HackCulture 2016

Emily Chen
Daniel George
Hui Lyu
Mark McCarthy

TALK2DATA

October 22, 2016

Overview

To bridge the gap between humanities and computer science, we introduce a new program called **TALK2DATA**, which allows anyone with **zero programming experience** to immediately start doing **data science**.

To use this program, all that is required is a web browser (*or a **Raspberry Pi***). Seriously, you don’t even need a keyboard! Just have a conversation with your computer (***in any language***) and have it create code to do all sorts of complex tasks for you. Even better, you can see or even **hear** both the code and the output, and **learn programming**!

This is the first system, ever built, that enables full-fledged voice-based programming using natural language in over 100 languages. Made possible by the recent breakthroughs in **machine learning** technology, particularly deep learning based on **neural networks**.

Powered by **Google** APIs, **Wolfram|Alpha**, and the **Wolfram Language** with numerous functions that use machine learning to smartly import any data format, find patterns in the data, and more!

We have preloaded our program with dozens of semantically structured datasets about **Champaign-Urbana** in order to encourage everyone to explore and gain more insight into issues that are relevant to our community. **Crowdsourcing** data analysis in this manner will raise awareness and lead to rapid innovations. New datasets can be freely added to our repository.

Features

1. **Makes everybody a data scientist:** If you know how to talk or write, then you're now a programmer and a data scientist! Freely available online for everyone.

2. **Teaches programming to kids and adults:** Code can be created with intuitive natural language input independent of any programming languages or paradigms. This encourages computational thinking rather than learning arbitrary syntax rules.

3. **Enhances accessibility**: Helps people with physical disabilities develop code and analyze data more easily. Audio based input and output, 103 supported languages.

4. **Preloaded with Champaign-Urbana community data:** Contains dozens of datasets related to Champaign and Urbana hosted on our web server. Anyone can start exploring and analyzing this data immediately.

5. **Smart import of any data format:** Our smart import, powered by the Wolfram Language, can automatically upload most types of datasets while recognizing their formatting and headers, using machine learning. This also preserves physical units within the datasets and automatically converts and combines units during later operations.

6. **Error correcting natural language processing:** You can type as sloppily as you want, and our text recognition and spell checking algorithms will correct your spelling for you and figure out what you intended to do.

7. **More functionality added everyday:** Thousands of different operations already recognized. We are working hard on adding more and more everyday.

8. **High performance cloud computing:** Analyze huge data sets and perform computationally intensive calculations through a browser on any device including Chromebooks, Raspberry Pis or smartphones (also works on the Wolfram Cloud app).

Specifications

The voice input is recognized with one of the most accurate voice recognition system that exists today: the Google Cloud Speech-to-Text engine, which is powered by deep neural networks and is constantly improving every minute. Next, the Google Translate API is used to convert the input string (supported for over 103 languages) into English.

Then our custom text-to-code parser converts the input sentences into Wolfram Language code. This code is executed partly on our servers and partly on the Wolfram Cloud. Unrecognized requests are outsourced to Wolfram|Alpha. Finally the results are displayed on the website.

Possible Questions to Ask for Example Datasets:

- 2014_Fire_Incident_Counts_by_Type :
  - Q1: What's the total number of incidents of all types in 2014?
- Fire___Rescue_Incidents_by_Day_and_Type:
  - Q1: During what time periods (based on alarm time) have most Fire incidents?
  - Q2: What's the proportion of False Alarm or False Call of all the incidents?
- Traffic_Accidents_since_2011__counted_by_Street_and_Block:
  - Q1: Most common street or block of most incidents since 2011?
- Bike_Crashes_from_2005_to_2014:
  - Q1: Plot locations (coordinates) on map of crashes with A-injury severity in 2013?
  - Q2: Most common traffic control for C-injury crashes in Champaign in 2014?
  - Q3: How many injury and fatality people in Urbana in 2010?

  - Q4: Does weather condition or road surface or traffic control has correlation with crash severity?
- Pedestrian_Crashes_from_2005_to_2014:
  - Q1: What's the difference of number of injury and crash severity between bike and pedestrian crashes? (conjunction of two datasets)
- City_of_Champaign_Most_Common_Violation_Tickets_1999-2011:
  - Q1: Histogram of the total number of violation tickets from 1999 to 2011? What trend?
  - Q2: Line / Scatter plot of underage alcohol offenses from 1999 to 2011? What trend?
  - Q3: Pie chart of types of violations in 2011?
  - Q4: Is there correlation / linear fit between Alcohol Possession on Public Property and Fighting?
  - Q5: Word cloud of violation tickets?
  - Q6: What are the top 5 violation tickets in Champaign during that time period? What's the answer from Wolfram Alpha (most common violation types in Champaign)? Is there overlapping?

Next steps

1. Develop mobile apps for Android/iOS and chatbots

Data science on the go, without having to type on tiny keyboards. All computations and data sets will be hosted for free on the cloud. We also plan to integrate handwriting recognition into the app. A prototype version with limited functionality is already available on the Wolfram Cloud app for iOS and Android.

2. Enable general-purpose programming (TALK2CODE)

The framework we built can be extended to allow general-purpose programming in any programming language as well as symbolic/numerical mathematical computation, using your voice. We are working on using machine learning techniques to automatically teach our system how to learn new commands.

Keywords

- Natural Language Processing, Sentence Recognition
- Neural Networks, Deep Learning, Machine Learning
- Smart Data Import, Data Science, Data Analysis, Wolfram Data Repository
- Wolfram|Alpha, Wolfram Language, Wolfram Cloud, Mathematica
- Google Cloud, Speech-to-Text, Speech Recognition, Voice-based Programming
- Accessibility, Computational Thinking, Code Literacy
- Champaign, Urbana, University of Illinois, HackCulture, Hackathon
- TALK2DATA, TALK2CODE
- Cloud Computing, Mobile Programming, Mobile Data Science