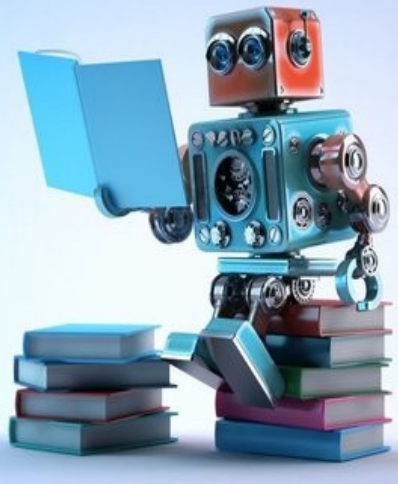
Машинное обучение

Вводная лекция



Калтович Артём
KaltovichArtyom@tut.by
Март 2018

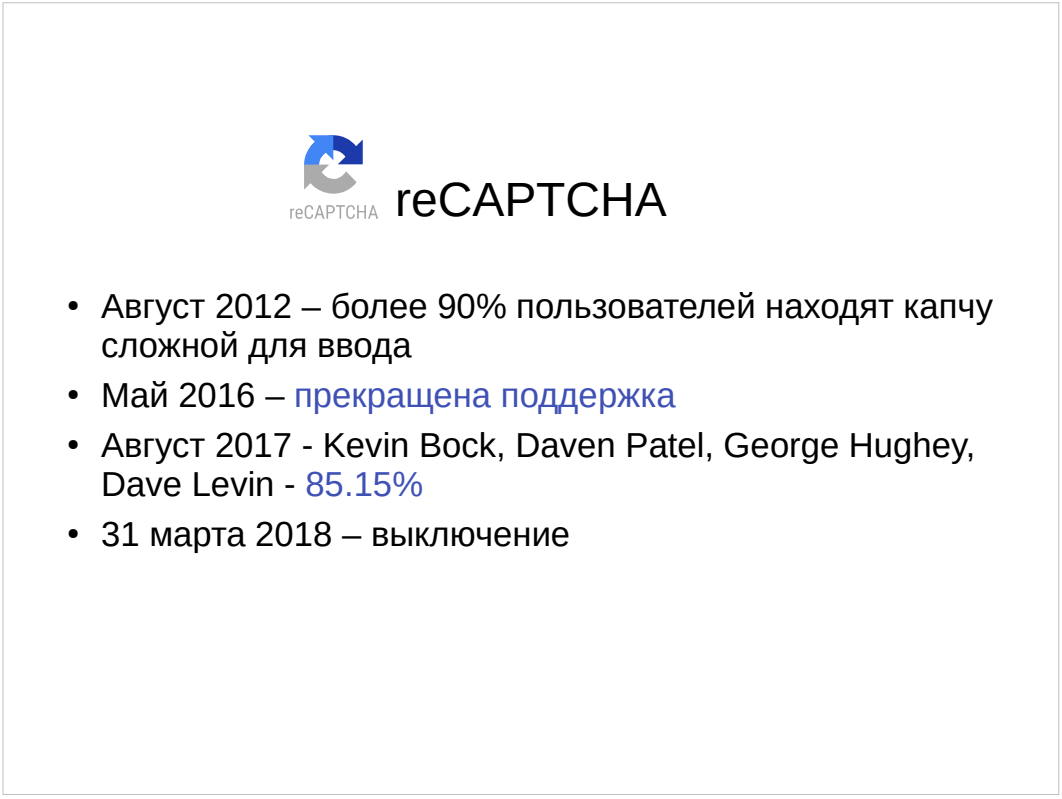
Source: 3 Fundamental Ways Machine Learning Will Change Business in 2018
https://www.entrepreneur.com/article/304945

reCAPTCHA

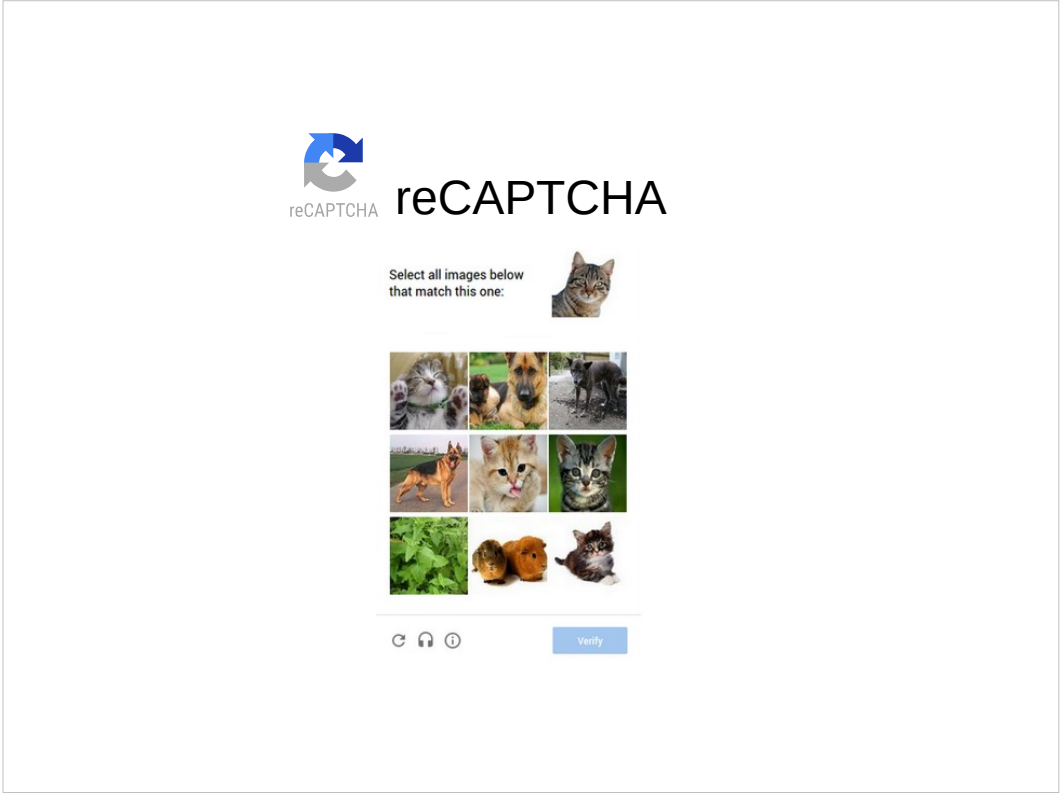
- 1 августа 2010 - Chad Houck – 31.8%
- 26 мая 2012 – Adam, C-P and Jeffball - 99.1% (аудио версия)
- За несколько часов до презентации Google выпустил обновление, точность упала до 60.95%
- 27 июня 2012 - Claudia Cruz, Fernando Uceda, and Leobardo Reyes - 82%

Иследователи утверждали, что Гугл постоянно менял свою капчу, при этом часто возвращаясь к предыдущей версии.

- 27 июня 2017 - Claudia Cruz, Fernando Uceda, and Leobardo Reyes – 82% - студенты из Мексики



Август 2017 – Университет Мериленда.



Август 2017 – Университет Мериленда.

Шахматы

- Февраль 1996 - Гарри Каспаров 4-2 Deep Blue
- Май 1997 - Гарри Каспаров 2½-3½ Deep Blue
- 2000 - программы Junior и Fritz матчи против Гарри Каспарова и Владимира Крамника – ничья
- Ноябрь-декабрь 2006 года - Владимир Крамник 2-4 Deep Fritz

В феврале 1996 года Гарри Каспаров победил шахматный суперкомпьютер Deep Blue со счетом 4-2. Этот матч выдающийся тем, что первую партию выиграл Deep Blue, автоматически став первым компьютером, победившим чемпиона мира по шахматам в турнирных условиях.

В мае 1997 года Deep Blue II выигрывает матч у Гарри Каспарова со счётом 3½ : 2½.

В 2000 году коммерческие шахматные программы Junior и Fritz смогли свести в ничью матчи против предыдущих мировых чемпионов Гарри Каспарова и Владимира Крамника.

В ноябре-декабре 2006 года чемпион мира Владимир Крамник играл с программой Deep Fritz. Матч закончился выигрышем машины со счётом 2-4.

Го

- Ноябрь 2015 - AlphaGo 4-1 Фань Хуэй
- Март 2016 - AlphaGo 4-1 Ли Седол
- Май 2017 AlphaGo 3-0 Эло Кэ Цзе

В октябре 2015 года программа AlphaGo, разработанная компанией DeepMind выиграла у трехкратного чемпиона Европы Фань Хуэя (2 профессиональный дан) матч из пяти партий со счётом 4—1. Это первый в истории случай, когда компьютер выиграл в го у профессионала в равной игре.

В марте 2016 года AlphaGo победила профессионала 9 дана Ли Седола в четырёх партиях из пяти.

В мае 2017 года на саммите «Future of Go Summit» AlphaGo выиграла три партии из трёх в мини-матче с одним из сильнейших игроков в мире, лидером мирового рейтинга Эло Кэ Цзе.

В чём же секрет такого успеха?

Машинное обучение

Титаник

Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
3	Panula, Mrs. Juha (Maria Emilia Ojala)	female	41.0	0	5	3101295	39.6875	NaN	Southampton
1	Penasco y Castellana, Mr. Victor de Satode 727	male	18.0	1	0	PC 17758	108.9000	C65	Cherbourg
2	Renouf, Mrs. Peter Henry (Lillian Jefferys)	female	30.0	3	0	31027	21.0000	NaN	Southampton
1	Taussig, Mr. Emil	male	52.0	1	1	110413	79.6500	E67	Southampton
3	Peduzzi, Mr. Joseph	male	NaN	0	0	A/5 2817	8.0500	NaN	Southampton

Пример данных
Спросить, какие правила предложат
Применить их
Попросить воспроизвести модель (входные данные - анализ – модель - новые данные – предсказание)
Нарисовать модель на доске
Спросить, можно ли автоматизировать

Машинное обучение

Очень похоже на то, что мы сделали сами, не находите?

Определение

Машинное обучение (англ. machine learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

Определение №2

Алгоритм A обучается с эффективностью E над данными D, если про росте мощности |D|, E проявляет тенденцию к увеличению.

Примеры применения

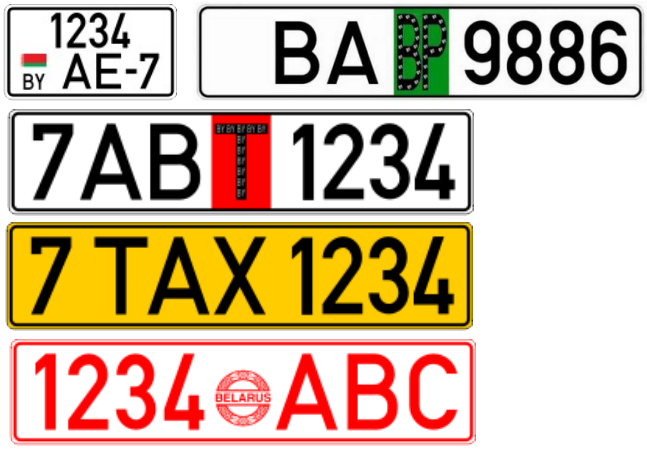
- Распознавание номеров автомобилей

Примеры применения

- Распознавание номеров автомобилей

Примеры применения

- Распознавание номеров автомобилей



Примеры применения

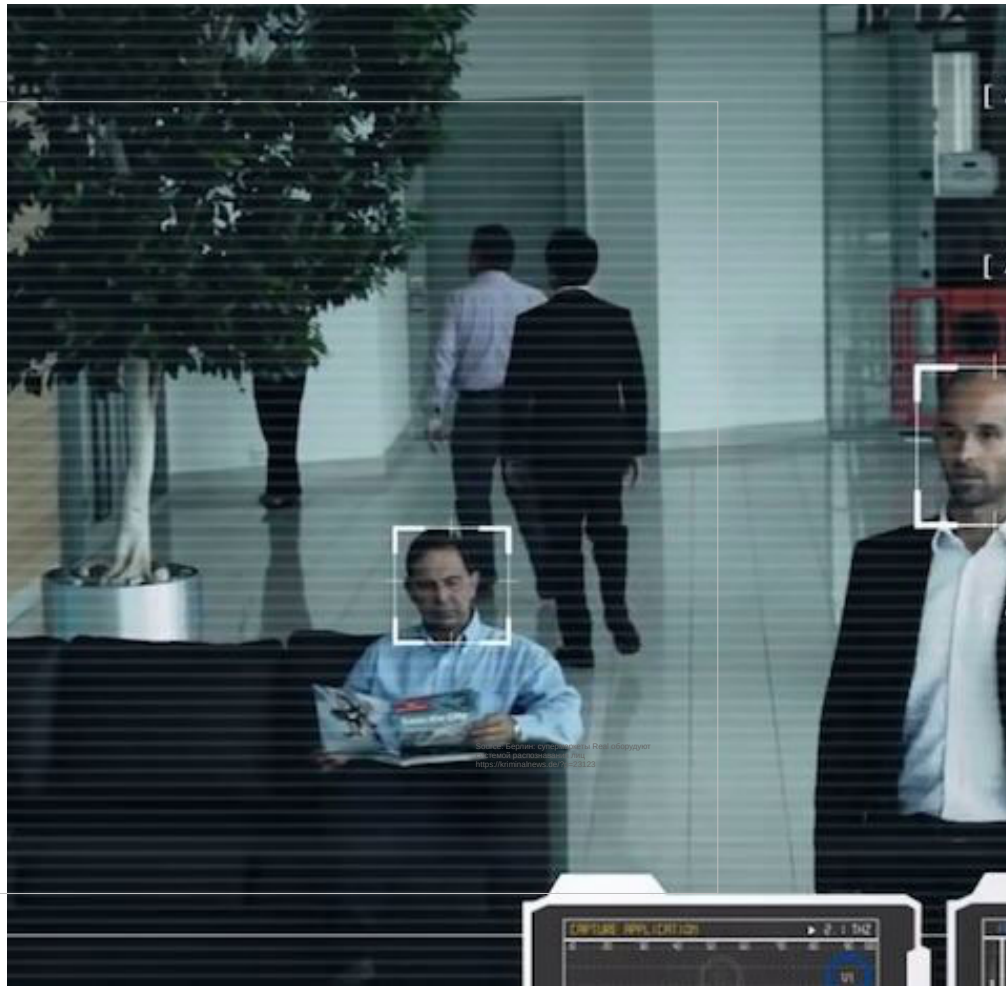
- Распознавание номеров автомобилей

Примеры применения

- Распознавание номеров автомобилей

Примеры применения

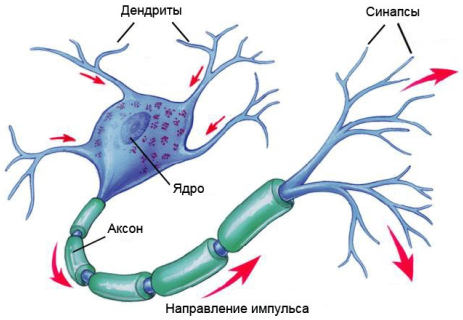
- Распознавание лиц



Примеры применения

- Распознавание изображений
- Перевод текстов
- Определение зловредных транзакций
- Фильтрация спама
- ИИ в компьютерных играх
- Высокоскоростная торговля на бирже
- Какие ещё?

Нейронные сети

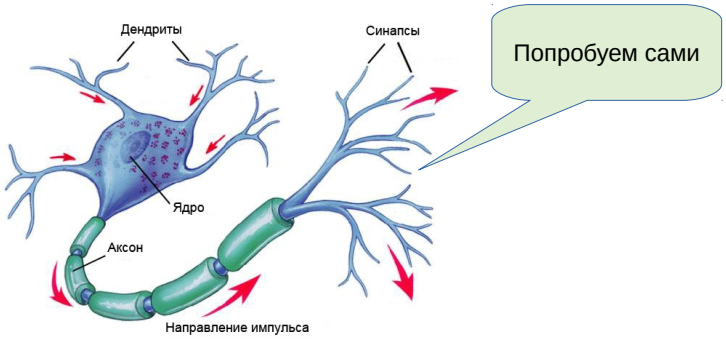


Source: ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ https://sites.google.com/site/anisimovkhv/learning/lis/lecture/tema16?tmpl=%2Fsystem%2Fapp%2Ftemplates%2Fprint%2F&showPrintDialog=1

Каждый нейрон имеет отростки нервных волокон двух типов - дендриты, по которым принимаются импульсы, и единственный аксон, по которому нейрон может передавать импульс. Аксон, который в конце разветвляется на волокна, контактирует с дендритами других нейронов через специальные образования - синапсы, которые влияют на силу импульса.

Можно считать, что при прохождении синапса сила импульса меняется в определенное число раз, которое мы будем называть весом синапса. Импульсы, поступившие к нейрону одновременно по нескольким дендритам, суммируются. Если суммарный импульс превышает некоторый порог, нейрон возбуждается, формирует собственный импульс и передает его далее по аксону. Важно отметить, что веса синапсов могут изменяться со временем, а значит, меняется и поведение соответствующего нейрона.

Нейронные сети



Source: ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ https://sites.google.com/site/anisimovkhv/learning/lis/lecture/tema16?tmpl=%2Fsystem%2Fapp%2Ftemplates%2Fprint%2F&showPrintDialog=1

Каждый нейрон имеет отростки нервных волокон двух типов - дендриты, по которым принимаются импульсы, и единственный аксон, по которому нейрон может передавать импульс. Аксон, который в конце разветвляется на волокна, контактирует с дендритами других нейронов через специальные образования - синапсы, которые влияют на силу импульса.

Можно считать, что при прохождении синапса сила импульса меняется в определенное число раз, которое мы будем называть весом синапса. Импульсы, поступившие к нейрону одновременно по нескольким дендритам, суммируются. Если суммарный импульс превышает некоторый порог, нейрон возбуждается, формирует собственный импульс и передает его далее по аксону. Важно отметить, что веса синапсов могут изменяться со временем, а значит, меняется и поведение соответствующего нейрона.

Обработка текста

Компьютер может работать только с числами?

Но как превратить слова в числа?

Мешок слов (или Bag of Words) это модель текстов на натуральном языке, в которой каждый документ или текст выглядит как неупорядоченный набор слов без сведений о связях между ними. Его можно представить в виде матрицы, каждая строка в которой соответствует отдельному документу или тексту, а каждый столбец — определенному слову. Ячейка на пересечении строки и столбца содержит количество вхождений слова в соответствующий документ.

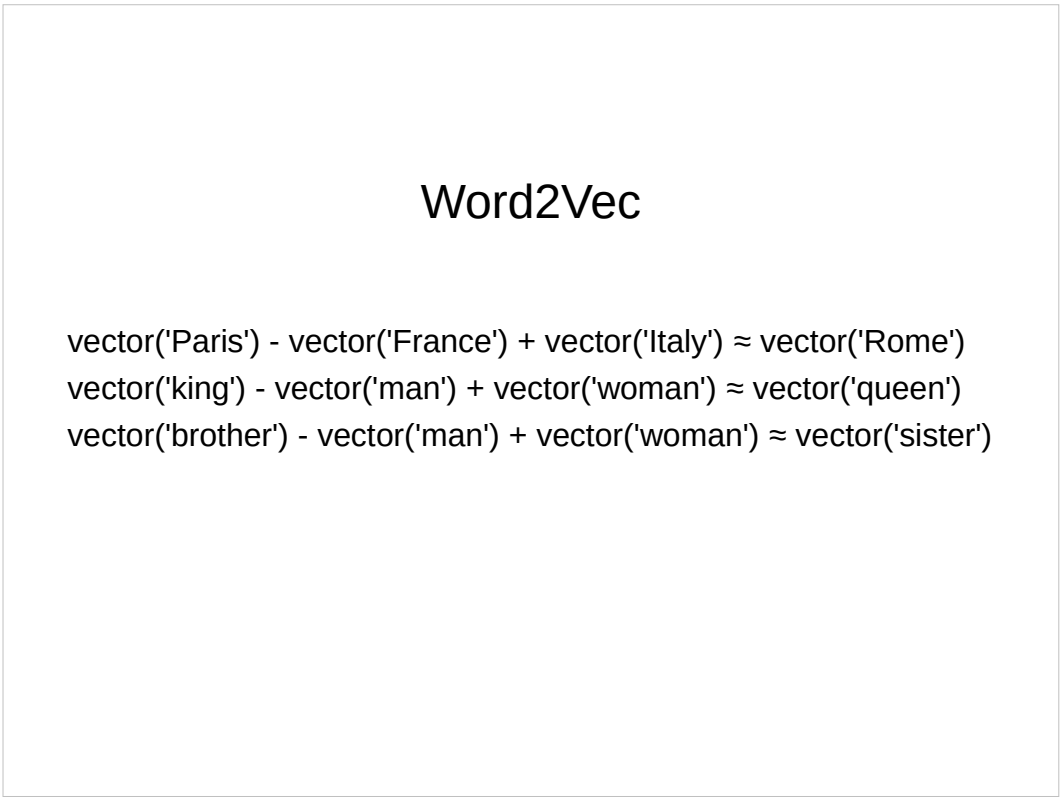
TF-IDF (от англ. TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

Word2Vec

Но что насчёт контекста?

Что если мы будем анализировать не только само слово, но и его соседей?

word2vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала он создает словарь, «обучаясь» на входных текстовых данных, 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 будут иметь близкие координаты векторов-слов. Полученные векторы-слова могут быть использованы для обработки естественного языка и машинного обучения.



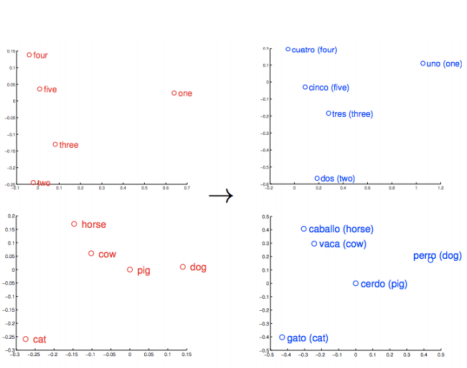
word2vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала он создает словарь, «обучаясь» на входных текстовых данных, а затем вычисляет векторное представление слов. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а следовательно, имеющие схожий смысл), в векторном представлении будут иметь близкие координаты векторов-слов. Полученные векторы-слова могут быть использованы для обработки естественного языка и машинного обучения.

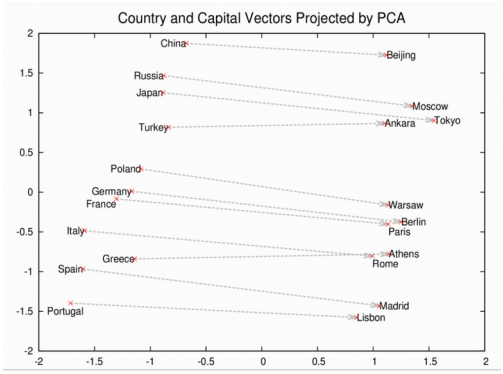
Word2Vec

- Geopolitics: Iraq - Violence = Jordan
- Distinction: Human - Animal = Ethics
- President - Power = Prime Minister
- Library - Books = Hall
- Analogy: Stock Market ≈ Thermometer

word2vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала он создает словарь, «обучаясь» на входных текстовых данных, 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 будут иметь близкие координаты векторов-слов. Полученные векторы-слова могут быть использованы для обработки естественного языка и машинного обучения.

Word2Vec





word2vec принимает большой текстовый корпус в качестве входных данных и сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Сначала он создает словарь, «обучаясь» на входных текстовых данных, 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 будут иметь близкие координаты векторов-слов. Полученные векторы-слова могут быть использованы для обработки естественного языка и машинного обучения.

Расстояние до Франции :)

spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130
switzerland	0.622323
luxembourg	0.610033
portugal	0.577154
russia	0.571507
germany	0.563291
catalonia	0.534176

Ресурсы

- http://scikit-learn.org/ - Документация, содержит много базовой информации и примеров
- http://playground.tensorflow.org/ - поиграться с нейронной сетью онлайн
- https://www.kaggle.com/ - сайт для соревнований и тренировок
- habrahabr.ru и geektimes.ru – русскоязычное сообщество
  - https://geektimes.ru/post/277088/ - статья про нейронные сети
- Конечно же, Википедия – лучше английская.