Rohini Vishwanathan
IST 332
Natural Language Processing

Lab Assignment #1- Text Preprocessing

Google colab- https://colab.research.google.com/drive/1_5vZGAGkuWDycsWyhbID9KxZ5QYI6WM7?usp=sharing

1) Extract Top 20 Reviews

Scrape top 20 reviews into a list

```
[12]
✓ 0s
```

```
single_page_review = []

# Slice [:20] to ensure "Top 20"
for review in reviews[:20]:
    username = review.find('a', class_='display-name').get_text(strip=True)
    publication = review.find('a', class_='publication').get_text(strip=True)
    review_date = review.find('span', {'data-qa':'review-date'}).get_text(strip=True)
    review_content = review.find('p', class_='review-text').get_text(strip=True)

    score = None
    score_tag = review.find("p", class_="original-score-and-url")
    if score_tag:
        m = re.search(r"Original Score:\s*([^|]+)", score_tag.get_text(strip=True))
        score = m.group(1).strip() if m else None

    single_review = [username, publication, review_date, review_content, score]
    single_page_review.append(single_review)


print(*single_page_review, sep="\n")
```

```
['Armond White', 'National Review', 'Sep 10, 2025', 'None of the couple's success, failure, or fame reflects the unsatisfying ups and downs of these times. Roach and McNamara cheat us of understanding ourselves. The Roses plays
['Neely Swanson', 'Beverly Hills Courier', 'Sep 10, 2025', 'The Roses is a very good movie, but what makes it exceptional are the two leads: Benedict Cumberbatch as Theo and Olivia Coleman as Ivy.', None]
['Sarah Vincent', 'Sarah G Vincent Views', 'Sep 8, 2025', 'It is like going to a party and ending on a down note. It is so strange to laugh for almost the entire movie then to leave as silent as the grave.', None]
['Susan Granger', 'SSG Syndicate', 'Sep 7, 2025', "How many ways can a critic warn: 'This picture is absolutely awful'? Ill-conceived and ineptly scripted, the bickering never stops. Resentment is redolent. And – bottom line – i
['Kelechi Ehenulo', 'Movie Marker', 'Sep 6, 2025', 'Instead of the sinking ship IP remake this could have easily been, it's the playfulness that keeps The Roses floating just above the waters. But in adding nothing new, its safe
['Whang Yee Ling', 'The Straits Times (Singapore)', 'Sep 5, 2025', 'This rancorous anti-romcom is a happy marriage of Hollywood studio entertainment and English drollery.', '3/5']
['Jared Rasic', 'Metro Times (Detroit, MI)', 'Sep 5, 2025', 'The Roses actually works better than the original, with Cumberbatch and Colman having such a gorgeous and homey chemistry that you genuinely want them to fix their pro
['Stephen Romei', 'The Australian', 'Sep 5, 2025', 'The dazzling script is thespian heaven for the stars, Olivia Colman and Benedict Cumberbatch, who put a wicked British spin on this remake of Danny DeVito's 1989 movie The War
['Maxwell Rabb', 'Chicago Reader', 'Sep 5, 2025', 'The pacing switches too suddenly between vigor and slackness, leaving the film without the propulsion its absurd premise demands, causing the conflict to drag on.', None]
['Josh Larsen', 'LarsenOnFilm', 'Sep 5, 2025', 'Colman and Cumberbatch leave dazzling trails of repartee as they zip along.', '2.5/4']
['Sandie Angulo Chen', 'Common Sense Media', 'Sep 5, 2025', 'This dark comedy benefits from brilliant stars and an amusing premise, but it eventually devolves into cringe-inducing over-the-top cruelty.', '3/5']
['Simon Foster', 'Screen-Space', 'Sep 5, 2025', 'It's occasionally fun, watching Ben and Livvy shriek at each other in that clipped British tone, but it never rises to what it ought have been.', '2.5/5']
['Jeffrey M. Anderson', 'Combustible Celluloid', 'Sep 5, 2025', "What's remarkable about 'The Roses' is how thoroughly different it is from DeVito's film; it could barely even qualify as a remake, which is refreshing.", '3.5/4']
['Glen Weldon', 'NPR', 'Sep 4, 2025', "It's a throwback film, the kind they don't make anymore — and given the sheer amount of Screenwriting 101 cliches that pile up — you can't help thinking that might be a good thing.", None
['Linda Holmes', 'Pop Culture Happy Hour (NPR Podcast)', 'Sep 4, 2025', 'This just ended up being like one million other things that I have seen about middle-aged couples trying to overcome the boring middle part of their marria
['Chloe Veltman', 'Pop Culture Happy Hour (NPR Podcast)', 'Sep 4, 2025', 'Having read the novel and seen the 1980s movie…I really adored how loving this movie actually is.', None]
['Leigh Paatsch', 'Daily Telegraph (Australia)', 'Sep 4, 2025', 'The acrimonious gymnastics of the Cumberbatch-Colman double act are often so impressive that you barely notice the movie's story is struggling to hit the same heig
['Jim Schembri', 'jimschembri.com', 'Sep 4, 2025', 'A likeable if admittedly mild remake of the savage Danny DeVito 1989 divorce comedy...sufficiently funny.', '3/5']
['Julian Wood', 'FILMINK (Australia)', 'Sep 4, 2025', '... very funny and the small observations yield a steady stream of chuckles that makes you largely look past the film's flaws.', '15/20']
['Louisa Moore', 'Screen Zealots', 'Sep 4, 2025', 'Takes a sharp marital comedy and turns it into a movie that's way more misery than fun.', None]
```

2. Write to CSV

Save to CSV

```
[13]
✓ 0s
```

```
df = pd.DataFrame(
    single_page_review,
    columns=["user_name", "publication", "review_date", "review", "rating"]
)
df.to_csv("the_roses_top20_reviews.csv", index=False)
df.head(3)
```

	user_name	publication	review_date	review	rating
0	Armond White	National Review	Sep 10, 2025	None of the couple's success, failure, or fame...	None
1	Neely Swanson	Beverly Hills Courier	Sep 10, 2025	The Roses is a very good movie, but what makes...	None
2	Sarah Vincent	Sarah G Vincent Views	Sep 8, 2025	It is like going to a party and ending on a do...	None

3. Contraction expansion, Tokenization, Misspellings

Contraction expansion before Tokenization

16] 0s

```
import contractions
import pandas as pd

def expand_contractions(text):
    if pd.isna(text):
        return ""
    words = []
    for w in str(text).split():
        words.append(contractions.fix(w))
    return ' '.join(words)

df['review_expanded'] = df['review'].apply(expand_contractions)
df[['review','review_expanded']].head(2)
```

	review	review_expanded
0	None of the couple's success, failure, or fame...	None of the couple's success, failure, or fame...
1	The Roses is a very good movie, but what makes...	The Roses is a very good movie, but what makes...

Tokenize

[17] 0s

```
from nltk.tokenize import wordpunct_tokenize

df['tokens_raw'] = df['review_expanded'].apply(lambda t: wordpunct_tokenize(t))
df[['user_name','tokens_raw']].head(3)
```

	user_name	tokens_raw
0	Armond White	[None, of, the, couple, ', s, success, ,, fail...
1	Neely Swanson	[The, Roses, is, a, very, good, movie, ,, but,...
2	Sarah Vincent	[It, is, like, going, to, a, party, and, endin...

Misspelling with Pyspellchecker

[18] 0s

```
from spellchecker import SpellChecker
spell = SpellChecker()

def list_misspellings(tokens):
    # keep alphabetic tokens to avoid punctuation/nums
    toks = [w for w in tokens if w.isalpha()]
    # pyspellchecker expects lowercased tokens
    toks = [w.lower() for w in toks]
    return sorted(spell.unknown(toks))

df['misspellings'] = df['tokens_raw'].apply(list_misspellings)
df[['user_name','misspellings']].head(10)
```

	user_name	misspellings
0	Armond White	[]
1	Neely Swanson	[cumberbatch, theo]
2	Sarah Vincent	[]
3	Susan Granger	[]
4	Kelechi Ehenulo	[ip]
5	Whang Yee Ling	[romcom]
6	Jared Rasic	[colman, cumberbatch]
7	Stephen Romei	[colman, cumberbatch, devito]
8	Maxwell Rabb	[]
9	Josh Larsen	[colman, cumberbatch]

4. Text normalization + Top 10 Tokens

Corpus and Top 10 most frequent tokens

```
# 3) Top-10 most frequent tokens in the corpus (Lab 1.5 FreqDist approach)
from nltk import FreqDist

fdist_reviews = FreqDist(corpus)  # frequency dictionary
# “Top 10” using the lab’s sorted-slice pattern
fdist_list = sorted(fdist_reviews.items(), key=lambda x: x[1], reverse=True)
top10_tokens = fdist_list[:10]
top10_tokens  # list of (token, count)
```

```
[('movie', 7),
 ('roses', 6),
 ('cumberbatch', 5),
 ('like', 4),
 ('remake', 4),
 ('...', 4),
 ('colman', 4),
 ('film', 4),
 ('funny', 3),
 ('devito', 3)]
```

5. What words appear with ‘Roses’ in a similar context

```
# Show context lines too
reviews_text.concordance('roses', width=80, lines=10)
```

```
Displaying 6 of 6 matches:
at us of understanding ourselves the roses plays like a hoax hollywood cannot g
ot go on making movies like this the roses is a very good movie but what makes
it is the playfulness that keeps the roses floating just above the waters but i
tertainment and english drollery the roses actually works better than the origi
 danny devito s movie the war of the roses the pacing switches too suddenly bet
ve been what is remarkable about the roses is how thoroughly different it is fr
```

6. Complete the following table for the first 10 reviews

Review Length (# of tokens from raw text)	Cleaned Review Length (# of tokens after text normalization)	Lexical Diversity
47	25	1,042
28	13	1,000
34	13	1,000
41	20	1,000
49	24	1,000
17	10	1,000

Review Length (# of tokens from raw text)	Cleaned Review Length (# of tokens after text normalization)	Lexical Diversity
38	16	1,000
38	20	1,000
28	15	1,000
13	8	1,000

7. Frequency Distribution plot for the top 20 most frequent tokens after text normalization

```
]
s
```

```python
from nltk import FreqDist

#  Build a single corpus from your normalized tokens (you already made tokens_clean in Q4)
corpus = [w for row in df['tokens_clean'] for w in row]

#  Frequency distribution (Lab 1.5)
fdist_reviews = FreqDist(corpus)

#  Plot the top 20 most frequent tokens
fdist_reviews.plot(20, cumulative=True)
```

```
<Axes: xlabel='Samples', ylabel='Cumulative Counts'>
```