

University of Texas at Arlington

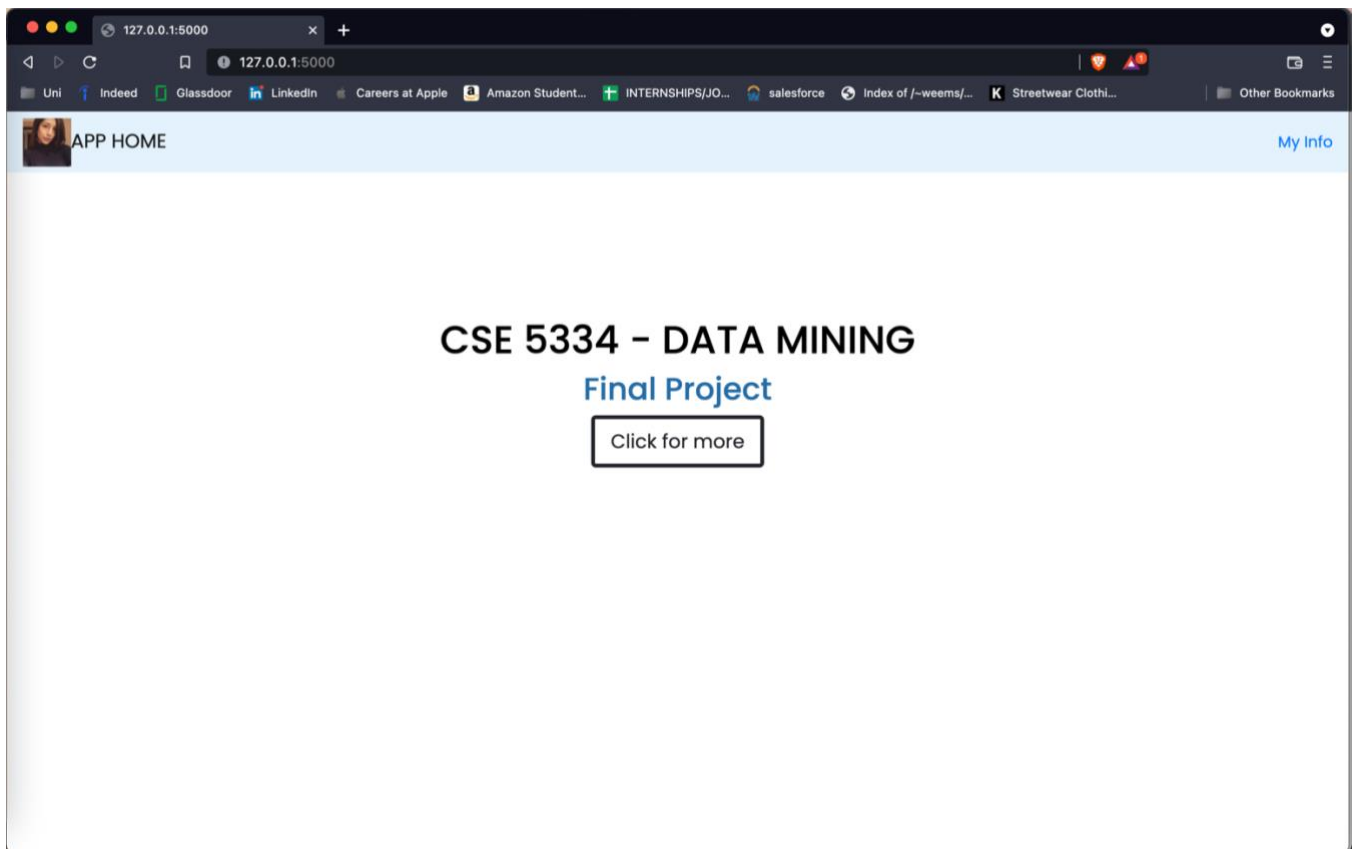
**CSE 5334 – Data Mining
Final Project Report**

Naïve Bayesian learning for classifying news text articles

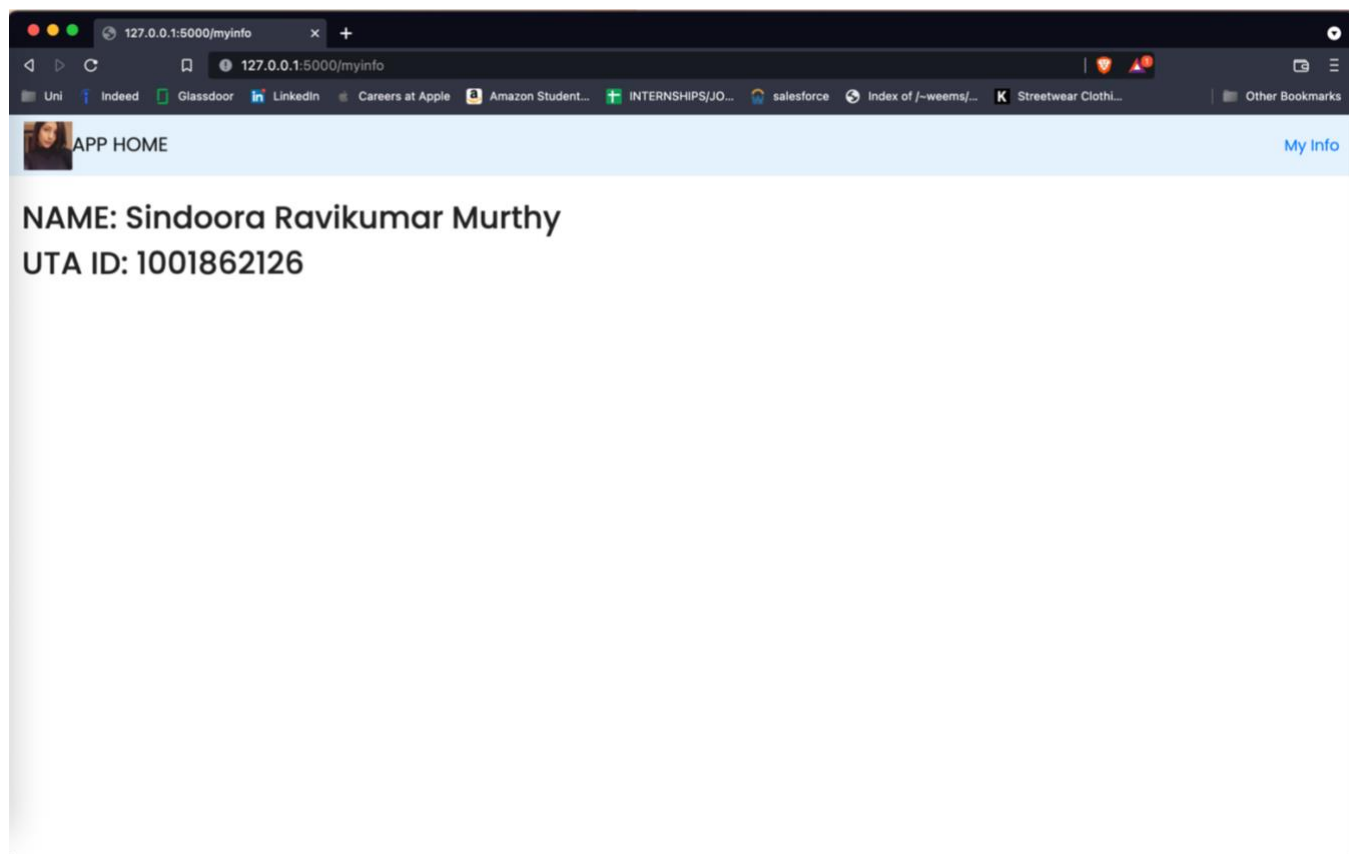
Sindoor Ravikumar Murthy
1001862126

SCREENSHOTS OF THE NBC APPLICATION

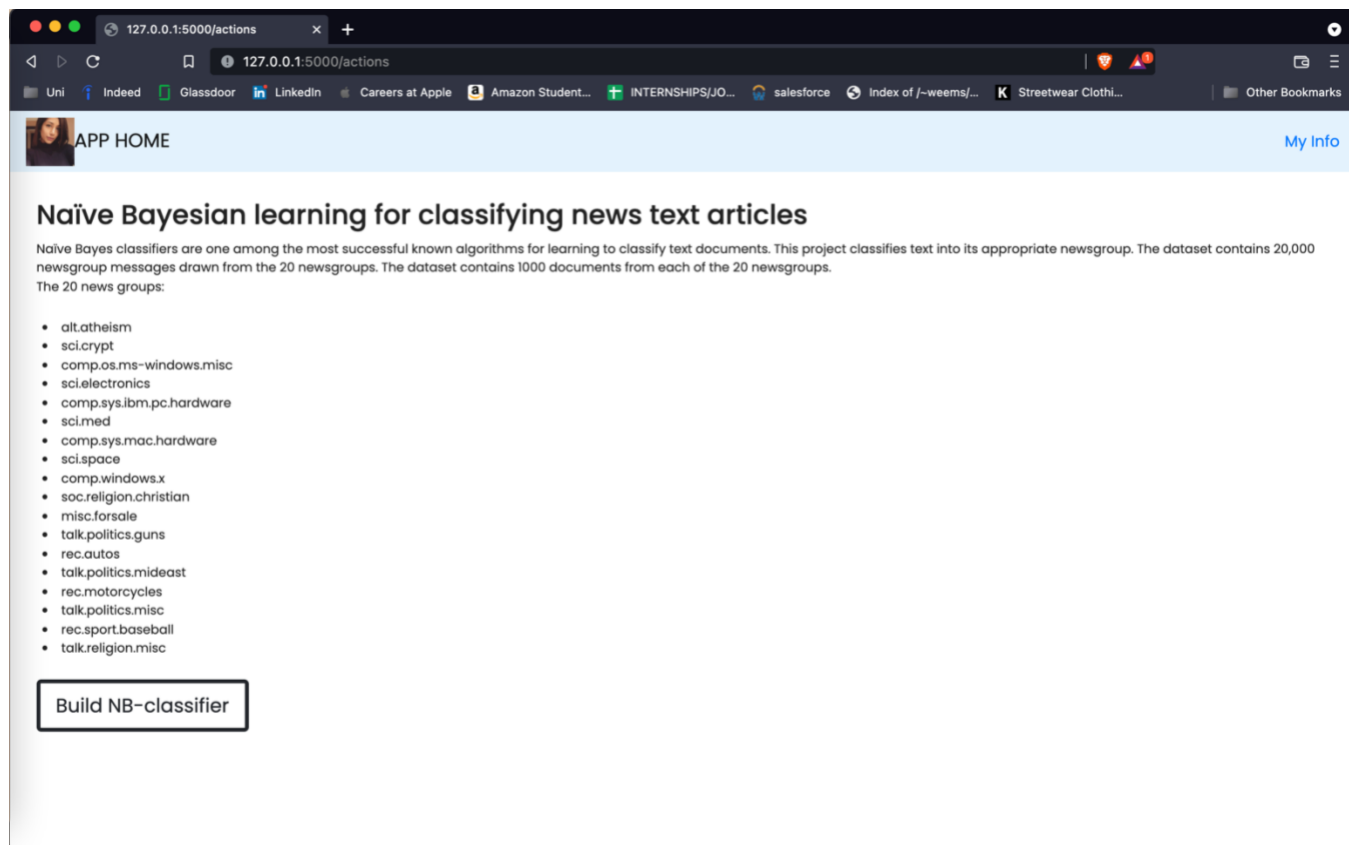
1. HOMEPAGE



2. MY INFO PAGE



3. NBC OVERVIEW



4. BUILD AND TEST MODEL

127.0.0.1:5000/index

APP HOME [My Info](#)

Steps in training the model

- For example, let us consider the following text -> "The microprocessor was made by AT&T, the wireless company that also has a big football stadium in Arlington, Texas"
- After data cleaning, the above sentence looks like this : "Microprocessor AT&T wireless company football stadium Arlington Texas"
- The words of a sentence are given a probability score, and the sentence is classified as a particular group according to the max probability score.

Microprocessor AT&T wireless company football stadium Arlington Texas

Word	Hardware	Technology	Sports	Geography
Microprocessor	23%	20%	2%	1%
AT&T	40%	36%	0%	2%
wireless	56%	20%	10%	10%
company	20%	50%	10%	10%
football	20%	45%	10%	5%
stadium	45%	10%	5%	90%
Arlington	45%	10%	5%	90%
Texas	45%	10%	5%	90%

Test the Model

(Please wait 4-5 minutes for results, do not close!)

5. RESULT OF TESTING MODEL & ACCURACY

127.0.0.1:5000/test

APP HOME [My Info](#)

Steps in training the model

- For example, let us consider the following text -> "The microprocessor was made by AT&T, the wireless company that also has a big football stadium in Arlington, Texas"
- After data cleaning, the above sentence looks like this : "Microprocessor AT&T wireless company football stadium Arlington Texas"
- The words of a sentence are given a probability score, and the sentence is classified as a particular group according to the max probability score.

Microprocessor AT&T wireless company football stadium Arlington Texas

Word	Hardware	Technology	Sports	Geography
Microprocessor	23%	20%	2%	1%
AT&T	40%	36%	0%	2%
wireless	56%	20%	10%	10%
company	20%	50%	10%	10%
football	20%	45%	10%	5%
stadium	45%	10%	5%	90%
Arlington	45%	10%	5%	90%
Texas	45%	10%	5%	90%

Accuracy of the trained NB classifier : **87.94262212859866%**

Time lapsed : 111 seconds