Lesson 37: Text Preprocessing

Data Science with Python – BSc Course

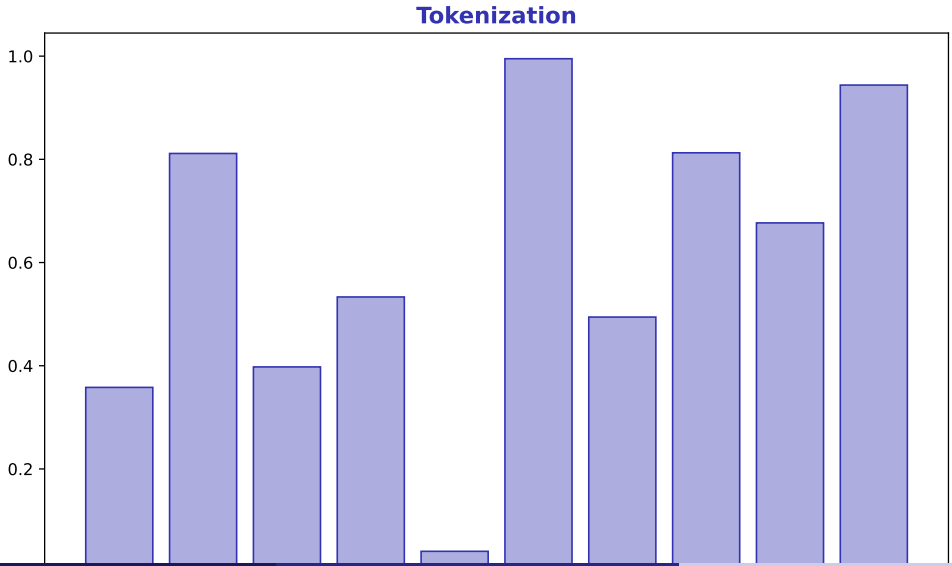
45 Minutes

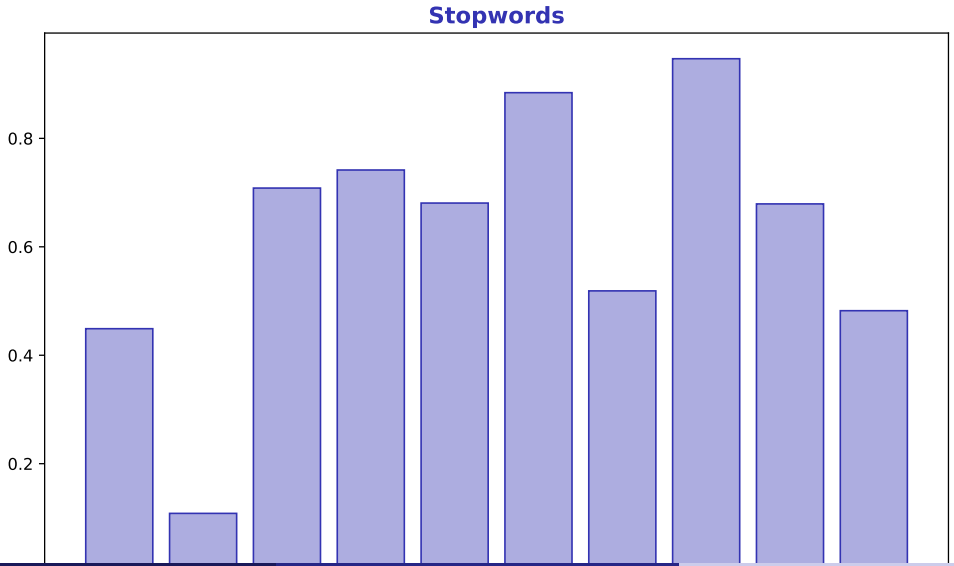
Learning Objectives

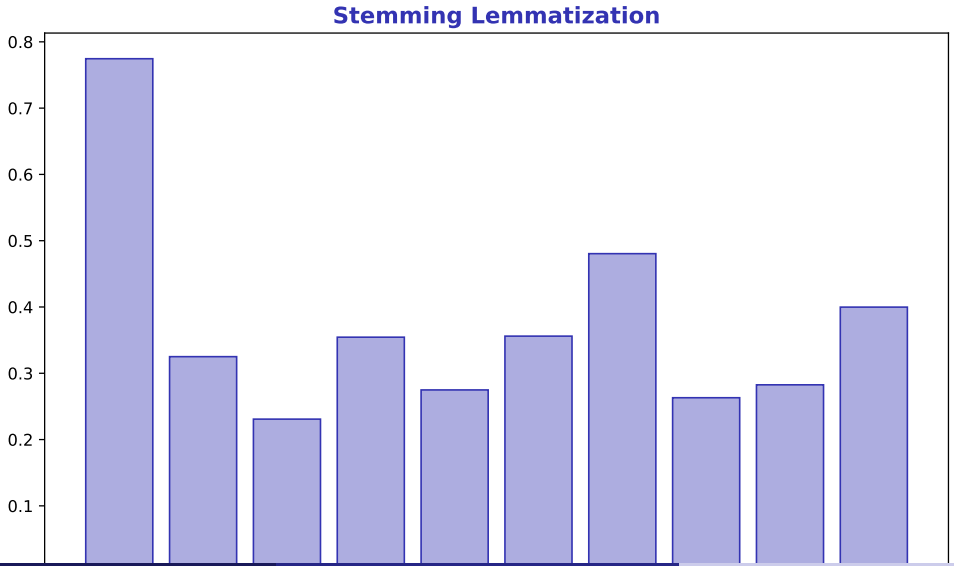
After this lesson, you will be able to:

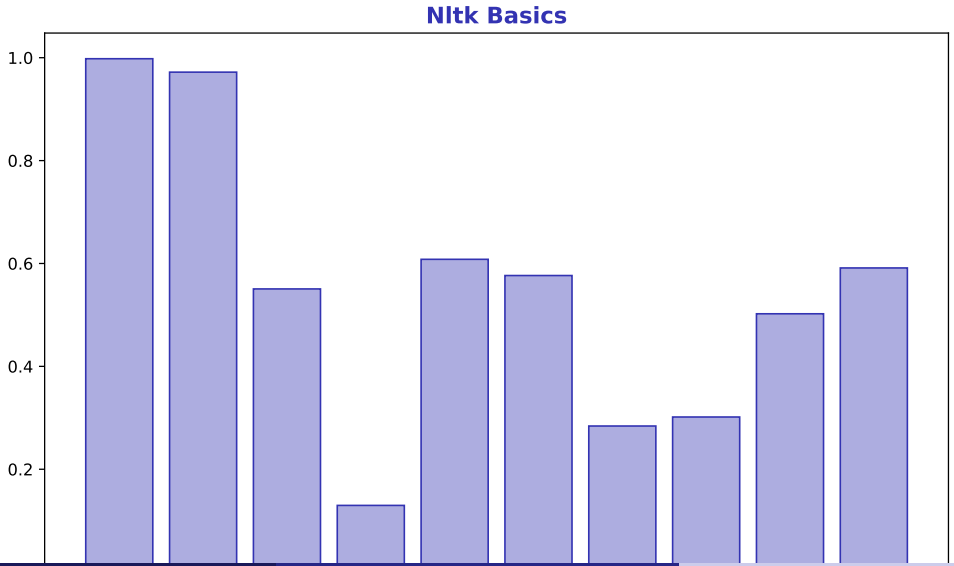
- Tokenize text
- Remove stopwords
- Apply stemming/lemmatization
- Clean financial text

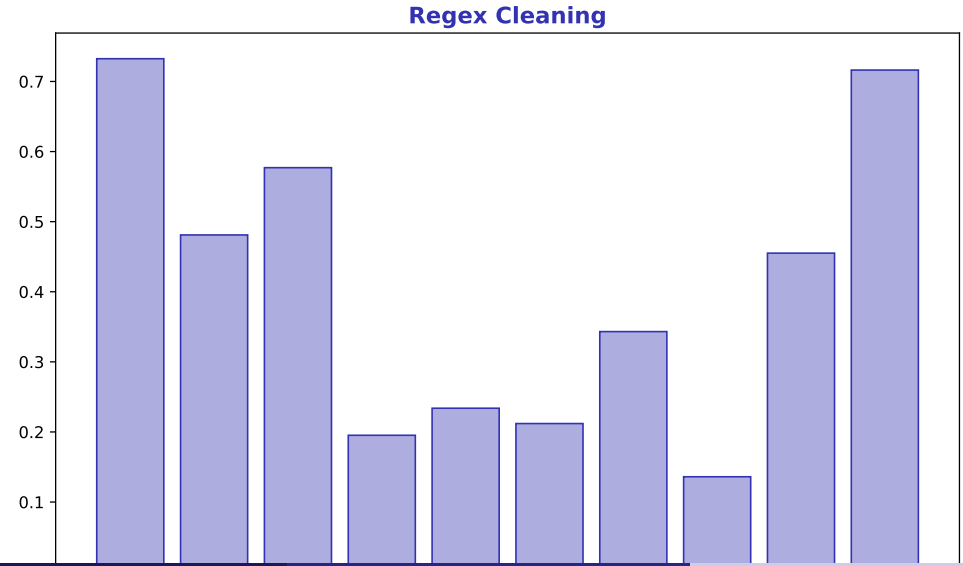
Building towards your final project

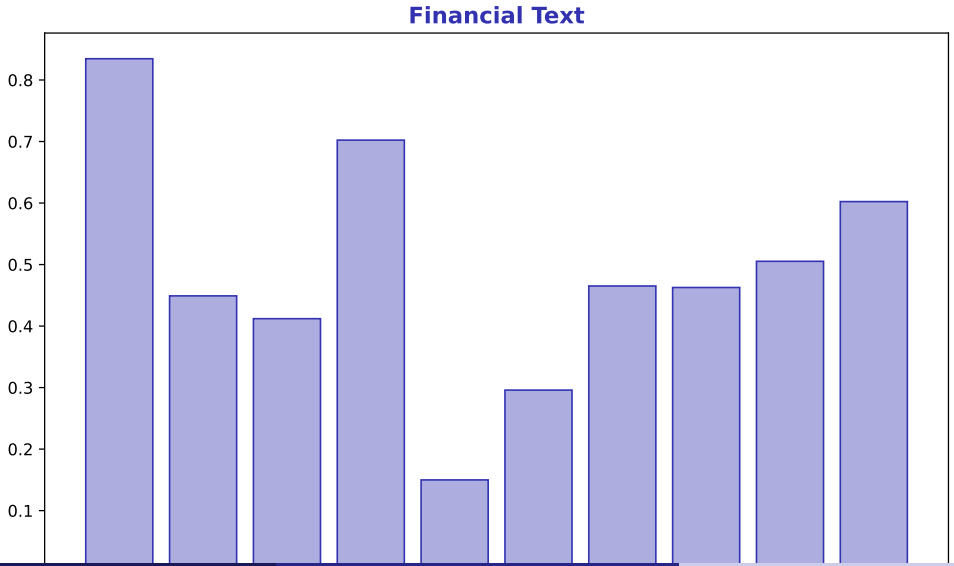

Tokenization
1.0
0.8
0.6
0.4
0.2

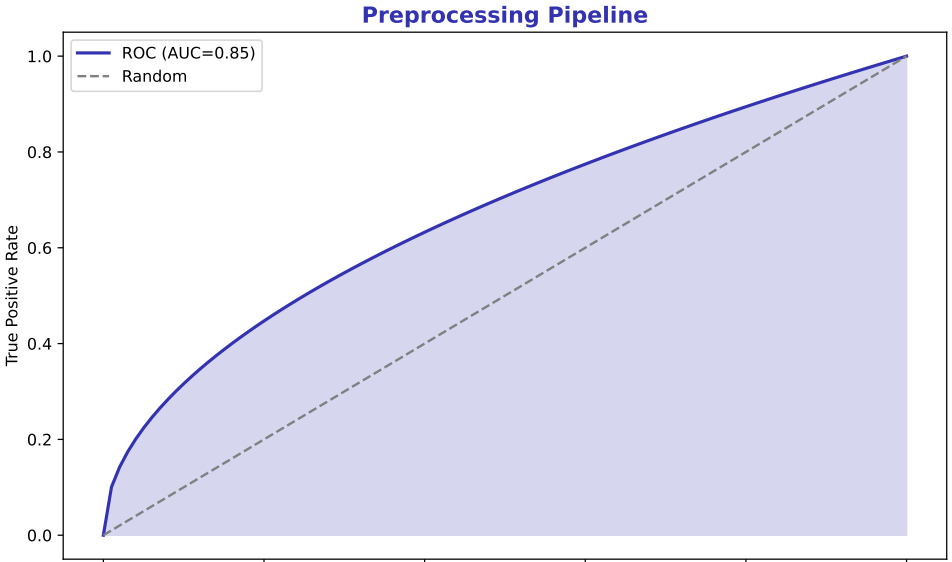

Stopwords
0.8
0.6
0.4
0.2

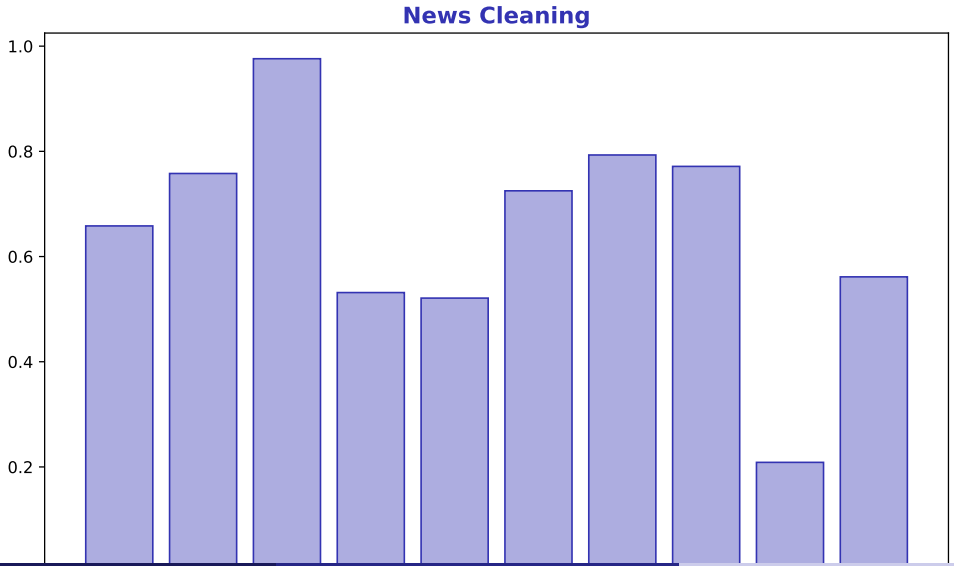

Stemming Lemmatization
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


NItk Basics
1.0
0.8
0.6
0.4
0.2


Regex Cleaning
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Financial Text
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


Preprocessing Pipeline
ROC (AUC=0.85)
Random
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0


News Cleaning
1.0
0.8
0.6
0.4
0.2

Lesson Summary

Key Takeaways:

- Tokenize text
- Remove stopwords
- Apply stemming/lemmatization
- Clean financial text

Apply these skills in your final project