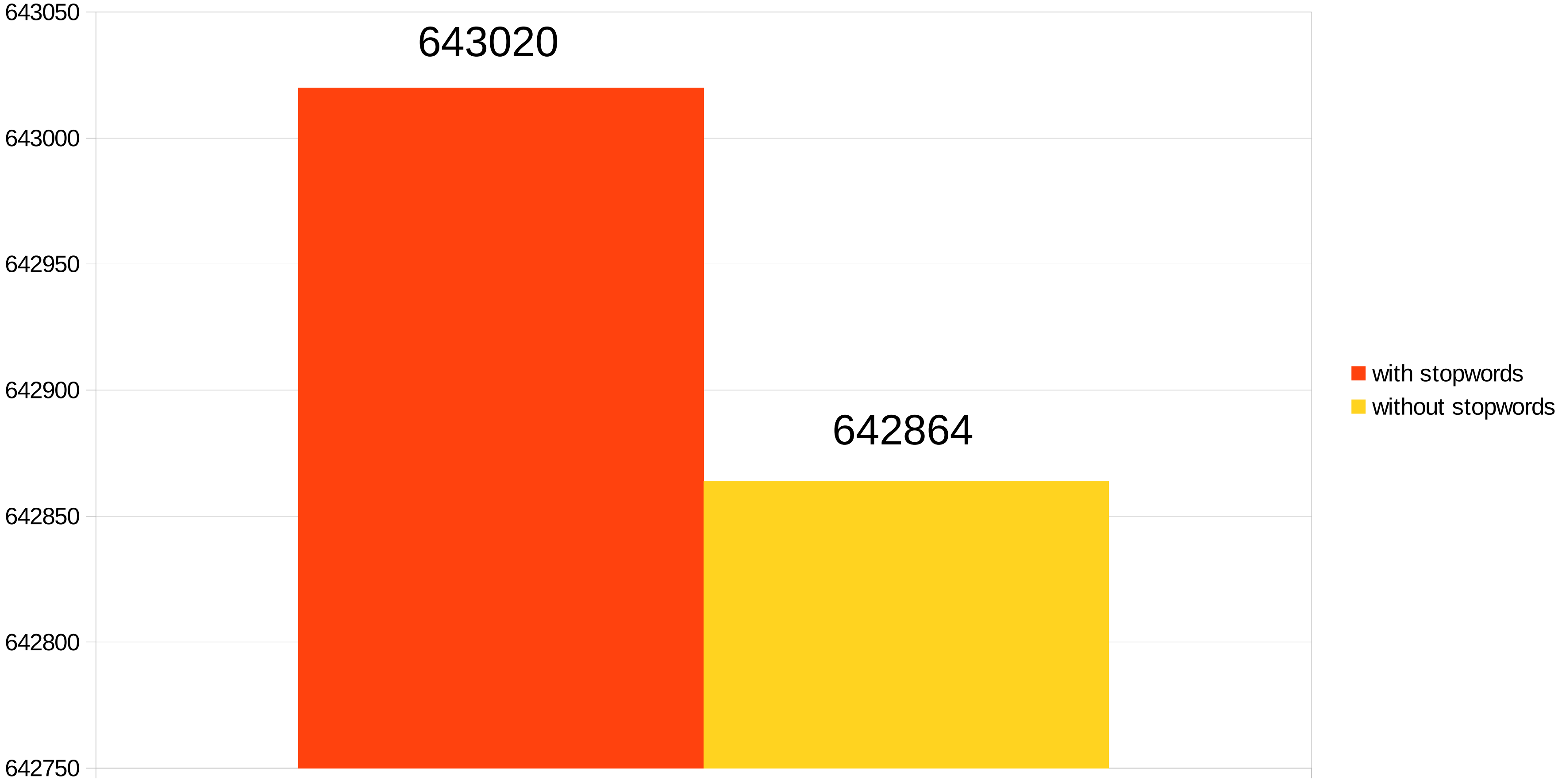
Touché task 1

How many words are in the data set after removing stop words?

Approach

- Concatenate conclusion and premise for every argument.
- Remove URLs beforehand using regex but keep them for possible later purposes.
- Use NLTK python package with its word_tokenizer.
- Use NLTKs built-in stop words.
- Remove all "words" that do not contain a digit or letter.
- Split words, that are of form "WORD.WORD".

Result


643050
643000
642950
642900
642850
642800
642750
643020
642864
with stopwords
without stopwords

Open problems

- Abbreviations (e.g. "n't") are interpreted as unique words.
- Some words are gibbrish (e.g. „jjjjjjjiiiiiiiiiiiiaa“)
- Some words contain mostly special characters (e.g. „==========pro“)