Session 2

Text Representation

April 15, 2024

Agenda

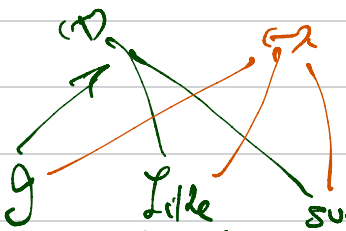
1. BOW recap + discussion
2. Use-case Medium Articles
3. Use BOW to generate recommendations
4. TF-IDF, use TF-IDF to generate recommendations

Stemming & Lemmetization → Text classification

→ use this or not

↓

Problem statement



I like sushi

For each word

find out freq

across entire D → (+)ve

(-)ve

Bag of Words

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

47K → daily → 200 M Visitor

↳ every month

→ Contractions

can't → can not

don't → do not

you'll → you will

Vocabulary → Set of all unique words

corpus → collection of all documents

good, nice, great

Problems

1. Sparse
2. doesn't Preserve similarity
3. Order

Step 2 - Each word in the sentence would be represented as below:

Word	Lecture	on	text	representation
Lecture	1	0	0	0
on	0	1	0	0
text	0	0	1	0
representation	0	0	0	1

Quiz time!

Time Left: 13s

What is a potential limitation of the Bag of Words model?

21 users have participated

- A It is computationally expensive. 14% → X
- B It does not consider the order of words. 76%
- C It requires labeled training data. 5% → X
- D It is only applicable to short documents. 5% → long ones

End Quiz Now

TF-IDF

TF → Term Frequency

IDF → Inverse document Frequency

Term Frequency → 10 / 100

apple → 10

docum → 100

IDF → apple → 10 → d_1 / 5 = 2

D → many docum

$d_1 - d_{10}$

D → d_1 → d_{10}

"Harry Potter" → 5 → d_1

→ $d_2, d_6, d_7, d_8, d_9, d_1$

$\log\left(\frac{5}{6}\right) \simeq 1$

$\simeq 0$

IDF → rarity of word

D → d_1 – d_{20}

"movie" → d_1 → 10

$d_2, d_4, d_5 \ldots d_{15}$ = $\log\left(\frac{10}{10}\right)$

10 other documents movie

$= \log(1)$

$= 0$

If word not like – d_2 – d_5

$\log\left(\frac{10}{1}\right) =$

How – rarity

Document matrix

↳ Relative Frequency count

TF × IDF

Time Left: 10s

In the TF-IDF formula, what does the Inverse Document Frequency (IDF) component measure? ?

22 users have participated

A	The frequency of a term in a document	0%
B	The importance of a term in a specific document	18%
C ✓	The rarity of a term in the entire collection of documents	82%
D	The total number of documents in the collection	0%

End Quiz Now

TF-IDF

↓

Sentiment