

# Title

## Demographic Analysis of Breast Cancer Awareness

Mayuri Kadam  
A20353234

# Problem

What ratio of population is aware of breast cancer and what region do they belong to?

# Approach

- Collect tweets from different users having mention of breast cancer related keys in their messages.
- Identifying the correct messages from the collected data and categorize whether relevant or not.
- Process relevant data for recognizing the gender and location of user, who tweets concern about the awareness programs.

# Data

Raw: User tweets

Required: Tweet text, coordinates, geo, user name, location and census data.

Collection : Twitter API

Processing: Tokenize raw data based on relevance

**Relevant :** .@TPSChrisBoddy and I unleashed our inner man today for a great cause: Breast Cancer Awareness. #pinkManiCURE

**Irrelevant:** NFL's generosity towards breast cancer awareness outpaced by its collection of player fines: <http://onforb.es/1Msu2lc>

# Timeline

Collect tweets from 10/10 to 11/10

Design tokenizer for available raw data: 10/31

Design classifier for gender and location : 11/07

Experiment and verify results: 11/14

Conclude and report: 11/20

Thank you!