```
# 5. Comparing tokenizers (optional for BS students I love bones
and also the Icelandic quotation marks!s email is johnny@hi.is
and his website is http://johnny_and_max.is.
"""

# Custom tokenizer


def custom_tokenizer(text):
   token_pattern = r"\b\w+\b|[.,!?;:/()-]"  # Matches words and
punctuation
   tokens = re.findall(token_pattern, text)
   return tokens


# Tokenize using NLTK word_tokenize
try:
   from nltk import download
   download("punkt")  # Ensure punkt is downloaded for
word_tokenize
   nltk_tokens = word_tokenize(text)
except:
   nltk_tokens = wordpunct_tokenize(text)  # Fallback if
word_tokenize fails

# Tokenize using NLTK's wordpunct_tokenize
wordpunct_tokens = wordpunct_tokenize(text)

# Tokenize using the custom tokenizer
custom_tokens = custom_tokenizer(text)

# Compare tokenized outputs
print("NLTK word_tokenize tokens:")
print(nltk_tokens)
print("\nNLTK wordpunct_tokenize tokens:")
print(wordpunct_tokens)
print("\nCustom tokenizer tokens:")
print(custom_tokens)

# Analyze forksjell


def compare_tokenizers(tokenizer_1, tokenizer_2):
   return set(tokenizer_1).symmetric_difference(set(tokenizer_2))
```

```
differences = compare_tokenizers(nltk_tokens, custom_tokens)
print("\nDifferences between NLTK and custom tokenizer:")
print(differences)
```