mobilenet.ipynb - Colab | nlp assignment - Colab

colab.research.google.com/drive/1T3LSvxUHmtW3-RIiZav3cWug4ZxeI3Wn#scrollTo=uiO0JIKVJR9z

nlp assignment

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text

RAM Disk Gemini

```
[1] !pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2024.5.15)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.4)
```

```
[12] import nltk
     from nltk import FreqDist, bigrams, trigrams
     from nltk.util import ngrams
     from nltk.corpus import reuters
     from collections import defaultdict, Counter
     nltk.download('reuters')
     nltk.download('punkt')
```

```
[nltk_data] Downloading package reuters to /root/nltk_data...
[nltk_data]   Package reuters is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

```
[13] import nltk
     from nltk.corpus import reuters
     from nltk import FreqDist
     from nltk.tokenize import word_tokenize
     from nltk.util import bigrams
     from nltk.util import trigrams
     from collections import defaultdict
```

✓ 0s completed at 9:17PM

29°C Mostly cloudy | Search | Screenshots | nlp assignment - Colab | train.py - project - Visua | ENG IN | 21:26 05-08-2024

mobilenet.ipynb - Colab | nlp assignment - Colab

colab.research.google.com/drive/1T3LSvxUHmtW3-RIiZav3cWug4ZxeI3Wn#scrollTo=P0us8rRrIQcM

nlp assignment

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text

RAM Disk Gemini

```
def get_unigrams(corpus):
    tokenized_text = [word.lower() for word in word_tokenize(corpus) if word.isalpha()]
    unigrams = FreqDist(tokenized_text)
    return unigrams

corpus = ' '.join(reuters.raw(fileid) for fileid in reuters.fileids()[:10])  # Sample corpus
unigrams = get_unigrams(corpus)

print("Unigrams:")
for unigram, freq in unigrams.items():
    print(f"{unigram}: {freq}")
```

```
cases: 2
these: 1
related: 2
trust: 1
occur: 1
focus: 1
bond: 1
dealing: 1
strengthening: 1
meiko: 1
holds: 2
optimistic: 1
entering: 1
here: 1
swiss: 1
universal: 2
subsidiary: 2
banca: 1
del: 1
gottardo: 2
finance: 1
grant: 1
licences: 1
subsidiaries: 2
banks: 3
soon: 1
```

✓ 0s completed at 9:17PM

29°C Mostly cloudy | Search | Screenshots | nlp assignment - Colab | train.py - project - Visua | ENG IN | 21:27 05-08-2024

mobilenet.ipynb - Colab | nlp assignment - Colab

colab.research.google.com/drive/1T3LSvxUHmtW3-RIiZav3cWug4ZxeI3Wn#scrollTo=Bmet2-c0IqWW

nlp assignment

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text

RAM Disk Gemini

```
def get_bigrams(corpus):
    tokenized_text = [word.lower() for word in word_tokenize(corpus) if word.isalpha()]
    bigram_list = list(bigrams(tokenized_text))
    bigrams_freq = FreqDist(bigram_list)
    return bigrams_freq

# Sample corpus
corpus = ' '.join(reuters.raw(fileid) for fileid in reuters.fileids()[:5])
bigrams_freq = get_bigrams(corpus)

print("Bigrams:")
for bigram, freq in bigrams_freq.items():
    print(f"{bigram}: {freq}")
```

```
('a', 'basis'): 1
('basis', 'followed'): 1
('followed', 'by'): 1
('by', 'oil'): 1
('oil', 'pct'): 1
('pct', 'and'): 3
('and', 'liquefied'): 1
('liquefied', 'natural'): 1
('gas', 'pct'): 1
('pct', 'they'): 1
('they', 'noted'): 1
('noted', 'thai'): 1
('thai', 'trade'): 1
('trade', 'deficit'): 2
('deficit', 'widens'): 1
('widens', 'in'): 1
('in', 'first'): 1
('first', 'quarter'): 4
('quarter', 'thailand'): 1
('thailand', 'trade'): 1
('deficit', 'widened'): 1
('widened', 'to'): 1
('billion', 'baht'): 3
```

✓ 0s completed at 9:17PM

29°C Mostly cloudy | Search | Screenshots | nlp assignment - Colab | train.py - project - Visua | ENG IN | 21:27 05-08-2024

mobilenet.ipynb - Colab | nlp assignment - Colab

colab.research.google.com/drive/1T3LSvxUHmtW3-RIiZav3cWug4ZxeI3Wn#scrollTo=VFzEmXfgIwiw

nlp assignment

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text

RAM Disk Gemini

```
def get_trigrams(corpus):
    tokenized_text = [word.lower() for word in word_tokenize(corpus) if word.isalpha()]
    trigram_list = list(trigrams(tokenized_text))
    trigrams_freq = FreqDist(trigram_list)
    return trigrams_freq

# Sample corpus
corpus = ' '.join(reuters.raw(fileid) for fileid in reuters.fileids()[:5])
trigrams_freq = get_trigrams(corpus)

print("Trigrams:")
for trigram, freq in trigrams_freq.items():
    print(f"{trigram}: {freq}")
```

```
('indonesia', 'the', 'world'): 1
('the', 'world', 'second'): 1
('world', 'second', 'largest'): 1
('second', 'largest', 'producer'): 1
('largest', 'producer', 'of'): 1
('producer', 'of', 'palm'): 1
('of', 'palm', 'oil'): 1
('palm', 'oil', 'after'): 1
('oil', 'after', 'malaysia'): 1
('after', 'malaysia', 'has'): 1
('malaysia', 'has', 'been'): 1
('has', 'been', 'forced'): 1
('been', 'forced', 'to'): 1
('forced', 'to', 'import'): 1
('to', 'import', 'palm'): 1
('import', 'palm', 'oil'): 1
('palm', 'oil', 'to'): 1
('oil', 'to', 'ensure'): 1
('to', 'ensure', 'supplies'): 1
('ensure', 'supplies', 'during'): 1
('supplies', 'during', 'the'): 1
('during', 'the', 'moslem'): 1
('the', 'moslem', 'fasting'): 1
('moslem', 'fasting', 'month'): 1
```

✓ 0s completed at 9:17PM

29°C Mostly cloudy | Search | Screenshots | nlp assignment - Colab | train.py - project - Visua | ENG IN | 21:27 05-08-2024

mobilenet.ipynb - Colab | nlp assignment - Colab

colab.research.google.com/drive/1T3LSvxUHmtW3-RIiZav3cWug4ZxeI3Wn#scrollTo=uiO0JIKVJR9z

nlp assignment

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Gemini

```
def get_bigram_probabilities(corpus):
    tokenized_text = [word.lower() for word in word_tokenize(corpus) if word.isalpha()]
    bigram_list = list(bigrams(tokenized_text))
    bigrams_freq = FreqDist(bigram_list)
    unigrams_freq = FreqDist(tokenized_text)

    bigram_prob = {}
    for (w1, w2), freq in bigrams_freq.items():
        bigram_prob[(w1, w2)] = freq / unigrams_freq[w1]
    return bigram_prob

# Sample corpus
corpus = ' '.join(reuters.raw(fileid) for fileid in reuters.fileids()[:5])
bigram_prob = get_bigram_probabilities(corpus)

print("Bigram Probabilities:")
for bigram, prob in bigram_prob.items():
    print(f"{bigram}: {prob:.4f}")
```

```
('could', 'possibly'): 0.5000
('possibly', 'increase'): 1.0000
('increase', 'its'): 0.5000
('its', 'international'): 0.2500
('international', 'market'): 0.3333
('market', 'share'): 0.5000
('share', 'indonesia'): 0.5000
('indonesia', 'the'): 0.2500
('world', 'second'): 0.3333
('second', 'largest'): 1.0000
('largest', 'producer'): 0.3333
('producer', 'of'): 1.0000
('of', 'palm'): 0.0200
('oil', 'after'): 0.1250
('after', 'malaysia'): 1.0000
('malaysia', 'has'): 0.3333
('been', 'forced'): 0.5000
('forced', 'to'): 1.0000
```

✓ 0s completed at 9:17PM

29°C Mostly cloudy | Search | Screenshots | nlp assignment - Colab | train.py - project - Visua | ENG IN | 21:27 05-08-2024

mobilenet.ipynb - Colab | nlp assignment - Colab

colab.research.google.com/drive/1T3LSvxUHmtW3-RIiZav3cWug4ZxeI3Wn#scrollTo=TeFgPNmCM-9l

nlp assignment

File Edit View Insert Runtime Tools Help All changes saved

Comment Share M

+ Code + Text

RAM Disk Gemini

```
def next_word_prediction(bigram_prob, word):
    candidates = {w2: prob for (w1, w2), prob in bigram_prob.items() if w1 == word}
    if not candidates:
        return "No prediction available"
    return max(candidates, key=candidates.get)

corpus = ' '.join(reuters.raw(fileid) for fileid in reuters.fileids()[:5])
bigram_prob = get_bigram_probabilities(corpus)

sentence = "The stock market"
last_word = sentence.split()[-1]

prediction = next_word_prediction(bigram_prob, last_word)

print(f"Sentence: '{sentence}'")
print(f"Next word prediction for '{last_word}': {prediction}")
```

```
Sentence: 'The stock market'
Next word prediction for 'market': accounting
```

0s completed at 9:17PM

29°C Mostly cloudy | Search | Screenshots | nlp assignment - Colab | train.py - project - Visua | ENG IN | 21:27 05-08-2024