18.05.22

OPENCAMPUS.sh

Transformers for Natural Language Processing and Beyond

THE DATASETS LIBRARY

- Quiz
- Literature Review
- Visualizing Sequences
- Dataset Characteristics

QUIZ

https://forms.office.com/r/MrRv71W2wC

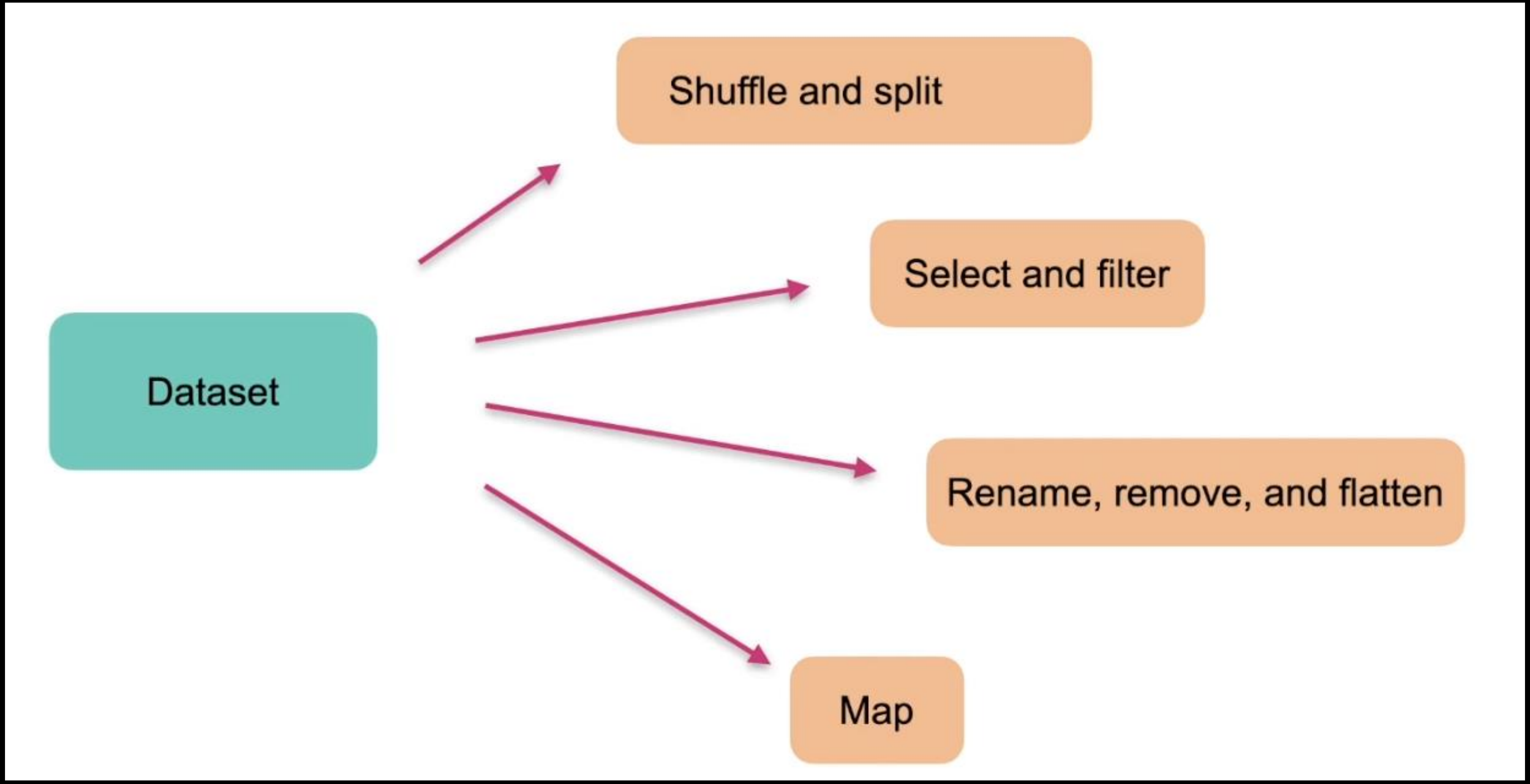
IMPORTING DATA

- **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



Dataset

Shuffle and split

Select and filter

Rename, remove, and flatten

Map

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=callbacks)

# 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?

PROJECT MILESTONES

- 11.05. Form project groups
- 18.05. Literature review
- 25.05. Dataset characteristics
- 01.06. Baseline model
- 08.06. Model & model evaluation (Joint Coding)
- 15.06. 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?

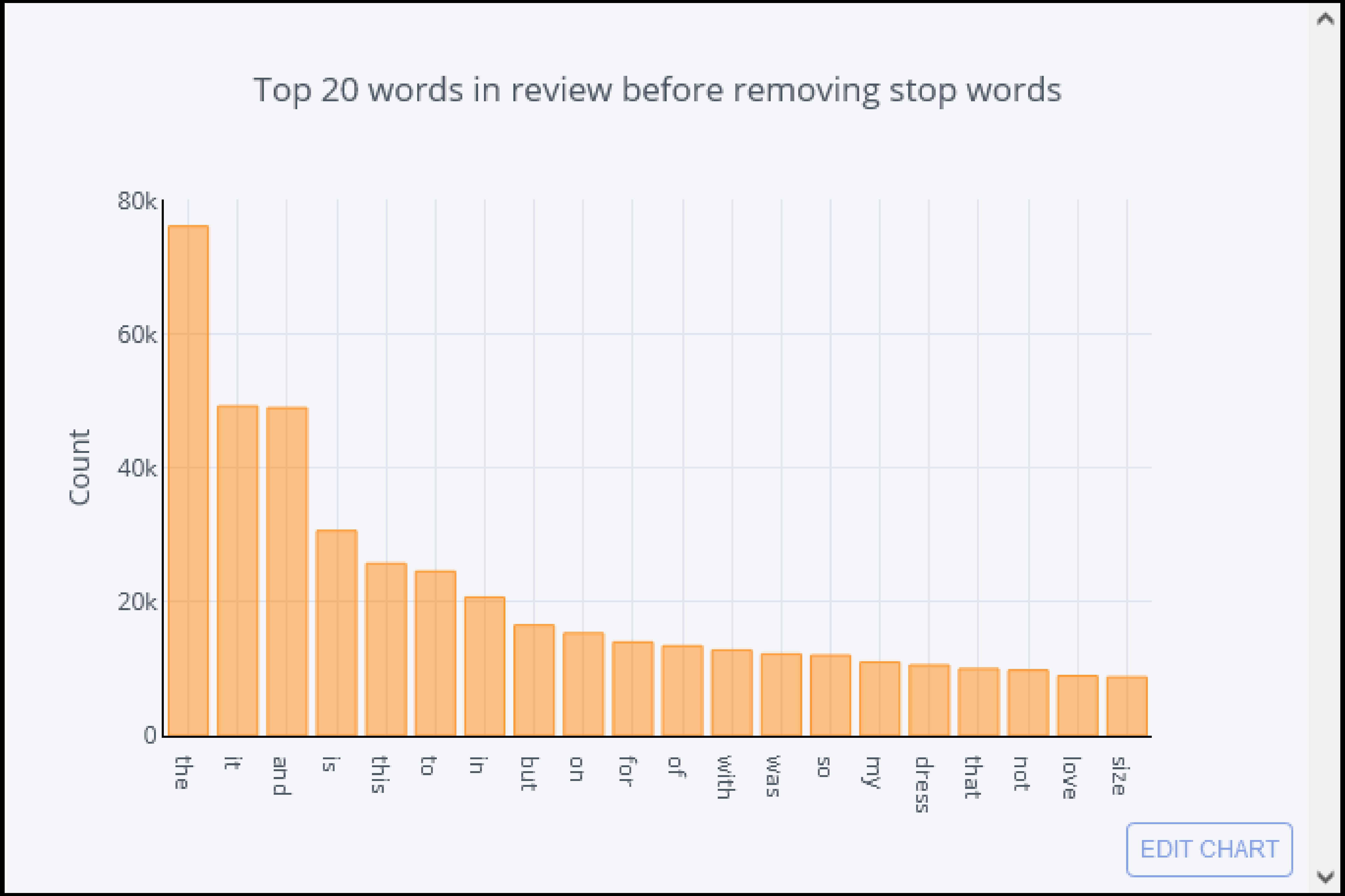
TODOS UNTIL NEXT WEEK

- **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

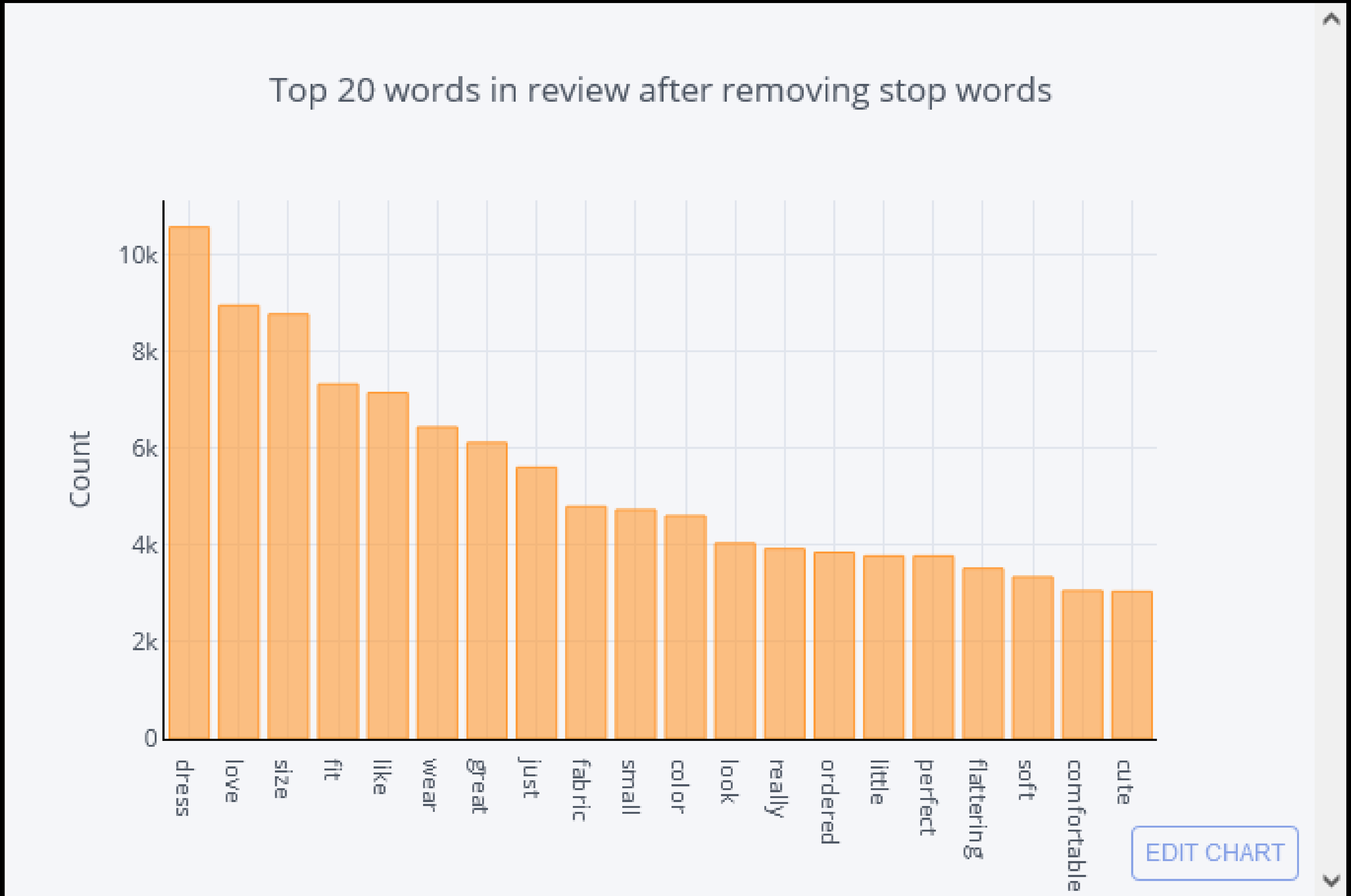
DATA VISUALIZATION

- 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

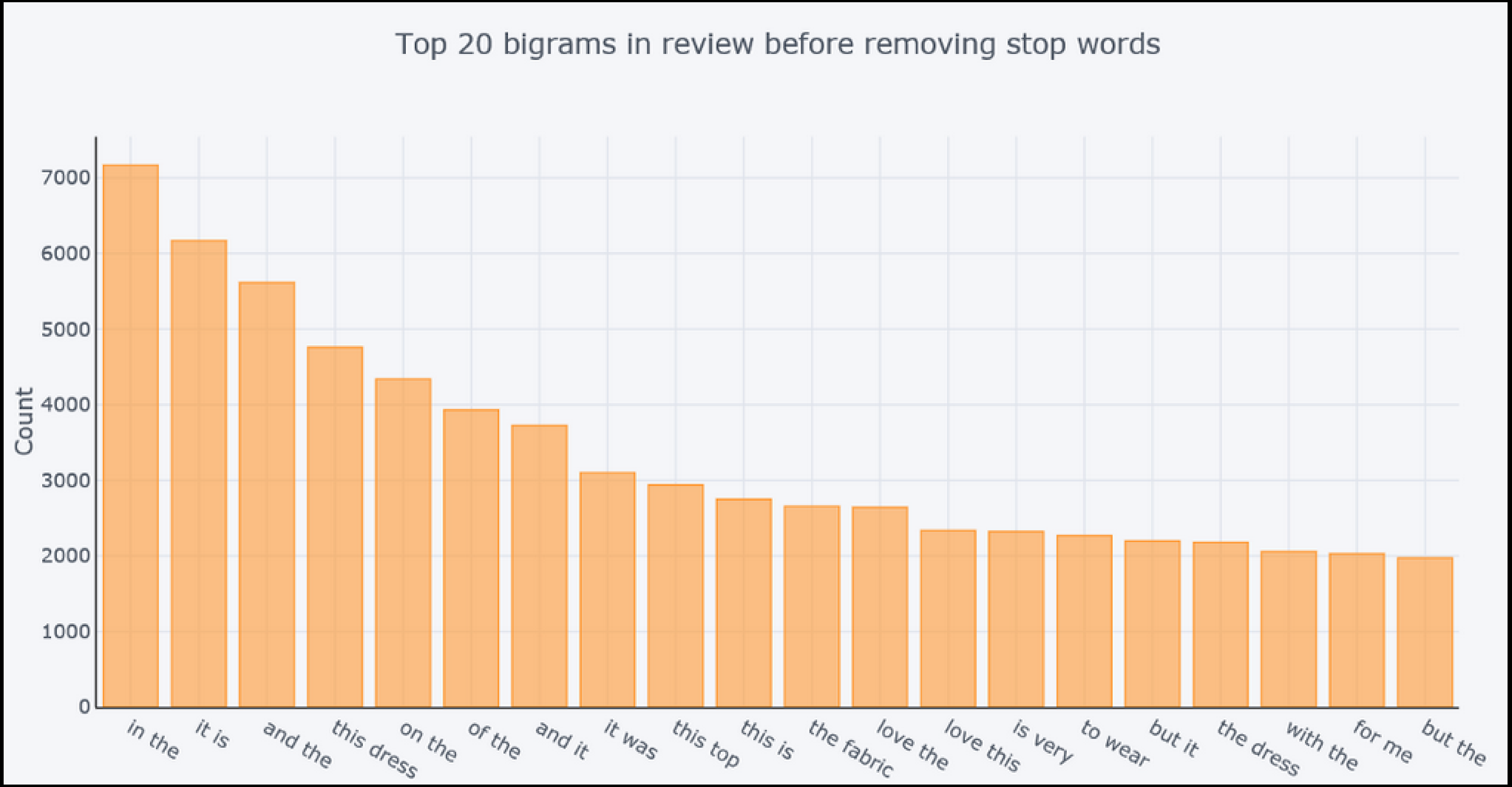
UNIGRAMS BEFORE REMOVING STOPWORDS



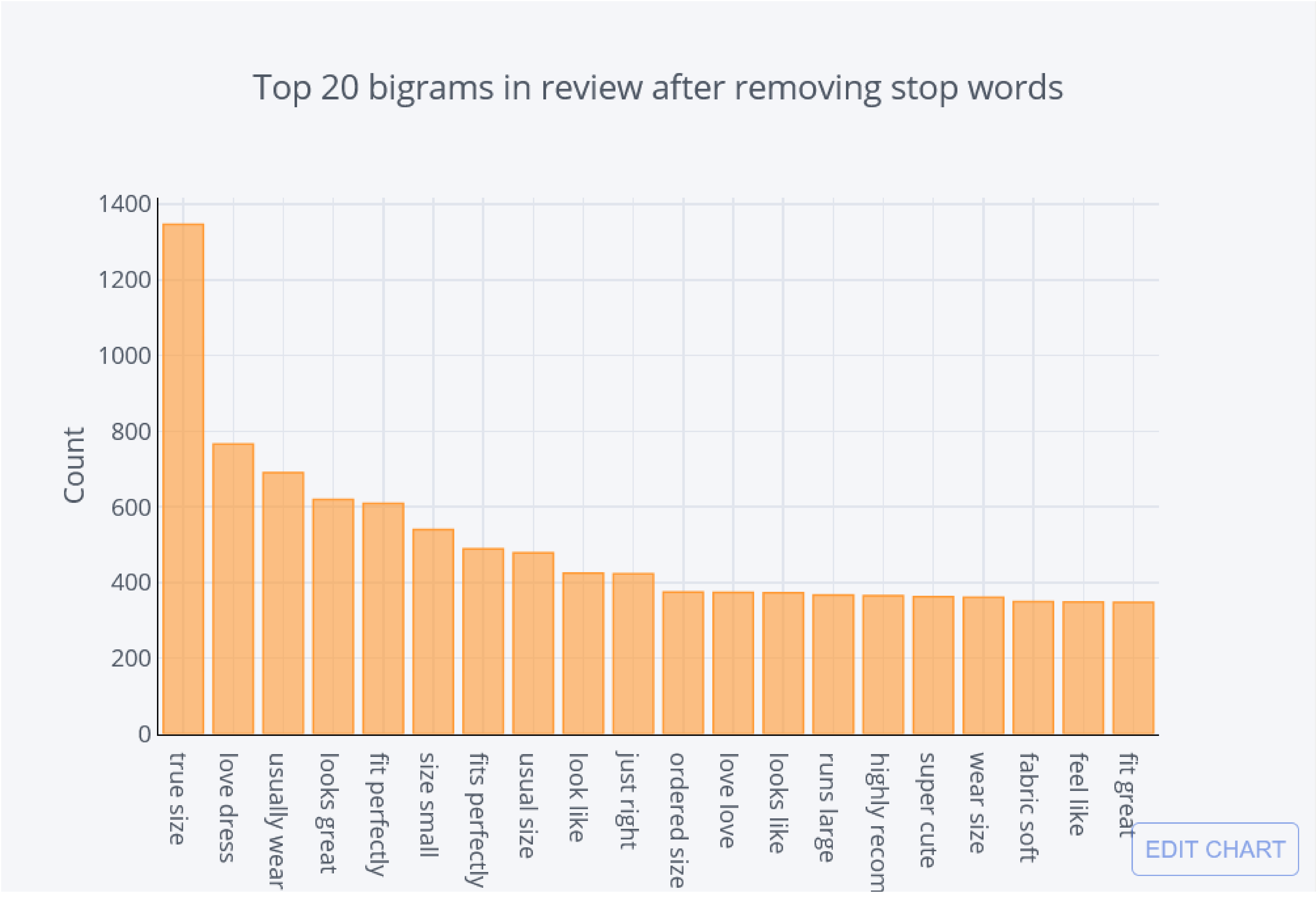
UNIGRAMS AFTER REMOVING STOPWORDS



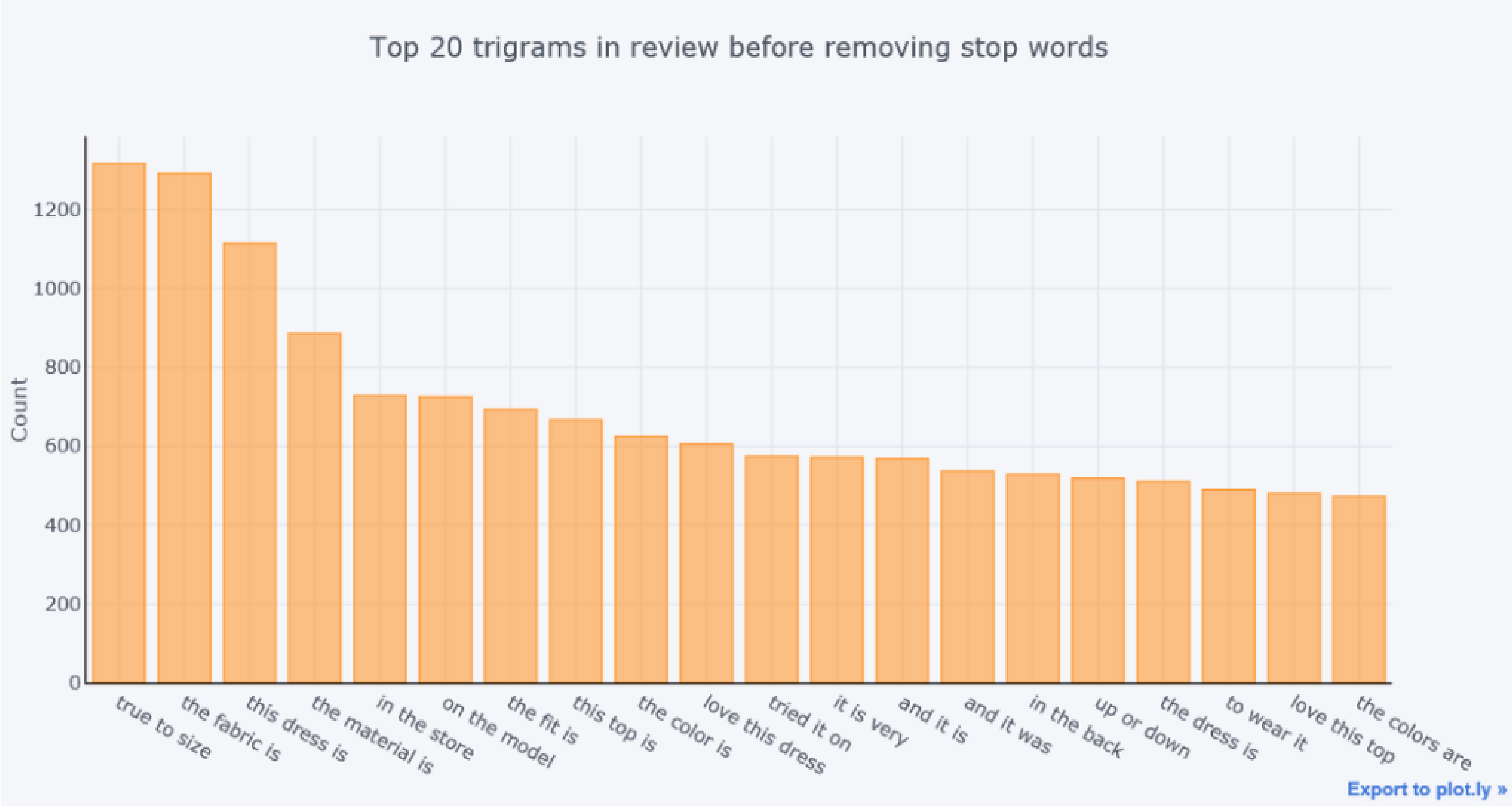
BIGRAMS BEFORE REMOVING STOPWORDS



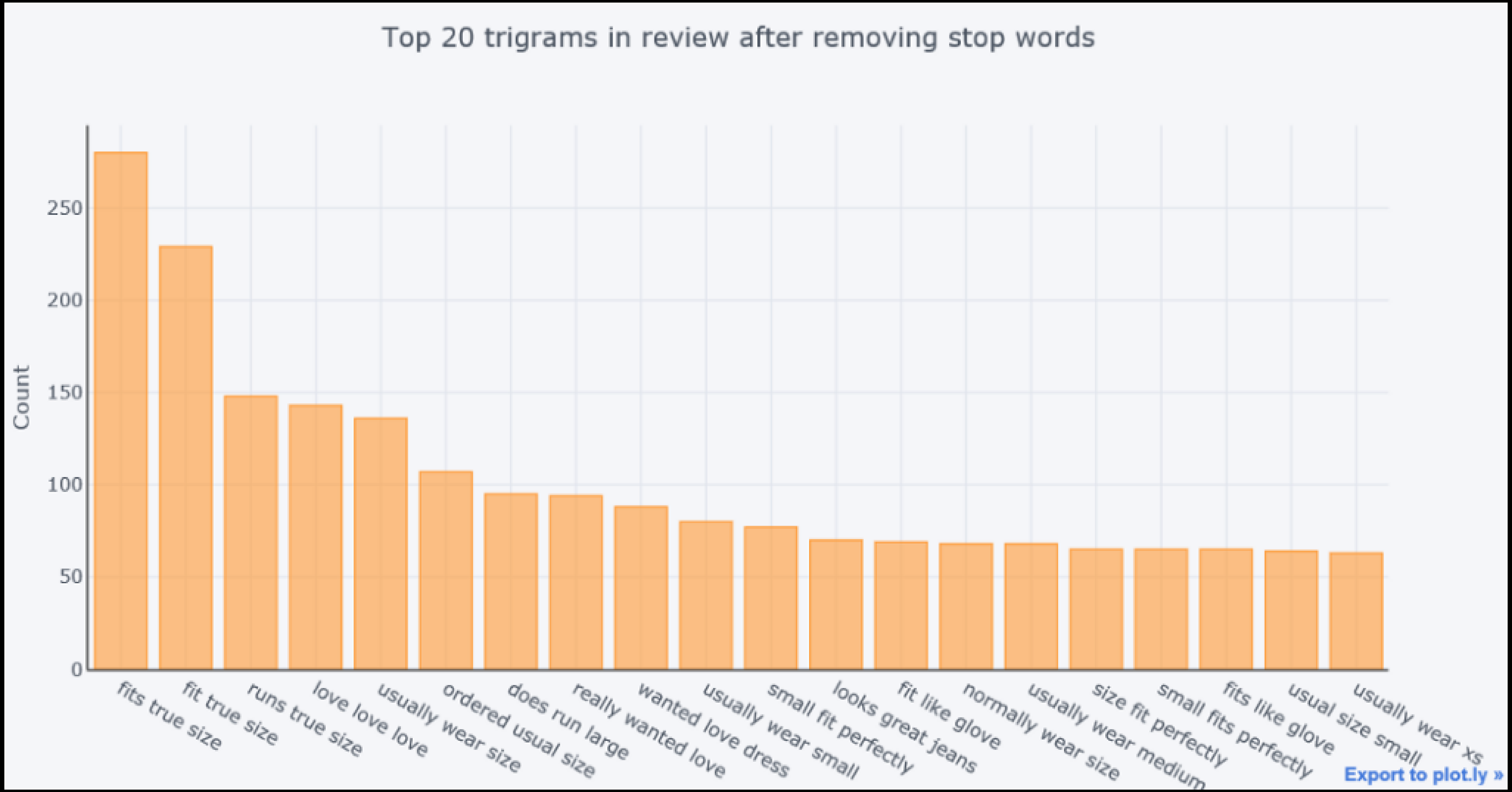
BIGRAMS AFTER REMOVING STOPWORDS



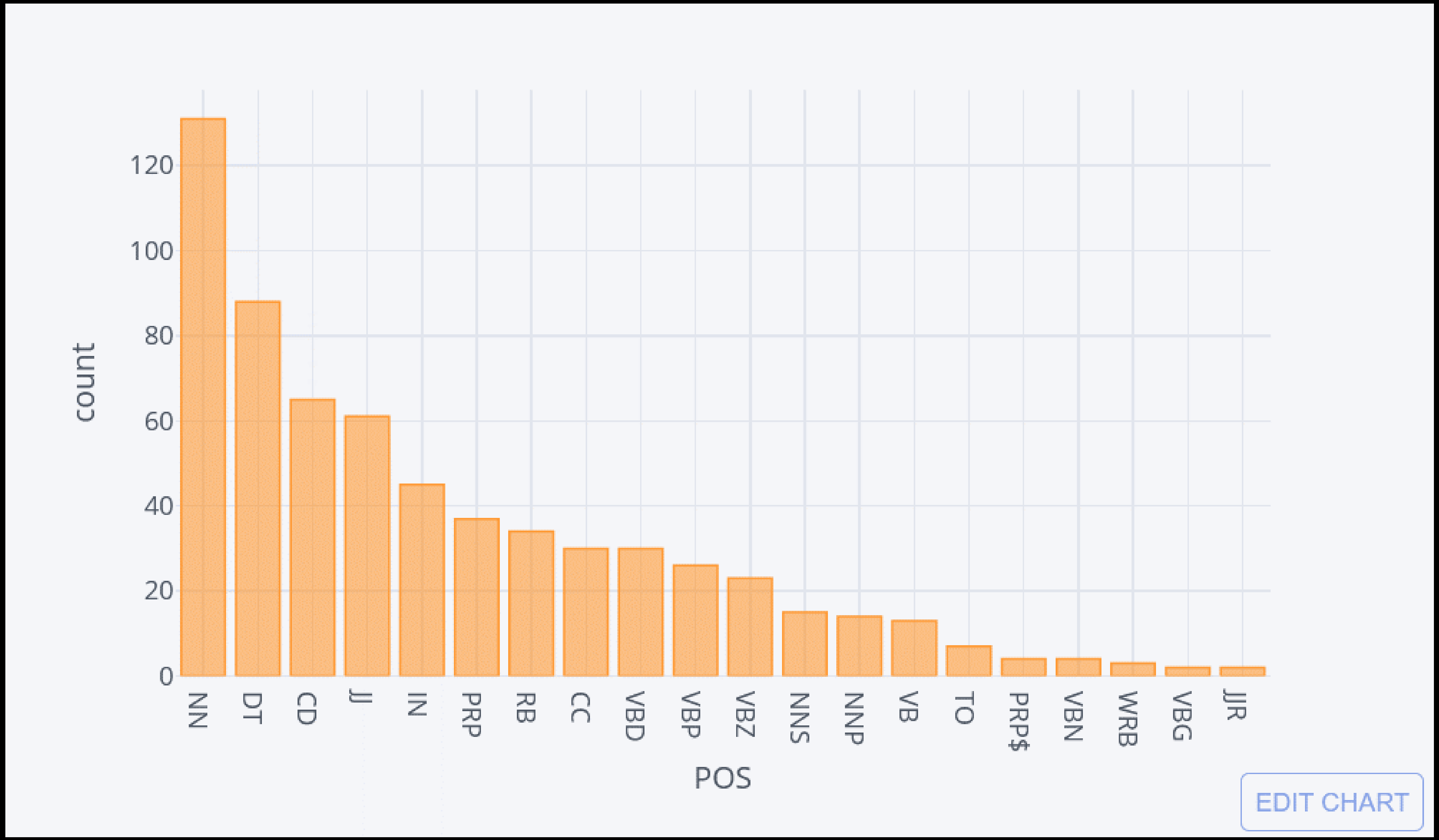
TRIGRAMS BEFORE REMOVING STOPWORDS



TRIGRAMS AFTER REMOVING STOPWORDS

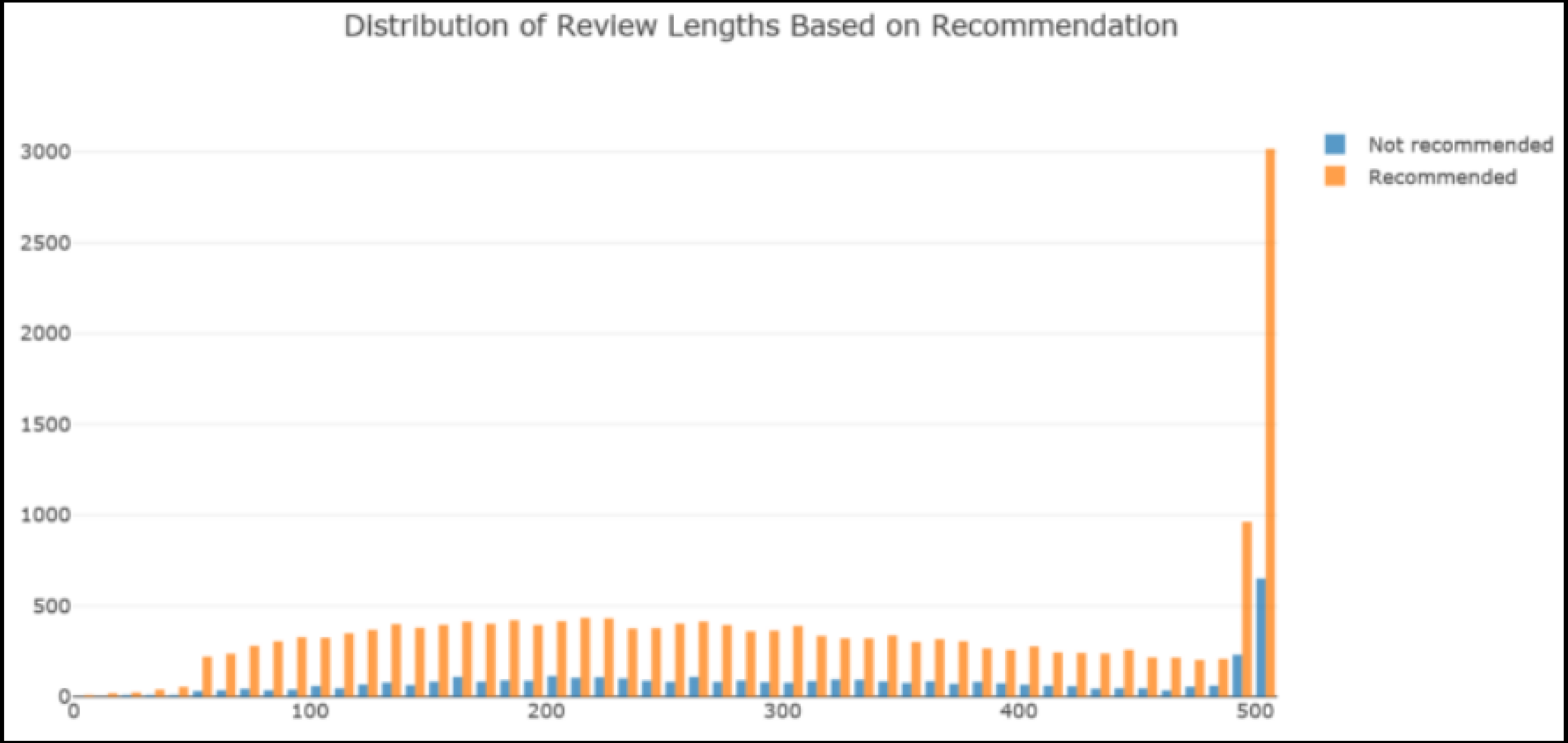


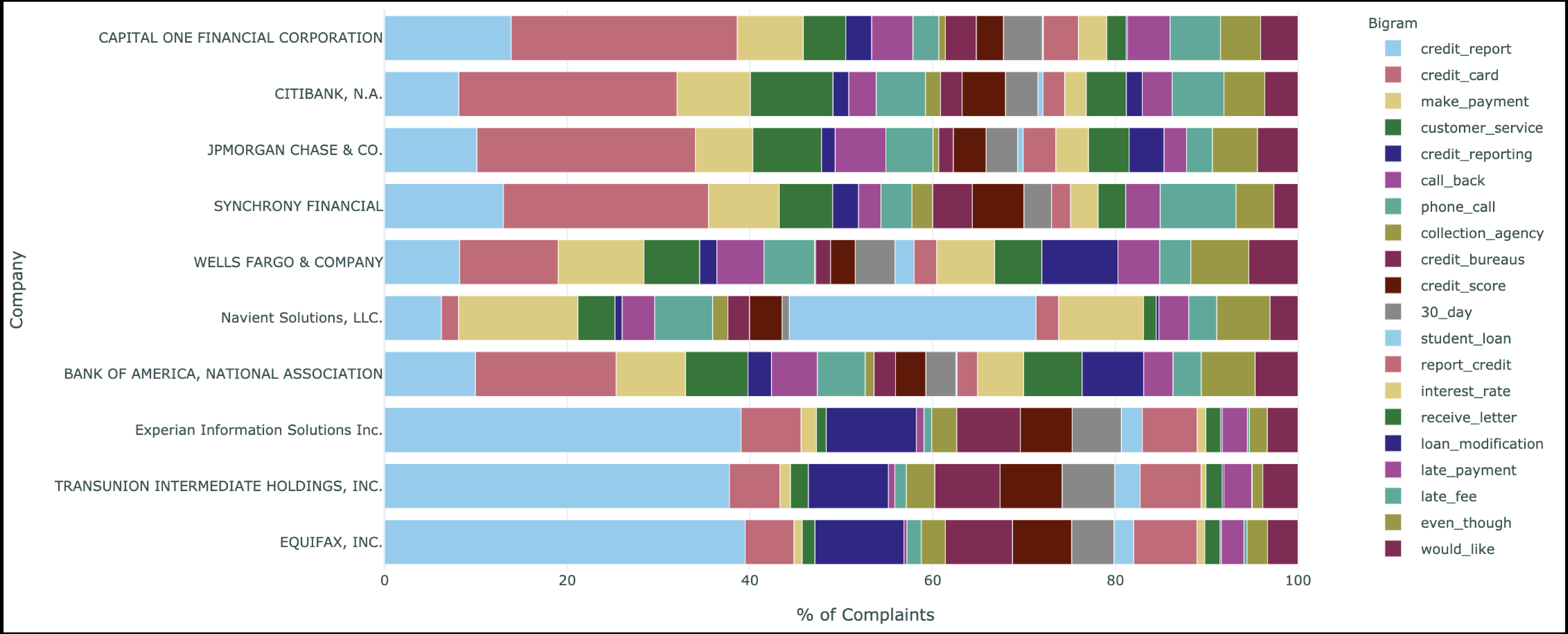
COUNTS OF PART-OF-SPEECH TAGS USING TEXTBLOB



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

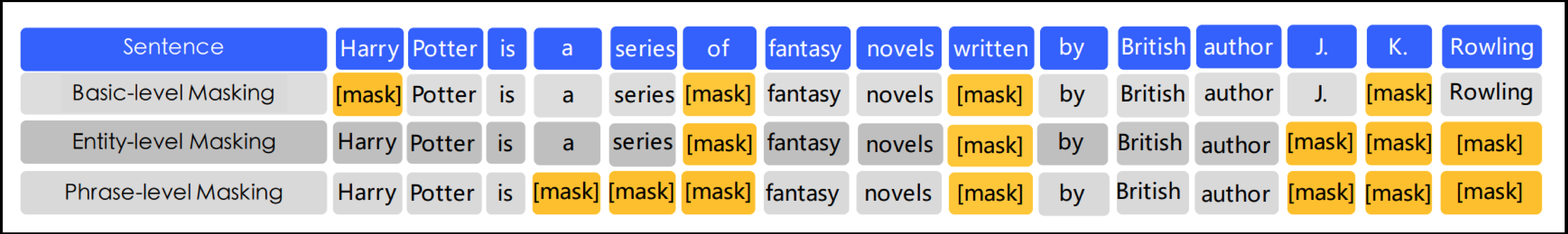
COMBINATION OF POSSIBLY RELEVANT VARIABLES





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

MASKING LEVELS



Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

Liu, B. (n.d.). NLP Pretraining—From BERT to XLNet. Retrieved December 15, 2020, from Title website: https://bangliu.github.io/survey/2019/07/01/NLP-Pretraining/