Work in Progress: Next-word Prediction with Approach Federated Learning

Bruno Aurélio Rozza de Moura Campos
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Brazil
brunocampos01@gmail.com

***Abstract*—When a user is typing text on a mobile device it can be useful to suggest the next word as this will optimize typing time and also avoid possible errors. However, this data has private information, which limits its movement to a centralized environment. With this scenario as motivation, in this paper we will demonstrate how to predict the next word while guaranteeing users privacy without moving the data around.**

***Index Terms*—Federated Learning, Text Generation, Data Privacy**

I. Introduction

This work has how goals to predict the next word, ensuring data privacy. Shakespeare pieces will be used as input data. These will be obtained, described, pre-processed and explored for a better understanding. Next, the federation environment will be created where each character in Shakespeare's works will be a participating user, and their speeches will be the input dataset in the prediction model. From this scenario, in which data is only found on mobile devices, Federated Learning will be used to perform the model training in a shared way without moving the data to a centralized environment. For this, steps will be carried out to load a pre-trained global model from a central server, data pre-processing and model training on the user's own device. Then the model parameters will be forwarded to the central server to update the federated average and train the global model.

II. Getting and Describe Data

Data were obtained from the Gutenberg project which has all Shakespeare play compiled into a single file. This text file is 12,788 lines, 1,134,708 tokens, 5,465,129 characters and 91 unique characteres. The data has many problems with formatting, in addition, in some pieces there is information about the Gutenberg project and usage license. This information is not valuable to predict the next word from the purpose of this work.

III. Preprocessing Text

To begin with, it is necessary to remove pieces unnecessary texts such as information about the Gutenberg project, terms of use license and poems. This data cleansing will help to get text with greater predictive value for the next-word on into Shakespeare's plays. Natural Language Processing (NLP) techniques for data processing will be applied.

A. Bag-of-Words

A bag-of-words (BoW) is a model maps a document into a vector as $v = [x1, x2, ..., xn]$, where xi denotes the occurrence of the ith word in basic terms [1]. In this case a BoW was generated in resulting a sparse array of data. To better visualize the BoW, only the twenty most common tokens in all parts were keep.

B. Term Frequency and Inverse Document Frequency

Term Frequency and Inverse Document Frequency (TF-IDF) determine the importance of a feature in a text document [2]. In this approach, instead of filling the document vectors with the raw count, as in the BoW approach, it is filled with a score. This technique comes in very handy for the next word prediction problem because it calculates the rarity of a word in a text.

C. Latent Dirichlet Allocation

As mentioned in [3] the latent Dirichlet allocation (LDA) is a topic model which discovers latent structures in data as a set of topics from a collection of texts. In this work, the ten most important topics containing twenty words each were calculated.

IV. Data Exploration

Using the BoW technique, the first twenty unigrams and bigrams of the text were calculated and how result, the figures 1 and 2 generated.

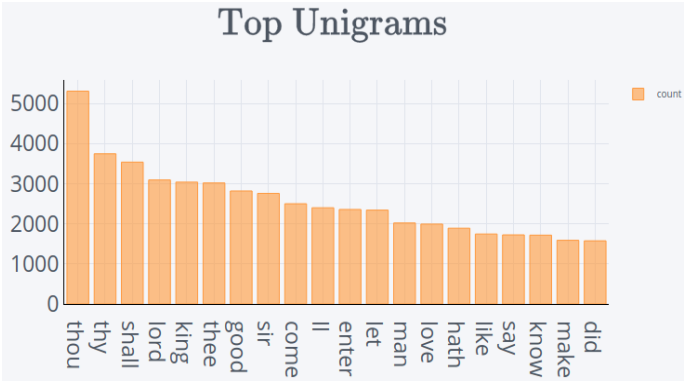

Fig. 1. The distribution of top unigrams after data cleansing.

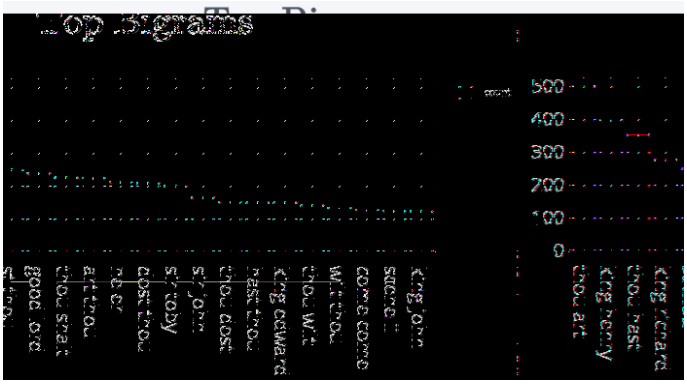

Fig. 2. The distribution of top bigrams after data cleansing.

A data representation in visual form that is easier to understand than text and tables is wordcloud hat provides a visual representation of words frequency [4]. In this work, the figure 3 was generated to visualize the main words in Shakespeare's plays.

Fig. 3. Wordcloud

estrutura de tópicos das peças

V. FEDERATION DESIGN

The Federation has 715 users, where each one corresponds to the speeches of the characters in Shakespeare's plays, that is, in the federation there is the user Romeo, from the tragedy Romeo and Juliet, who has all of his own speeches.

REFERENCES

[1] D. Yan, K. Li, S. Gu and L. Yang, "Network-Based Bag-of-Words Model for Text Classification," in IEEE Access, vol. 8, pp. 82641-82652, 2020, doi: 10.1109/ACCESS.2020.2991074.

[2] N. S. Mohd Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," in IEEE Access, vol. 9, pp. 52177-52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[3] C. Karmakar, C. O. Dumitru, G. Schwarz and M. Datcu, "Feature-Free Explainable Data Mining in SAR Images Using Latent Dirichlet Allocation," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 676-689, 2021, doi: 10.1109/JSTARS.2020.3039012.

[4] M. F. A. Bashri and R. Kusumaningrum, "Sentiment analysis using Latent Dirichlet Allocation and topic polarity wordcloud visualization," 2017 5th International Conference on Information and Communication Technology (ICoIC7), 2017, pp. 1-5, doi: 10.1109/ICoICT.2017.8074651.