BrainStation
The Digital Learning Company™

William Nankap

Capstone Final Submission

April 2023

NLP ClassifyChat

An innovative text classification project based on natural language processing. It leverages conversations from the freeCodeCamp chat to provide an accurate and efficient solution **for automatically categorizing messages and extracting relevant information from large volumes of textual data.**

Our project brings added value to the company by enabling:

- Classify the dominant topics in discussions between students on the platform
- Detect difficulties encountered by students in learning or using the platform
- Recommend content tailored to participants' interests
- Note and evaluate the impact of socialization among learners

Our dataset contains around 5 million rows and 24 columns. It was retrieved from Kaggle and subsequently cleaned and pre-processed.

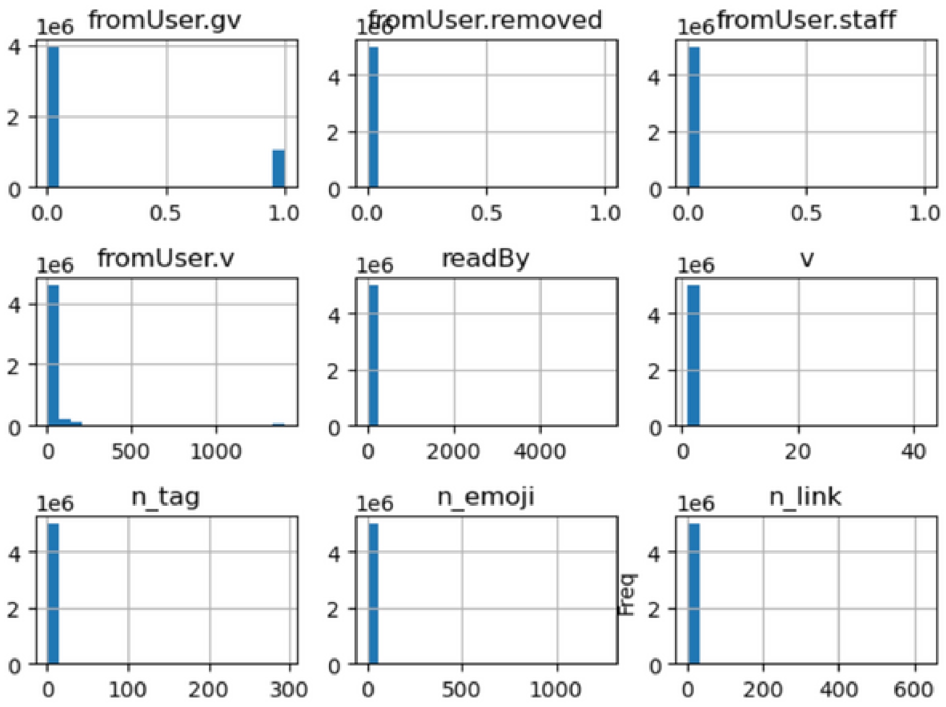
The cleaning and pre-processing process consisted in:

- Modify the type of certain columns
- Remove noise from our text column

William Nankap

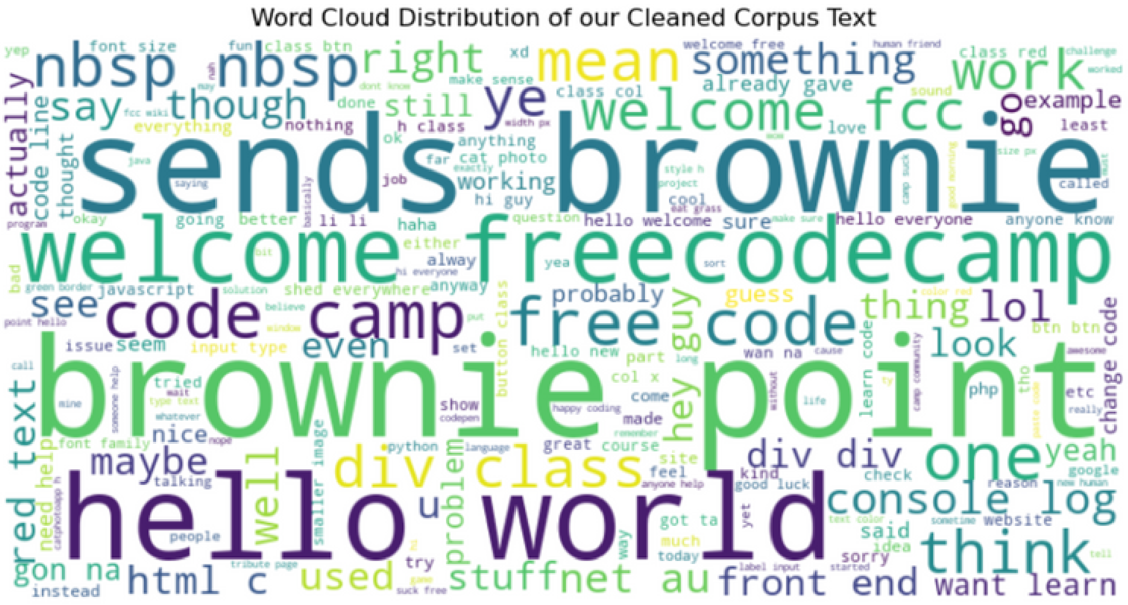
Bixi Project Deliverable 2

Insight



Distribution of my numeric columns in the dataset, we notice a badly balanced distribution.

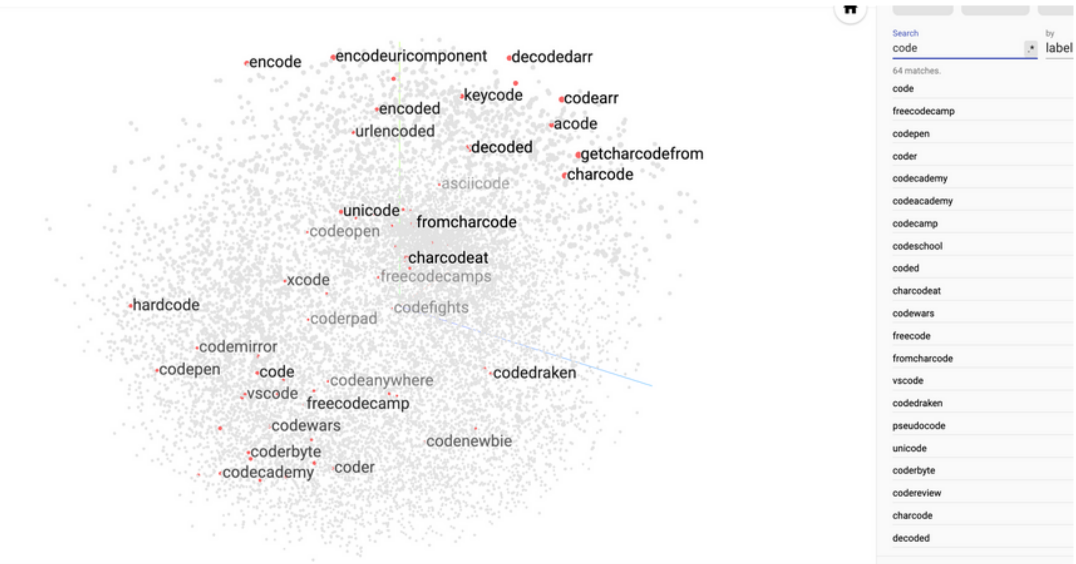
They won't be very useful for creating a model, a logistic regression test was carried out but the predictions were very weak, requiring no further investigation.



Word distribution in my text corpus, the size is proportional to the number of occurrences in the dataset.

Observations:

- **Brownie point: refers to gifts exchanged between users**
- **Hello world: a form of greeting familiar to this environment (very often used in programming for a new project)**
- **Freecodecamp: the working platform**
- **Code: html, js,**
- **Leisure**



Search: code

by: label

64 matches.

- code
- freecodecamp
- codepen
- coder
- codecademy
- codeacademy
- codecamp
- codeschool
- coded
- charcodeat
- codewars
- freecode
- fromcharcode
- vscode
- codedraken
- pseudocode
- unicode
- coderbyte
- codereview
- charcode
- decoded

- **Vectorization of words, calculation of the distance between groups of words, similarities between groups of words**

Conclusion

In conclusion, the various models created enabled us to identify 4 main themes addressed by the learners:

- **Code**: students shared html or js code, which means they were probably learning these languages at the time.
- **Help**: students asked each other a lot of questions, which shows a strong and solid community.
- **Gifts**: students shared rewards and encouraged each other to progress.
- **Social leisure:** they enjoyed sharing joy and humor with each other

The project meets 70% of my initial expectations, given the origin of the dataset, I had assumed that the students would talk about programming (html, css), which is the case, but I wanted to have much more detail on the different programming topics they tackled. This may take even longer, as the dataset contained links shared between students, which may give a clue to the difficulties they were going through.