ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ

Маша Шеянова, masha.shejanova@gmail.com

October 26, 2018

НИУ ВШЭ

КОРПУСА

что это?

В общем виде — собрание текстов.
Обычно имеют в виду **размеченные** тексты:

- с лемматизацией
- с тэггингом
- с грамматическим анализом
- (редко) с синтаксисом
- (ещё реже) с семантикой
- с чем-то ещё своим — например, тональностью

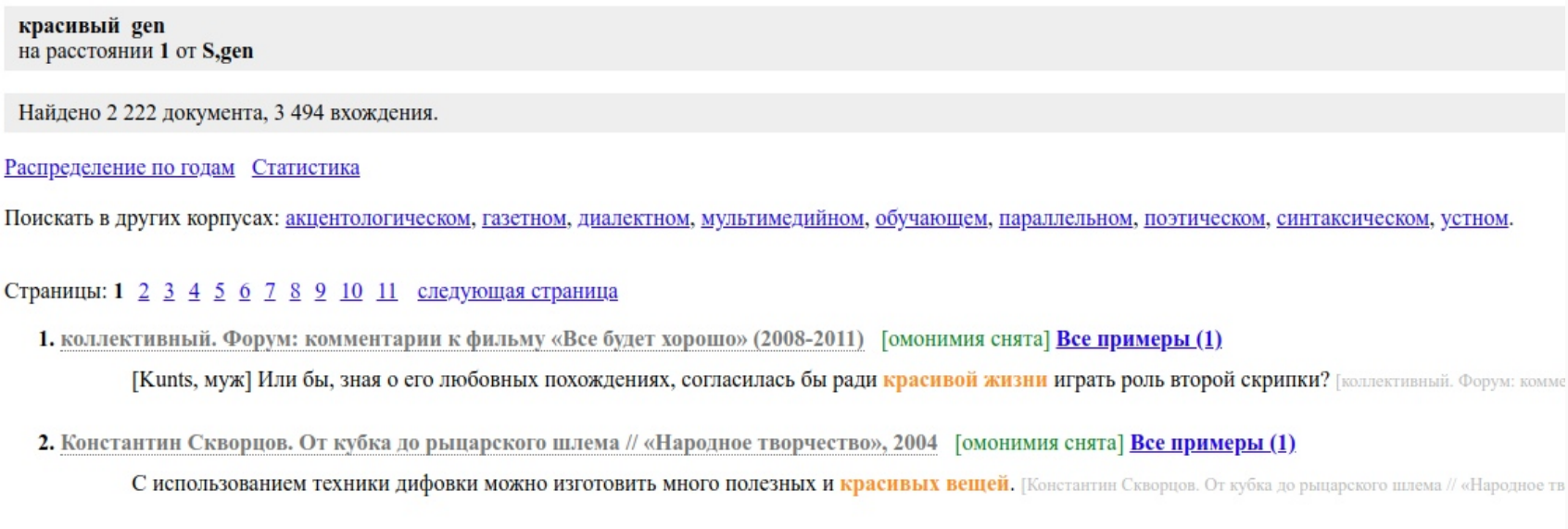
ПАРАЛЛЕЛЬНЫЕ КОРПУСА

Это нечто отдельное, и очень полезное!

Английский	Японский
How much is that red umbrella?	Ano akai kasa wa ikura desu ka.
How much is that small camera?	Ano chiisai kamera wa ikura desu ka.

Используется, в первую очередь, для машинного перевода и кросс-языковых лингвистических исследований.

http://ruscorpora.ru — национальный корпус русского языка



красивый gen
на расстоянии **1** от **S,gen**

Найдено 2 222 документа, 3 494 вхождения.

Распределение по годам Статистика

Поискать в других корпусах: акцентологическом, газетном, диалектном, мультимедийном, обучающем, параллельном, поэтическом, синтаксическом, устном.

Страницы: **1** 2 3 4 5 6 7 8 9 10 11 следующая страница

1. **коллективный. Форум: комментарии к фильму «Все будет хорошо» (2008-2011)** [омонимия снята] **Все примеры (1)**

   [Kunts, муж] Или бы, зная о его любовных похождениях, согласилась бы ради **красивой жизни** играть роль второй скрипки? [коллективный. Форум: комме

2. **Константин Скворцов. От кубка до рыцарского шлема // «Народное творчество», 2004** [омонимия снята] **Все примеры (1)**

   С использованием техники дифовки можно изготовить много полезных и **красивых вещей**. [Константин Скворцов. От кубка до рыцарского шлема // «Народное тв

Сбалансированный по жанрам корпус с морфологической и семантической разметкой. Умеет очень много.

RUSVECTORES

Векторная семантика: 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 картина 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

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)

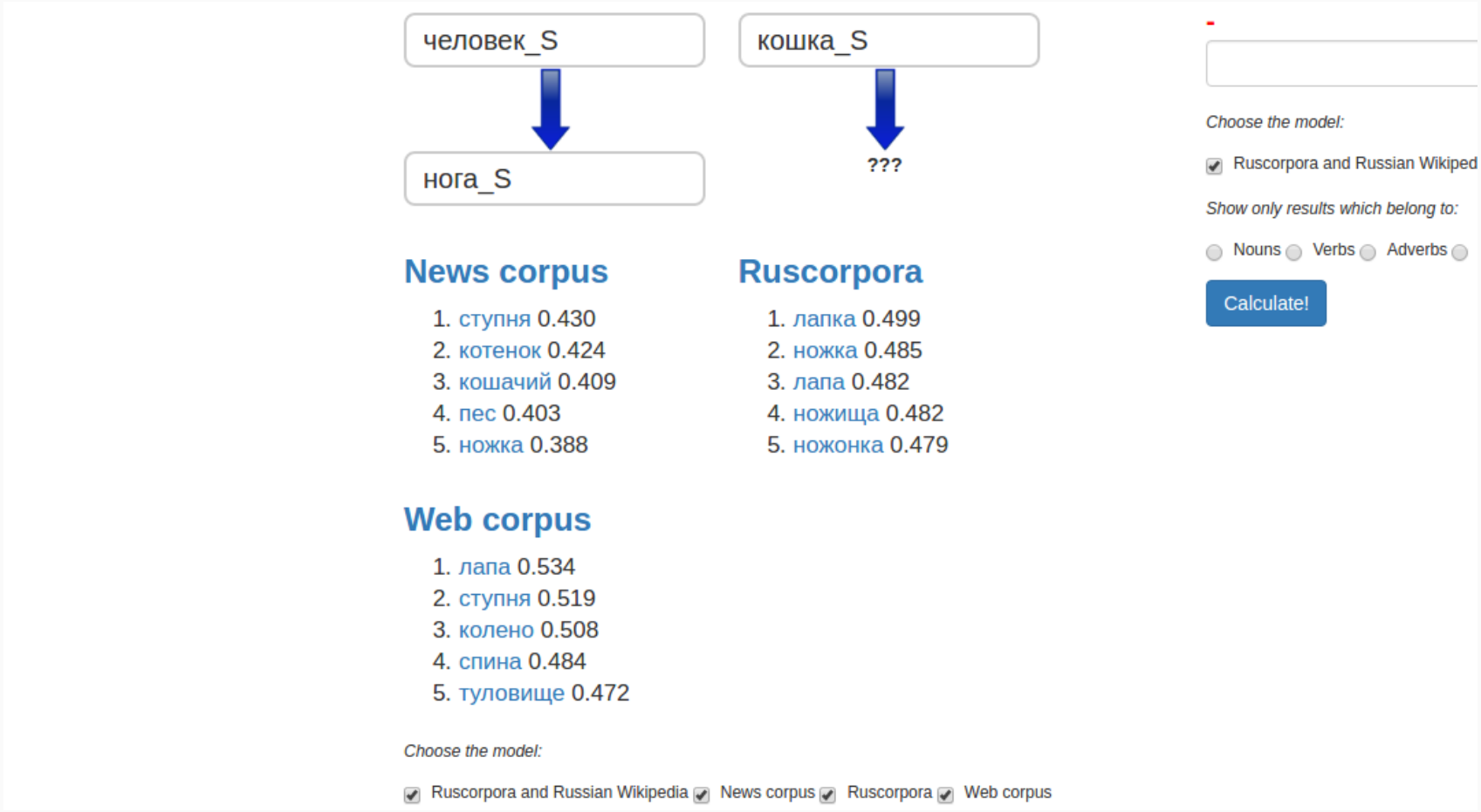
НКРЯ и Wikipedia

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
нога_S
кошка_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 Wiped
Show only results which belong to:
Nouns
Verbs
Adverbs
Calculate!

Новостной корпус

Визуализировать в TensorFlow Projector

ЗАДАНИЕ

ЗАДАНИЕ

Разбейтесь на команды по 2 человека. Каждая команда должна изучить один из онлайн ресурсов:

- mystem+: http://web-corpora.net/wsgi/mystemplus.wsgi/mystemplus/ (особое внимание — вкладке Тэггеры)
- UDPipe: http://lindat.mff.cuni.cz/services/udpipe/
- Главред: https://glvrd.ru/
- tatoeba: https://tatoeba.org/eng/
- ГИКРЯ: http://www.webcorpora.ru/
- AdaGram: http://adagram.ll-cl.org

... а потом рассказать всем, что он делает и зачем нужен.
Стоит упомянуть: для чего ресурс; если получилось разобраться — как он работает; какие шаги из лингвистического пайплайна использует.