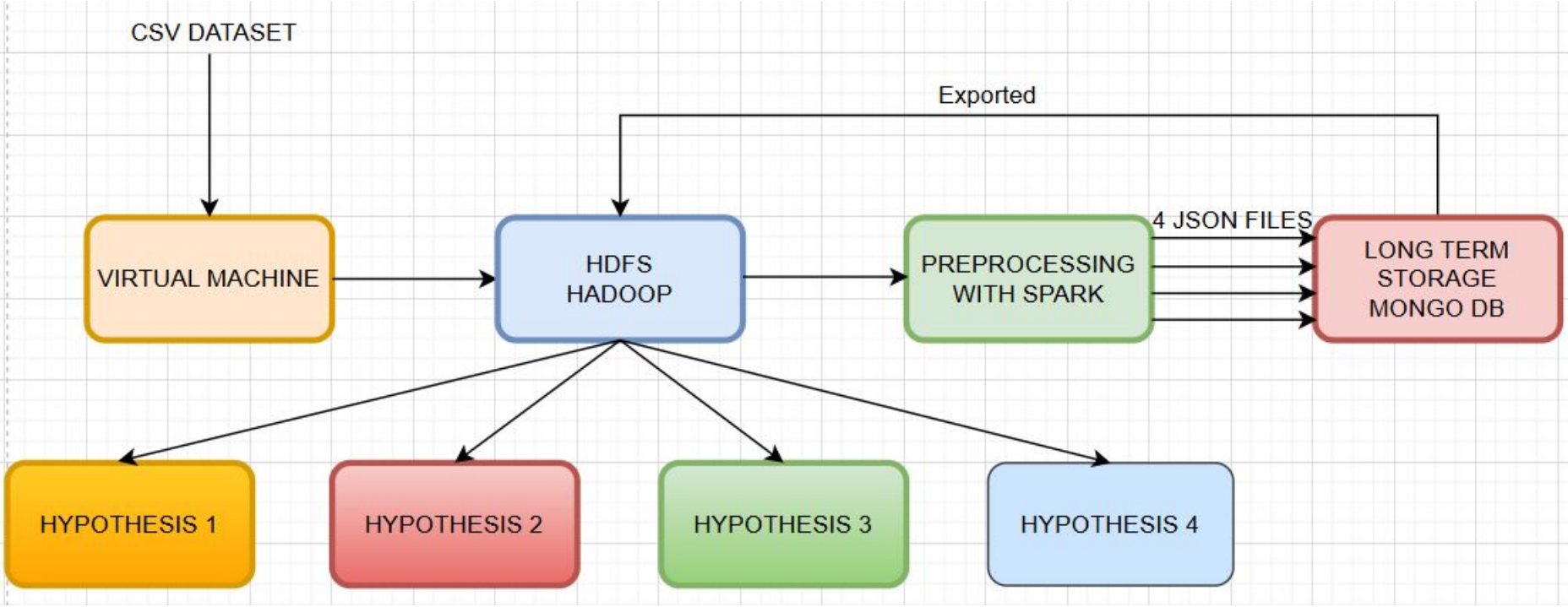
OKCupid Profiles

Structure of our project



Preprocessing: cleaning and transformation

-Handling NaNs → put a point in empty cells

-Essay Consolidation → merging all essays from 0 to 9 → full_profile

-Keyword Tokenization and Counting

-Stored in MongoDB

GENERAL DATA

Total females: 24.117 (40,23%)

Total males: 35.829 (59,77%)

Categories with non empty cells	Categories with high percentage of null values
Age, sex, orientation, …	Body_shape, religion, education, …

Hypothesis:

1)“Women write longer profile descriptions than men.”

2)“Homosexual and bisexual people mention artistic activities (music, film, art) more frequently than heterosexual people.”

3)“Women mention more frequently that they want to have children than men.”

4)“People with higher educational levels identify less frequently with organized religions.”

“Women write longer profile descriptions than men.”

-Query to retrieve only 2 fields from each user: the gender and the full profile text

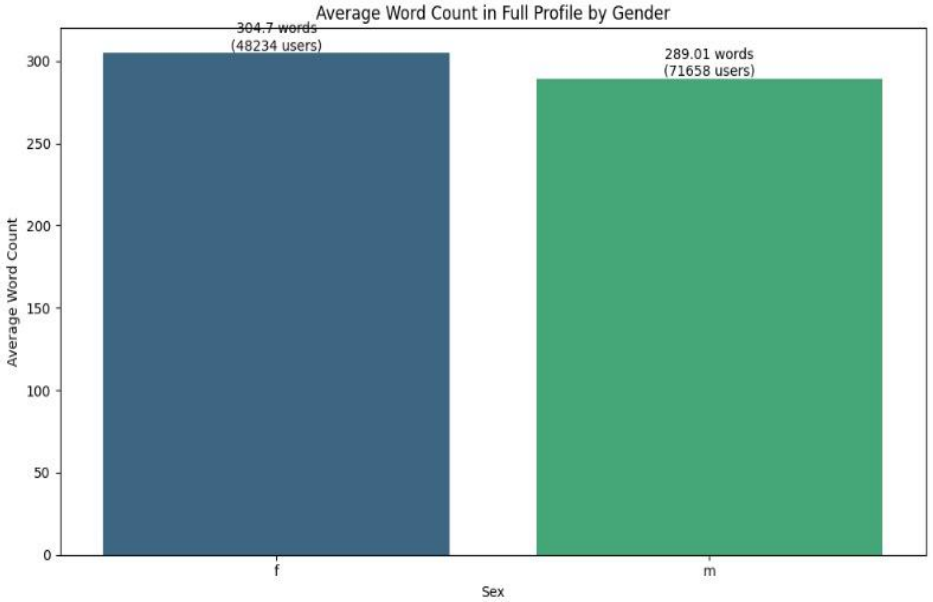
-Insert data into a Pandas DataFrame

-We removed punctuation marks

-Grouped the data by gender and calculated the average number of words per profile

-Compare whether, on average, women write longer profiles than men

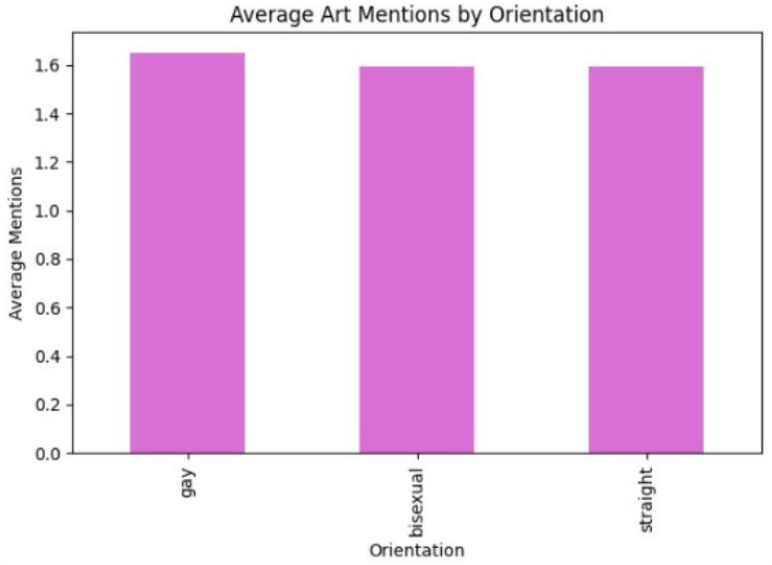
Our hypothesis is **TRUE**



“Homosexual and bisexual people mention artistic activities (music, film, art) more frequently than heterosexual people.”

- Count art keywords in the comments
- Perform average of number of keywords of each person
- **TRUE**

```
Average Art Mentions by Orientation:
orientation
gay          1.652630
bisexual     1.593455
straight     1.593124
Name: art_mentions, dtype: float64
```



“Women mention frequently that they want to have children”

- Create a MapFunction with the keywords: ‘kids’, ‘children’ and all their possible variants.
- Filter children mentions by gender.
- Total number of mentions (could be more than 1 for the same person) so we calculate the ratio.
- TRUE

```
Mentions of children by gender:
f: 11100 mentions
m: 10734 mentions
Normalized values:
f: 0.1151
m: 0.0749
```

“People with higher educational levels identify less frequently with organized religions”

A lot of people: 40452 people don’t want to say their religion.
We used a MapReduce job to test whether people with higher education identify less often with organized religions.

The map function grouped profiles into three education levels (higher, general, lower).
It then checked if the religion field contained any structured religion keywords (e.g., catholicism, islam, judaism).

The reduce step aggregated counts per group to compare how often each education level mentioned religion.

Observing the graph: TRUE