Shikhar Sharma (110739968)
Gopi Chaganti (110739515)

INTRODUCTION

What are the broad range of topics, people of different age groups and genders are interested in? How age and gender influence the topics people care about. Can we visualize this information in an interactive manner?
Generally, it makes intuitive sense that people of different ages and gender will have interests in different topics and tend to write blogs on these topics. Words usage and writing style will also be different for different age and gender groups. For example, kids tend to be more interested in cartoons and games and tend to write articles about their favorite super hero's. They use simple and childish words like “cool”, “lol”, etc unlike old people. Some might think that older people have a complicated vocabulary since they have more experience. Another hypothesis might be that males use more action oriented words than females, who in turn, use more descriptive words. So final inference from discussion is that, kind of words people use in their writing represents the topic and reflects their age.

GOAL

Our goal in this project is to make this intuition concrete by analyzing blog posts by people of different ages and gender to try to find out what these differences are, what their nature is, and then visualizing it.

APPROACH

Our Approach consists of three main steps. First Collecting the blog posts data by crawling the web. Fetching the metadata associated with the posts. Then cleaning and analysing the data. And the final step involves visualising the data in an interactive way.

Data Collection:

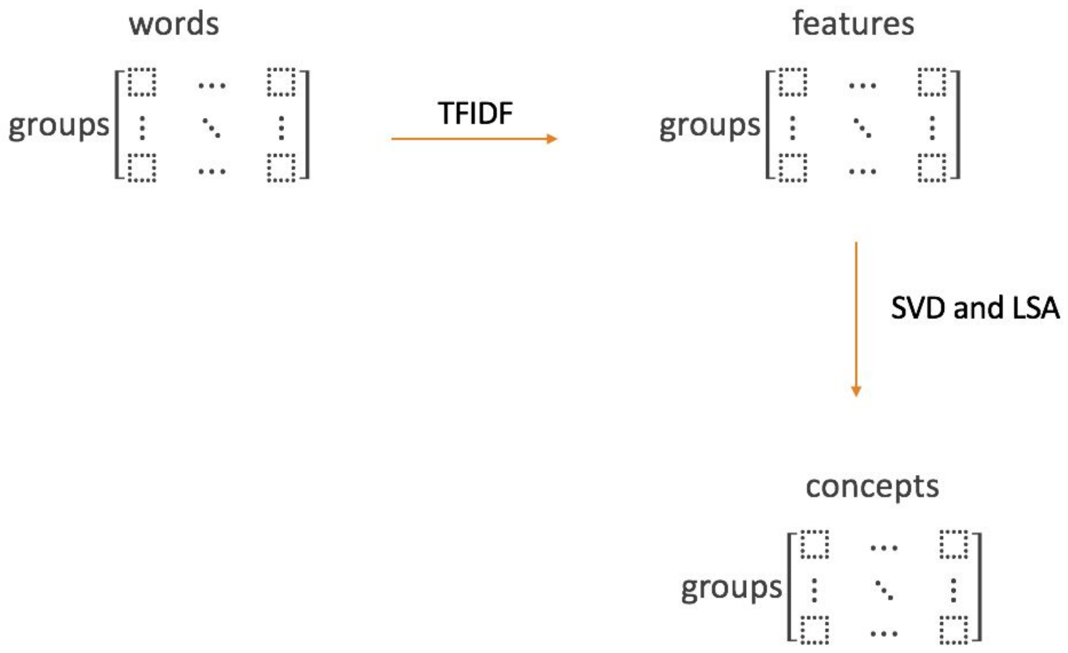
To study the writing patterns and interests of different age and gender groups we have decided to analyse their blog posts as it reflects both their interests and word usage. So we have crawled the web for blog posts and collected over 35,000 users blogs of different people who fall under different age groups and genders from blogger.com website. On an average each blog has 150 posts. And each post has approximately 2300 words. We also have collected the metadata for the posts like user name, user age, post created date, user gender. One difficulty we faced during the process is, user grows age over a period of time so associated each blog post with the user age based on the created date of the post rather than keeping the user current age for all his blog posts.

Data Cleaning, Tokenization and Grouping:

We have used adaptive sampling technique to make sure that the analysis data sample represents all groups of people in appropriate proportion. We then tokenized the posts by using regular expression pattern matching. Then we have grouped the tokens(words) based on age groups and gender. So our processed dataset looks like below.
Dataset = [{age: X, gender: Y, words: []}]

Data Analysis:

We have constructed TFIDF matrix for our data, which will assign weights to the words according to its presence in the data. Then we have passed this matrix to Singular value decomposition method to extract useful words from each group. We then performed Latent Semantic Analysis to derive concepts from the groups. These are the concepts people of that group are actively talking about or interested in. In this process we made sure that selected topics represents the most of the variance in the original data by taking appropriate number of concepts into account.

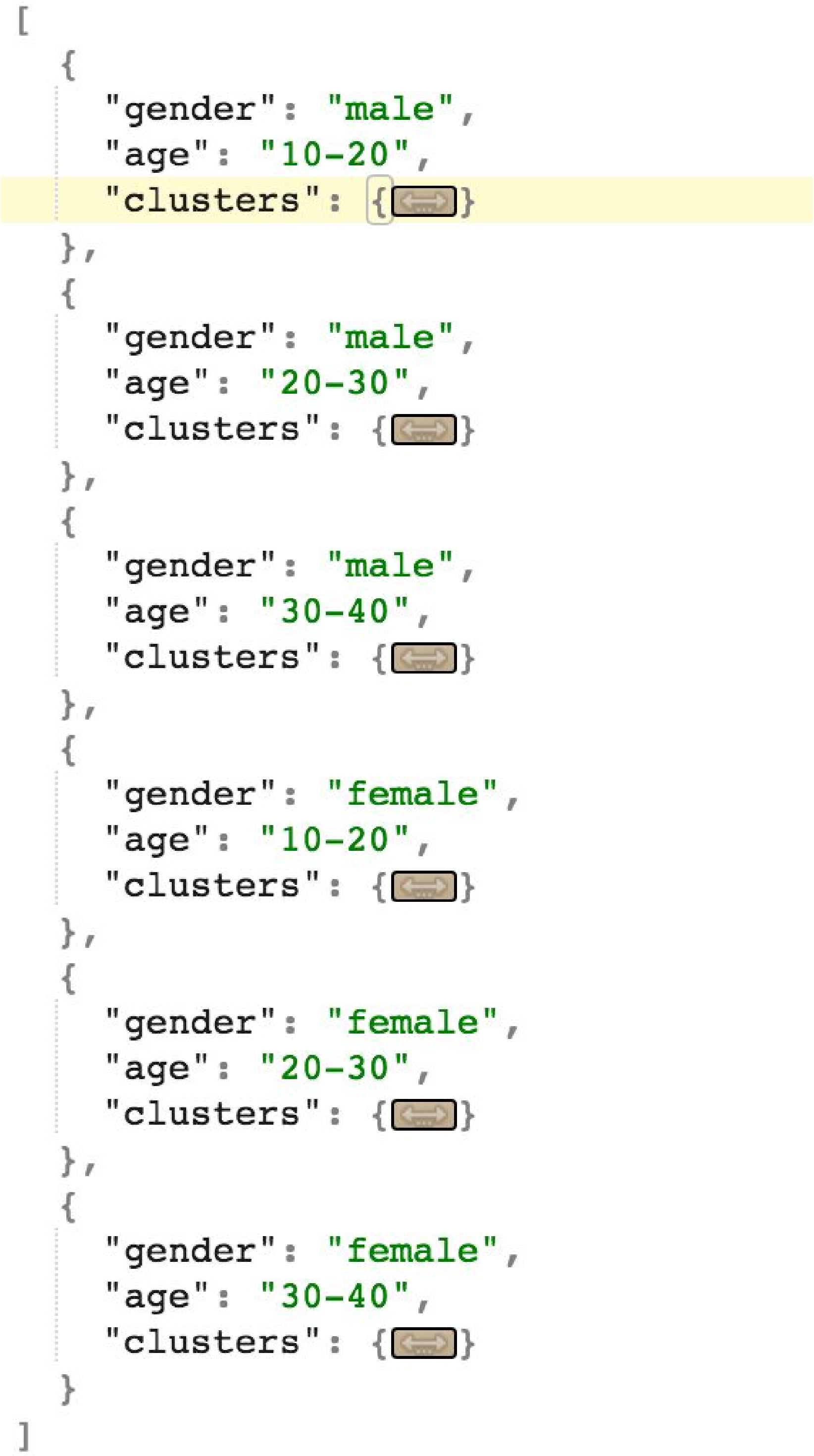


We then used K-means clustering to categorise similar concepts in the same group. Then these concepts were projected into words space. Each word in the cluster was weighted based on its distance to the center. Finally, we selected top 50 words per each cluster according to weightage. Here each cluster represents a topic and can be further explored to see the words that represent the cluster.

So, essentially we have found the topics people of different age groups and gender are interested in.

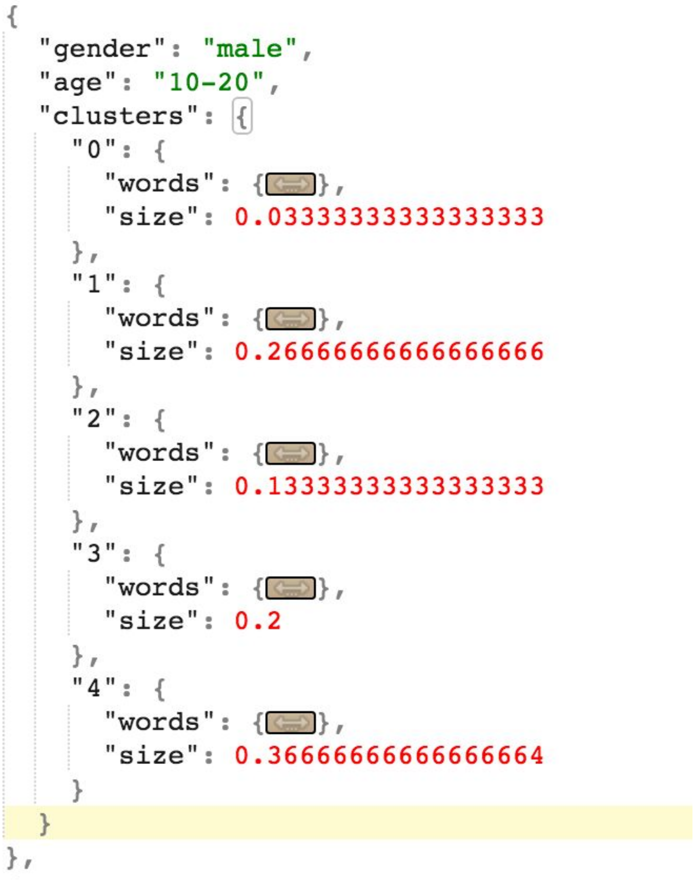
Data obtained after analysis:

1) Topics(clusters) for each group



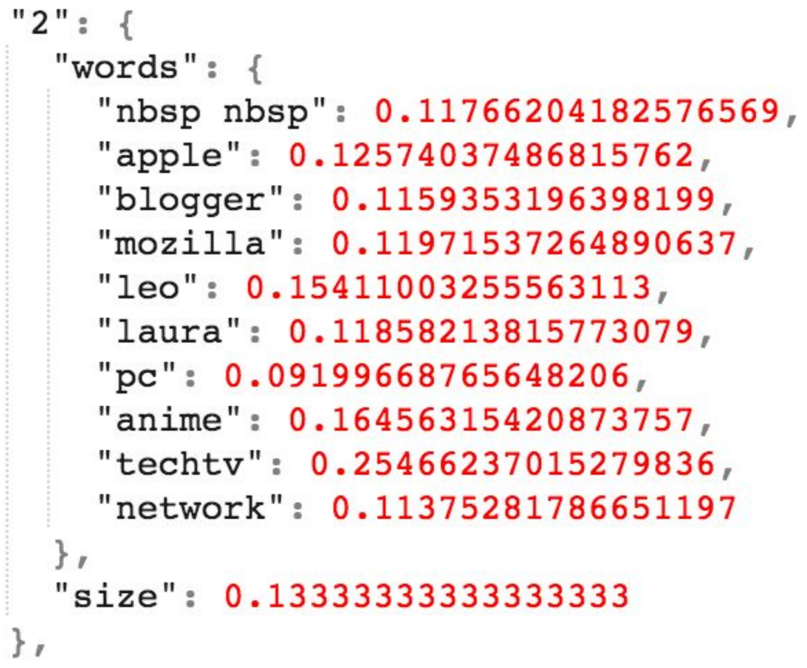
```
[
  {
    "gender": "male",
    "age": "10-20",
    "clusters": {↔}
  },
  {
    "gender": "male",
    "age": "20-30",
    "clusters": {↔}
  },
  {
    "gender": "male",
    "age": "30-40",
    "clusters": {↔}
  },
  {
    "gender": "female",
    "age": "10-20",
    "clusters": {↔}
  },
  {
    "gender": "female",
    "age": "20-30",
    "clusters": {↔}
  },
  {
    "gender": "female",
    "age": "30-40",
    "clusters": {↔}
  }
]
```

2) Cluster data for an age group



```
{
  "gender": "male",
  "age": "10-20",
  "clusters": {
    "0": {
      "words": {↔},
      "size": 0.03333333333333333
    },
    "1": {
      "words": {↔},
      "size": 0.26666666666666666
    },
    "2": {
      "words": {↔},
      "size": 0.13333333333333333
    },
    "3": {
      "words": {↔},
      "size": 0.2
    },
    "4": {
      "words": {↔},
      "size": 0.36666666666666664
    }
  }
},
```

3) words that represent a cluster (topic)



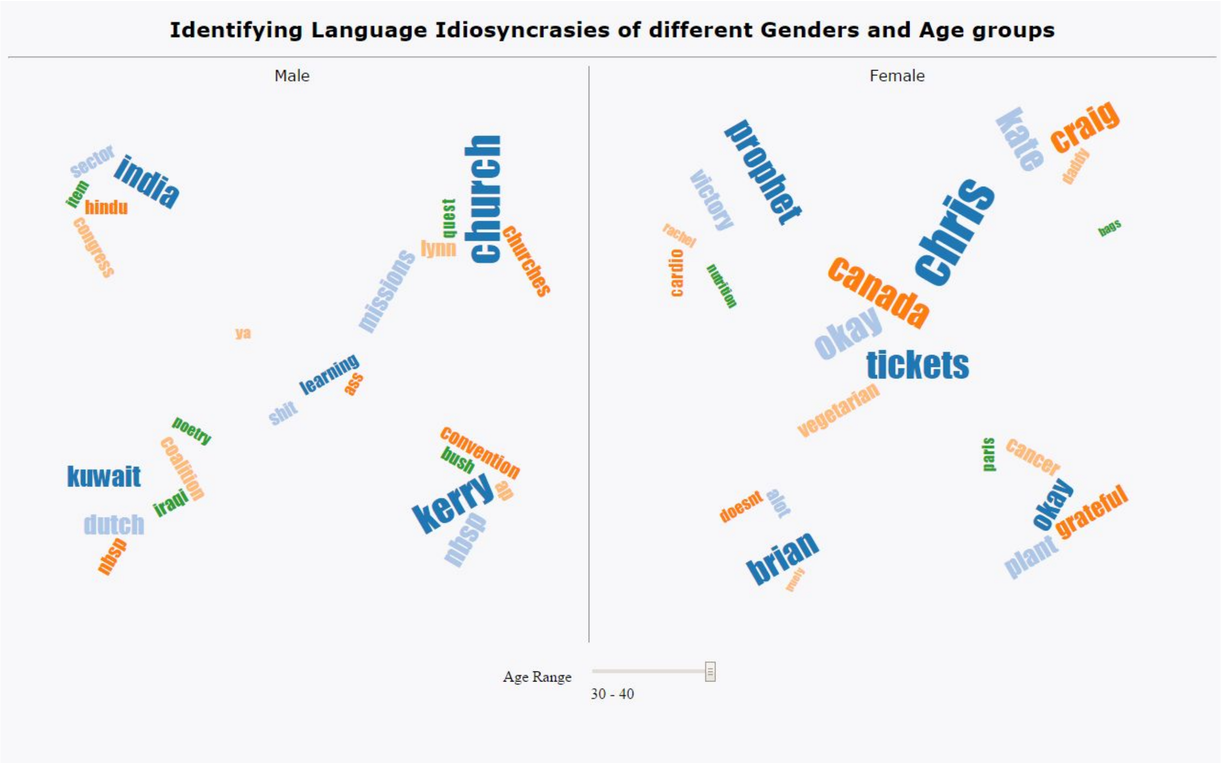
```
"2": {
  "words": {
    "nbsp nbsp": 0.11766204182576569,
    "apple": 0.12574037486815762,
    "blogger": 0.1159353196398199,
    "mozilla": 0.11971537264890637,
    "leo": 0.15411003255563113,
    "laura": 0.11858213815773079,
    "pc": 0.09199668765648206,
    "anime": 0.16456315420873757,
    "techtv": 0.25466237015279836,
    "network": 0.11375281786651197
  },
  "size": 0.13333333333333333
},
```

Visualization and Interactions:

URL: http://allv28.all.cs.stonybrook.edu/shiksharma/vis/final/

We plot the data as word cloud clusters, with different cluster space for males and females. As you can see in the figure, the male cluster space is on the left, and the female cluster space is on the right. Each cluster represents similar concepts (topics).

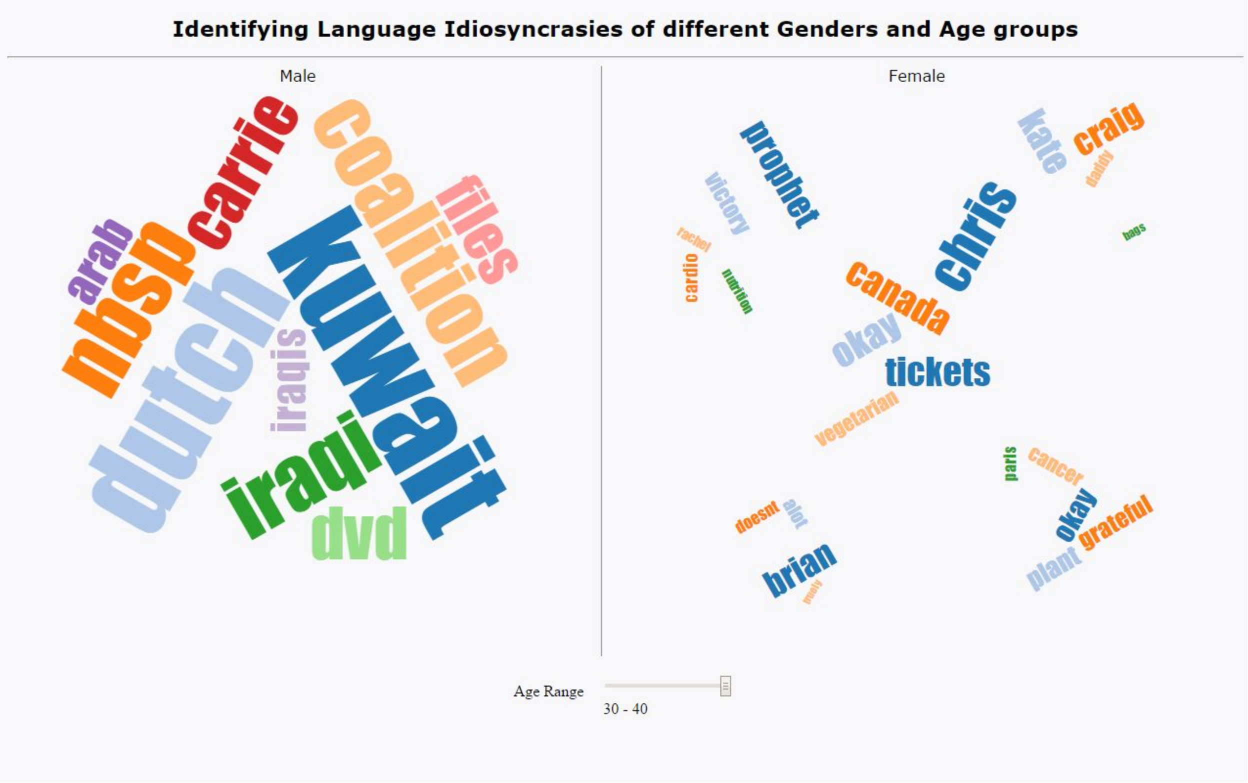
The size of the clusters and the contained words is based on the strength of the word in the analysis. You can see in the figure that the words have coalesced into clusters.



Identifying Language Idiosyncrasies of different Genders and Age groups

Pop up Clusters:

You also have the ability to zoom in on a particular cluster.
If you click on any individual cluster, that cluster pops up and you see more words in the zoomed-in cluster view.
(See figure that follows)



Identifying Language Idiosyncrasies of different Genders and Age groups

Male

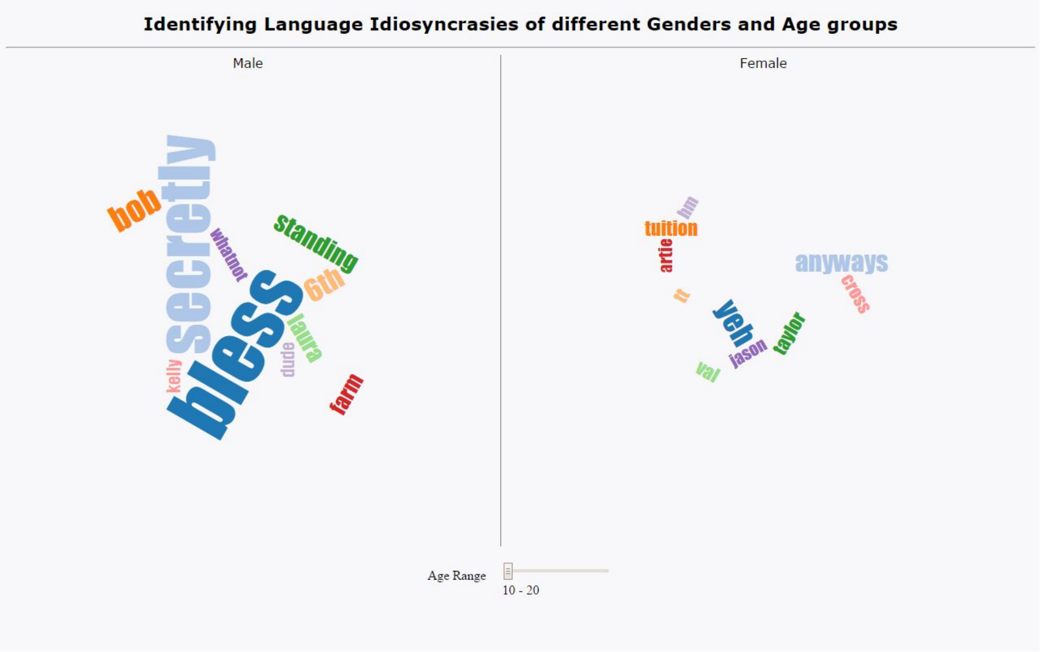
Female

Age Range

30 - 40

Age Intervals:

On the bottom of the page, you have the option of selecting an age interval for both genders. When you select a particular age interval, the word clusters update to reflect the change.



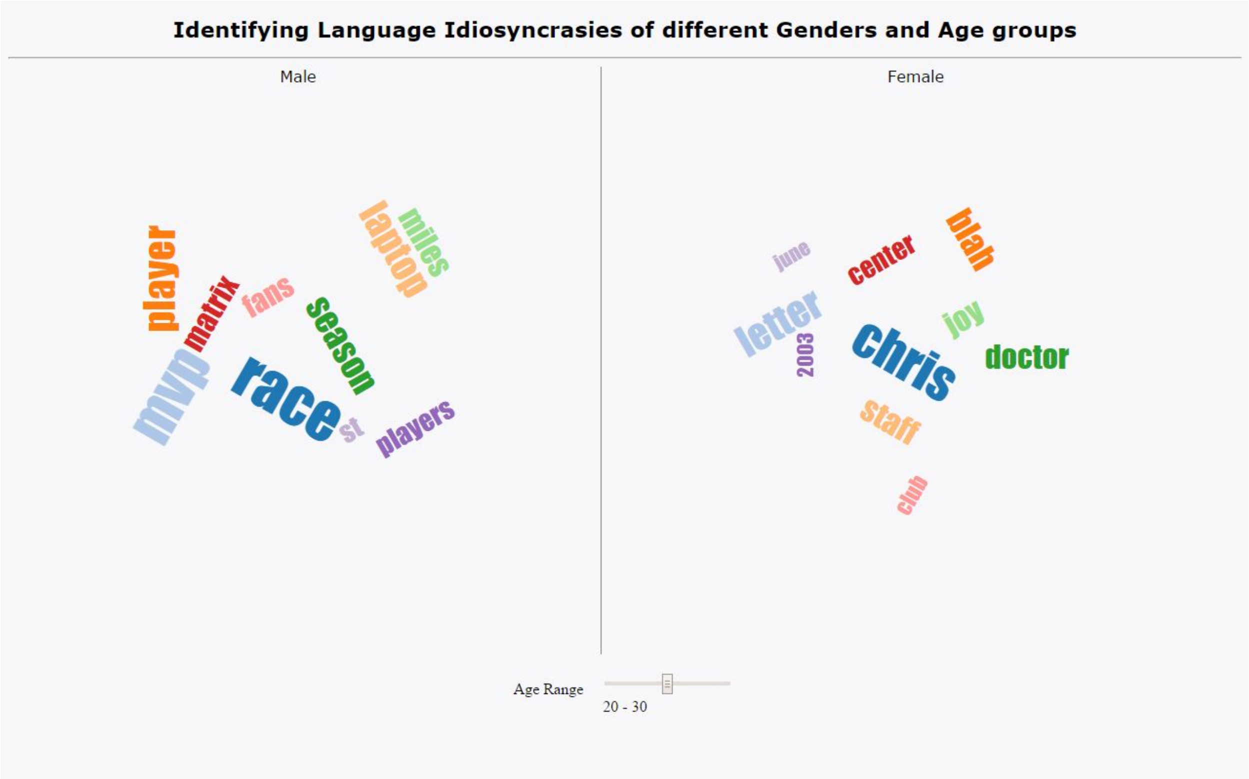
Identifying Language Idiosyncrasies of different Genders and Age groups

Male

Female

Age Range

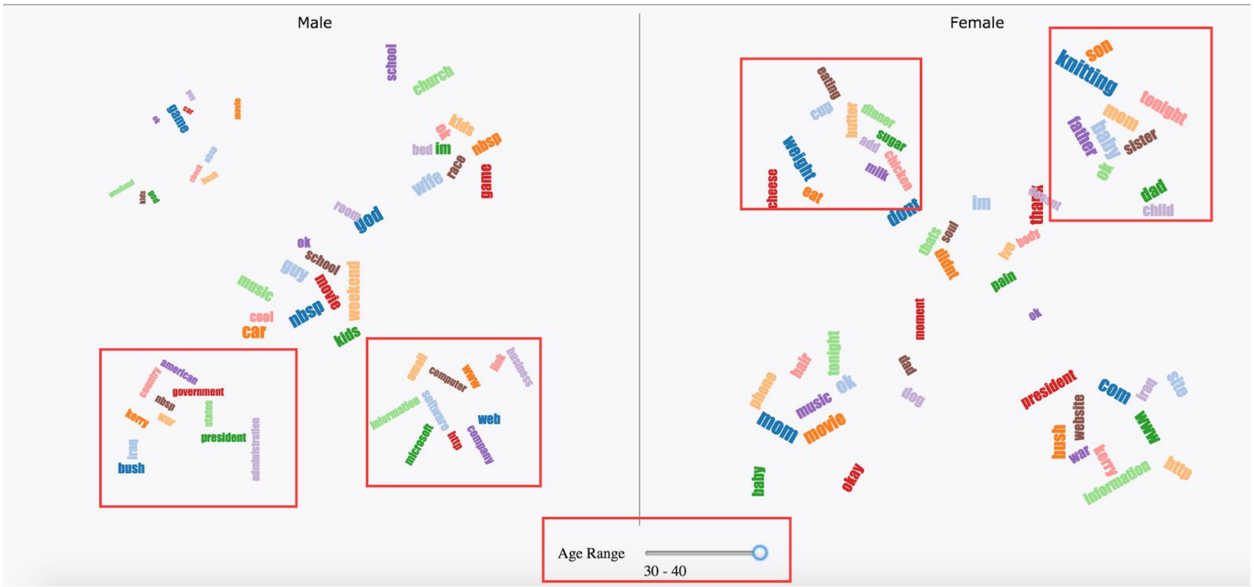
10 - 20



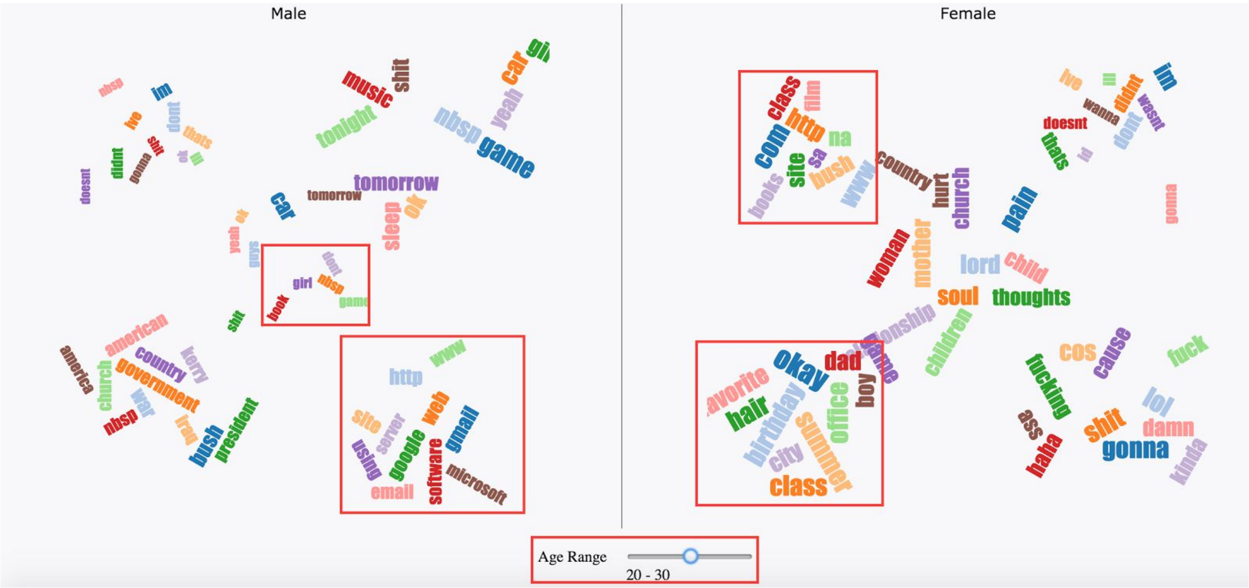
Identifying Language Idiosyncrasies of different Genders and Age groups

Results:

Our results show that women in the 30-40 age group are more conscious about eating habits and family while men in the same age group care about jobs, politics, technology, hobbies.



Women in the 20-30 age group talk more about parties, studies, boys while men in the same age group talk about music, girls, technology.



Girls in the 10-20 age group talk more about music, weekend, photos while men in the same age group talk about games, internet.

Conclusion:

It is clear from this project that age and gender have absolute impact on the topics people choose and blog. We can use this analysis to build a model which can estimate the age and gender of author given some text.