BIGDATA TEAM

Natural Language Processing и продуктивизация моделей

Драль Алексей, study@bigdatateam.org
CEO at BigData Team, http://bigdatateam.org
https://www.facebook.com/bigdatateam

MEGAFON

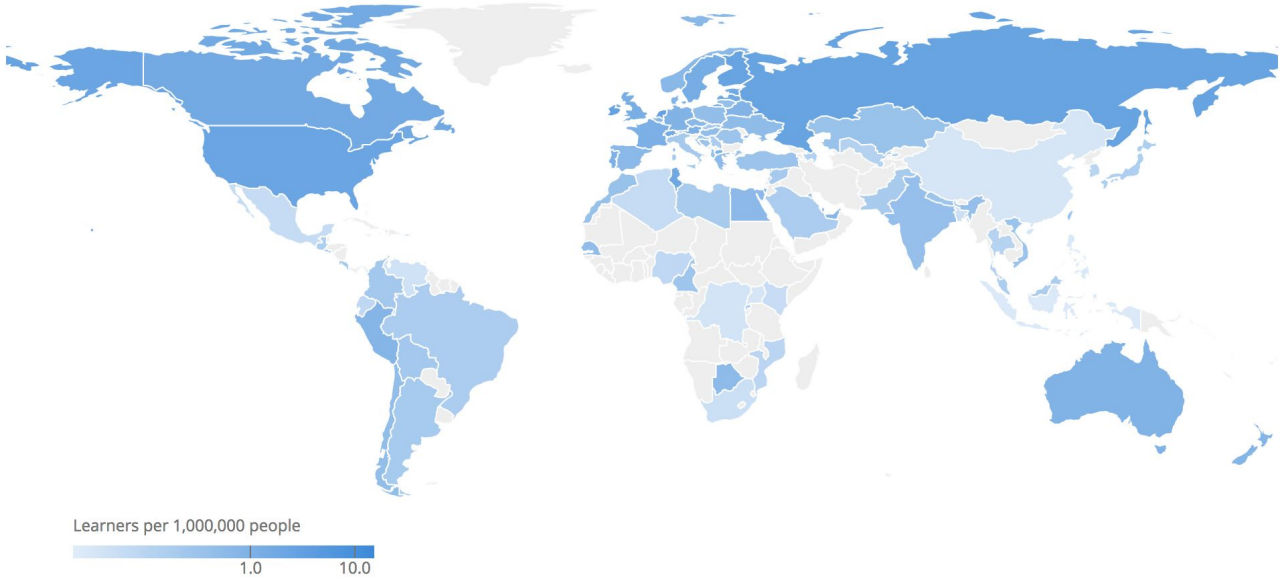
09.12.2020, online

BIGDATA TEAM

Проводим обучение по Big Data, Machine Learning, Python для:

BIGDATA TEAM

Образовательные партнеры Яндекс по курсам Big Data на Coursera



Partners:

coursera

Yandex

https://bigdatateam.org/big-data-engineering (50+ тысяч слушателей)

BIGDATA TEAM

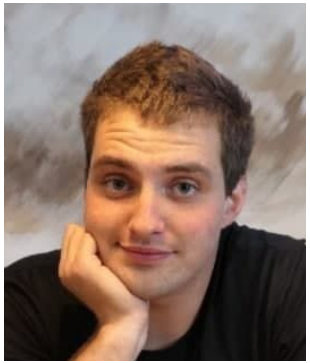
Команда курса NLP

Алексей Драль

Максим Рябинин

Радослав Нейчев

Ася Ройтберг



Арсений Ашуха

Виктория Брынза

Наталья Корепанова

Анастасия Янина

Дмитрий Игнатов

Обо мне

- СУНЦ МГУ
- МГУ
- ШАД Яндекса
- Рамблер
- Яндекс
- Amazon AWS
- МФТИ
- Сбербанк
- BigData Team

BIGDATA TEAM

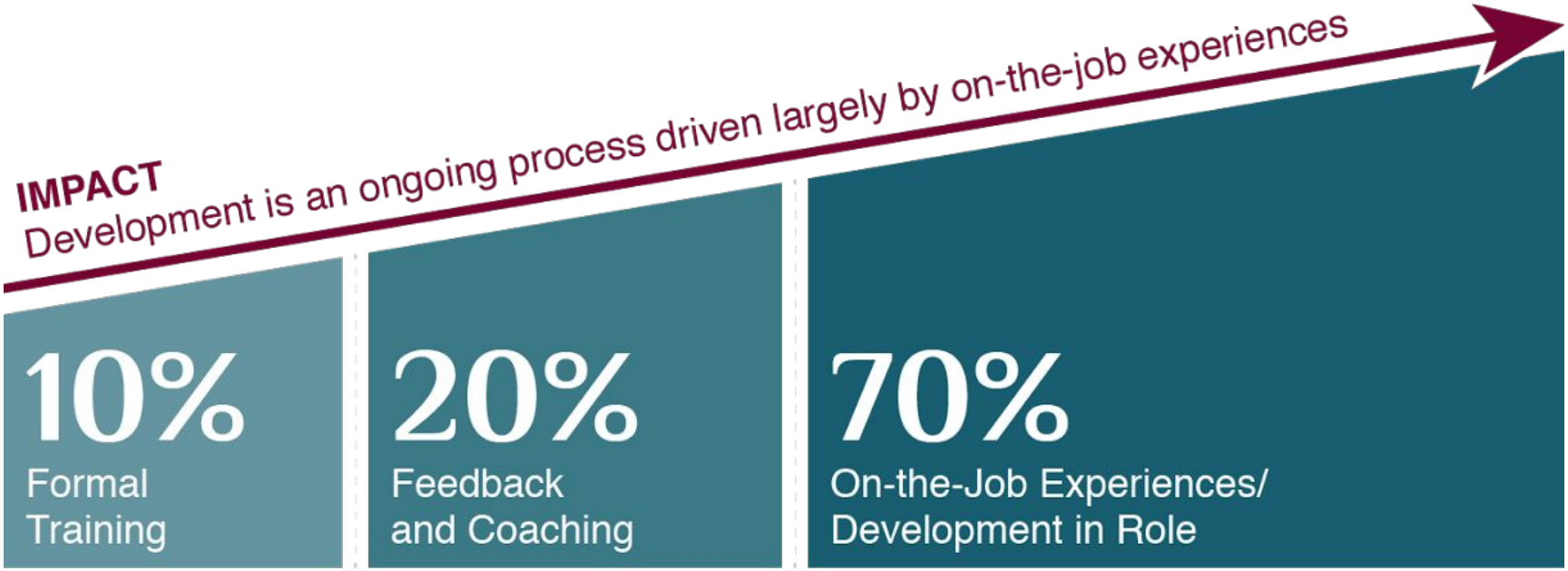
Программа курса

1. Введение в embeddings + HMM
2. Компьютерная лингвистика и разметка
3. Классификация текстов и работа со звуком
4. Тематическое моделирование
5. Языковые модели (Language Models)
6. Преобразование последовательностей (seq2seq, attention)
7. Transfer Learning
8. Приглашенная лекция от исследователя/практика NLP
9. Self-Supervised Learning (самообучение)

10-12. Продуктивизация моделей (TBD)

BIGDATA TEAM

Наш подход



- 1 ДЗ на каждый учебный модуль с нагрузкой 6 часов
- итоговое тестирование

BIGDATA TEAM

Метрики F2F обучения

до

20-30%

после

70-80%

HDFS	MR	MR	MR	MR	Hive	Hive	Hive	Hive	Spark	Spark	Spark	Spark	Spark	RT	RT	RT	RT	NoSQL	NoSQL
0	0	0.75	0	0	0	0	0	1	1	0	0	1	0	0	0.66	0	0	0	0
0.5	0	1	0	0.6	0	1	0	1	0.75	0	0	1	0	0	0.66	0.66	1	1	1
0	0	1	0	0	0	0	0	1	0.5	0	0	0	0.5	0	0	0	0	0	0
0	1	0.75	0	0.4	0	0	0	1	0.75	1	0	0	0	1	0	0	0	0	0
0.75	0	1	0	0.8	0	0	0	1	0.75	1	1	1	1	0	0	0	0	0	0
0.25	0	1	0	0	0	1	0	1	0.5	1	0	0	0.5	0	0.66	0.66	0	0	0
0.75	0	1	0	0.8	0	0	0	1	0.75	0	0	0	0	0	0.99	0.66	1	0	0
0.75	1	1	1	0.8	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0
0	0	0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	0	0.5	0	0	0	0	0	1	1	0	1	0	0.75	0	0.33	0.66	1	1	0
0.5	0	0.75	0	1	0	0	1	1	0.75	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	1	0	0	0	1	1	1	1	0	0.75	0	0.99	0.33	1	0	0
0.75	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0
0.25	0	0.75	1	0	0	0.75	0	1	0.75	1	1	1	1	0	0	0	0	0	1
1	0	0.5	0	0	0	0	0	0	0.75	0	0	0	0	0	0	0	0	0	1
0.75	0	1	0	0	0	0	0	1	0.75	0	0	0	1	1	0.66	0	0	0	0
0.5	0	1	0	1	1	0	1	1	0	1	1	0	1	0	0	0	0	0	0

HDFS	MR	MR	MR	MR	Hive	Hive	Hive	Hive	Spark	Spark	Spark	Spark	Spark	RT	RT	RT	RT	NoSQL	NoSQL
1	0	1	1	0.8	1	1	1	1	1	1	1	1	0.6	1	1.00	1.00	1	0.25	0.75
0.75	1	1	1	0.6	1	1	1	0.75	0.5	1	0.5	1	0.6	1	1.00	1.00	1	1	1
1	0	1	1	1	1	0	1	0.5	0.5	0	0	1	1	0	0.67	1.00	0	0.5	0.75
1	0	1	1	1	1	0	1	0.5	0.5	0	0	1	1	0	0.67	1.00	0	0.5	0.75
0.5	1	1	1	1	1	1	1	0.75	0.25	1	0.75	1	0.8	1	1.00	1.00	0	0.75	1
1	1	1	0	0.2	1	0	1	0.75	1	0	1	0	1	1	1.00	0.50	0	0.75	0.75
0.5	1	0	0	0.8	0	0	0	0.5	0.75	0	0.75	1	0.6	0	0.67	0.50	0	0.5	0.75
0.5	1	0	1	0.8	1	1	0	0.75	0.75	1	0.5	1	0.8	1	1.00	0.50	1	0.5	1
0.75	1	1	1	1	1	0	0	1	0.5	0	0.5	1	0.8	1	0.33	0.83	0	0.75	0.5
1	1	1	1	0.8	1	1	1	0.75	1	0	1	1	0.6	1	1.00	1.00	0	0.75	1
1	0	1	1	1	1	0	0	1	0.5	0	1	1	0.6	1	1.00	0.50	1	0.75	1
0.75	0	1	1	1	1	1	1	0.75	1	0	0.5	0	0.8	1	1.00	1.00	1	0.5	1
0.75	1	1	1	0.6	1	1	1	0.75	0	0	0.5	1	0	0	0.33	0.00	1	1	1
1	0	1	1	0.8	1	0	1	0.75	0.25	0	0.5	1	1	1	0.67	1.00	1	1	1
0.5	1	1	1	0.8	1	0	0	1	1	1	0.5	1	0.8	0	1.00	0.67	0	0.75	0.25

BIGDATA TEAM

План на сегодня

- 19:15-20:00 Введение в embeddings, Радослав Нейчев
- 20:00-20:55 Практика по embeddings, Максим Рябинин
- -- перерыв (10 мин)
- 21:05-22:00 НММ, Наталья Корепанова
- 🔥 обратная связь (2 мин)

Thank you! Any questions?

Feedback:
http://rebrand.ly/mfnlp2020q4_feedback_01_embhmm

Dral Alexey, study@bigdatateam.org
CEO at BigData Team, http://bigdatateam.org
https://www.linkedin.com/in/alexey-dral
https://www.facebook.com/bigdatateam