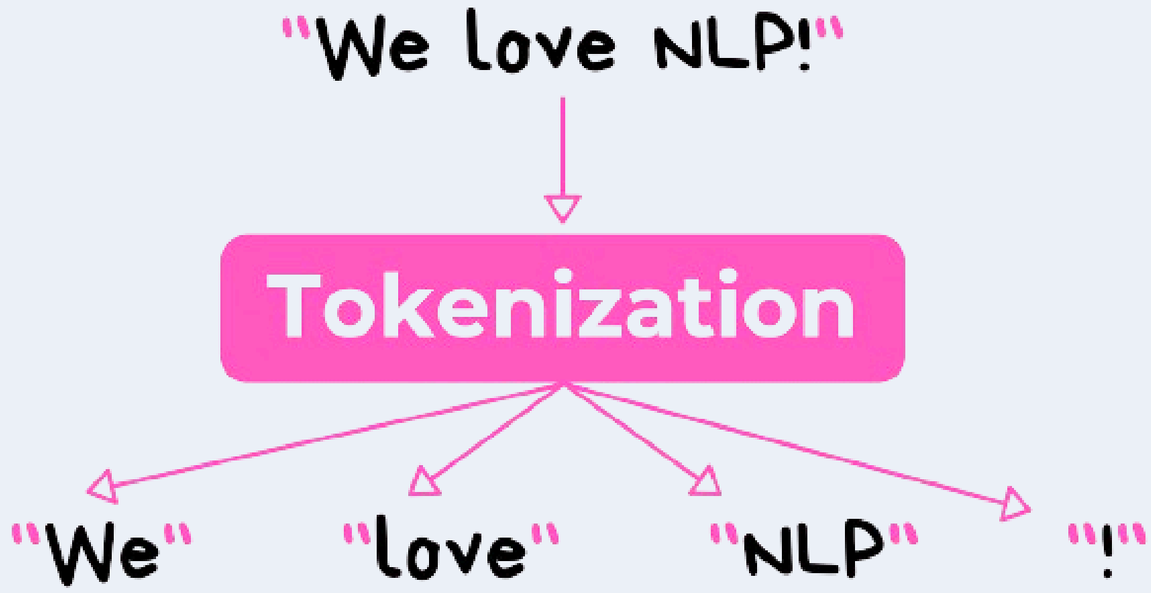
ARMENIAN TOKENIZER

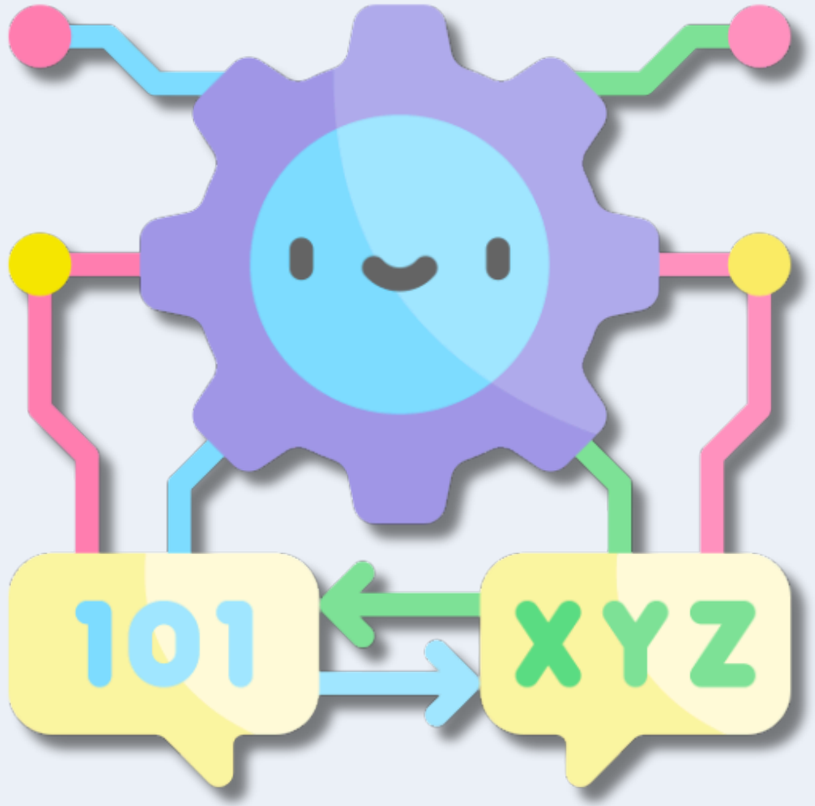
A tokenization system for the Armenian language



Presented by: Anna Shaljyan, Naira Maria Barseghyan

OUR PROJECT'S OBJECTIVE

- This project aims to develop a tokenization system for the Armenian language and train it using a corpus derived from Armenian Wikipedia.
- The project will explore various tokenization methods that are suitable for the nuances of the Armenian language.
- The ultimate goal is to create a robust tool that can accurately segment Armenian text into tokens.



TIKTOKEN

- cl100k_base
- p50k_base

→ 1. LOAD MODEL

2. ENCODE

3. DECODE

- SimpleBytePairEncoding

CORPUS

VOCAB_SIZE

PAT_STR

Tiktoken library For Tokenization In OpenAI API

```
['T', 'ik', 'token', ' library', ' For', ' Token', 'ization', ' In',
' Open','AI', ' API']
```

OpenAI

CUSTOM WORDPIECE

- 30,000 Vocabulary size
- 300,000 Wikipedia Articles

PREPROCESSING

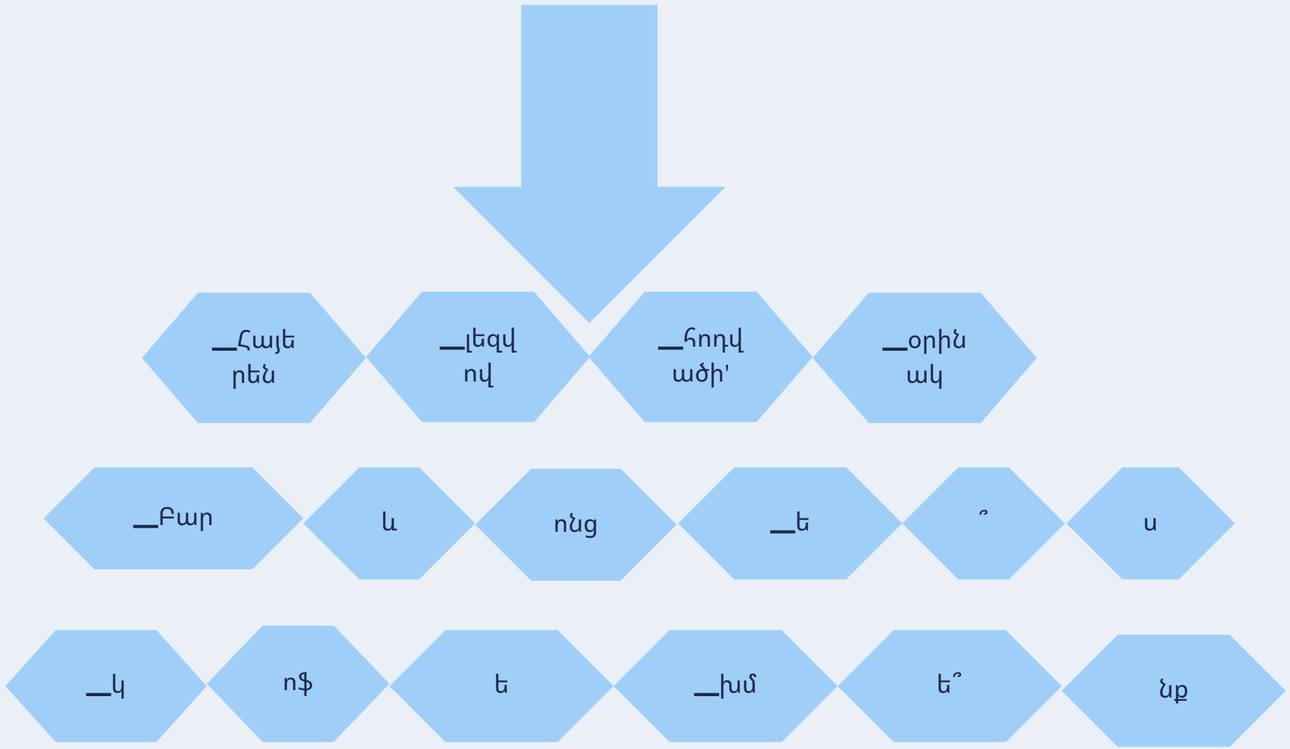
- Text Standardization: Lowercasing whole text
- Whitespace Management
- In word tokens Management

SENTENCEPIECE

- BPE Model
- 30,000 Vocabulary size
- 150,000 Wikipedia Articles
- LLAMA 2 Training setup

Հայերեն լեզվով հոդվածի օրինակ:

Բարև ոնց ե՞ս՝ կոֆե խմե՞նք:



CUSTOM BPE TOKENIZER

- 30,000 Vocabulary size
- 300,000 Wikipedia Articles
- Minimal number of OOV tokens

PREPROCESSING

- Text Standardization and addition of special token to save information.
- HTML Cleanup: Removes HTML entities
- Whitespace Management

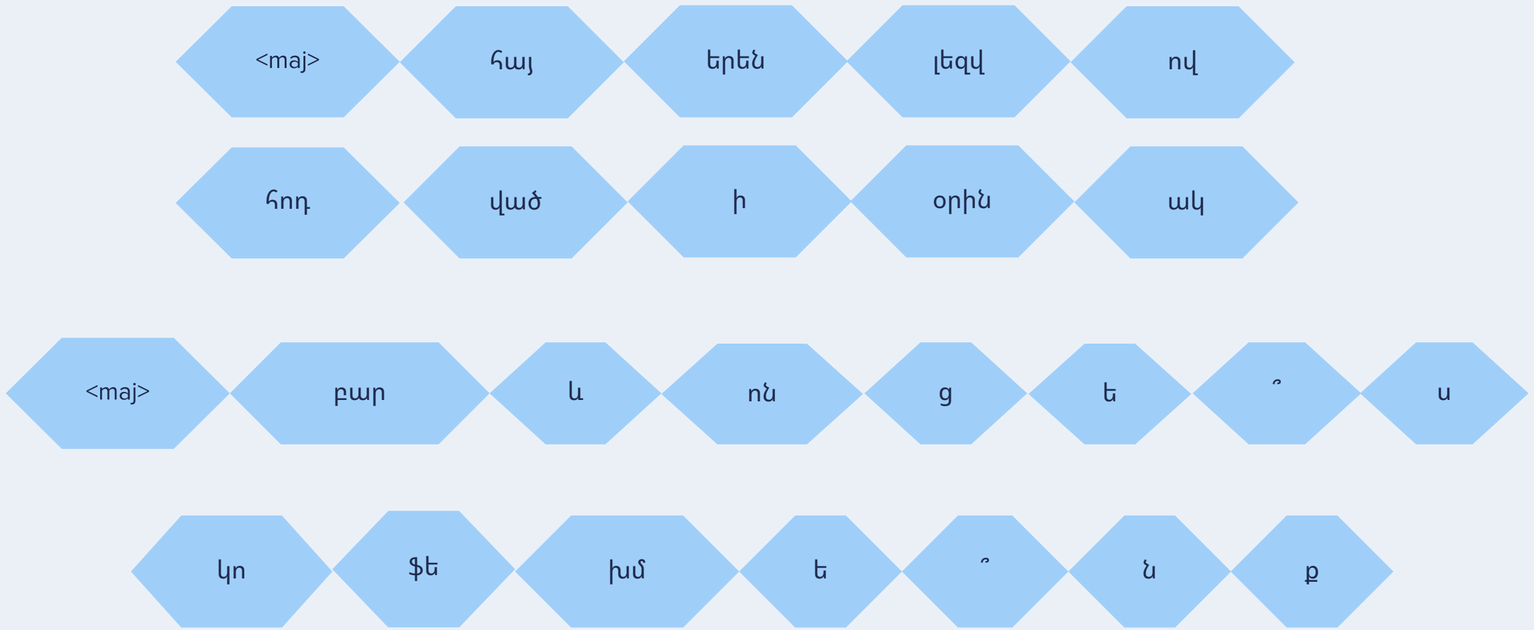
INTERESTING FINDINGS

- All Armenian Prefixes and Suffixes
- Some Armenian Words
- Lack of “Casual” words

CUSTOM BPE TOKENIZER

Հայերեն լեզվով հոդվածի օրինակ։

Բարև ոնց ե՞ս՝ կոֆե խմե՞նք։

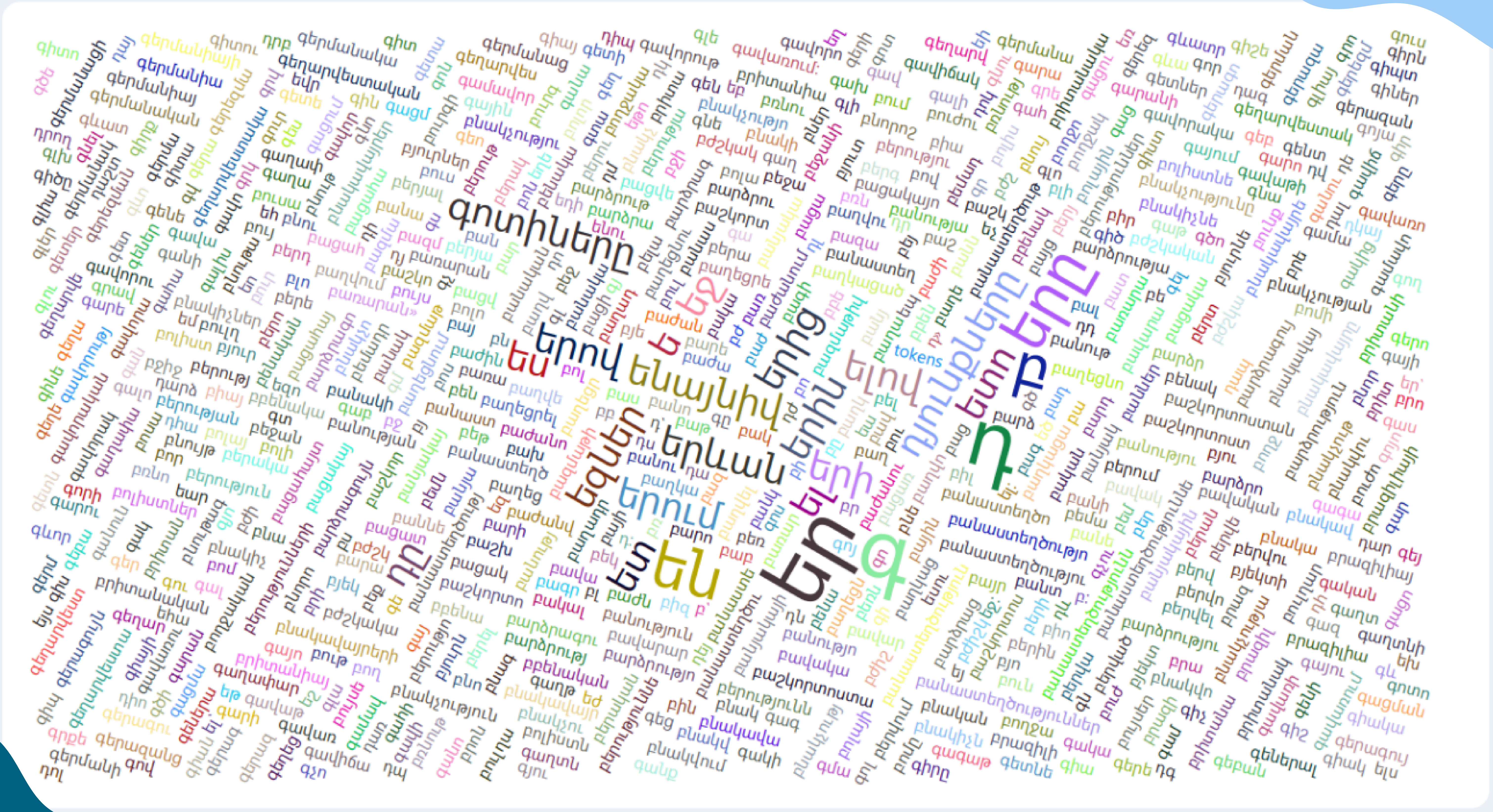


BPE WORDCLOUDS

BPE WORDCLOUDS

WORDPIECE WORDCLOUDS

WORDPIECE WORDCLOUDS



THANK YOU FOR LISTENING!

ANY QUESTIONS?