COGS 108 Week 8 A04/A07

Nov 20, 2023

Me after fixing my code for 2349234 time and finding another error1

Slides Adapted from Previous Quarter

AGENDA FOR TODAY

1 LOGISTICS

2 TF-IDF

3 DISCUSSION
LAB 7

LOGISTICS

DUE DATES

- Q5 due TONIGHT Nov 20, 11:59PM
- Data Checkpoint will be available by Wednesday Nov 22
  - Refer to feedback make revisions on the next checkpoint!
  - If you successfully addressed the issues on data checkpoint feedback, you will get points back!
- A3 due Wednesday, Nov 22, 11:59PM

TF-IDF

TF-IDF

- What is TF-IDF?
  - term frequency-inverse document frequency.
  - a statistical measure that evaluates how relevant a word is to a document in a collection of documents.

TF-IDF

- How does it work?
  - Multiplies two terms:
  - how many times a word appears in a document
  - the inverse document frequency of the word across a set of documents.
  - So, if the word is very common and appears in many documents, this number will approach 0. Otherwise, it will approach 1.

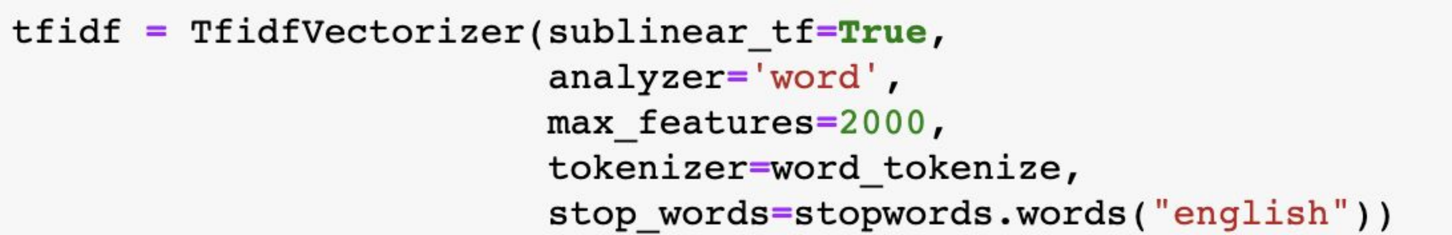
TF-IDF

- How does it work?

$$\text{TF-IDF}(t, d, D) = \underbrace{\log(1 + \text{frequency of t in d})}_{\text{Term frequency}} \quad \underbrace{\log\left(\frac{\text{Total \# of documents}}{\text{\# of documents that contain the word t}}\right)}_{\text{Inverse document frequency}}$$

TF-IDF

- Implementing in python



```
tfidf = TfidfVectorizer(sublinear_tf=True,
                        analyzer='word',
                        max_features=2000,
                        tokenizer=word_tokenize,
                        stop_words=stopwords.words("english"))
```

Sublinear_tf: Apply sublinear tf scaling, i.e. replace tf with 1 + log(tf).

max_features: If not None, build a vocabulary that only consider the top max_features ordered by term frequency across the corpus. Otherwise, all features are used.

DISCUSSION LAB 7: TEXT

THANKS!

Questions on Campuswire or office hours

Office hours: Tue/Thu, 4-5 PM

WHAT GIVES PEOPLE FEELINGS OF POWER
MONEY
STATUS
USING THE TERMINAL IN FRONT OF NON-PROGRAMMERS
@iamnotanartist_