j}^N s_{ij} b_i b_j}},$$

Оценка качества embedding'ов

- Существует два типа оценки: внутренняя (intrinsic) и внешняя (extrinsic).
- Внутренняя основана на внутренних свойствах, то есть насколько хорошо они в целом отражают смысл. **Быстро, но оторвано от практических задач.**
- При использовании внешней предполагается оценка на реальной задаче. Необходимо обучать модель несколько раз: по одной модели каждого вида. Затем можно сравнить на качество этих моделей. **Сильно дольше.**

Внутренняя оценка качества (intrinsic)

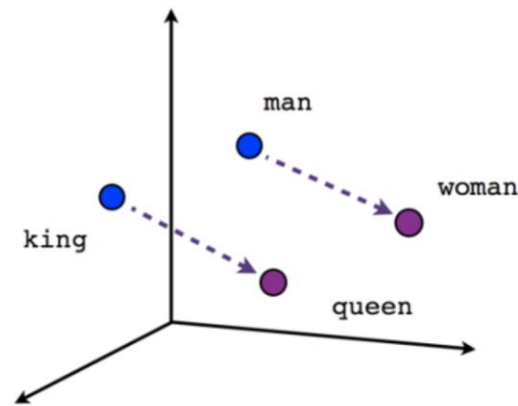
Хорошие эмбединги образуют семантическое пространство, в котором близкие точки обычно имеют схожее значение



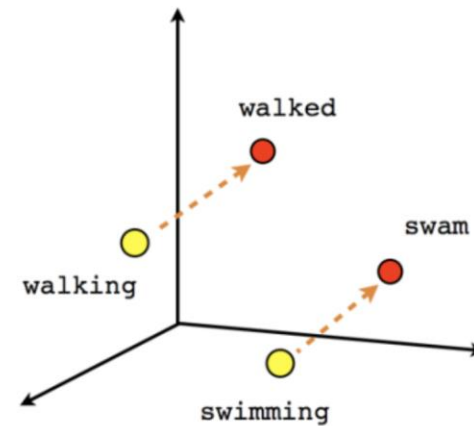
Выборка ближайших соседей (по косинусному или евклидову расстоянию) — один из методов оценки качества эмбедингов.

Внутренняя оценка качества (intrinsic)

Еще одно интересное свойство - многие семантические и синтаксические отношения между словами линейны в векторном пространстве слов.



Male-Female



Verb tense

Использование векторов

- Поскольку каждому тексту поставлен соответствие вектор в семантическом пространстве, мы можем вычислить расстояние между любыми двумя текстами. Имея расстояние между текстами, можно использовать алгоритм kMeans для проведения кластеризации или классификации.
- Также семантические вектора фактически являются векторами признаков, которые можно использовать в любой модели классификации

