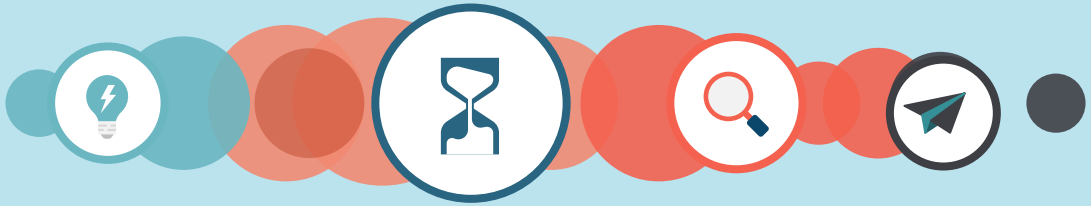
Presentation-4
October, 5



Novel Recipe Generation

Indraprastha Institute of Information Technology

Submitted By:

Adarsh Singh Kushwah

Niharika

Parul Sikri

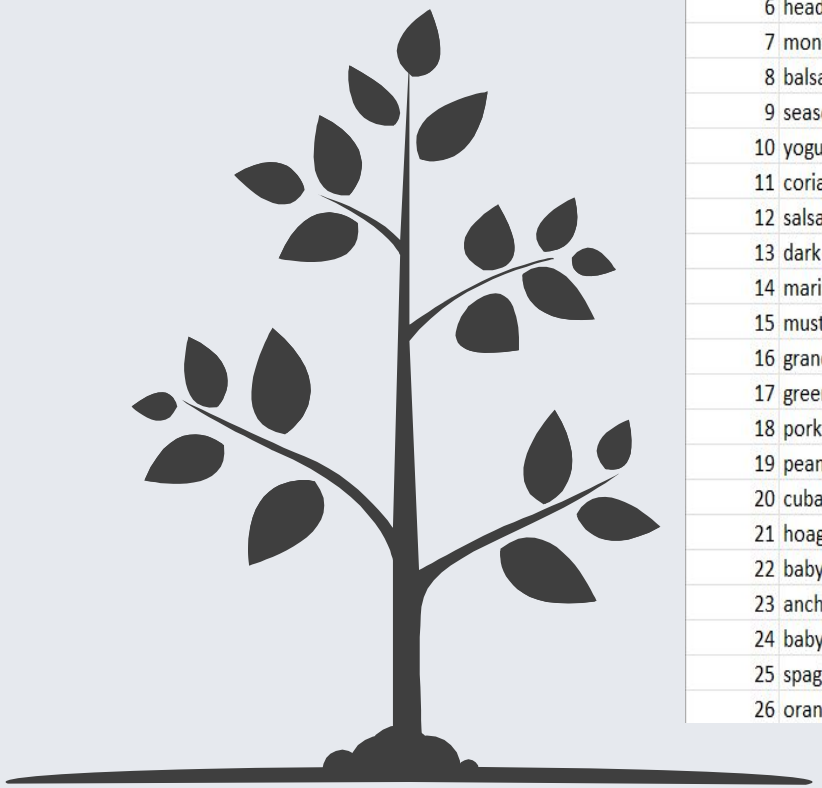
Shrey Rastogi

Table of Contents

TASKS GIVEN

- Find out the count of recipes which have their sizes equal to 1
- Find out the time taken to generate 100 recipes.
- Generate 500 recipes and send them to the Turing test team.
- Study about the GPT2 model.

1. 30 recipes with size equal to 1.
2. Time taken to generate 100 recipes = 58-60 minutes



S.No	Ingredients	Novel Recipe
1	yellow pea,apple,lemon juice zest,brown rice,ancho chile,active yeast;	- 1 1/2 cups flour, sifted
2	white bean,green chili;	- 2 1/2 teaspoons baking powder
3	cocoa,orange,feta cheese,sherry wine,bacon,white rice,lard;	- 1 1/4 cups sugar
4	chickpea flour,squash,parsley sprig,green chili,vegetable,baking potato,bu	- 1 1/4 cups milk, lukewarm
5	celery salt,self rising flour;	- 1/4 cup canola oil
6	head green cabbage,lime peel,lasagna noodle,coffee granule,raspberry,ha	- 6 large eggs
7	monterey jack pepper cheese,onion soup mix,cherry tomato;	- 2/3 cup milk
8	balsamic vinegar,thyme,lovage;	- 1 cup almonds, sliced or 2/3 cup almonds
9	seasoning salt,orange flower water,pork chop,candy,cabbage,nutmeg,pin	- 1 teaspoon almond extract
10	yogurt,graham cracker crumb,chili pepper flake,taco seasoning mix,italiar	- 2/3 cup sugar
11	coriander,celery,red lentil,self rising flour,baking powder,kaffir lime leaf,c	- 1/2 cup almonds, whole
12	salsa,tomato puree,chicken liver,saffron;	- 1 large egg
13	dark chocolate,liquid smoke,star anise,cheese curd;	- 2 teaspoons baking powder
14	marinara sauce,filet beef,vegetable oil cooking spray,cake flour,malt vine	- 1/4 - 1/2 cup sesame seeds, toasted
15	mustard seed,pork loin,ginger juice;	## Cooking instructions ##
16	grand marnier,apricot jam,chutney,egg white,lettuce;	- Preheat oven to 375 degrees
17	green lentil,chicken wing,pancetta,cayenne,flour,baking soda;	- Place almonds on a baking sheet and lightly toast on each side, about 2 minutes per side
18	pork tenderloin,ladyfinger,kale,lemonade,shrimp paste,red pepper;	- Put almonds onto large cookie sheet with sides
19	peanut,tamarind juice,egg white;	- Bake for 7 8 min, cool
20	cuban bread,sherry wine,hard egg,parsley flake,red pepper flake,dill pickle	- Stir sugar and almond extract in large bowl
21	hoagie,chile powder,salmon,pesto sauce,peppercorn;	- Bring milk, milk, almonds and extract to boil and pour over to flour mixture in a slow stream, whisking constantly
22	baby artichoke,beef roast,oyster,purple onion,leg lamb,celery root;	- When smooth, remove from heat and add almonds
23	anchovy paste,tomato green chile pepper,chicken drumstick,zucchini,kiell	- Cool 10 minutes
24	baby spinach,chicken bouillon cube,mushroom,currant,oregano leaf,cilan	- Whisk egg in a large bowl with baking powder until thick and lemon colored, about 5 minutes
25	spaghetti,flour tortilla,wonton skin,tomato paste,head cabbage,red chili p	- Mix in sugar and almond mixture
26	orange,black olive,chicken stock,raisin,beef rib,pork skin;	- Add

Generative Pre-trained Transformer (GPT)-2

1. GPT2 is basically a machine learning language model that is able to look at part of a sentence and predict the next word (single token).

2. The GPT-2 was trained on a massive 40GB dataset called WebText that the OpenAI researchers crawled from the internet as part of the research effort.

3. The smallest variant of the trained GPT-2, takes up 500MBs of storage to store all of its parameters. The largest GPT-2 variant is 13 times the size so it could take up more than 6.5 GBs of storage space.

4. GPT2 is built using transformer decoder blocks.

5. GPT2 outputs one token at a time.

6. GPT2 is auto-regressive in nature: each token in the sentence has the context of the previous words. After each token is produced, that token is added to the sequence of inputs. And that new sequence becomes the input to the model in its next step. This is an idea called "auto-regression".

1. GPT-2 has a parameter called top-k that we can use to have the model consider sampling words other than the top word (which is the case when top-k = 1).

2. Gpt 2 model is trained on over 1.5 billion parameters while gpt 3 model is trained on 175 billion parameters.

3. GPT-2 was known to have poor performance when given tasks in specialized areas such as music and storytelling. GPT-3 can now go further with tasks such as answering questions, writing essays, text summarization, language translation.

Overview of feeding in text and generating a single token:

1. **Tokenization** – Take some words, break them up into their common pieces. Take those common pieces and replace them with a number. Tokenization is necessary because computers only work with numbers. This also represents words efficiently.

→The cats played with the yarn.

→ The | cat | s | play | ed | with | the | y | arn |.|

→- -1- -|- -2- | 3 |- -4- | 5- |- - 6- - |- -1-| 7 |- 8- -|9|

2. **Embedding with time signal** — Take that string of numbers and convert each number to a vector. This captures the position of words relative to one another and allows words to take value from other words associated with them, e.g.

"The boy ran through the woods, and he surely had not stolen the cherry pie for which they were chasing him."

In this sentence "he" should clearly tie a lot of its meaning to "boy".

The same is true for " the" and "boy". The definite article carries a lot of meaning in the larger context.

3. Decoder Block —

The pieces are self-attention blocks, feedforward neural nets, and Normalization.
Self-attention blocks identify which words to focus on.

In the sentence, "Jimmy played with the burning bush, and then went around to the next bush," the words "Jimmy," "played," "burning," and "bush" capture a high proportion of the meaning in that sentence.

This idea that certain words and phases capture more meaning and thus should be given more "attention" is the intuition of self-attention blocks.

4. Linear Layer —

Prior to the tokenization process, a vocabulary size will be decided upon and a vocabulary will be setup.

The vocabulary is just a list of all the possible tokens (numbers) that can be produced and which letter or group of letters the tokens are equal to.

The linear layer takes the output of the last decoder block and converts it to a vector whose dimensions are vocabulary size by 1.
In short, it takes a lot of inputs and produces a list where each spot represents a token. The higher the number in the spot the better the chance that that token is the best pick.

5. Softmax —

Converts the output of the linear layer to a probability distribution. The output of the linear layer tells you information about which tokens are the best picks, but it is hard to use. The values range from very small values(huge negative values) to very large values and their meaning is in relation to all the other values.

To make them easier to use apply the Softmax function which converts the vector to a probability distribution. This means each number represents the probability that that token is the correct one.

6. Pick a token —

Choose the method to pick the next token from the probability distribution of tokens, and use that method to pick the token.

There are various methods to do so including, greedy, temperature sampling, nucleus sampling, and top-k sampling.

Convert Token to a word piece using the vocabulary.

THANK YOU