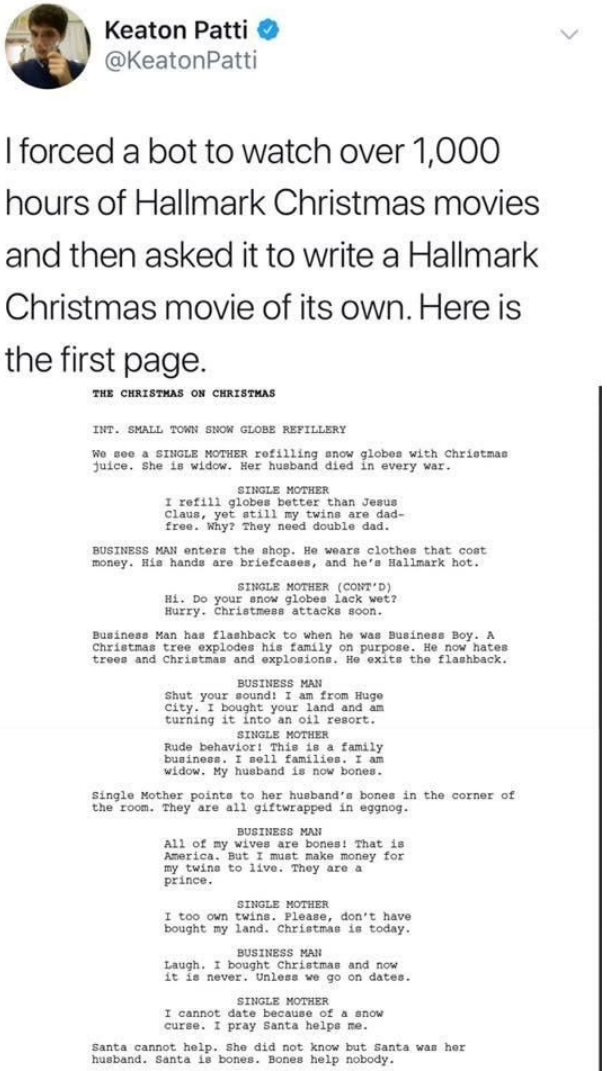
Natural Language Processing

Angel Burr, Neilly Herrera Tan, Zhan Shi

Overview

The aim of our project is to learn about natural language processing.

We want to explore this by creating a fake text generator that takes some input corpus, and generates a similar output to that text file.



Keaton Patti
@KeatonPatti

I forced a bot to watch over 1,000 hours of Hallmark Christmas movies and then asked it to write a Hallmark Christmas movie of its own. Here is the first page.

THE CHRISTMAS ON CHRISTMAS

INT. SMALL TOWN SNOW GLOBE REFILLERY

We see a SINGLE MOTHER refilling snow globes with Christmas juice. She is widow. Her husband died in every war.

SINGLE MOTHER
I refill globes better than Jesus Claus, yet still my twins are dad-free. Why? They need double dad.

BUSINESS MAN enters the shop. He wears clothes that cost money. His hands are briefcases, and he's Hallmark hot.

SINGLE MOTHER (CONT'D)
Hi. Do your snow globes lack wet? Hurry. Christmess attacks soon.

Business Man has flashback to when he was Business Boy. A Christmas tree explodes his family on purpose. He now hates trees and Christmas and explosions. He exits the flashback.

BUSINESS MAN
Shut your sound! I am from Huge City. I bought your land and am turning it into an oil resort.

SINGLE MOTHER
Rude behavior! This is a family business. I sell families. I am widow. My husband is now bones.

Single Mother points to her husband's bones in the corner of the room. They are all giftwrapped in eggnog.

BUSINESS MAN
All of my wives are bones! That is America. But I must make money for my twins to live. They are a prince.

SINGLE MOTHER
I too own twins. Please, don't have bought my land. Christmas is today.

BUSINESS MAN
Laugh. I bought Christmas and now it is never. Unless we go on dates.

SINGLE MOTHER
I cannot date because of a snow curse. I pray Santa helps me.

Santa cannot help. She did not know but Santa was her husband. Santa is bones. Bones help nobody.

Data sources:

Text files from Project Gutenberg

POEMS

Gerontion

Thou hast nor youth nor age
But as it were an after dinner sleep
Dreaming of both.

Here I am, an old man in a dry month,
Being read to by a boy, waiting for rain.
I was neither at the hot gates
Nor fought in the warm rain
Nor knee deep in the salt marsh, heaving a cutlass,
Bitten by flies, fought.
My house is a decayed house,
And the jew squats on the window sill, the owner,
Spawned in some estaminet of Antwerp,
Blistered in Brussels, patched and peeled in London.
The goat coughs at night in the field overhead;
Rocks, moss, stonecrop, iron, merds.
The woman keeps the kitchen, makes tea,
Sneezes at evening, poking the peevish gutter.

I an old man,
A dull head among windy spaces.

Signs are taken for wonders. "We would see a sign":
The word within a word, unable to speak a word,
Swaddled with darkness. In the juvescence of the year
Came Christ the tiger

In depraved May, dogwood and chestnut, flowering Judas,
To be eaten, to be divided, to be drunk
Among whispers; by Mr. Silvero
With caressing hands, at Limoges
Who walked all night in the next room;
By Hakagawa, bowing among the Titians;
By Madame de Tornquist, in the dark room
Shifting the candles; Fraulein von Kulp
Who turned in the hall, one hand on the door. Vacant shuttles
Weave the wind. I have no ghosts,
An old man in a draughty house
Under a windy knob.

Use cases:

1. For anyone who is interested in generating a wall of content
2. Generating poems in the style of an author: create a nonsensical or educational Twitter bot that posts these poems
3. Useful for type completion or chatbot scenarios as well

User requirements:

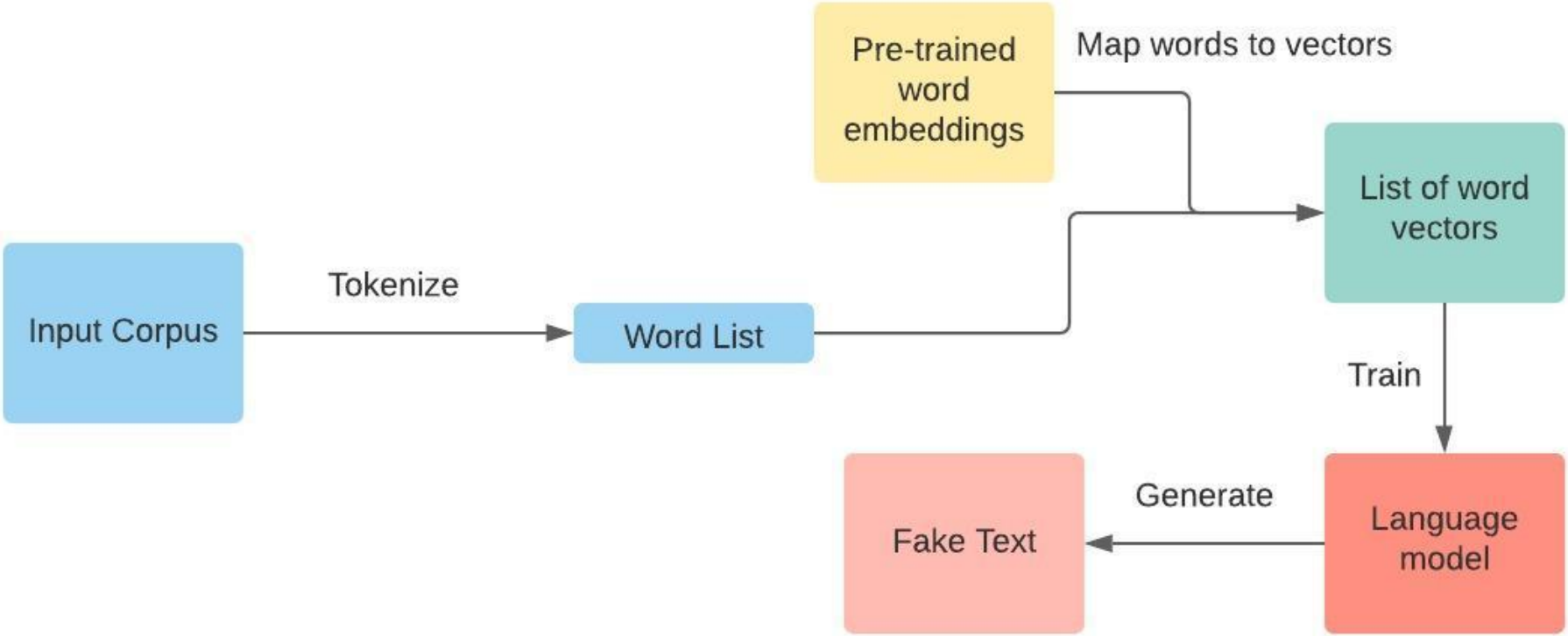
1. Have a basic knowledge of Python (would be able to run a script), and who want to learn more about NLP

Use scenario:

Sally is an avid poetry fanatic who loves Emily Dickinson's work. However, she's read all of her work already, and wants to read more.

Even though Emily Dickinson is long gone, Emily can generate new poems in her writing style that she can live vicariously through with this bot.

Structure


Pre-trained word embeddings
Map words to vectors
List of word vectors
Input Corpus
Tokenize
Word List
Train
Fake Text
Generate
Language model

Things learned:

1. How to use nltk to for word processing e.g. tokenizing a text file
2. How to build a complete neural net using MxNet.
3. How the vanilla RNN/LSTM + greedy decoder not work

```
Out[64]: 'he is the the the the the the the'
```

4. How annoying and inefficient it is to write tests after finishing the coding
5. How to resolve merge conflicts

Thank you!