ТЕХНОСФЕРА

Занятие № 7

Работа с текстовыми данными

Евгений Некрасов

План занятия

1. Особенности текстовых данных
2. Подходы к моделированию текстов
3. Особенности матричного представления текстов
4. Тематическое моделирование
5. Векторное представление слов
6. Об итоговых проектах

Извлечение признаков

Сампл (**пример**) - это вектор чисел.

Извлечение признаков - представление реального или цифрового объекта в виде вектора чисел.

Объект - **текстовый документ**

Как построить векторное представление?

Особенности текстовых данных

Что усложняет работу с текстами:

- Омонимы (одинаковое написание разное значение)
- Синонимы (разное написание, похожее значение)
- Опечатки и ошибки
- Неологизмы
- Пунктуация
- Порядок слов
- Контекст

Омонимы

Пример:

- "эфир" в смысле органического вещества
- "эфир" в смысле "радиовещание и телевидение"
- "эфир" в смысле криптовалюта

Синонимы

Пример:

- учиться, выучивать, изучать, научаться, разучивать, твердить, зубрить, штудировать, усваивать, грызть гранит науки

Опечатки, неологизмы

Примеры:

- "президент" → "перзидент"
- "искусство" → "исскуство"
- "гуглить"
- "молоткастосерпастый"

Пунктуация

Пример:

- "Казнить, нельзя помиловать!"
- "Казнить нельзя, помиловать!"

Порядок слов

Пример:

- "Бытие определяет сознание."
- "Сознание определяет бытие."

Локальный контекст

Анафоры - зависимость интерпретации некоторого выражения от другого выражения.

Пример:

- "Мы отдали бананы обезьянам, потому что они были голодные."
- "Мы отдали бананы обезьянам, потому что они были перезрелые."

Глобальный контекст

Прием, белка-6. Седой покинул орлиное гнездо.

При в определенных ситуациях слова полностью меняют смысл.

Модель bag of words (мешок слов)

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Предобработка текста

- Приведение текста к одному регистру
- Удаление знаков пунктуации
- Удаление стоп-слов
- Использование лингвистических онтологий
- Стемминг

Лингвистические онтологии, проект WordNet

WordNet - семантическая сеть для английского языка.

Базовой словарной единицей в **WordNet** является не отдельное слово, а так называемый синонимический ряд "**синсет**"

WordNet, структура

Синсеты в **WordNet** связаны между собой различными семантическими отношениями:

гипероним (завтрак → прием пищи)

гипоним (прием пищи → обед)

has-member (факультет → профессор)

member-of (пилот → экипаж)

мероним (стол → ножка)

антоним (лидер → последователь)

Стемминг

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

До:

"я помню чудное мгновенье передо мной явилась ты как мимолетное виденье как гений чистой красоты"

После:

"я помн чудн мгновен перед мно яв ты как мимолетн виден как ген чист красот"

Стемминг

До:

"лошадь лошади лошадей лошадью лошадки кони конный конная конюшня конюх"

После:

"лошад лошад лошад лошад лошадк кон кон кон конюшн конюх"

Snowball Stemmer

TF-IDF, не все слова равнозначны

TF-IDF - статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. (TF-IDF = TF*IDF)

TF (term frequency) - отношение числа вхождений некоторого слова к общему числу слов документа.

IDF (inverse document frequency) - логарифм обратной частоты, с которой некоторое слово встречается в документах коллекции.

N-граммы слов

Текст: "я помню чудное мгновенье"

тип	N-граммы
1-граммы	я помню чудное мгновенье
2-граммы	я помню помню чудное чудное мгновенье
3-граммы	я помню чудное помню чудное мгновенье

N-граммы символов

Текст: "япомнючудноемгновенье"

тип	N-граммы
3-граммы	япо, пом, омн, мню, нюч, ючу, чуд, удн, дно, ное, оем, емг, мгн, гно, нов, ове, вен, ень, нье

Такой подход:

- Обладает некоторой устойчивостью к опечаткам
- Понимает родственные слова

Задача

Дано:

- 1M документов
- 1M возможных слов

Размер матрицы признаков:

4 *10^6 * 10^6 = 4 Тб

Как с этим работать?

Sparse матрицы

- Когда только очень небольшая часть значений в матрице отлична от 0, разумно хранить информацию только об этих значениях.

- Модуль scipy.sparse реализует работу с разряженными матрицами аналогично numpy

- Часть моделей из sklearn поддерживает работу с матрицами scipy.sparse

Другие подходы

- Текст как последовательность символов или слов
- Представление текста в виде тензора
- Искусственные нейронные сети как модель

Тематическое моделирование

Тематическая модель (topic model) - модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции.

Вход:

Коллекция текстовых документов.

Выход:

Вектор, составленный из оценок степени принадлежности документа каждой из тем.

Зачем нужно тематическое моделирование

- ранжировать документы по степени релевантности заданной теме (тематический поиск)
- построить тематический каталог коллекции документов
- определить, как темы изменялись со временем
- определить тематику различных сущностей, связанных с документами (журналов, конференций, организаций, авторов)
- разбить документ на тематически однородные фрагменты

Латентное размещение Дирихле

Латентное размещение Дирихле (LDA, Latent Dirichlet allocation) - наиболее популярная тематическая модель.

Вход: коллекция документов, количество тем

Выход: описание тем через слова коллекции документов, представление документа в виде смеси тем

LDA модель

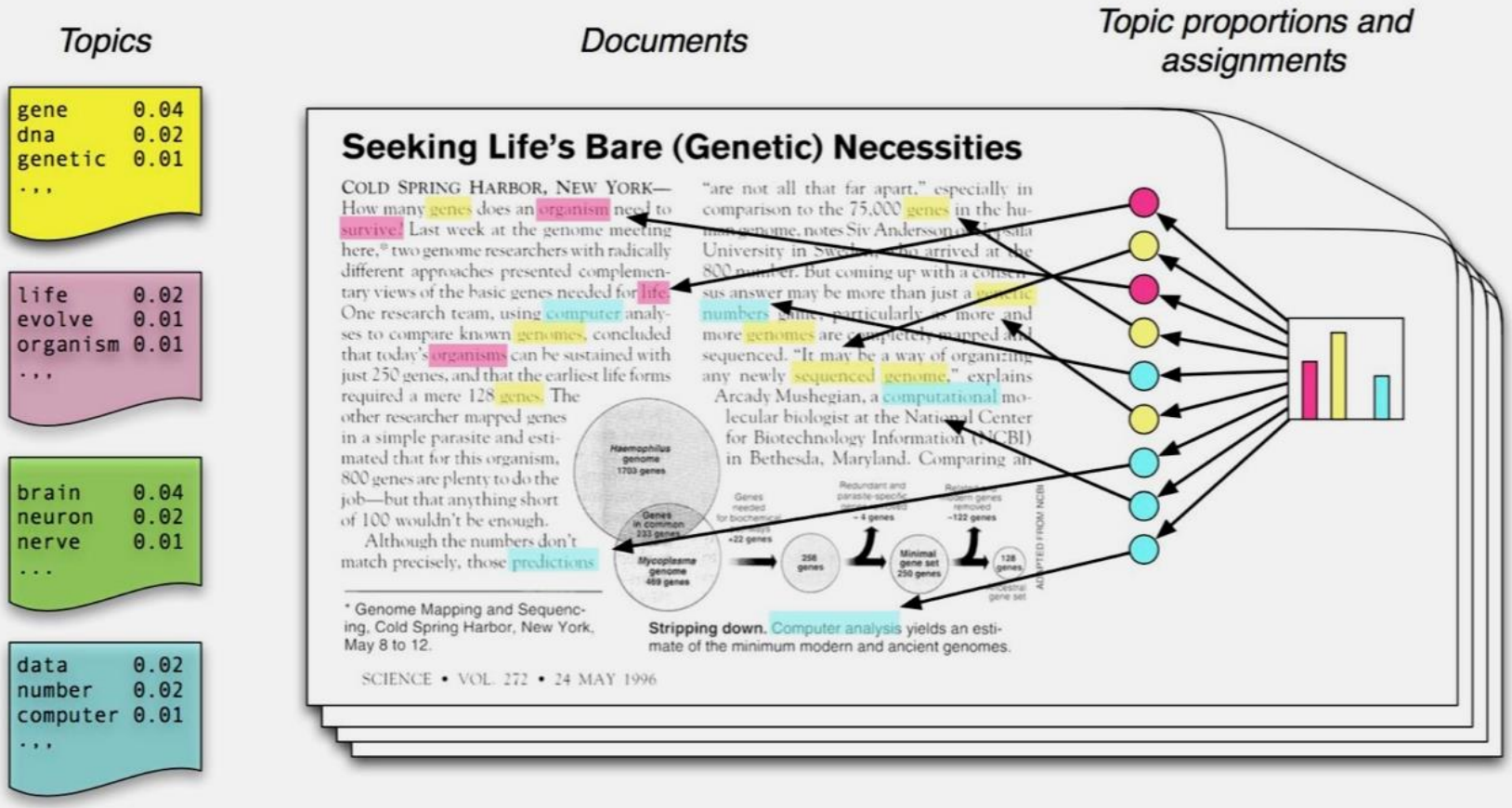
- порядок документов в коллекции не важен
- документ - это мешок слов
- коллекцию документов можно рассматривать как простую выборку пар «документ-слово» (d,w)
- каждая тема t из T описывается неизвестным распределением p(w|t) на множестве слов w из W
- каждый документ d из D описывается неизвестным распределением p(t|d) на множестве тем t из T

LDA модель

- векторы документов (p(t1|d), p(t2|d) ...) порождаются одним и тем же вероятностным распределением Дирихле на нормированных T - мерных векторах
- векторы тем (p(w1|t), p(w2|t) ...) порождаются одним и тем же вероятностным распределением Дирихле на нормированных W - мерных векторах

$$p(d,w) = \sum_{t \in T} p(d)p(w|t)p(t|d)$$

LDA модель


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Векторное представление слов

Векторное представление слова (word embedding) — вещественный вектор в пространстве с фиксированной невысокой размерностью (порядка десятков-сотен). Векторные представления для семантически близких слов похожи.

Зачем нужно векторное представление слов

Сжатые векторные представления слов

1. полезны сами по себе, например, для поиска синонимов или опечаток в поисковых запросах
2. используются в качестве признаков для задач машинного обучения

Word2vec

word2vec — группа алгоритмов для получения векторных представлений слов

Вход:

большой текстовый корпус

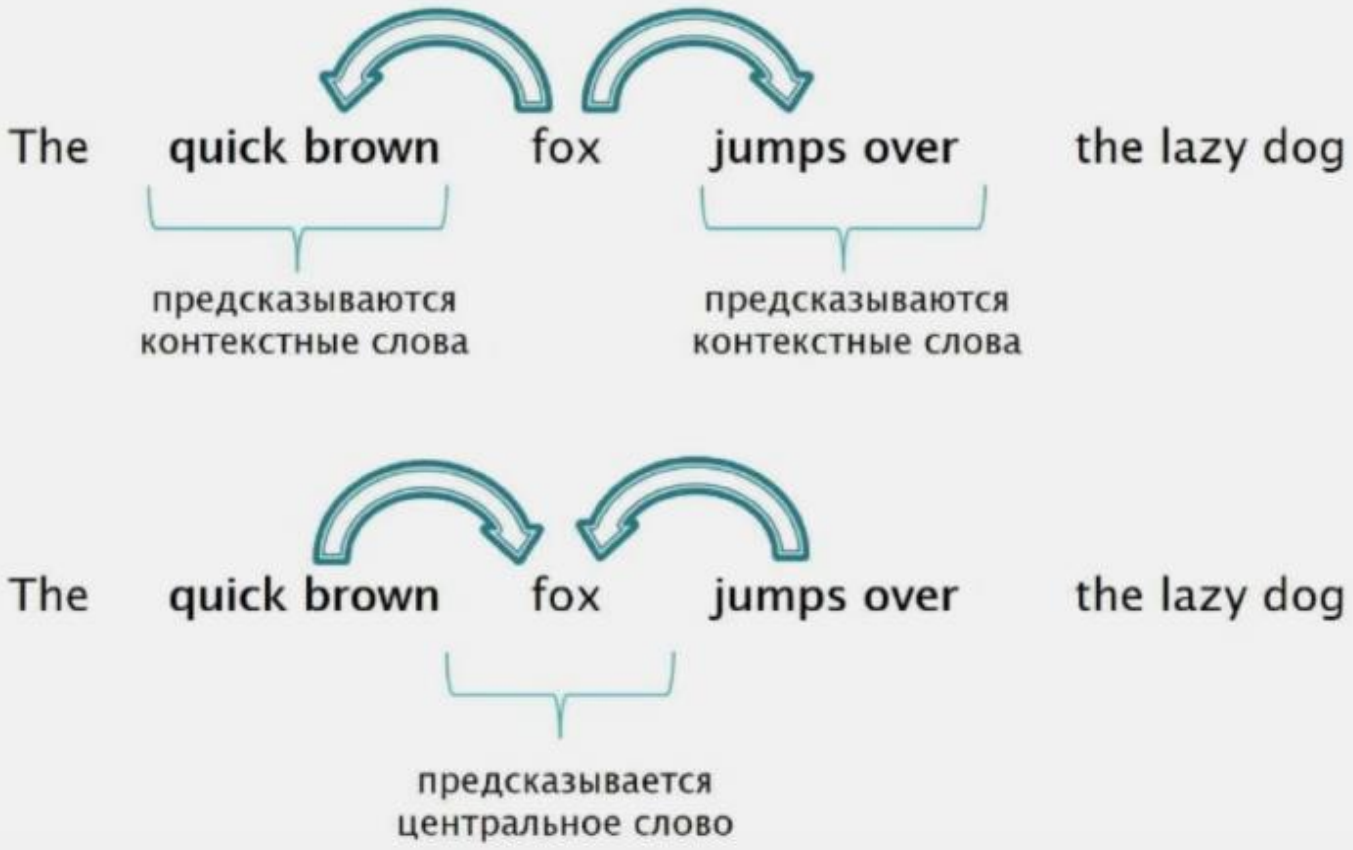
Выход:

векторное представление слов

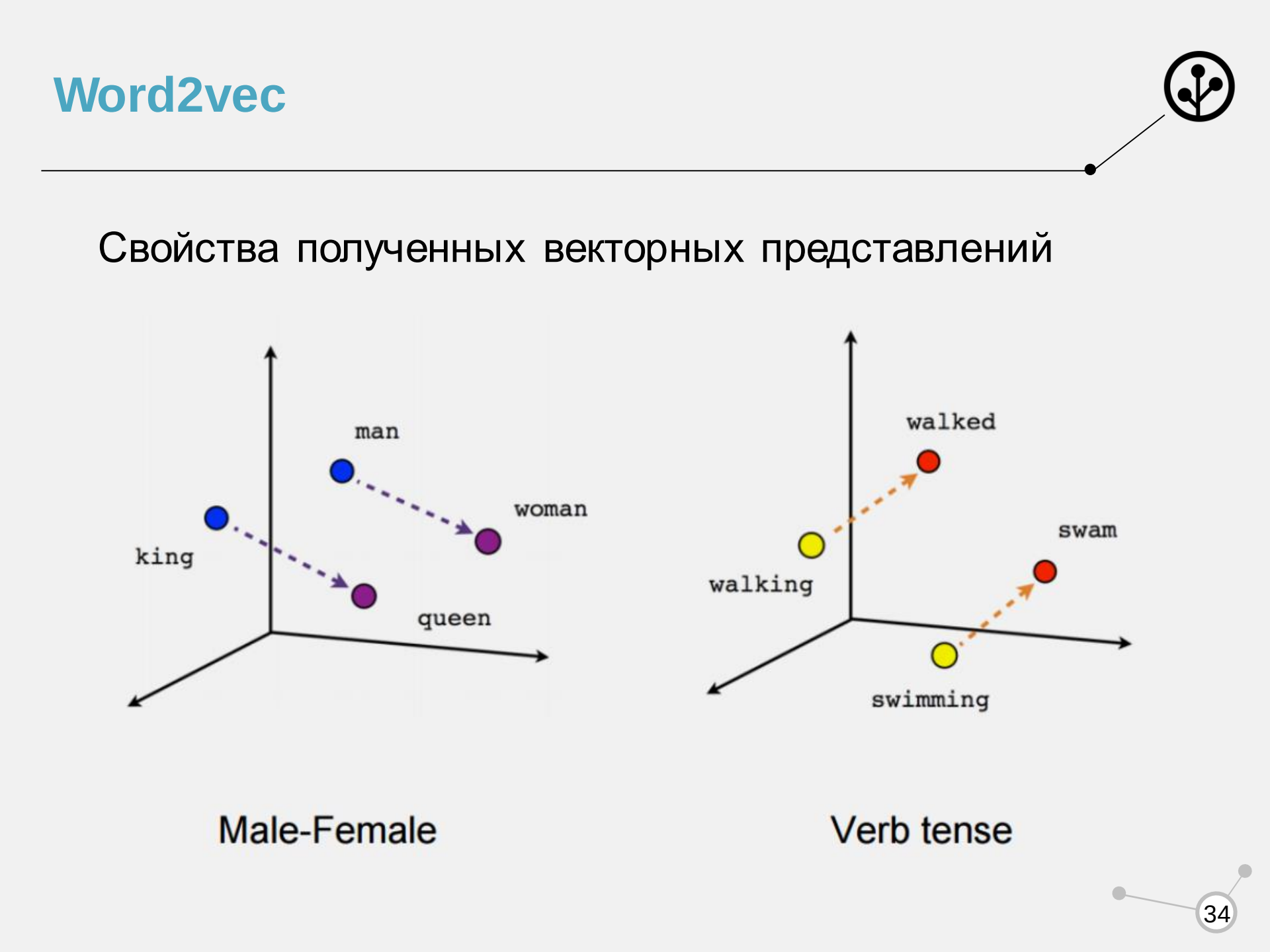
Томаш Миколов, 2013

Word2vec

Идея алгоритма, Skip-gram и CBOW



Word2vec



Свойства полученных векторных представлений

Male-Female

Verb tense

Где взять векторные представления слов?

https://radimrehurek.com/gensim/

https://fasttext.cc/

http://rusvectores.org/

ТЕХНОСФЕРА

Семинар

Итоговый проект, что это

1. Проект делает команда из 1-4 студентов
2. Проект должен быть посвящен анализу данных
3. Проект должен включать машинное обучение
4. Проект должен быть представлен на последнем занятии в виде презентации

Итоговый проект, регламент

- Презентация 5-7 слайдов
- 10 минут на доклад
- 5 минут на вопросы
- Каждый участник команды должен рассказать не менее 1 слайда

Итоговый проект, план презентации

1. Название проекта, состав команды, кто что делал
2. Описание данных и постановка задачи
3. Описание проблем, которые возникли
4. Описание решения
5. Результат

Итоговый проект, как согласовать

1. Выбрать проект из предложенных или придумать
2. Ввести в форму название и описание проекта
3. Получить статус утвержден

Форма:

https://docs.google.com/spreadsheets/d/1mFYA2j1WHuUZpUHKsZRh93jgMQcE-Fj5st6A1JLiVyU/edit?usp=sharing

Итоговый проект, тайминг

- 3 недели на реализацию проекта
- 1 неделя на подготовку презентации

Согласование проекта может занять до 2 недель!!!

Итоговый проект, варианты

Обязательно должен включать исследование набора данных с использованием машинного обучения

Где взять данные:

1. Собрать набор данных из интернета
2. https://www.kaggle.com/datasets

Пример проекта на открытых данных:

https://www.kaggle.com/fabiendaniel/film-recommendation-engine/notebook

Итоговый проект, варианты

Примеры датасетов:

https://www.kaggle.com/mczielinski/bitcoin-historical-data

https://www.kaggle.com/hugomathien/soccer

https://www.kaggle.com/kemical/kickstarter-projects

https://www.kaggle.com/nsharan/h-1b-visa

https://www.kaggle.com/san-francisco/sf-police-calls-for-service-and-incidents

https://www.kaggle.com/zynicide/wine-reviews

Итоговый проект, варианты

Соревнования:

https://www.kaggle.com/c/two-sigma-financial-news

https://www.kaggle.com/c/pubg-finish-placement-prediction

https://www.kaggle.com/c/competitive-data-science-predict-future-sales

https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

Домашнее задание №#7

- Выбрать проект
- Собрать команду
- Утвердить проект

Срок сдачи

12 ноября 2017

ТЕХНОСФЕРА

Спасибо за внимание!

Евгений Некрасов

e.nekrasov@corp.mail.ru