ДИСТРИБУТИВНАЯ СЕМАНТИКА

Маша Шеянова, masha.shejanova@gmail.com

August 2, 2018

НИУ ВШЭ

INTRO

что это?

Что мы хотим:

- уметь считать расстояние между словами
- учитывая только **значения слов** (насколько слова близки друг к другу по значению)
- делать это автоматически

Пример: лампа и светильник — ближе, чем лампа и лавка.

КАК К ЭТОМУ ПОДСТУПИТЬСЯ?

Дистрибутивная гипотеза: значения слов определяются их контекстами. Слова с похожими типичными контекстами имеют схожее значение.

> You shall know a word by the company it keeps! (J.R.Firth)

КАК ЭТО РАБОТАЕТ?

Нам нужно:

- много текстов, чтобы картинка была репрезентативной
- посчитать в этих текстах взаимную встречаемость слов друг с другом
- найти слова, которые могут заменить друг друга и слова, у которых нет общих контекстов

Готово! Мы прекрасны и можем

- находить слова, близкие по значению к данному
- строить семантические пропорции
- строить семантические визуализации

Примеры: RusVectores

Что такое RUSVECTORES?

На rusvectores можно найти слова, наиболее близкие к данному, построить семантическую пропорцию и многое другое.

БЛИЗКИЕ СЛОВА

Семантические аналоги для *спокойный* (ALL)

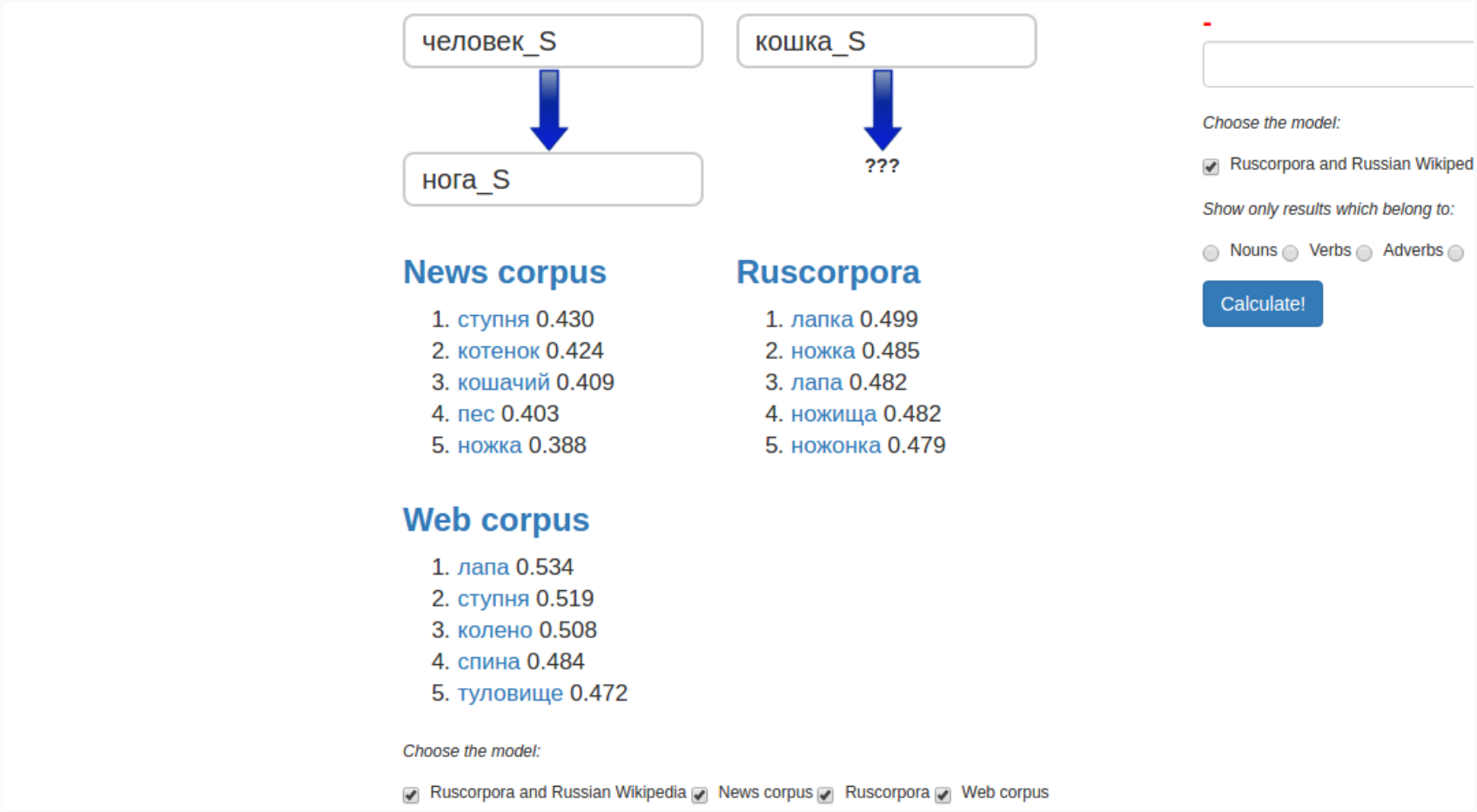
НКРЯ и Wikipedia

1. невозмутимый 0.69
2. безмятежный 0.68
3. спокойный 0.67
4. -спокойный 0.66
5. несуетливый 0.65
6. умиротворенный 0.65
7. умиротворять 0.63
8. раздумчивый 0.63
9. неторопливый 0.62
10. кроткий 0.62

Новостной корпус

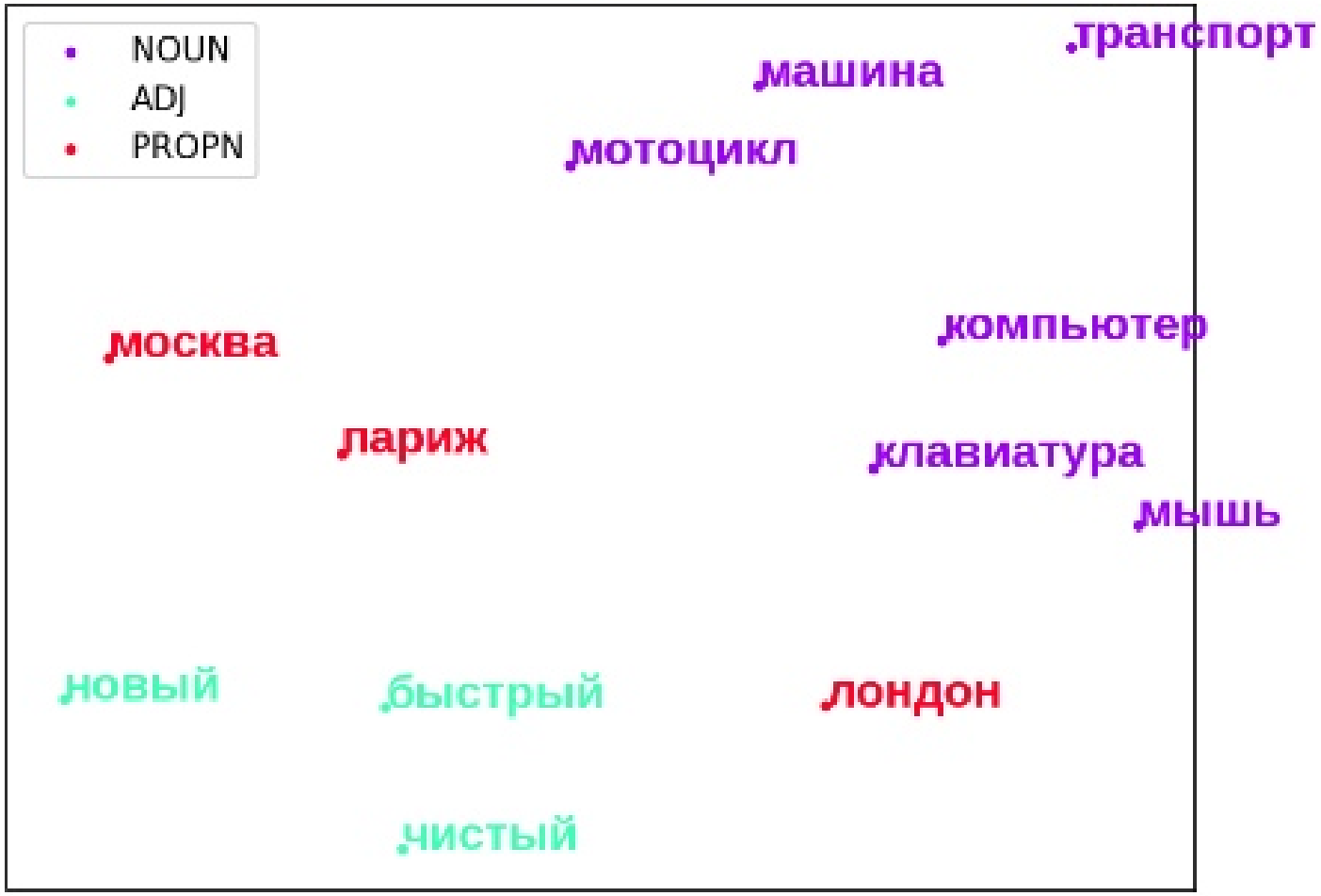
1. умиротворенный 0.52
2. размеренный 0.50
3. безмятежный 0.50
4. неспокойный 0.50
5. уравновешенный 0.49
6. расслабленный 0.47
7. беспокойный 0.47
8. неторопливый 0.45
9. доброжелательный 0.45
10. дружелюбный 0.44

ПРОПОРЦИИ


человек_S
кошка_S
нога_S
???
News corpus
1. ступня 0.430
2. котенок 0.424
3. кошачий 0.409
4. пес 0.403
5. ножка 0.388
Ruscorpora
1. лапка 0.499
2. ножка 0.485
3. лапа 0.482
4. ножища 0.482
5. ножонка 0.479
Web corpus
1. лапа 0.534
2. ступня 0.519
3. колено 0.508
4. спина 0.484
5. туловище 0.472
Choose the model:
Ruscorpora and Russian Wikipedia
News corpus
Ruscorpora
Web corpus
Choose the model:
Ruscorpora and Russian Wikiped
Show only results which belong to:
Nouns
Verbs
Adverbs
Calculate!

Новостной корпус

Визуализировать в TensorFlow Projector



КАК ЭТО РАБОТАЕТ

РАССТОЯНИЯ МЕЖДУ СЛОВАМИ

Во введении я говорила, что мы считаем близость на основе контекстов. Но как именно?

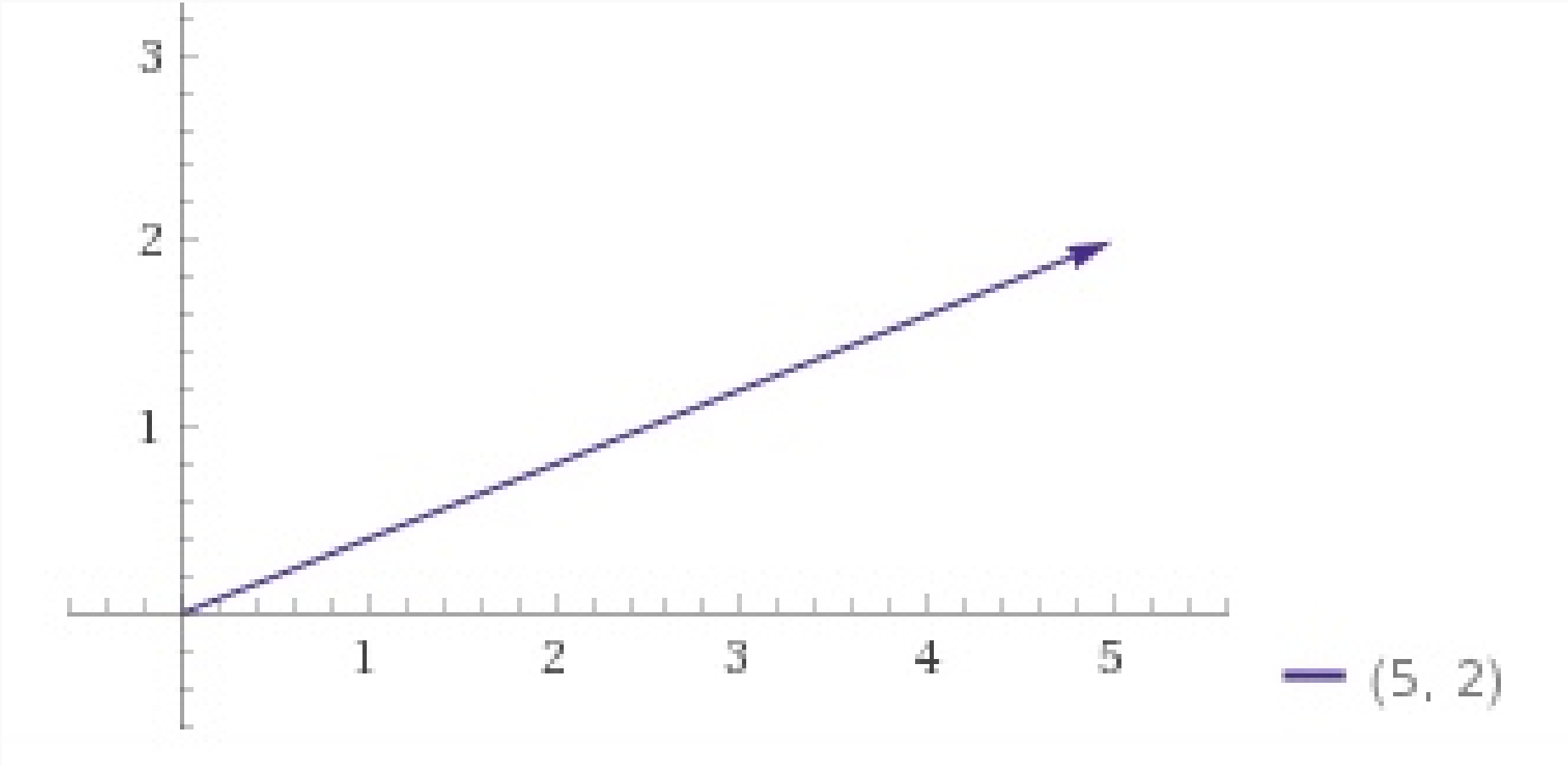
Превращаем слова в векторы
и измеряем косинусное расстояние[1] между ними.

[1] Есть, впрочем, и другие метрики.

Что такое вектор?

Что такое вектор

В школе нас обычно учат, что вектор — это стрелочка в системе координат.



Ту же стрелочку можно представить как набор чисел: $(5, 2)$.

Что такое вектор

А что если простанство 3-мерное? 4-мерное? 100500-мерное? В 3D стрелочку представить ещё можно. А в 100500-мерном придётся обходиться числами.

Слова и вектора. Счётные модели

Но **как** мы векторизуем слово? Как уже было сказано, по контекстам.

можно ли это настраивать?

Да. Например, можно поиграться с размером окна.
Можно считать все вхождения слов в окне от 5 до нашего слова до 5 после:

туда [пришла. Потом мы начали смотреть **мультики** и до двух ночи не] ложились …

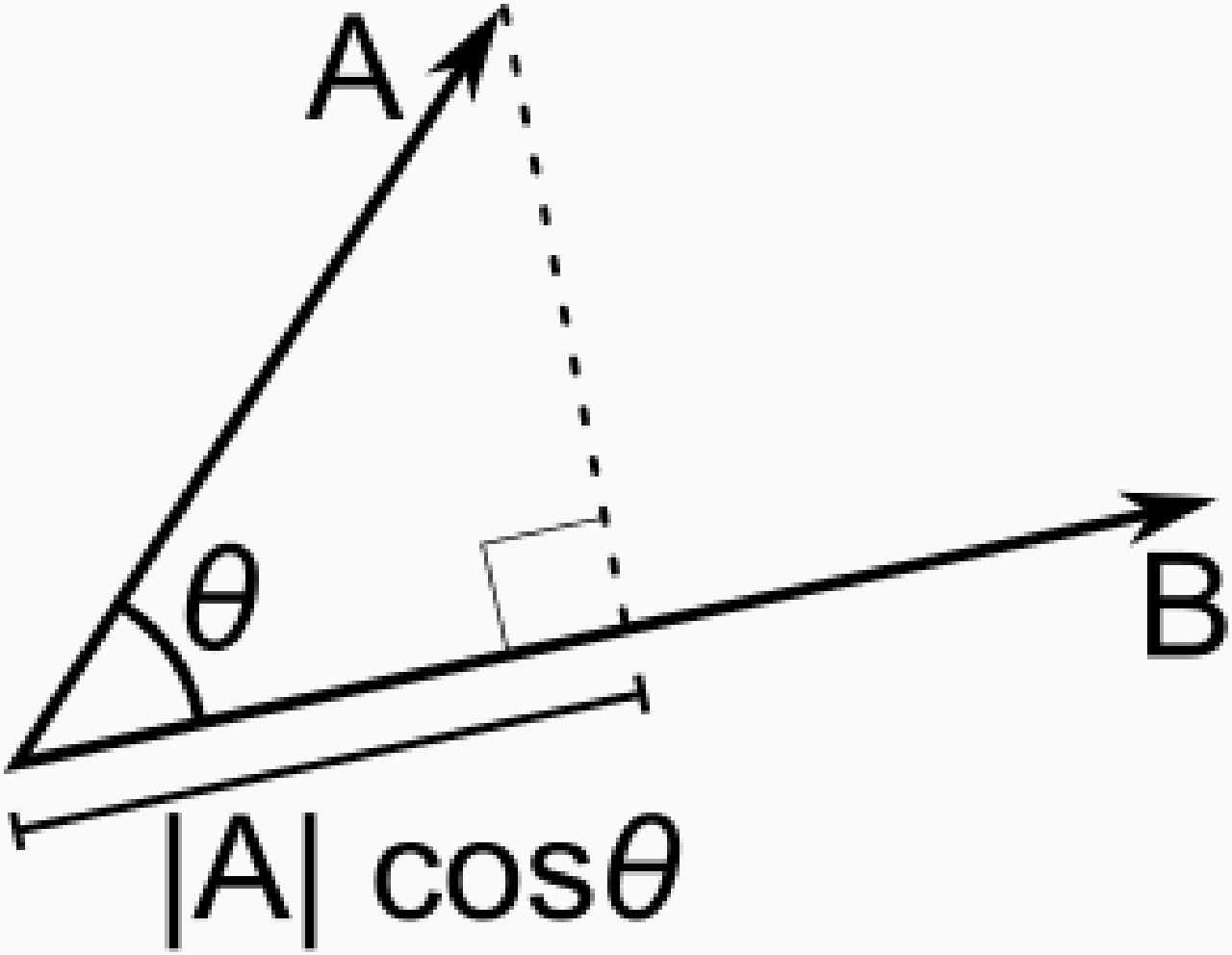
А можно — от -3 до +3:

туда пришла. Потом [мы начали смотреть **мультики** и до двух] ночи не ложились …

От этого будет зависеть, в каких отношениях находятся близкие, согласно нашей можели, слова.

КОСИНУСНАЯ БЛИЗОСТЬ

Что это? Да просто косинус угла между веторами!



Чем косинус угла ближе к единице, тем ближе слова друг к другу, чем ближе к 0 — тем дальше.

КОСИНУСНАЯ БЛИЗОСТЬ

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

НЕСОВЕРШЕНСТВО СЧЁТНЫХ МОДЕЛЕЙ

У счётных моделей есть глобальные недостатки:

- Размер векторов получается огромным (в общем случае равен объёму лексикона).
- Это очень замедляет операции сравнения векторов.
- Мы не знаем точно, что в наших векторах нужная информация, а что мусор. Они просто взяты из корпуса.

НЕЙРОСЕТОЧКИ!

Как быть? Используем нейросеточки для **предсказания** векторов.

Мы пытаемся для каждого слова найти такой вектор, чтобы он был максимально схож с векторами типичных соседей и максимально отличался от векторов слов, которые соседями данному слову не являются.

ИНСТРУМЕНТЫ

GENSIM, WORD2VEC

В 2013 году исследователь Tomas Mikolov из Google с соавторами разработали word2vec, который позволяет тренировать нейронные языковые модели на больших корпусах.

Сейчас для многих языков (например, для русского) есть готовые обученные модели модели!

doc2vec

А что если нам надо векторизовать целый текст?
doc2vec умеет делать это с помощью усреднения векторов каждого слова в документе!
Цитируя статью doc2vec,

> In Doc2Vec, we represent each document as a simple average of the word embeddings of all the words in the document.

ПРИМЕНЕНИЕ

Окей, мы умеем считать расстояние между словами. Что дальше?

ПРИМЕНЕНИЕ

- поиск синонимов и вообще похожих слов
- снятие семантической омонимии (Word Sense Disambiguation)
- признаки для машинного обучения с текстами
- … и туча всего другого!

Спасибо за внимание!

Вопросы?