Компьютерная лингвистика: вводная лекция

Ключевые задачи компьютерной лингвистики

Обработка естественного языка (NLP)

понимание и генерация текста.

Машинный перевод

автоматический перевод текста между языками.

Вопросно-ответные системы

взаимодействие с пользователем через вопросы и ответы.

Распознавание речи

преобразование аудиосигналов в текст.

Синтез речи

генерация голоса из текста.

Анализ тональности

анализ эмоциональной окраски текстов.

https://colab.research.google.com/drive/1zuMJ7d8nAZsWM9G6mREt3EWBIdsHW90n?usp=sharing

https://colab.research.google.com/drive/1zuMJ7d8nAZsWM9G6mREt3EWBIdsHW90n?usp=sharing

Токен — это минимальная текстовая единица, с которой работает модель.

Генерация текста — это последовательное предсказание следующего токена.

Язык может быть представлен как последовательность дискретных единиц.

Модель учится вероятностному распределению этих единиц.

Значит, мы занимаемся формализацией языка через математику.

https://colab.research.google.com/drive/1zuMJ7d8nAZsWM9G6mREt3EWBIdsHW90n?usp=sharing

Но язык — это не только последовательность токенов.
У него есть структура.

Лингвистика формализует язык как систему структурных отношений.

Это значит, что мы можем посмотреть: морфологию (формы слов), синтаксис (связи между словами), семантику (значение).

Морфология описывает, какая у слова форма и какие у неё грамматические признаки (падеж, число, род, время…).

Синтаксис описывает структуру предложения: связи между словами и их роли (подлежащее, сказуемое, дополнение…).

Семантика — это значение. В вычислительном виде его часто приближают через “близость контекстов”: похожие по смыслу фразы имеют близкие векторы.

Вектор — (это упорядоченный набор чисел) — это числовое представление текста.
Откуда взялись векторы?
Наша LLM-модель преобразует последовательность токенов в вектор фиксированной длины: текст → токены → эмбеддинги слов → трансформер → агрегированный вектор

Что значит “близкие векторы”

$$\cos(v1, v2)$$

Если:

- cos ≈ 1 → смысл похож
- cos ≈ 0 → не связаны
- cos < 0 → противоположные направления

Векторное представление — это отображение текста из дискретного пространства символов в непрерывное евклидово пространство $\mathbb{R}^n$.

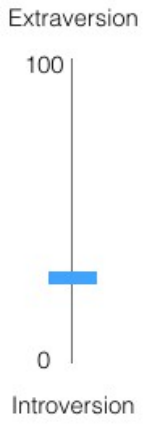
В сухом остатке: любая модель обучена так, что смысловая близость отражается в геометрической близости.

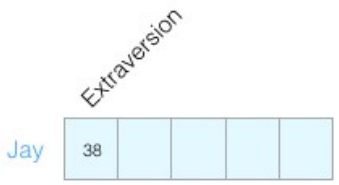
King − Man + Woman = ?

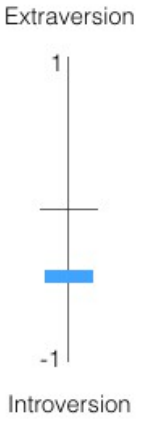
Распределённое представление — это способ представить языковую единицу (слово, фразу, документ) как вектор чисел,
где значение “распределено” по множеству координат.
В английском варианте статей и учебников — embeddings.

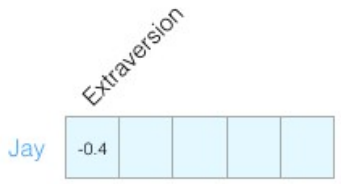
Word2vec — метод создания эмбедингов, разработанный в 2013 году. В большинстве случаев речь идет о кодировании слов.

Уйдем от математики и посмотрим на соседнюю область —
тест «большая пятерка»

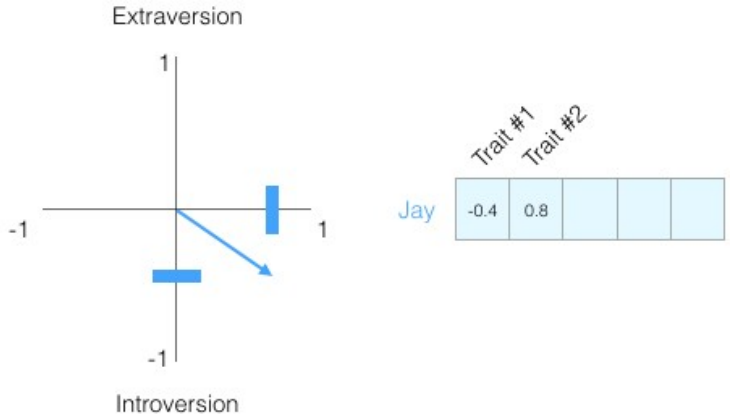








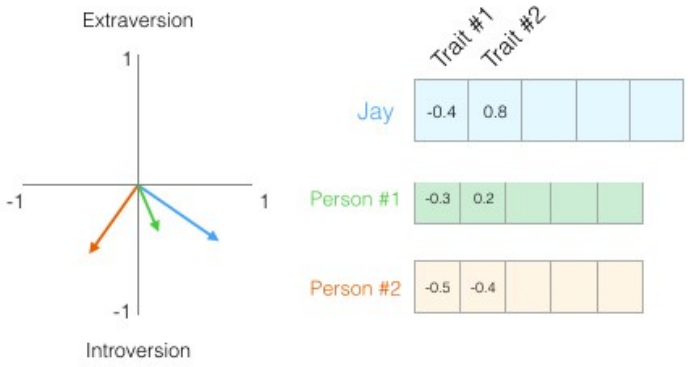
Одной характеристики мало для любого суждения,
добавим ещё одно измерение: ещё одну характеристику из теста.



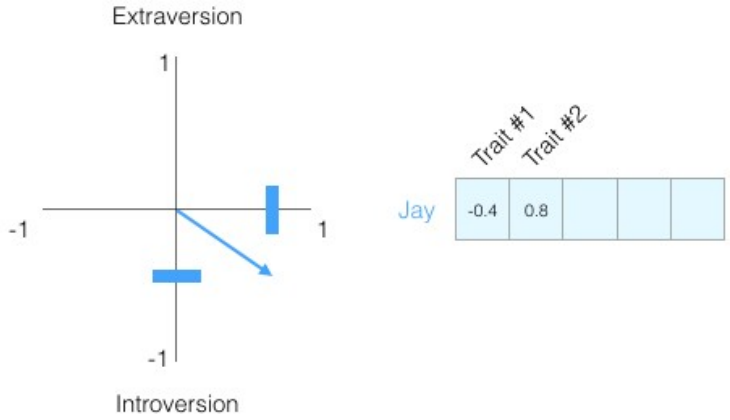
Можно представить эти два измерения как точку на графике или,
но лучше — как **вектор** от начала координат до этой точки.

вектор частично (но подробнее, чем одна точка) отражает личность

Допустим, Jay выбыл из игры,
кто на него похож больше?

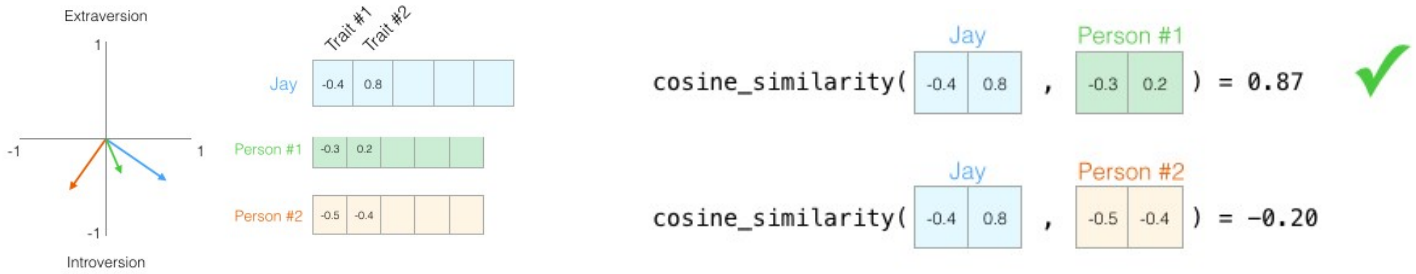


Одной характеристики мало для любого суждения,
добавим ещё одно измерение: ещё одну характеристику из теста.

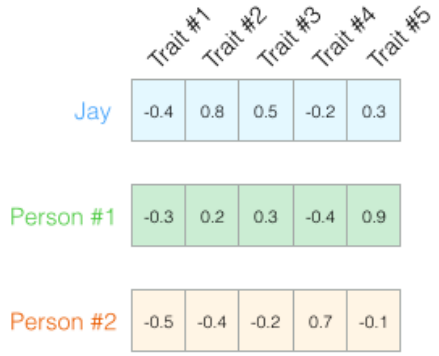


Можно представить эти два измерения как точку на графике или,
но лучше — как **вектор** от начала координат до этой точки.

вектор частично (но подробнее, чем одна точка) отражает личность

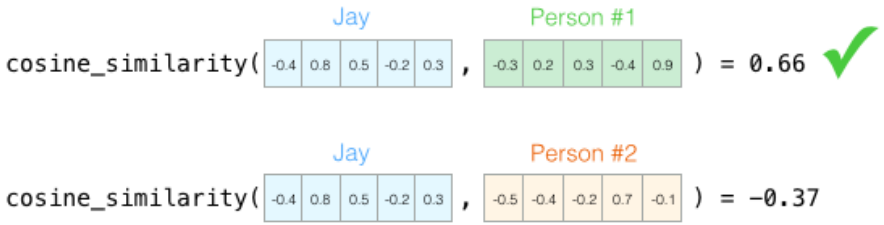


Косинусное сходство (cosine similarity) — мера сходства между двумя векторами, которая вычисляется на основе косинуса угла между ними. Чем меньше угол, тем больше сходство.



увы, аккуратные стрелки в 2D больше не для нас, капитанские выводы больше не доступны.

Но! косинусное сходство нормально считается и для многомерного пространства.



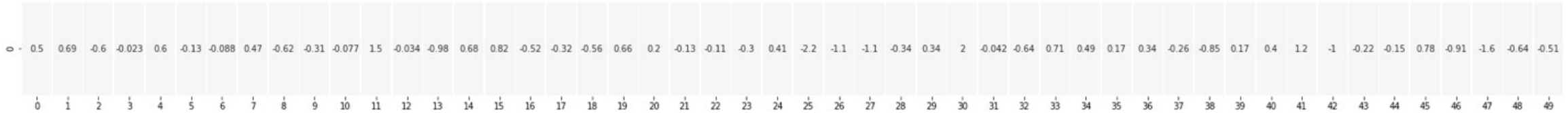
Более того: по пяти измерениям результат гораздо точнее

Мораль:
1. Людей (и другие объекты) можно представить в виде числовых векторов, и это отлично подходит для машин!
2. Мы можем легко вычислить, насколько похожи векторы.

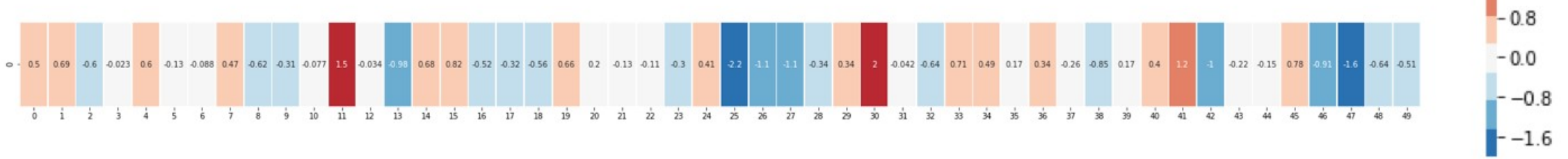
Вернемся к королю

Эмбединг для слова «король»:

[0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 , -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042]



Очень интересно, но не понятно.
Раскрасим ячейки по их значениям:
красный для близких к 2, белый для близких к 0, синий для близких к −2).



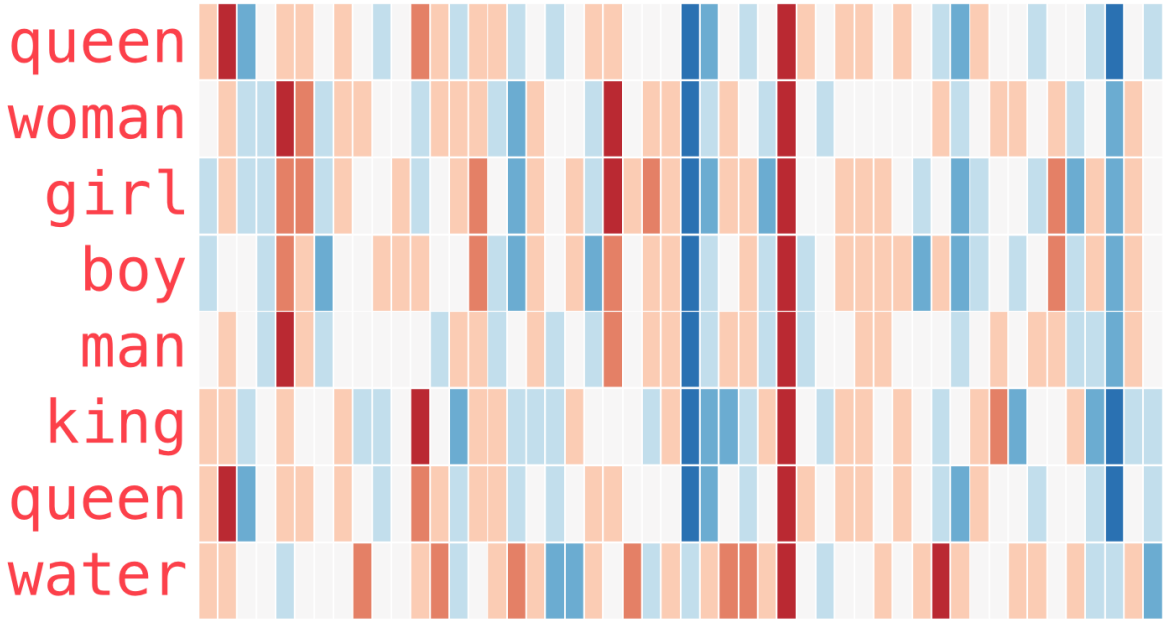
“king”

“Man”

“Woman”

«мужчина» и «женщина» гораздо ближе друг к другу, чем к «королю»

Мораль.
Векторные представления захватывают довольно много информации/значения/ассоциаций этих слов.



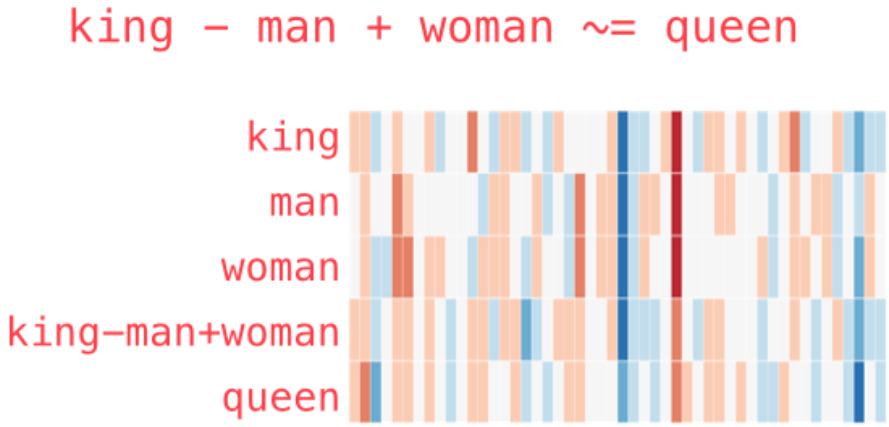
1) через все слова проходит одна красная колонка. То есть эти слова похожи в этом конкретном измерении (и мы не знаем, что в нём закодировано).
2) «женщина» и «девушка» во многом похожи, а «мужчина» с «мальчиком».
3) «мальчик» и «девочка» тоже похожи в некоторых измерениях, но отличаются от «женщины» и «мужчины». мб, удалось закодировать контекст молодости?
4) есть чёткие измерения, где «король» и «королева» похожи друг на друга и отличаются от всех остальных. мб, закодирована расплывчатая концепция королевской власти?
5) что с водой?

Мы можем складывать и вычитать векторы слов, получая интересные результаты.
Самый известный пример — формула **«король − мужчина + женщина»**:

```
model.most_similar(positive=["king","woman"], negative=["man"])
```

```
[('queen', 0.8523603677749634),
 ('throne', 0.7664333581924438),
 ('prince', 0.7592144012451172),
 ('daughter', 0.7473883032798767),
 ('elizabeth', 0.7460219860076904),
 ('princess', 0.7424570322036743),
 ('kingdom', 0.7337411642074585),
 ('monarch', 0.721449077129364),
 ('eldest', 0.7184862494468689),
 ('widow', 0.7099430561065674)]
```

коэффициент геометрического сходства



«королева» не синоним, но это наиболее близкий результат из 400 000 вложений слов в наборе данных

Резюме

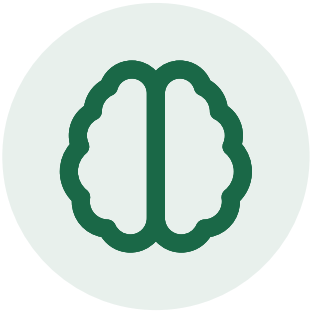
язык — это не только «текст»,
а система структур:

Морфология

Изучение структуры слов, например, морфемный анализ. Инструменты: морфологические разборщики.

Синтаксис

Разбор грамматических связей и построение деревьев зависимостей. Инструменты: парсеры.



Семантика

Анализ значений слов и фраз. Инструменты: векторные модели, семантические сети.

Резюме

Морфология формализуется как признаки

наукой → NOUN | Case=Ins | Gender=Fem | Number=Sing

Синтаксис формализуется как граф

есть ROOT, есть роли (nsubj, obj и т. д.), есть структура

язык = структура

Семантика формализуется как геометрия

предложение → вектор; близость смыслов → cos между векторами

значение можно представить численно

Язык можно представить как:

систему признаков,
систему зависимостей,
систему векторов в пространстве.

а еще… как **систему контекстов и намерений**

Прагматика

Это еще один уровень лингвистического анализа. Изучение контекста и намерений коммуникации.

Итог

Компьютерная лингвистика

междисциплинарная область, изучающая
язык как формальную систему
(признаков, структур, значений и контекстов)
и использующая алгоритмы и
математические модели
для анализа и автоматической обработки
естественного языка

Истоки компьютерной лингвистики

середина XX века:
задачи автоматического перевода,
формального описания грамматики,
синтаксического анализа.

В дальнейшем...

область включила статистические методы,
векторные модели значений,
машинное обучение
и нейросетевые модели,
приближаясь к моделированию
семантики и прагматики.

Что такое язык?

Методы и подходы в компьютерной лингвистике