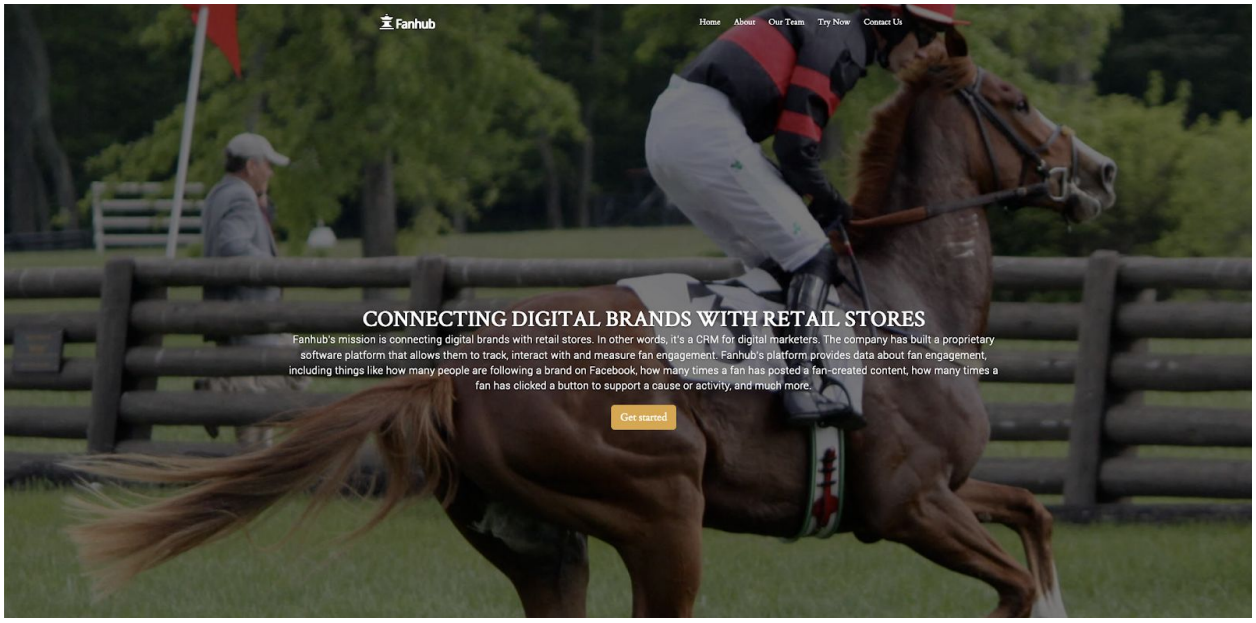
Machine Learning for the Arts
UCSD FALL 2019
**FINAL PROJECT**

# Recuria: The OpenAI Random Startup Generator




Winson Luk

## DESCRIPTION

Every startup claims to be disrupting an industry or changing the world. Most startup ideas are destined to fail, but some truly change the world. By training on thousands of startup taglines, articles, and interviews, this project aims to generate a lot of bad startup ideas, and a few good ones.

**Concept**:
Most startup ideas can be summarized in just one paragraph. The tagline describes the overarching concept (e.g., "Uber is finding you better ways to move, work, and succeed"), and the next few sentences can provide a more detailed description of the product, as well as context on the startup's history, people, and industry.

The tagline can be created by finetuning GPT-2 with a dataset of startup taglines (https://github.com/winsonluk/gpt_pitches), and the subsequent sentences can be generated by feeding this tagline as a prefix into a other GPT-2 models finetuned with startup descriptions and company analyses (https://github.com/winsonluk/gpt_descriptions and https://github.com/winsonluk/gpt_summaries).

The ideas generated have been fairly realistic (most are bad, some are good), and I incorporated them into faux startup website similar to https://tiffzhang.com/startup, with a few thousand permutations of ideas. The value of these ideas depend solely on the reader's interpretation (see reader-response theory (https://en.wikipedia.org/wiki/Reader-response_criticism)), but hopefully some of these ideas are cohesive enough to serve as inspiration.

**Technique**:

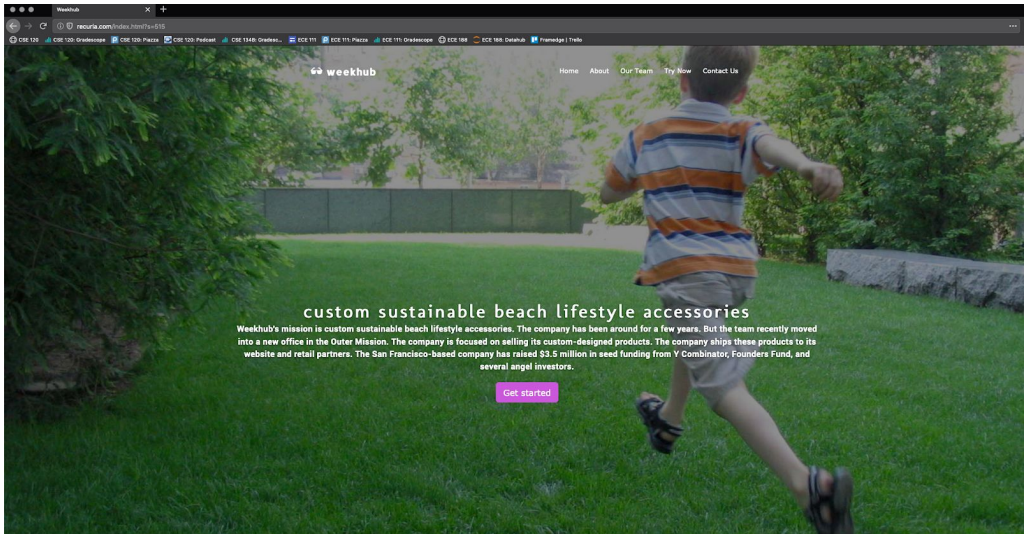
- I created three models finetuned with the gpt-2-simple library.
- The first model is trained on startup taglines from Startups List (https://www.startups-list.com/). I used Rick Hennessy's scraped dataset (https://data.world/rickyhennessy/startup-names-and-descriptions). Download: https://winsonluk.com/assets/ideas.zip
- The second model is from the previous dataset, but with startup descriptions rather than taglines. Download: https://winsonluk.com/assets/gpt_summaries.zip
- The third model is trained on TechCrunch posts, which focus on the latest developments in technology. The scraped data is from Kaggle (https://www.kaggle.com/thibalbo/techcrunch-posts-compilation). Download: https://winsonluk.com/assets/gpt.zip
- The website is hosted on AWS, with the generated pitches stored in a DynamoDB database and available as a serverless Lambda application.

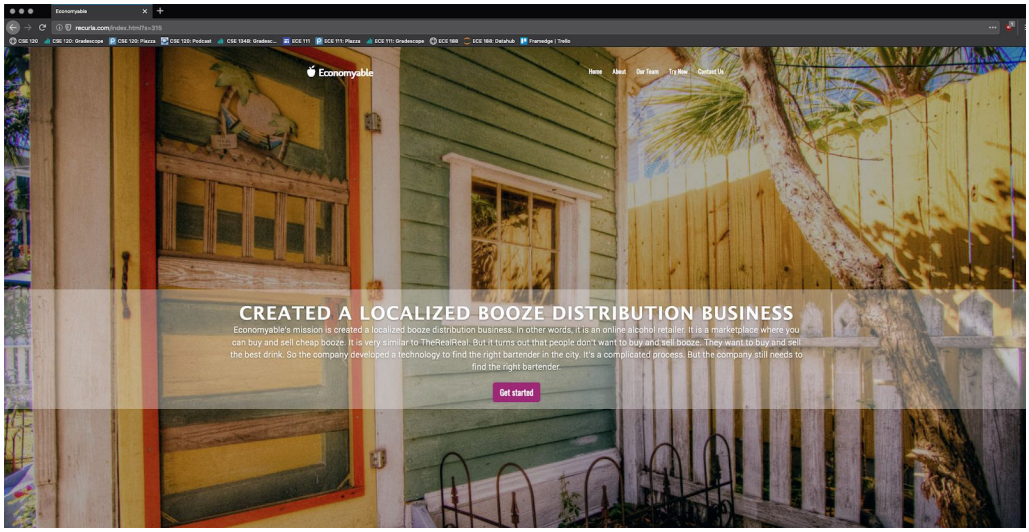
**Process**:

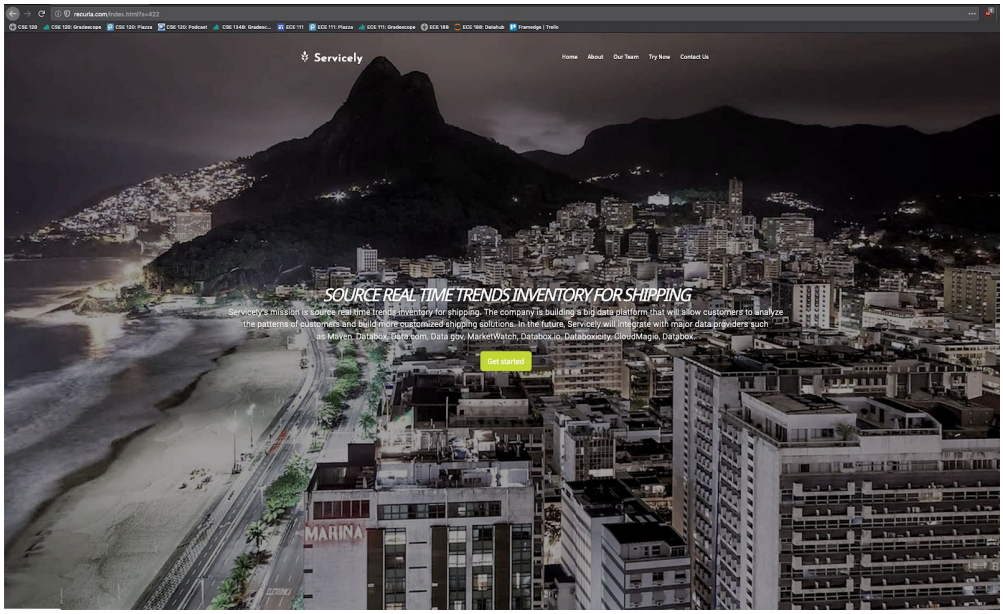
- The GPT-2 774M model doesn't fit into the memory of any single GPU, so the multi-gpu fork of gpt-2-simple (https://github.com/huntrontrakkr/gpt-2-simple) needs to be installed to train with the 774M model. This fork was merged to master in v0.7 (I actually helped fix a bug in the pull request), so now a simple pip install gpt-2-simple v0.7+ should be sufficient.
- I used 4 x Tesla V100 GPUs and 16 GB of RAM on Vast.ai (https://vast.ai) to train the models. Training will fail with single GPUs or less than 16 GB of RAM. After training, generation can be performed with a single GPU, though 16 GB of RAM is still necessary. I used Vast.ai for training and generation because the datasets and outputs took too long to train on Datahub.
- The startup tagline and description models are finetuned to a loss of around 0.1, while the larger TechCrunch model is finetuned to a loss of 1.8.
- I sampled all models with temperature ranges from 0.2 to 2.0 and top-p from 0.1 to 1.0 (higher values translate to more "creativity" in the text) to find the optimal parameters for realistic text generation.
- I sequentially generated 6000+ startup taglines, descriptions, and discussions. On AWS, I set up a DynamoDB instance loaded with these descriptions. I hooked it up to an Amazon API Gateway service with Lambda to serve the descriptions as a serverless application available through REST. I modified the startup generator website to send AJAX calls to this backend every time the page is reloaded, fetching a random entry from the database.

**Result**:

- Website: http://recuria.com
- Taglines only: https://github.com/winsonluk/gpt_pitches/blob/master/io/pitches.txt
- Taglines with descriptions: https://github.com/winsonluk/gpt_summaries/blob/master/io/summaries.txt
- Taglines with TechCrunch commentary: https://github.com/winsonluk/gpt_descriptions/blob/master/io/descriptions.txt
- Taglines with descriptions *and* TechCrunch commentary: https://github.com/winsonluk/gpt_descriptions/blob/master/io/descriptions_with_summaries.txt


weekhub
Home
About
Our Team
Try Now
Contact Us
custom sustainable beach lifestyle accessories
Weekhub's mission is custom sustainable beach lifestyle accessories. The company has been around for a few years. But the team recently moved into a new office in the Outer Mission. The company is focused on selling its custom-designed products. The company ships these products to its website and retail partners. The San Francisco-based company has raised $3.5 million in seed funding from Y Combinator, Founders Fund, and several angel investors.
Get started


Economyable
Home
About
Our Team
Try Now
Contact Us
CREATED A LOCALIZED BOOZE DISTRIBUTION BUSINESS
Economyable's mission is created a localized booze distribution business. In other words, it is an online alcohol retailer. It is a marketplace where you can buy and sell cheap booze. It is very similar to TheRealReal. But it turns out that people don't want to buy and sell booze. They want to buy and sell the best drink. So the company developed a technology to find the right bartender in the city. It's a complicated process. But the company still needs to find the right bartender.
Get started


Servicely
Home
About
Our Team
Try Now
Contact Us
SOURCE REAL TIME TRENDS INVENTORY FOR SHIPPING
Servicely's mission is source real time trends inventory for shipping. The company is building a big data platform that will allow customers to analyze the patterns of customers and build more customized shipping solutions. In the future, Servicely will integrate with major data providers such as Maven, Databox, Data.com, Data.gov, MarketWatch, Databox.io, Databoxionly, CloudMagic, Databox.
Get started
MARINA

**Reflection**:

This project provoked me to reconsider the relationship between humans and machines. While my original goal was to produce a website that could aid entrepreneurs in their creative pursuits, the plausibility of the results beg the question of whether a website can replace entrepreneurs in their capacity as creative thinkers. I shared this website with actual entrepreneurs on Reddit, and the reactions ranged from impressed to cynically dismissive. Many people identified the potential for ML to creatively discover and showcase uncontested business opportunities. One commented, “I’d bet 300 ideas generated from this little tool produce 30 decent ones and 3 great ones” (/u/biggerobrothero). However, others maintained that since purpose precedes action, startups catered to humans will always need to be created by humans. “A person with a good idea is [...] in tune with problems actual people face in the real world and ways we might conquer that” (/u/ianperera).

I agree that technology will never obsolesce human creativity. But this project convinced me that creativity is no longer a human monopoly. When one person creates something for another - whether a startup or a song or a piece of artwork - creativity arises from the desire to please an audience in a unique way. This course has shown that ML art can please, and even dazzle, audiences through unexpected mediums and methods. And if we consider desire as not just an urge by “actual people [...] in the real world,” but simply as the pursuit of an objective, ML models undoubtedly expresses a desire to please its audience (us). So by my definition, when ML aims to produce an appealing output and does so successfully, creativity is forged. But unfortunately, this isn’t enough creativity to start a company on its own.

**REFERENCE**:


- Papers
  - https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Repositories
  - https://github.com/huntrontrakkr/gpt-2-simple
  - https://github.com/tiffz/startup
- Blog posts
  - https://minimaxir.com/2019/09/howto-gpt2/
  - https://towardsdatascience.com/how-to-sample-from-language-models-682bceb97277


**CODE:**

- github.com/ucsd-ml-arts/ml-art-final-wluk
- https://github.com/winsonluk/startup
- https://github.com/winsonluk/gpt_pitches
- https://github.com/winsonluk/gpt_descriptions
- https://github.com/winsonluk/gpt_summaries

**RESULT:** recuria.com