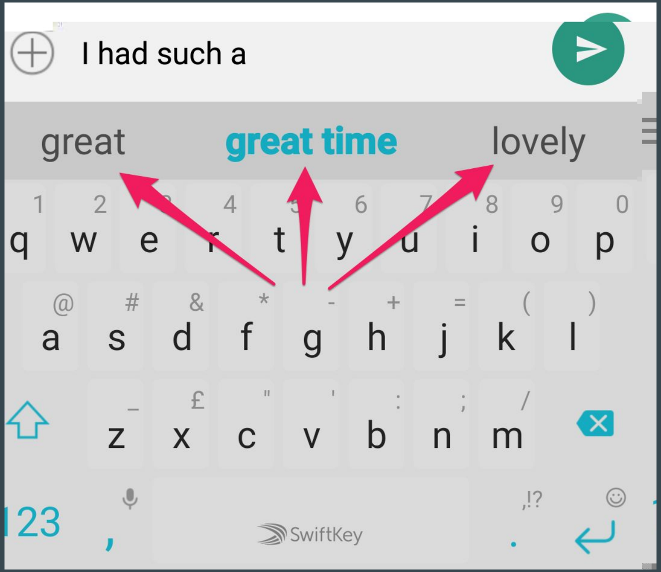
Next-word Prediction with Approach Federated Learning

Bruno A. R. M. Campos

Larissa Peluso Rozza

Business Understanding

Predict the next word ensuring data privacy



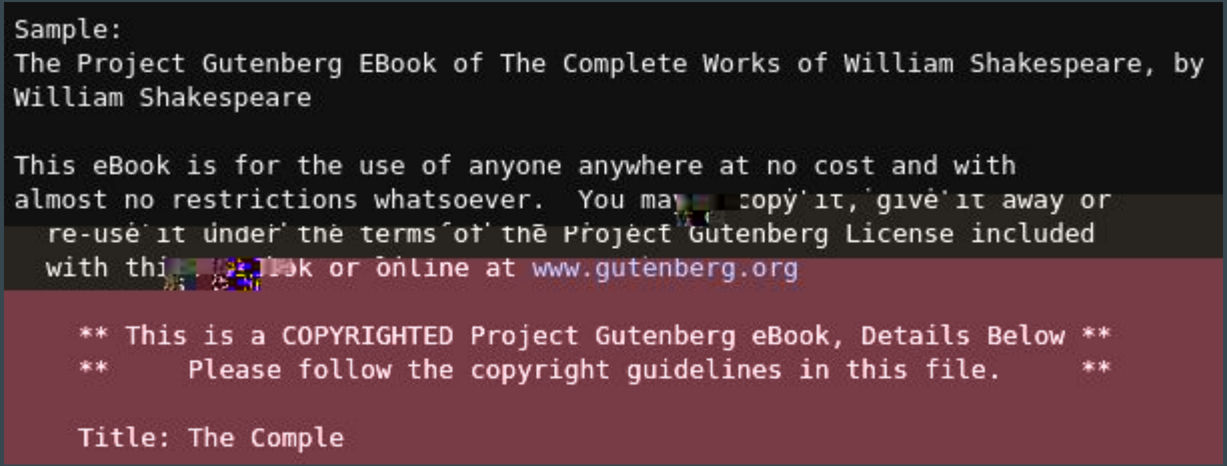
I had such a

Collect Initial Data

- The data consists in Shakespeare’s plays
- Hosted in Project Gutenberg
- Unique file with all plays
- txt format
- Script to download

Describe Data

- Total rows: 124.788
- Total tokens: 1.134.708
- Total characters: 5.465.129
- Total unique character: 91
- Sample, first 500 characters:

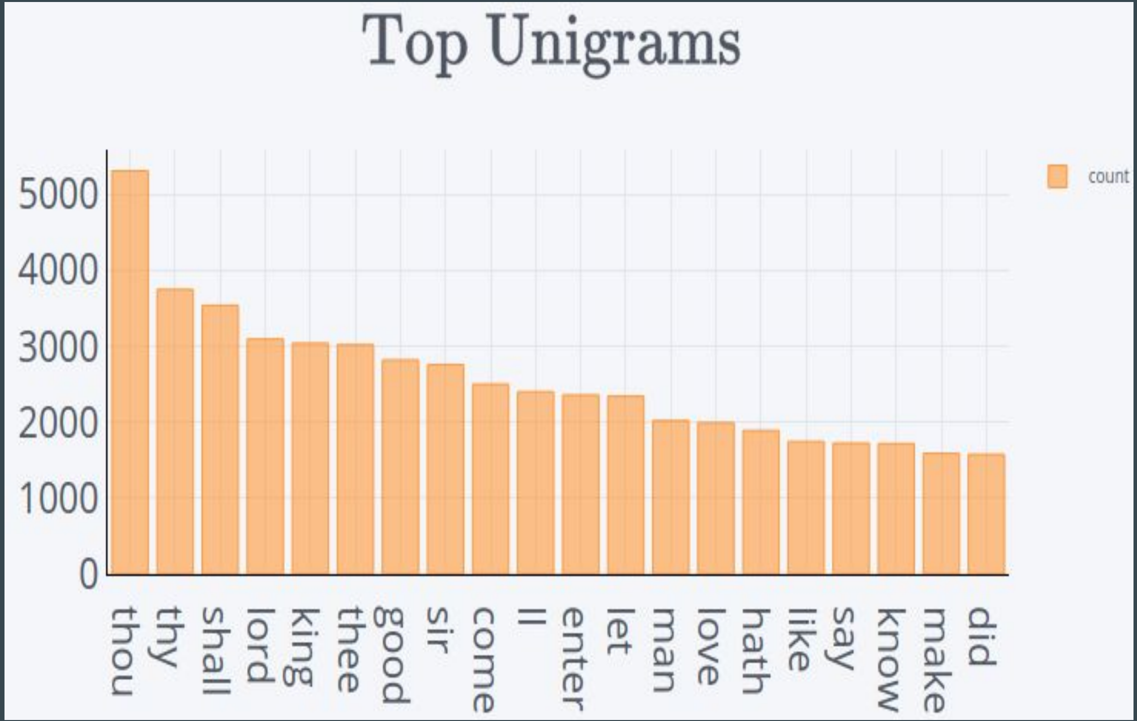


```
Sample:
The Project Gutenberg EBook of The Complete Works of William Shakespeare, by
William Shakespeare

This eBook is for the use of anyone anywhere at no cost and with
almost no restrictions whatsoever.  You may copy it, give it away or
re-use it under the terms of the Project Gutenberg License included
with this eBook or online at www.gutenberg.org

    ** This is a COPYRIGHTED Project Gutenberg eBook, Details Below **
    **     Please follow the copyright guidelines in this file.     **

    Title: The Comple
```

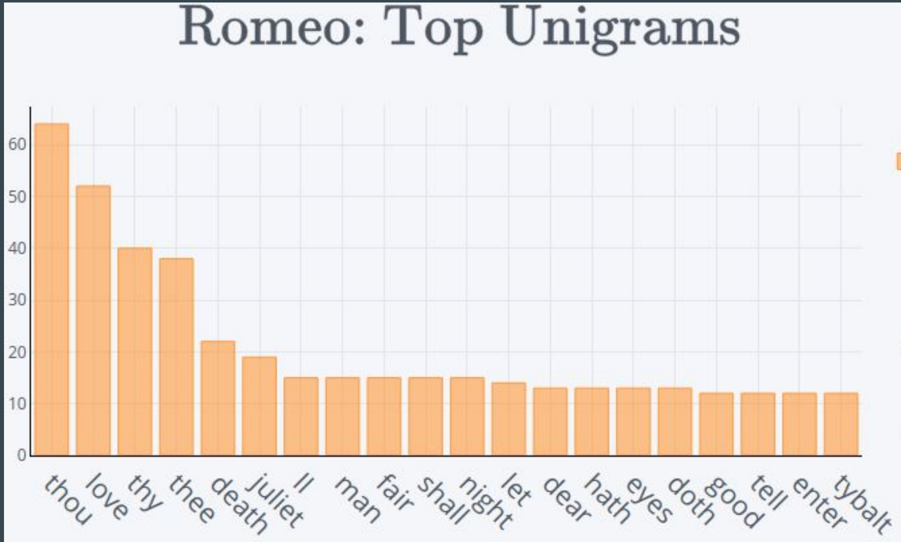
Preprocessing

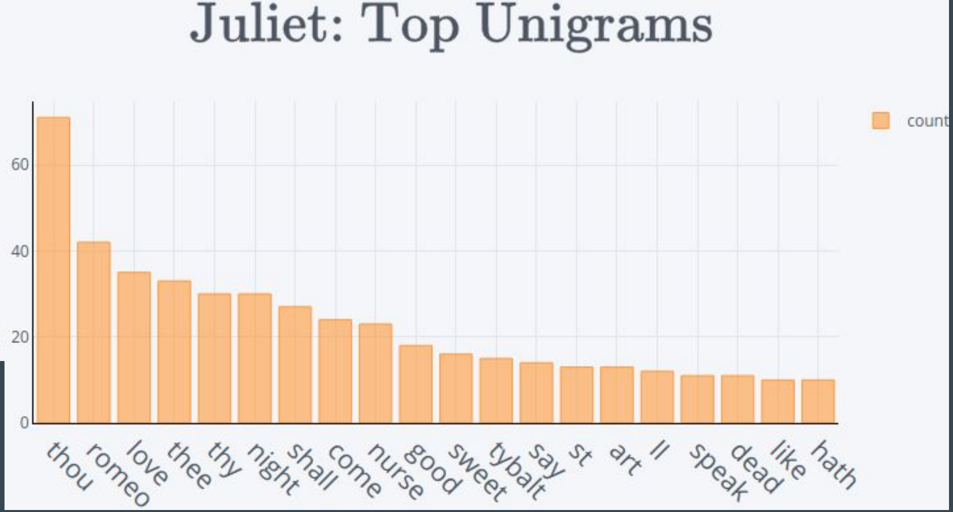
- Function to remove unnecessary text
- Bag-of-Words
- Term Frequency and Inverse Document Frequency (TF-IDF)
- Latent Dirichlet Allocation (LDA)

Explore Data (n-grams)

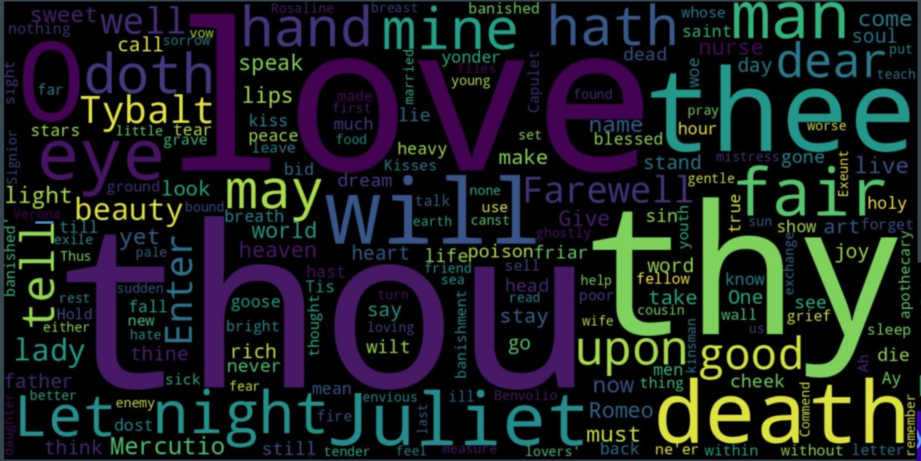


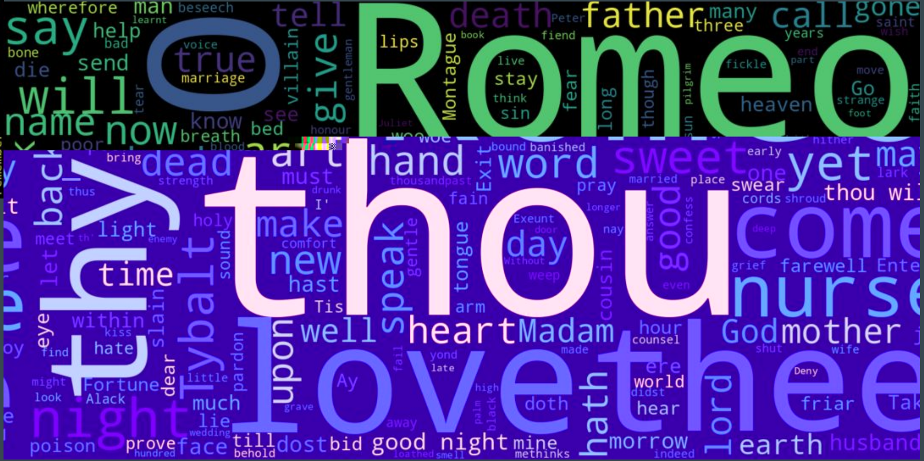
Explore Data (n-grams): Romeo and Juliet





Explore Data (WordCloud): Romeo and Juliet

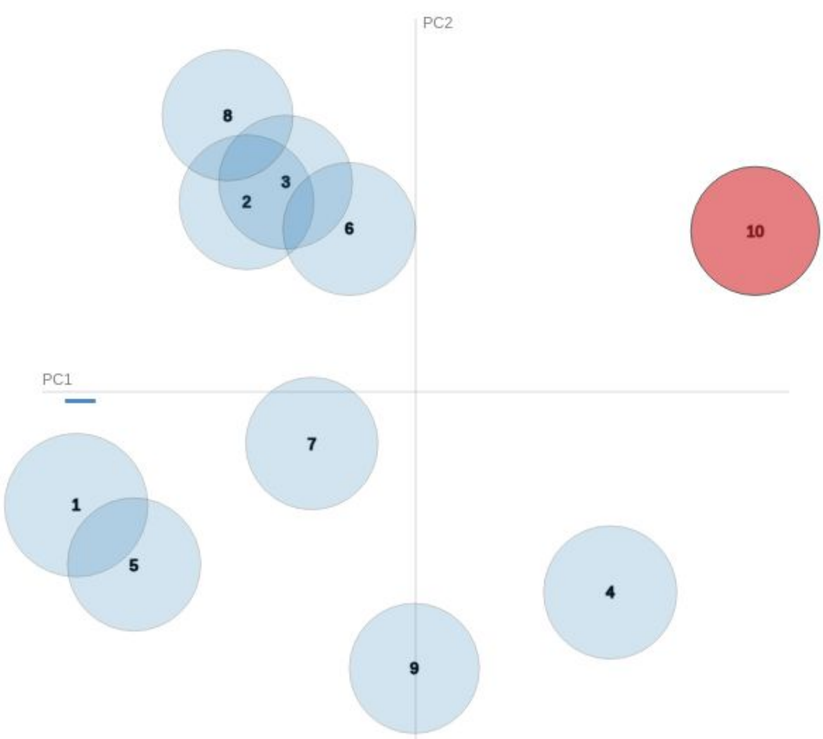




Word Topic Clusters

By analyzing the words, topics can be classified:

- Topic 1: Family Duel
- Topic 2: Actions of a king
- Topic 3: Romantic dispute
- Topic 4: Farewell (Despedida)
- Topic 5: Relationship between queen, mother, prince and gentleman



NOTE: Each bubble represents a topic

Federation Design

- To simulate the federation of clients, the module will be used tff.simulation
- 715 users where each example corresponds to a data set of spoken by the character in a given play.

Valeu

•••

/brunocampos01

/LarissaPeluso