

Wolfram Summer School 2019: статистическое моделирование Wolfram Language

Константин Корилов
Санкт-Петербургский политехнический университет



Присоединяйтесь к разговору #WolframRussiaConf

Содержание

1. Что такое Wolfram Summer School?
2. Задача проекта
3. Решение: сбор данных
4. Решение: ML-модели
5. Результаты
6. Перспективы

Что такое Wolfram Summer School?

In[]:=



Задача проекта

Проблема

Насколько семантически схожи встроенные функции Wolfram Language? Построить карту, где точки — функции, а расстояние между ними отражает семантическое расстояние между функциями.

Out[]:=

Series
Integrate

ListPlot
Plot

Задача проекта

Как решать?

Дистрибутивная гипотеза в лингвистике

Out[]:=

Слова с похожим смыслом встречаются в схожем контексте.

Задача: построить языковую модель

Найти распределенное представление встроенных функций языка (embeddings): $\{f_i\} \rightarrow \mathbb{R}^d$.

Z.Chen and M.Monperrus, "A Literature Study of Embeddings on Source Code," ArXiv190403061 Cs Stat, Apr.2019.

Решение: сбор данных

Источник №1: справочная система Wolfram Mathematica

Всего элементов в Wolfram Language:

In[]:=

```
WolframLanguageData[]//Length
```

Out[]:=

5436

Доступ к примерам из документации:

In[]:=

```
EntityValue[WolframLanguageData["Integrate"],"DocumentationExampleInputs"]//First
```

Out[]:=

```
BasicExamples -> {{Integrate[x^2 + Sin[x], x]}, {Integrate[1/(x^3 + 1), {x, 0, 1}], Plot[1/(x^3 + 1), {x, 0, 1}, Filling -> Axis]},  
{Sqrt[x + Sqrt[x]] dx, TraditionalForm[%]}, {Integrate[Log[x] Exp[-x^2] dx]}
```

Далее сделать примеры запускаемыми:

- Извлечь Expressions из Cells.

Решение: сбор данных

Источник №2: системные файлы Wolfram Mathematica

Всего системных файлов:

In[]:=

```
FileNames[{"*.wl","*.m"},$InstallationDirectory,Infinity]//Length
```

Out[]:=

5561

Далее сделать примеры запускаемыми:

- Отфильтровать файлы только с данными.
- Удалить строки из файлов с Package из-за проблемы с функцией ToExpression.

Решение: сбор данных

Источник №3: GitHub

In[]:=

The screenshot shows the GitHub search interface with the query 'language:Mathematica'. The search results are sorted by 'Best match' and show 4,362 available repository results. The left sidebar lists various filters: Repositories (4K+), Code (0), Commits (0), Issues (19K), Packages (1K), Marketplace (839), Topics (542K), Wikis (0), and Users (3K). The main results list includes:

- InterImm/exoplanets** (系外行星) - Mathematica, 13 stars, updated on 3 Mar 2018.
- hpfem/hermes-examples** - Mathematica, 11 stars, updated on 5 Apr 2017.

- Таймаут для репозиторий (скачали ~57Gb).
- Удалить файлы, которые не связаны с Wolfram Language (~317 Mb).
- Использовать логику обработки системных файлов Wolfram Mathematica.

Решение: сбор данных

Основные проблемы:

- Надо использовать HoldComplete везде.
- Сильная связь языка с фронтендом.

464 991 Wolfram Language Expressions

In[]:=

code	source	index
HoldComplete[TypeGlobalQ[SchemaBase64Binary] ^_ True;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	306
HoldComplete[TypeNamespace[SchemaBase64Binary] ^_ "http://www.w3.org/2001/XMLSchema";]	/SystemFiles/Links/XMLSchema/XMLSchema.m	307
HoldComplete[TypeLocalName[SchemaBase64Binary] ^_ "base64Binary";]	/SystemFiles/Links/XMLSchema/XMLSchema.m	308
HoldComplete[TypeQ[SchemaBase64Binary] ^_ True;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	309
HoldComplete[Protect[SchemaBase64Binary];]	/SystemFiles/Links/XMLSchema/XMLSchema.m	310
HoldComplete[TypeSymbol["http://www.w3.org/2001/XMLSchema", "dateTime"] = SchemaDateTime;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	311
HoldComplete[TypeGlobalQ[SchemaDateTime] ^_ True;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	312
HoldComplete[TypeNamespace[SchemaDateTime] ^_ "http://www.w3.org/2001/XMLSchema";]	/SystemFiles/Links/XMLSchema/XMLSchema.m	313
HoldComplete[TypeLocalName[SchemaDateTime] ^_ "dateTime";]	/SystemFiles/Links/XMLSchema/XMLSchema.m	314
HoldComplete[TypeQ[SchemaDateTime] ^_ True;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	315
HoldComplete[Protect[SchemaDateTime];]	/SystemFiles/Links/XMLSchema/XMLSchema.m	316
HoldComplete[TypeSymbol["http://www.w3.org/2001/XMLSchema", "date"] = SchemaDate;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	317
HoldComplete[TypeGlobalQ[SchemaDate] ^_ True;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	318
HoldComplete[TypeNamespace[SchemaDate] ^_ "http://www.w3.org/2001/XMLSchema";]	/SystemFiles/Links/XMLSchema/XMLSchema.m	319
HoldComplete[TypeLocalName[SchemaDate] ^_ "date";]	/SystemFiles/Links/XMLSchema/XMLSchema.m	320
HoldComplete[TypeQ[SchemaDate] ^_ True;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	321
HoldComplete[Protect[SchemaDate];]	/SystemFiles/Links/XMLSchema/XMLSchema.m	322
HoldComplete[TypeSymbol["http://www.w3.org/2001/XMLSchema", "time"] = SchemaTime;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	323
HoldComplete[TypeGlobalQ[SchemaTime] ^_ True;]	/SystemFiles/Links/XMLSchema/XMLSchema.m	324
HoldComplete[TypeNamespace[SchemaTime] ^_ "http://www.w3.org/2001/XMLSchema";]	/SystemFiles/Links/XMLSchema/XMLSchema.m	325

K < showing 464672-464691 of 464699 > X

Решение: сбор датасета

Очистка данных

- Заменить на специальные токены вместо литералов:
 - \$DATA\$, \$INT\$, \$REAL\$, \$STR\$, \$VAR\$
- Удалить глубокие выражения. Отсечка по размеру выражений (< 30 000).

In[]:=

$$a^3 + b \cdot c^3 // \text{Unevaluated} // \text{ByteCount}$$

Out[]:=

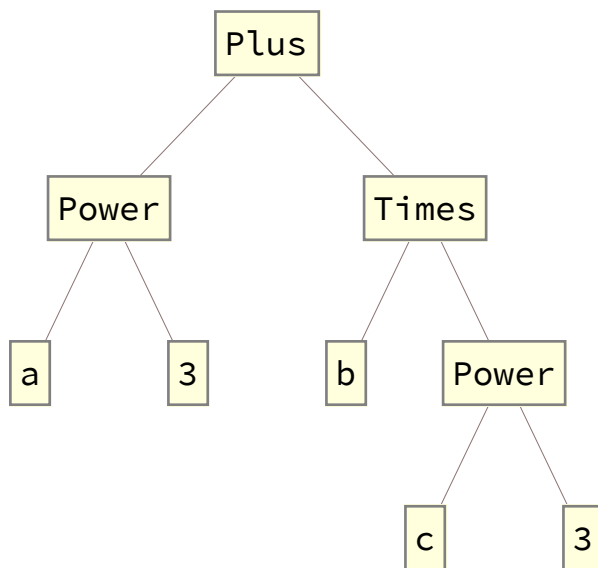
256

- Заменить дерево токенов на последовательность.

In[]:=

a³+b c³//TreeForm

Out[]:=TreeForm=

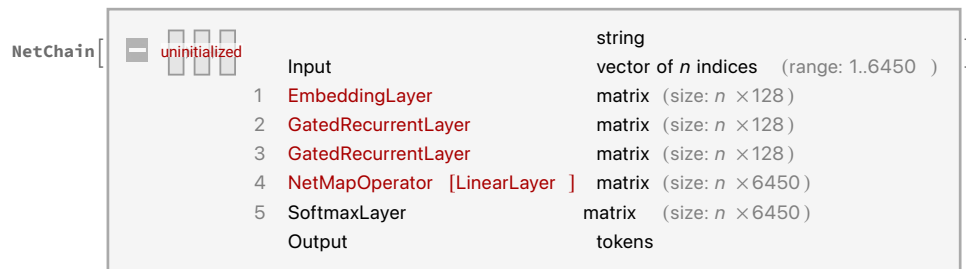


Решение: ML-модели

RNN (=рекуррентная нейронная сеть)

■ Архитектура:

In[]:=



Решение: ML-модели

t-SNE (=t-distributed Stochastic Neighbor Embedding)

Высокоразмерная модель

$$p_{j|i} = \frac{\exp\left[\frac{-L[x_i - x_j]^2}{2\sigma_i^2}\right]}{\sum_{k \neq i} \exp\left[\frac{-L[x_i - x_k]^2}{2\sigma_i^2}\right]}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Низкоразмерная модель

$$q_{ij} = \frac{(1 + L[y_i - y_j])^{-1}}{\sum_{k \neq l} (1 + L[y_k - y_l])^{-1}}$$

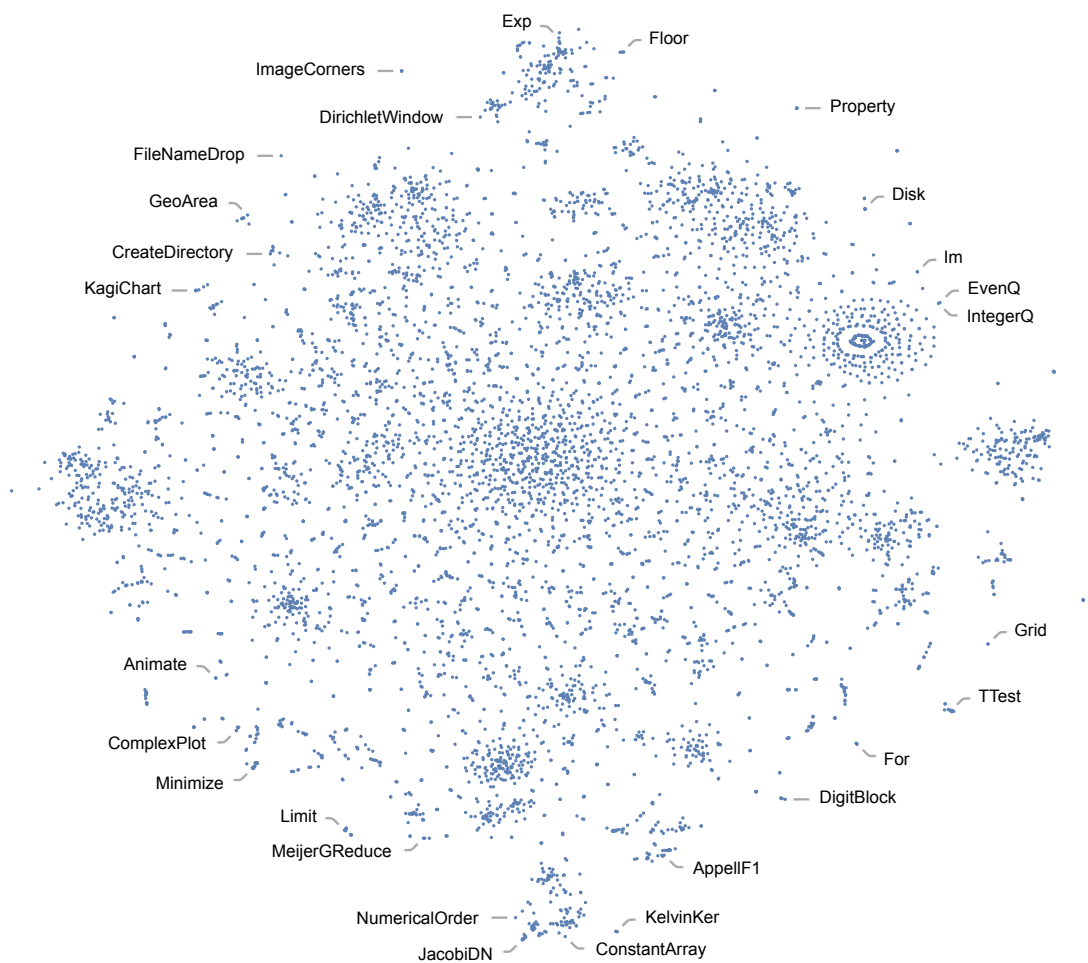
Минимизация

$$KL = \sum_{ij} p_{ij} \log \left[\frac{p_{ij}}{q_{ij}} \right]$$

Результаты

Карта функций Wolfram Language

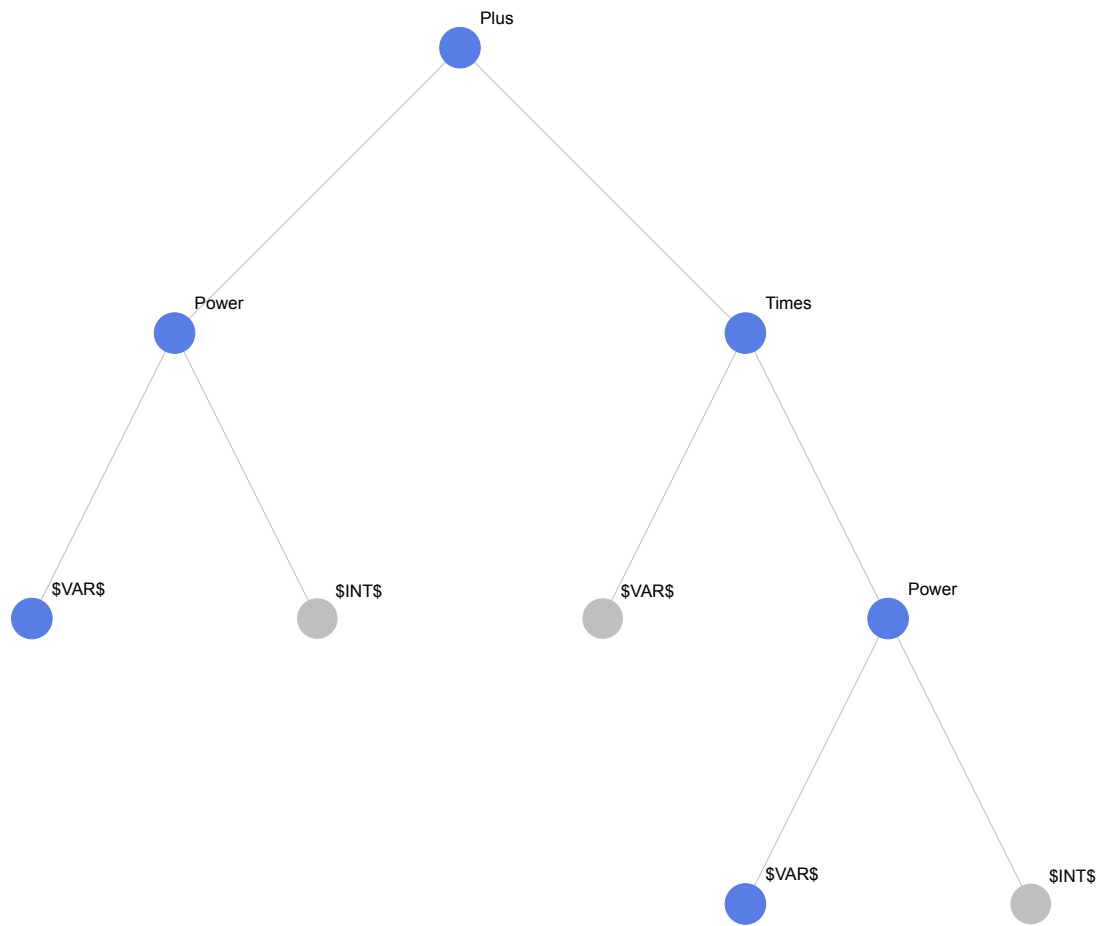
In[]:=



Перспективы

code2vec

U. Alon, M. Zilberstein, O. Levy, and E. Yahav, "code2vec: learning distributed representations of code," *Proc. ACM Program. Lang.*, vol. 3, no. POPL, pp. 1–29, Jan. 2019.



Перспективы

- Опубликовать данные в The Wolfram Data Repository.
- Исследовать модели из современных статей.
- Потенциальные приложения:
 - — умное автодополнение для Wolfram Language,
 - — умный рефакторинг: генераторы имен переменных и функций,
 - — проверка на плагиат,
 - — анализ на ошибки.

Спасибо за внимание!

Out[]=



Out[]=

<https://github.com/ckorikov/WSS-19>