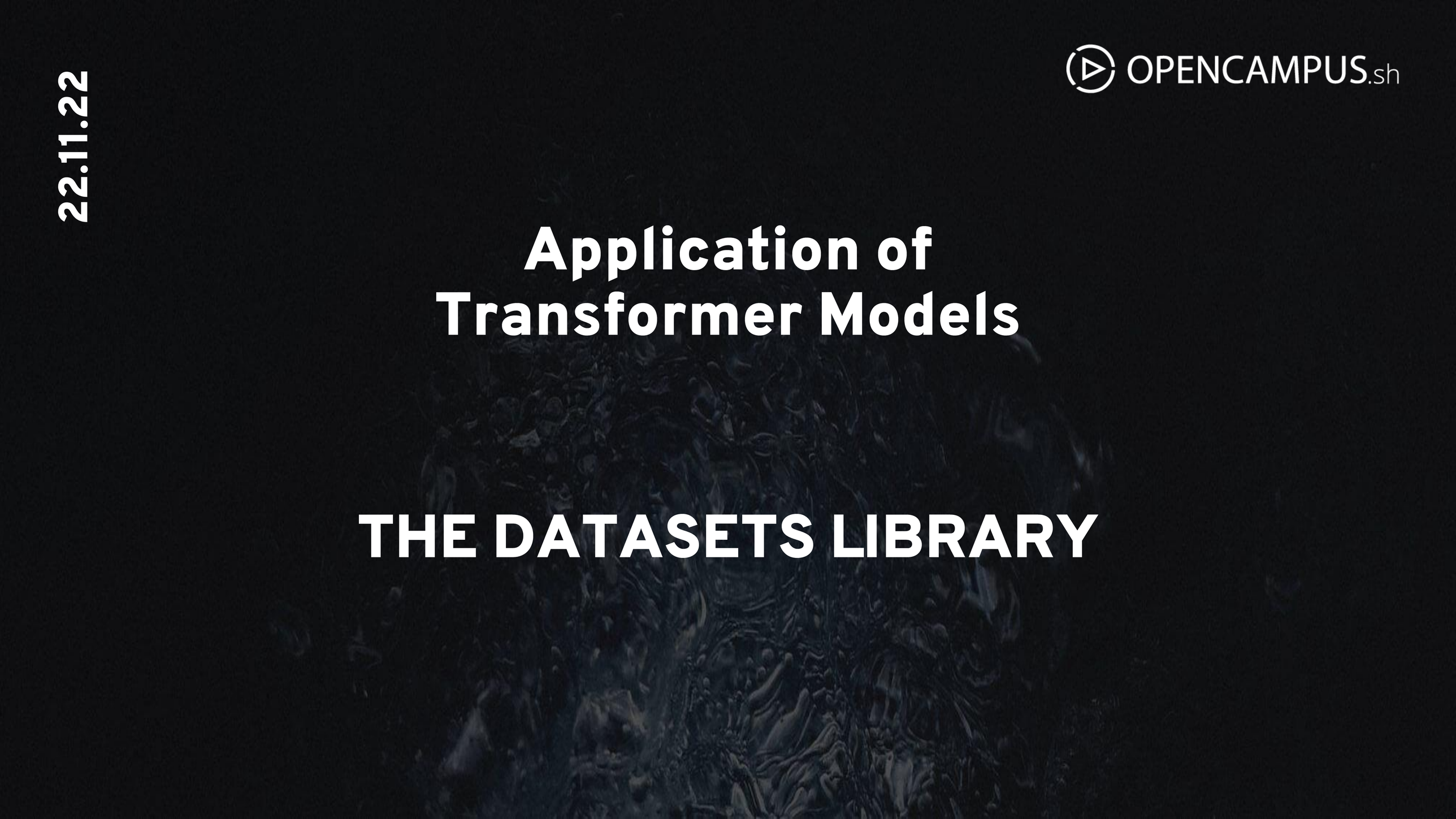
22.11.22

OPENCAMPUS.sh

Application of Transformer Models

THE DATASETS LIBRARY



- Quiz
- Literature Review
- Visualizing Sequences
- Dataset Characteristics

QUIZ

https://forms.office.com/r/hNg8zKfLEf

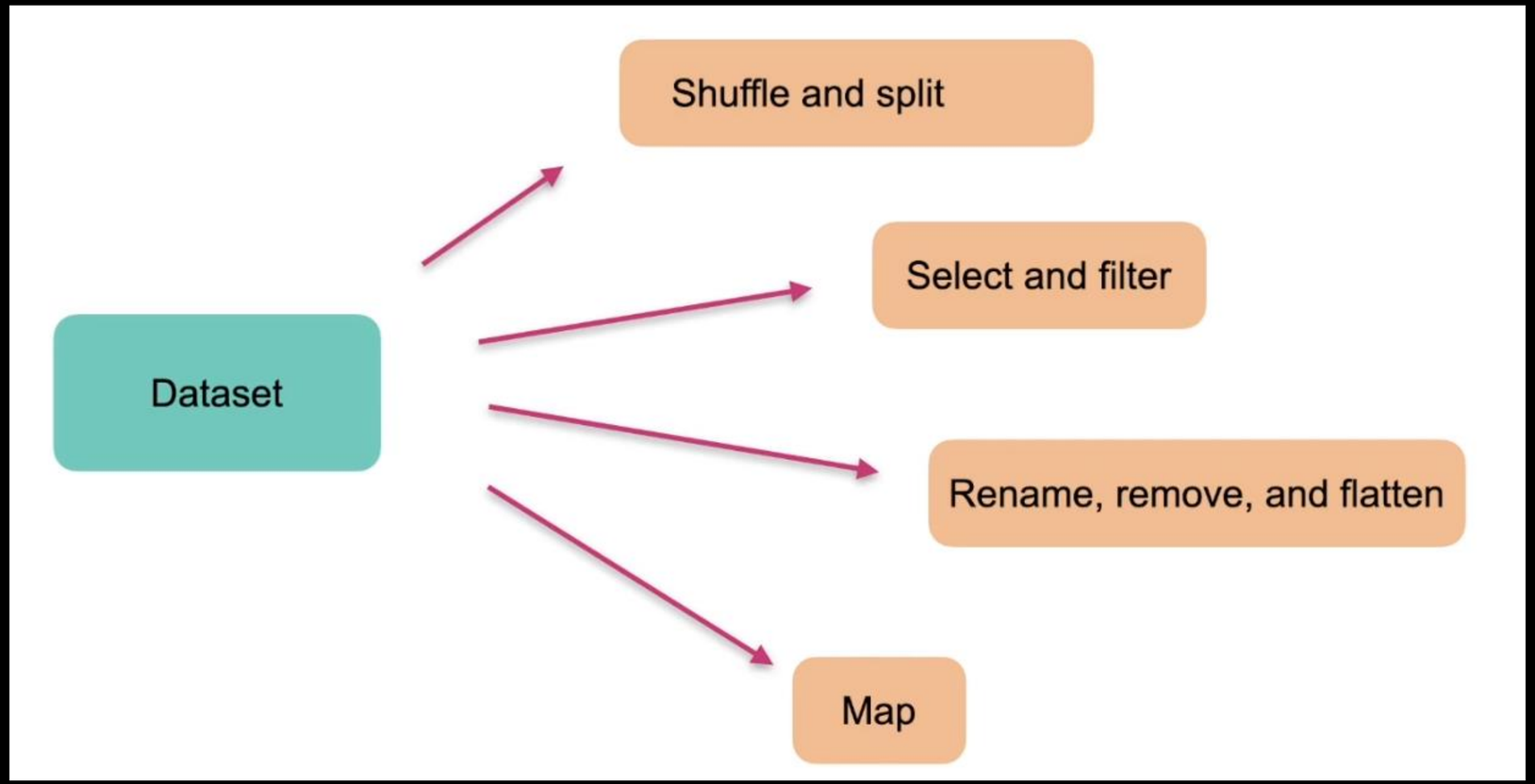
IMPORTING DATA

- **From CSV**
  ```
  load_dataset("csv", data_files="my_file.csv")
  ```

- **From JSON**
  ```
  load_dataset("json", data_files="my_file.jsonl")
  ```

- **From Pandas (Pickle)**
  ```
  load_dataset("pandas", data_files="my_dataframe.pkl")
  Dataset.from_pandas(my_dataframe)
  ```

DATASET METHODS



SAVING MODELS

GitHub, GitLab, Bitbucket, or a similar service using

- git and git LFS

Hugging Face Hub using

- huggingface_hub library (based on git and git FLS)
- push_to_hub API

HUGGING FACE HUB LIBRARY

```
# authentication
from huggingface_hub import notebook_login
notebook_login()

# saving via callback method
from transformers import PushToHubCallback
callback = PushToHubCallback(
    "bert-finetuned-mrpc", save_strategy="epoch", tokenizer=tokenizer
)
model.fit(train_dataset, epochs=2, callbacks=callback)

# saving manually
model.push_to_hub("bert-finetuned-mrpc, commit="End of training")
```

LITERATURE REVIEW

- Search for transformer models applied to similar problems
- Focus on the structure of the input and of the output
- Are there pretrained models that you can use?
- Which type of model is best suited?
- Do you need tokenization?
- Do you need a type of embedding layer?

FEEDBACK GROUPS

Group 1:

- Jonathan/Julian: Arguments Mining / NER Task on data already collected
- Max: Q&A model
- Khan: Classification of activity descriptions according to keywords

Group 2:

- Saif/ Emmanuel/ Kristian/ Atul: Time Series Prediction Financial/Climate
- Manpreet: Unsupervised training of log data to predict user behavior
- Laura/ Janosch/ Valentin : Training a model to produce text written in different authors' style

Group 3:

- Benjamin/ Malte/ T.-Niklas: Speech to speech models including translation
- Jeremy/ Veit/ Christian: Transcribing and summarizing Podcasts

PROJECT MILESTONES

16.11.	Form project groups
23.11.	Literature review
30.11.	Dataset characteristics
04.01.	Baseline model
11.01.	Project presentations

DATASET CHARACTERISTICS

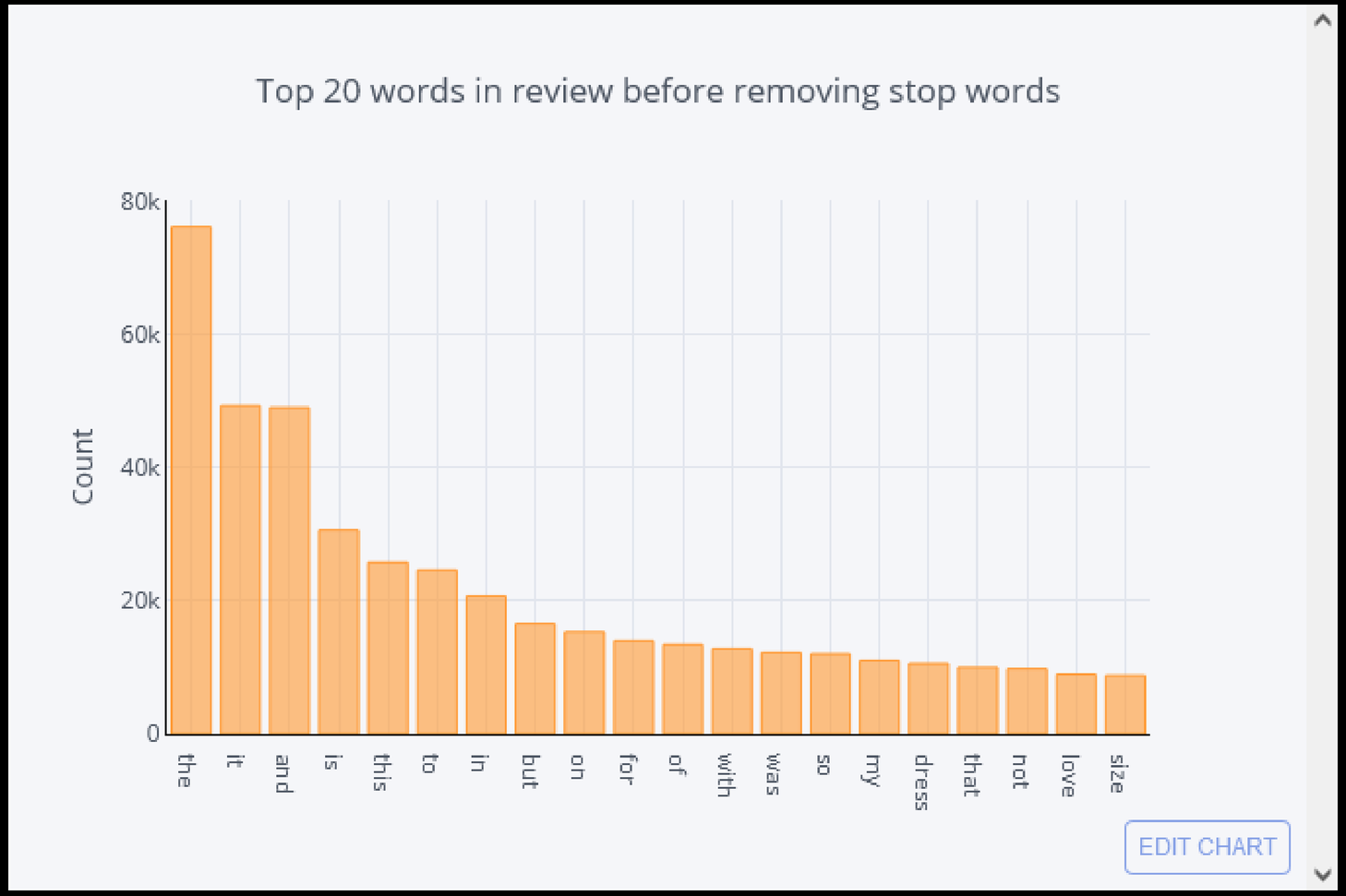
- Is your collection of samples possibly biased?
- How must the data be collected to be used with your model?
- For classification problems:
  - Is your sample balanced across all classes?
  - If not, how will you deal with it?
- For generation problems:
  - Define a set of prompts and investigate the completions according to their bias.

DATA VISUALIZATION

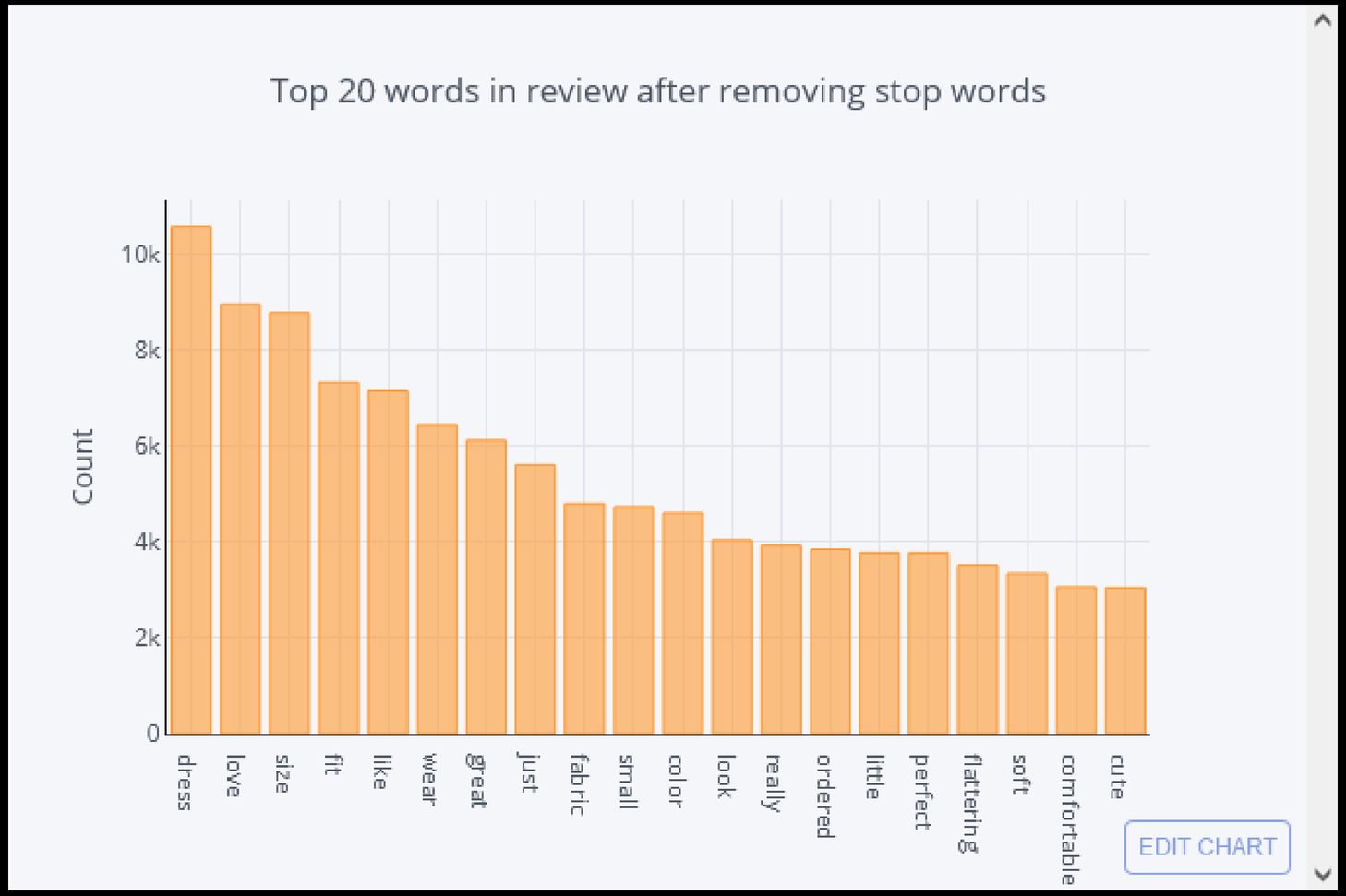
- Li, S. (2019, April 27). *A Complete Exploratory Data Analysis and Visualization for Text Data*. Medium. https://towardsdatascience.com/a-complete-exploratory-data-analysis-and-visualization-for-text-data-29fb1b96fb6a

- Example data using E-commerce reviews on cloth

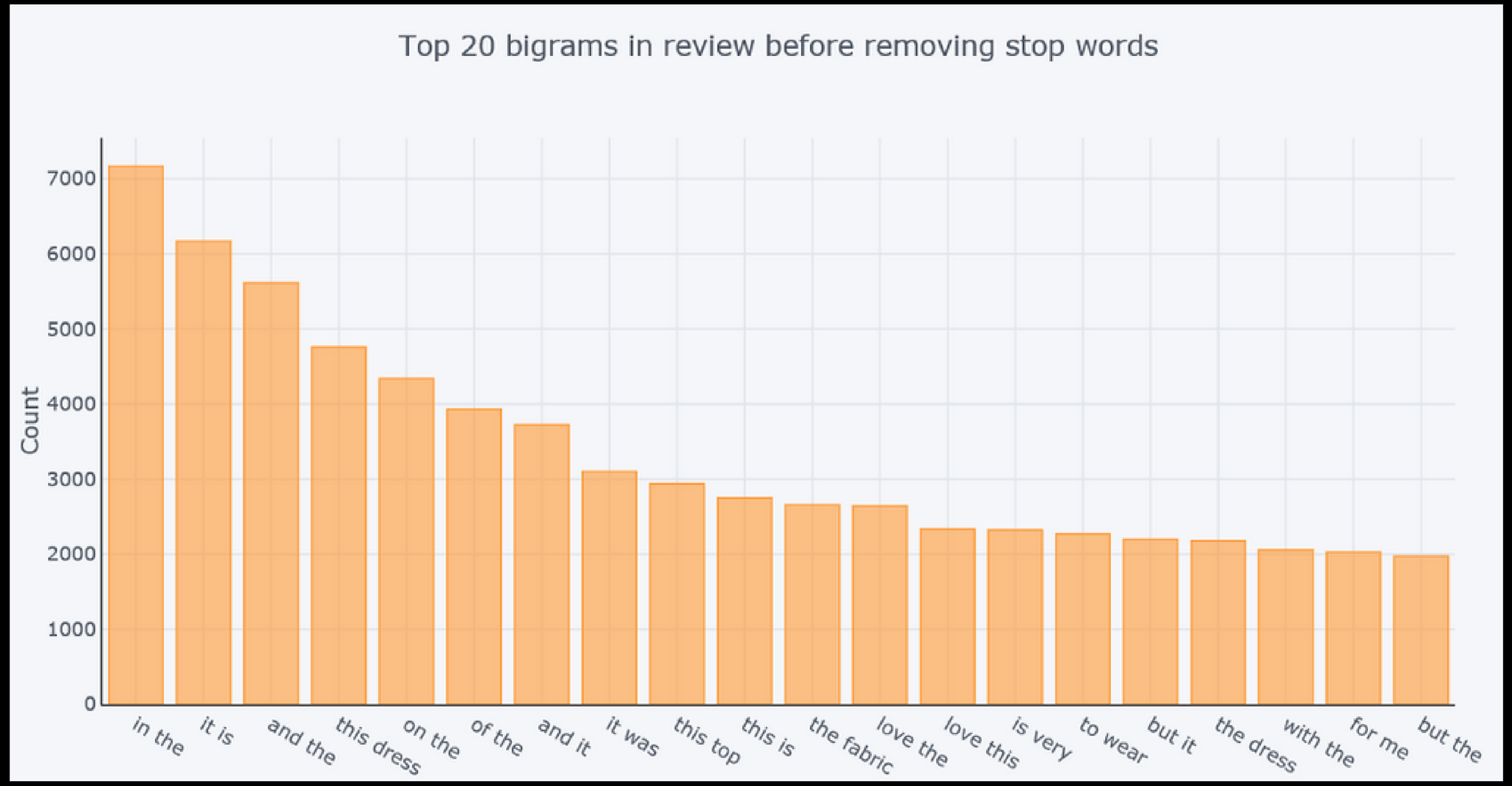
UNIGRAMS BEFORE REMOVING STOPWORDS



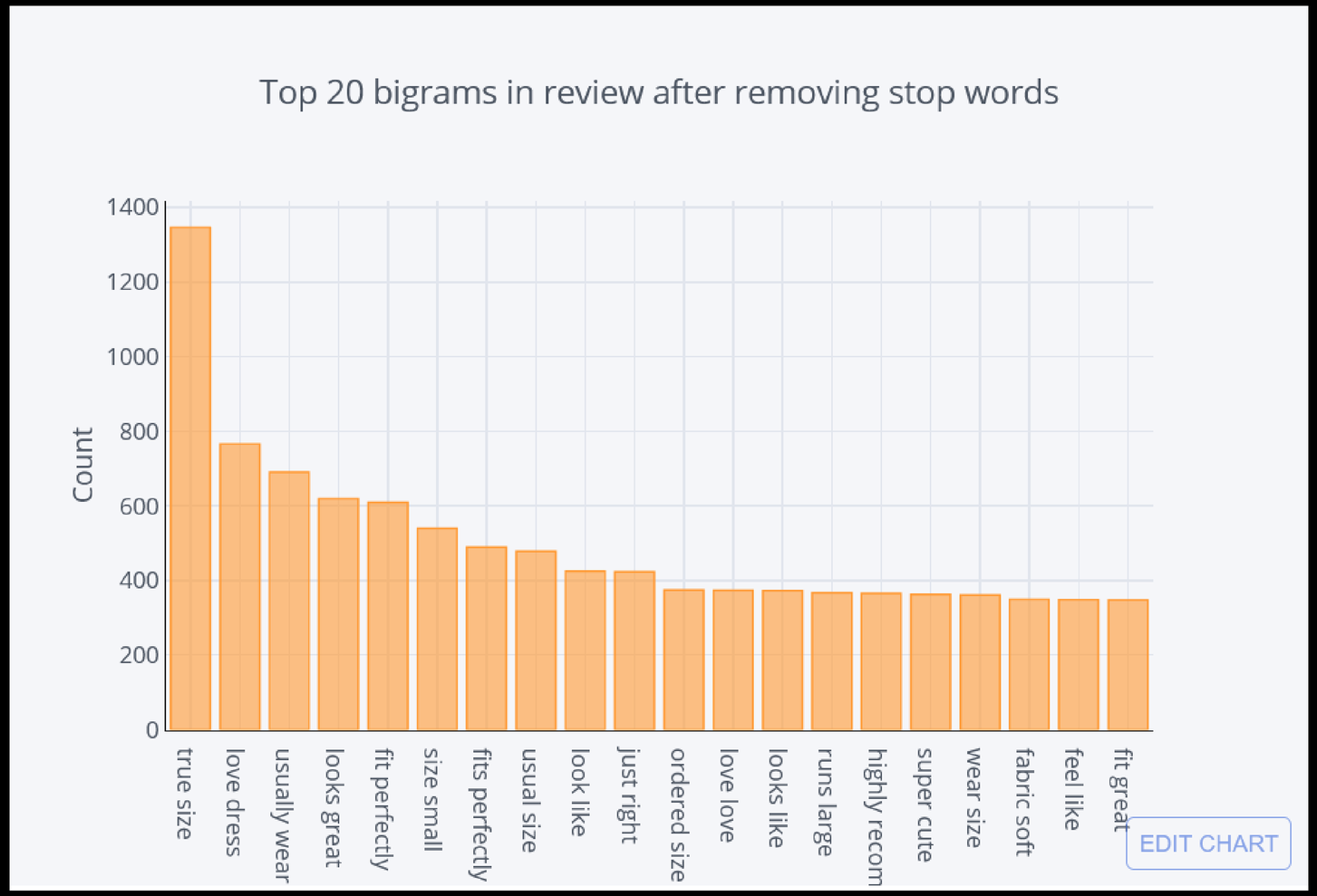
UNIGRAMS AFTER REMOVING STOPWORDS



BIGRAMS BEFORE REMOVING STOPWORDS

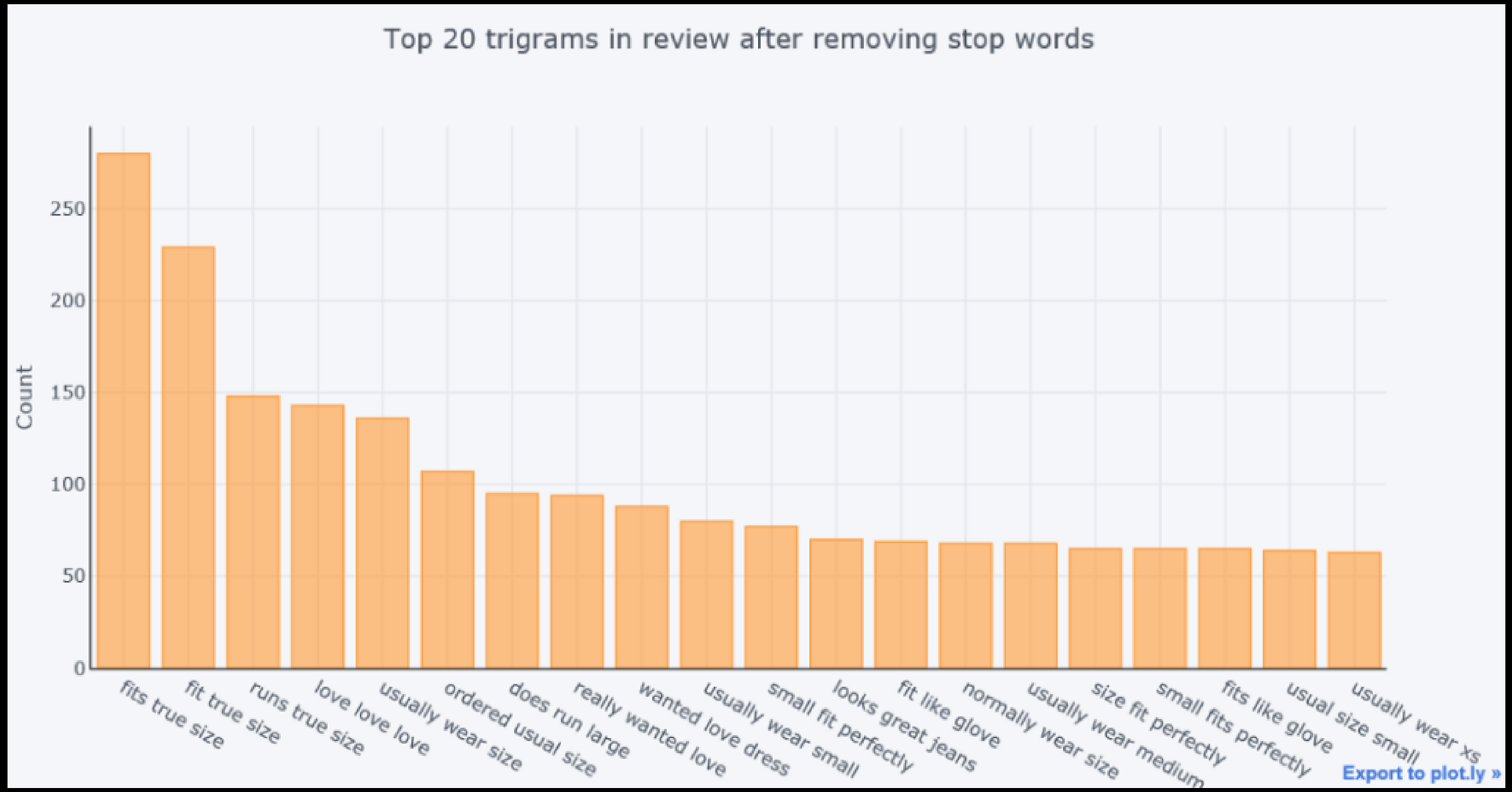


BIGRAMS AFTER REMOVING STOPWORDS

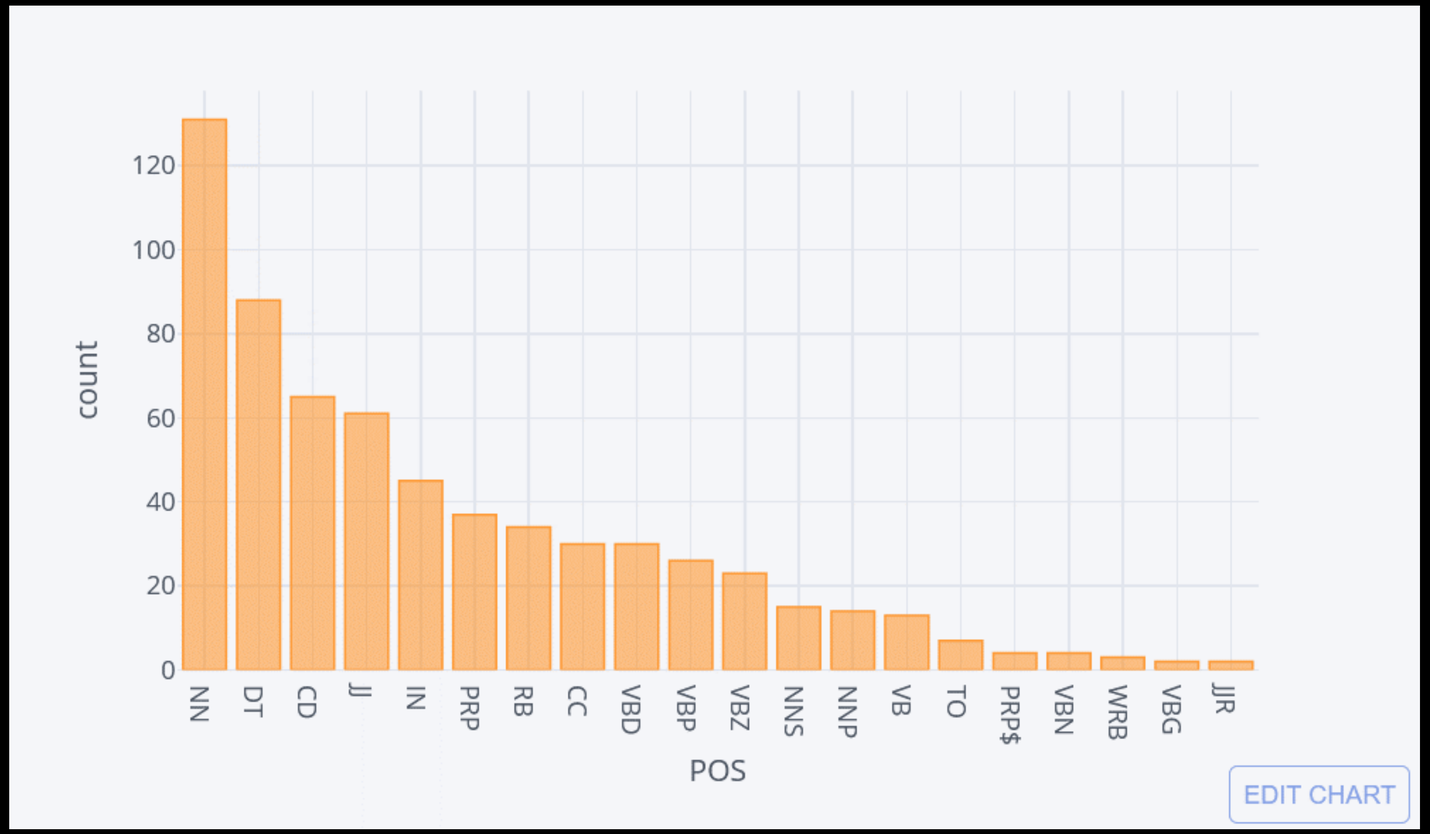


TRIGRAMS BEFORE REMOVING STOPWORDS

TRIGRAMS AFTER REMOVING STOPWORDS

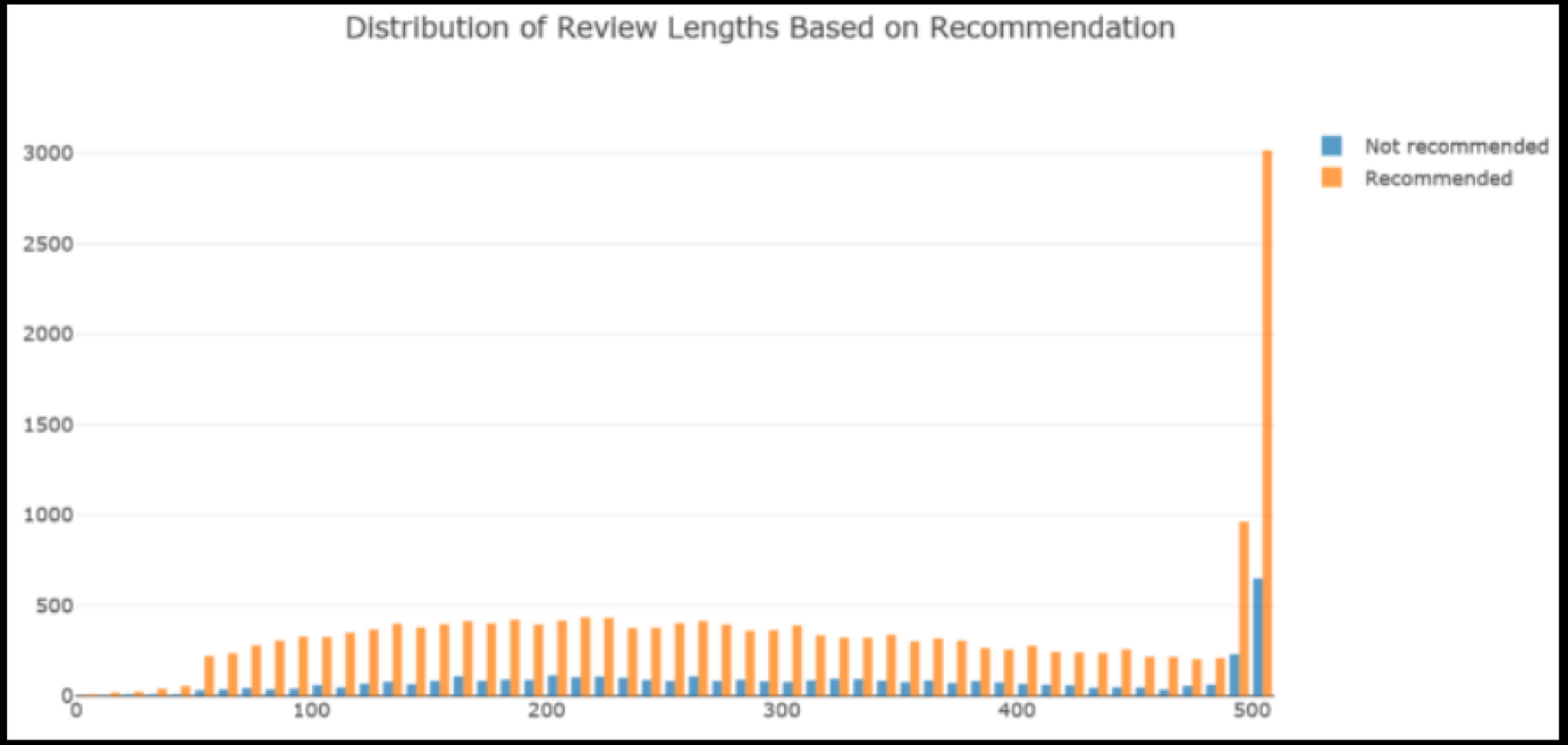


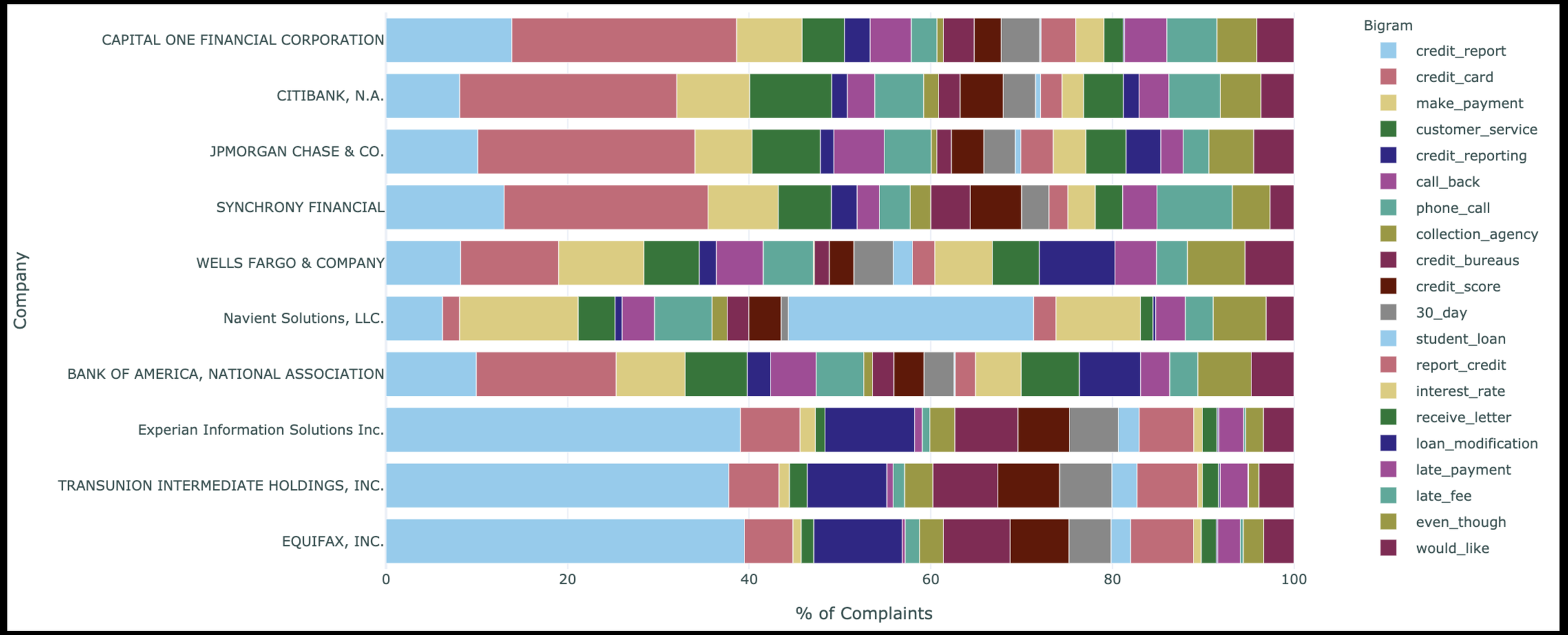
COUNTS OF PART-OF-SPEECH TAGS USING TEXTBLOB



See also:
https://textblob.readthedocs.io/en/dev/quickstart.html

COMBINATION OF POSSIBLY RELEVANT VARIABLES





Hwang, J. P. (2020, March 30). NLP visualisations for clear, immediate insights into text data and outputs. *Plotly*. https://medium.com/plotly/nlp-visualisations-for-clear-immediate-insights-into-text-data-and-outputs-9ebfab168d5b

TODOS UNTIL NEXT WEEK

- Complete chapter 6 (The Tokenizers Library) of the Hugging Face course

- Dataset Characteristics:
  Write down the specifics of how your data was collected and investigate potential biases, imbalance, or outliers in your data