CAPSTONE PROJECT

ARABIC POETRY NLP

Munera AlRajhi

Table of Contents

- INTRODUCTION

  Project Scope, Idea, and Motivation.

- DATA COLLECTION AND PROCESSING.

  Web Scraping, EDA, and NLP Processing.

- DATA CLASSIFICATION

  Multi-class classification and Topic Modeling.

- CONCLUSION

  Challenges and Technical Issues, Future Work, Acknowledgment

INTRODUCTION

WHY ARABIC POETRY?

Something I'm passionate about

Challenging Topic

Something New!

PROJECT SCOPE

WHAT ARE OUR MAIN QUESTIONS?

- HOW TO CLASSIFY POEMS BASED ON LABELS?

-HOW TO PREDICT LABELS BASED ON WORDS CLUSTERING?

DATA COLLECTION & PROCESSING.

DATA COLLECTING

HOW?

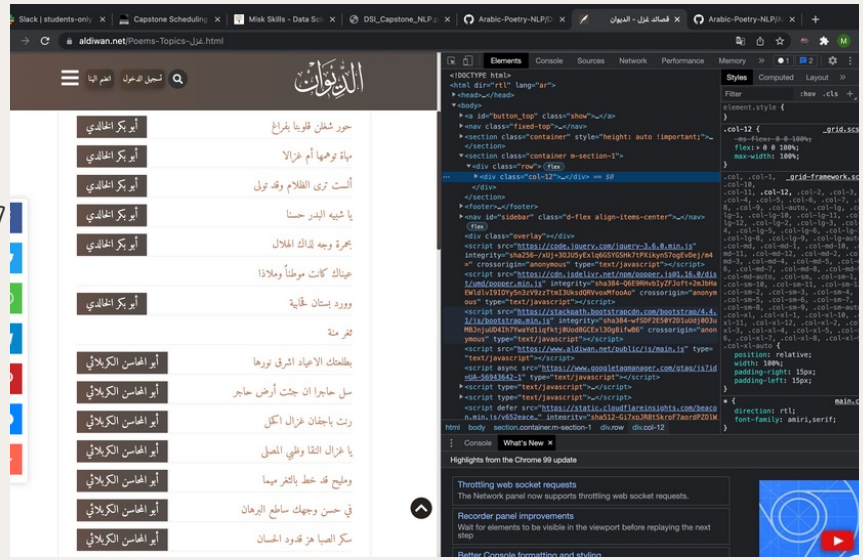
Web Scraping.

FROM WHERE?

AlDiwan for Arabic Poetry Webstibe: https://www.aldiwan.net/

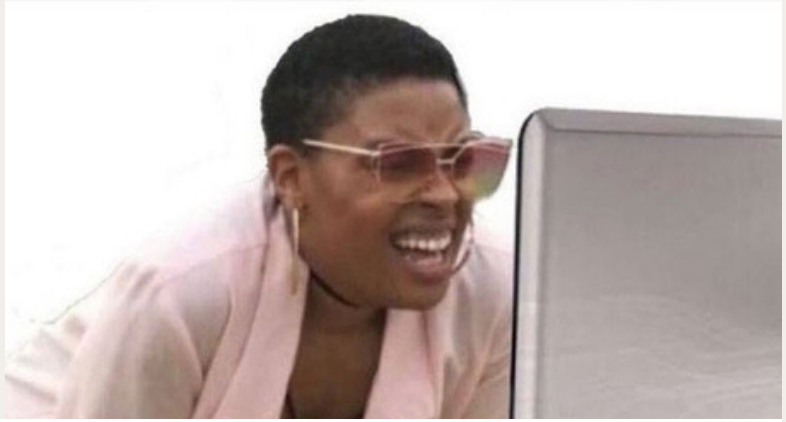
YES, LET'S DO IT!!!!

VS



MY EXPECTATION

REALITY



Sitting there and trying to understand someone else code, be like:

WEB SCRAPING

```
for page in range(1,32):
    url = requests.get(f"https://www.aldiwan.net/Poems-Topics-%D8%BA%D8%B2%D9%84.html?page={page}")
    soup = BeautifulSoup(url.text, 'html.parser')
    results = soup.find_all('div', attrs={'class':'record col-12'})

    for result in results:
        href = result.find('a')['href']
        GazalPages.append((href))


for page in GazalPages:

    url = requests.get(f"https://www.aldiwan.net/{page}")
    soup = BeautifulSoup(url.content, 'html.parser')
    poem = str(soup.find('div', attrs={'class':'bet-1 row pt-0 px-5 pb-4 justify-content-center'}).text.strip()).rep
    VersusNumber= soup.find('p', attrs={'class':'d-inline-block px-2 mt-0 mb-1 poem-control main-color'})
    poet = soup.find('h2', attrs={'class':'text-center h3 mt-3 mb-0'}).text.strip()
    century = soup.find('p', attrs={'class':'main-color text-center mb-0 ltr'}).text.strip()

    sss = soup.find_all('div', attrs={'class':'col-6 col-md-3'})
    dummy = []

    for heading in sss:
        info = heading.find('a')
        dummy.append(info.text.strip())

    GazalPoems.append((poem,VersusNumber,poet,century,dummy[0],dummy[1],dummy[2]))
    print(page)

GazalPoems=pd.DataFrame(GazalPoems,columns=['Poem','Number of Versus','Poet',"Century","Label","Type",'Metre'])
```

life lessons!!

- **ALWAYS** KEEP YOUR CODE NEAT AND UNDERSTANDABLE, SO OTHER PEOPLE WON'T MAKE MEMES ABOUT YOU!

- **BE PATIENT**, DON'T JUMP TO CONCLUSIONS AND ONLY FOCUS ON YOUR NEXT STEP

life lessons!!

SAYING **YES** TO **LUJAIN** IS LIKE ORDERING FROM STARBUCKS, BOTH WILL COST YOU SOMETHING!! EITHER EXTRA WORK OR EXTRA MONEY!!

FINAL DATASET

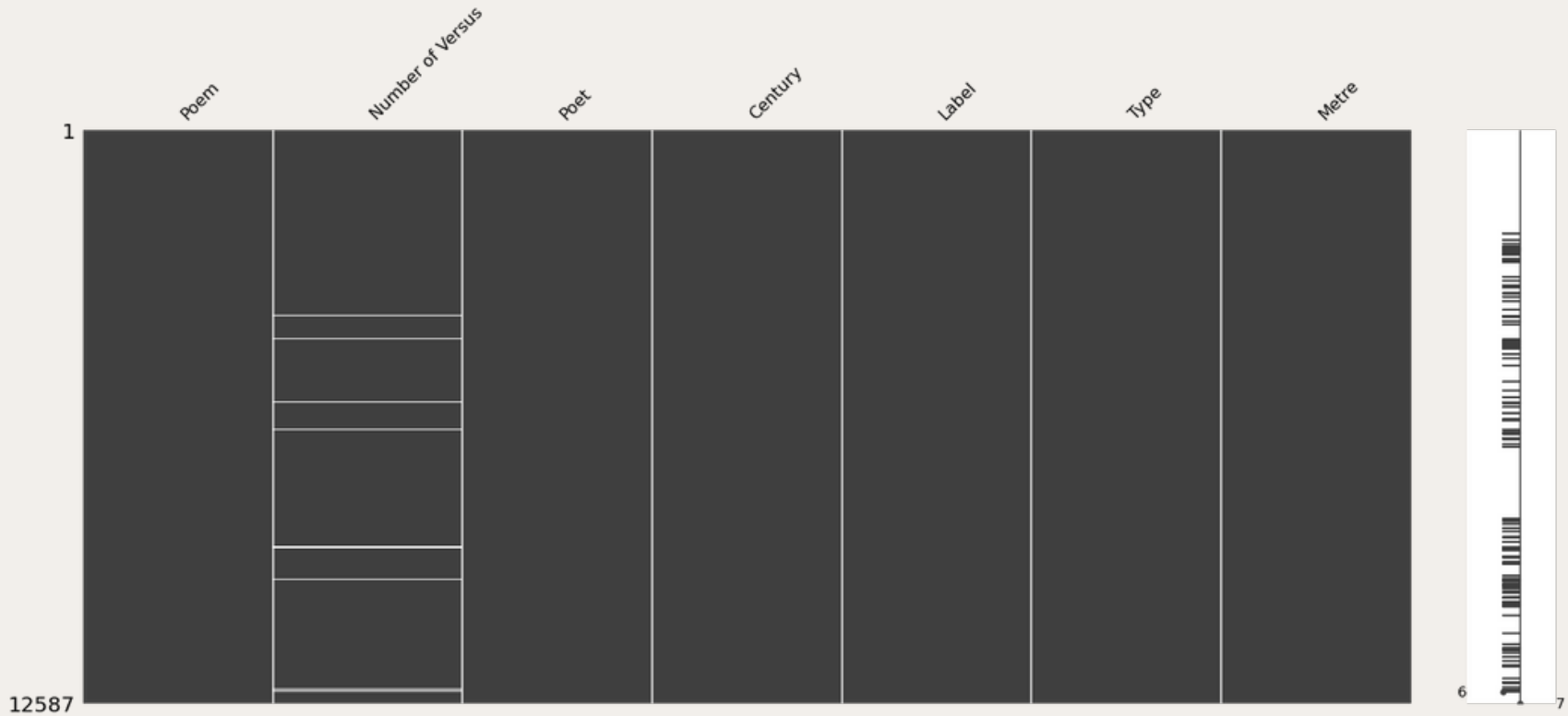
FINALLY!!

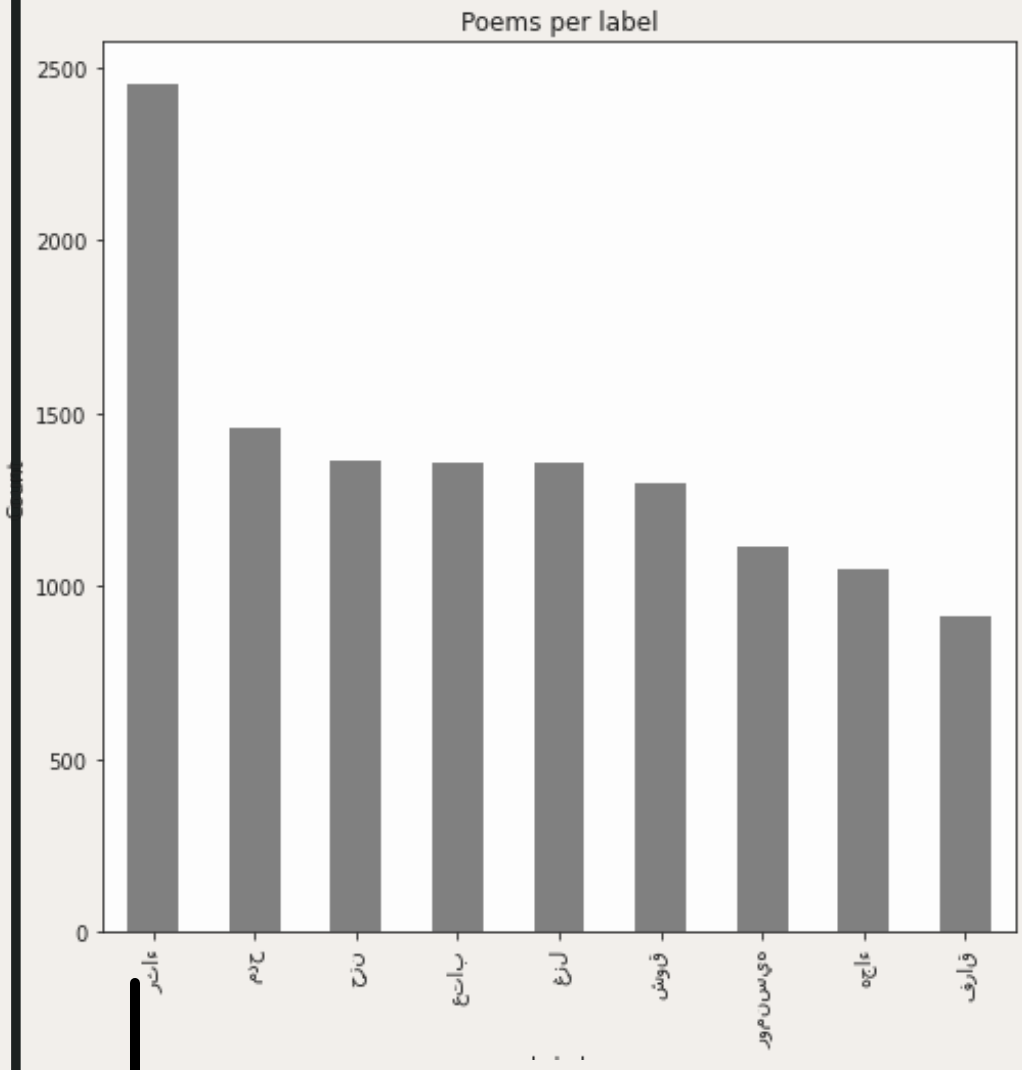
we now have a complete dataset, with a 12587 poem along with 7 features.

	Poem	Number of Versus	Poet	Century	Label	Type	Metre
12226	أتلك رياض أم خدود نواعم وفيها أقاح أم ثغور بوا...	82	الأرجاني	العصر الأندلسي	غزل	عموديه	بحر الطويل
10107	لم يبق بينك من جمالك منسما لمن الذهاب وفي طلاب...	24	سليمان الصولة	سوريا	شوق	عموديه	بحر الكامل
1753	اسقني واسق يوسفا مزة الطعم قرقفا دع من العيش ك...	7	ابو نواس	العصر العباسي	فراق	عموديه	بحر مجزوء الخفيف
4984	جاورت ربك يا أبا العباس وتركت شبلك رحمة للناس ...	24	سليمان الصولة	سوريا	رثاء	عموديه	بحر الكامل
6075	سقى الله تلك الدار هامية القطر مدى الدهر ما نا...	19	العشاري	العصر العثماني	رثاء	عموديه	بحر الطويل
2247	إذا وهى الحب فالهجران يقتله وإن تمكن فالهجران ...	2	خليل مطران	لبنان	فراق	عموديه	بحر البسيط
7836	على وجهه بالكاس صرف سلافة سقاني بيمناه فطابت ب...	2	المفتي عبداللطيف فتح الله	لبنان	رومنسيه	عموديه	بحر الطويل
8799	أرى جسمي تحط به البلايا وما شارفت معترك المناي...	2	المحبي	العصر العثماني	حزن	عموديه	بحر الوافر
4190	سلام يفوق الروض فيه الأزاهر ويزري بعقد الدر في...	23	المفتي عبداللطيف فتح الله	لبنان	مدح	عموديه	بحر الطويل
4641	كساك الصوم أعمار الليالي وأعقبك العنيمة في الم...	8	ابن بابك	العصر العباسي	مدح	عموديه	بحر الوافر

DATA CLEANING AND PREPARATION

Handling Missing Values

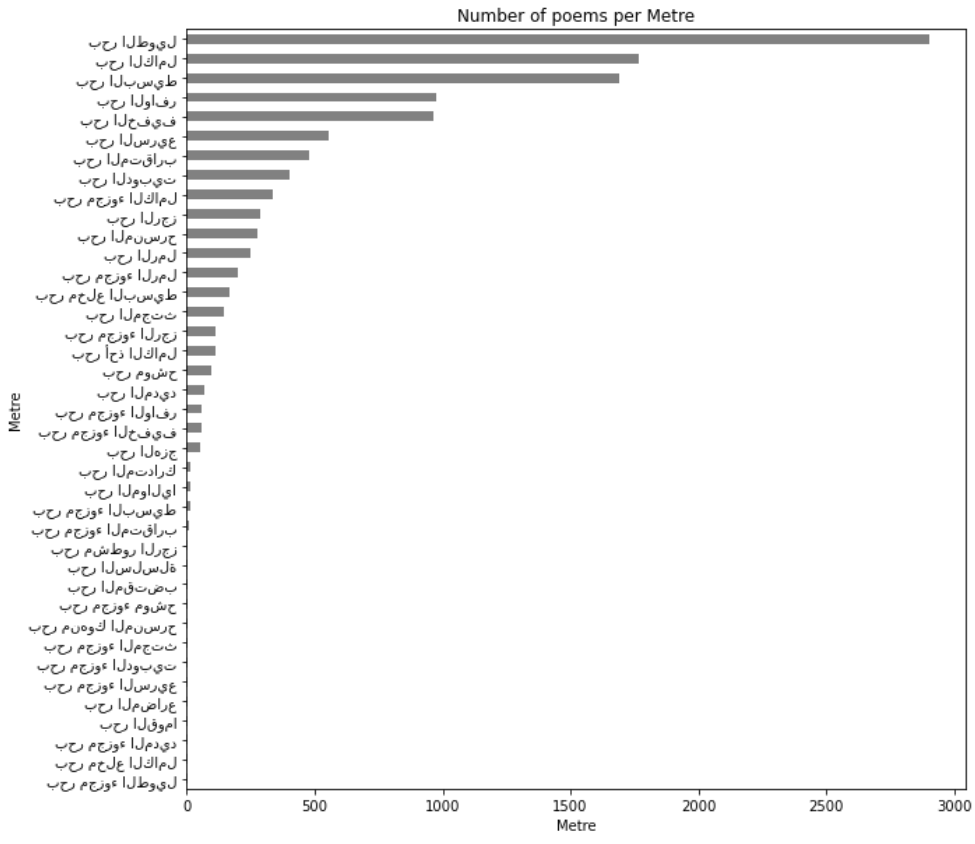




WE HAVE 9 LABELS

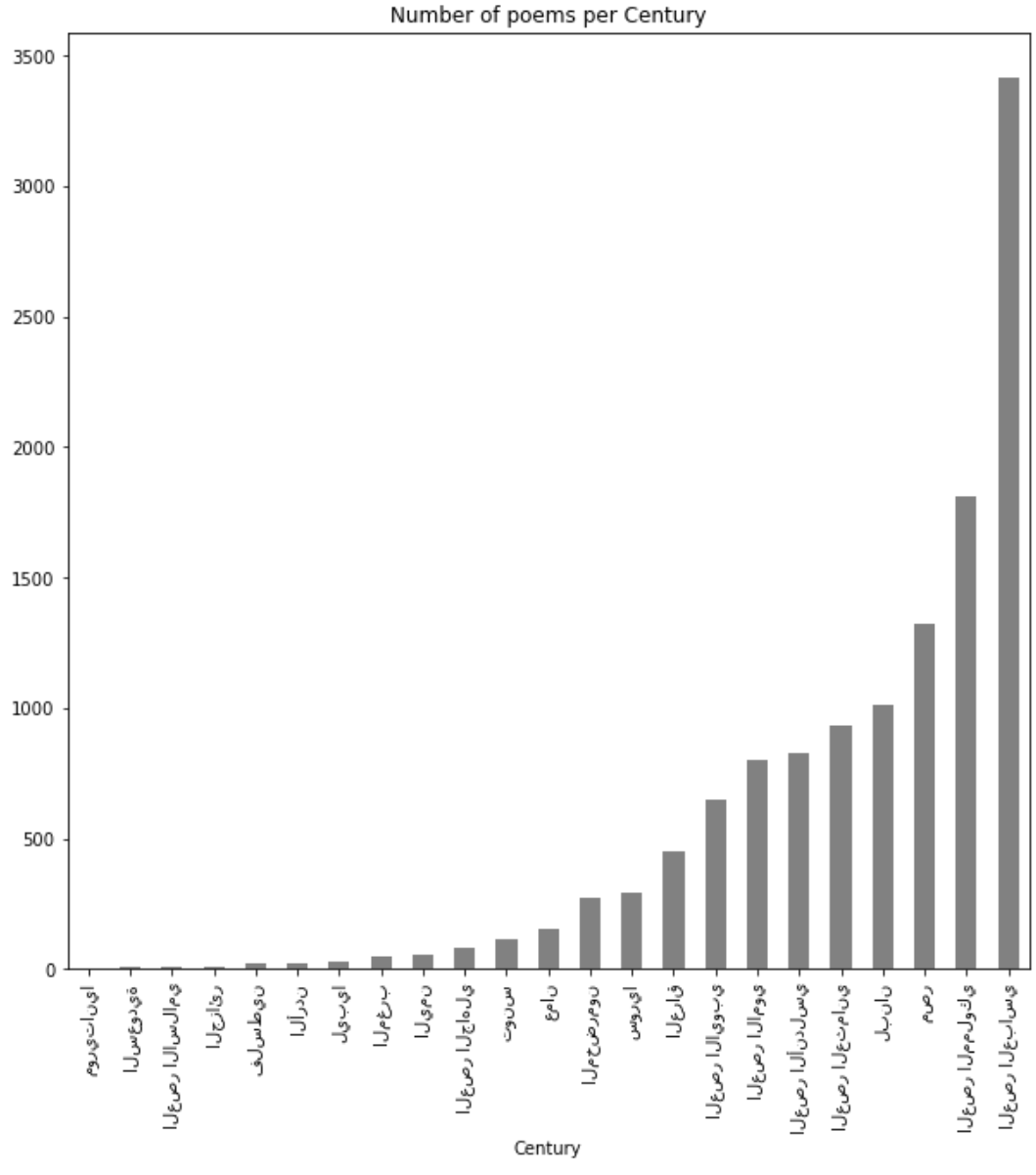
'عتاب', 'فراق', 'هجاء', 'مدح', 'رثاء', ''
'رومنسيه', 'حزن', 'شوق',
'غزل'

رثاء have the most poems with 2500 poem



بحر الطويل have the most poems with 2901poem

WE HAVE 66 METRE



العصر العباسي have the most poems with 3417 poem

WE HAVE 23 CENTURY

DATA PROCESSING

TOKENIZATION

['', أخ, 'على', 'عتبت', 'وإذا', 'فاستبقه', 'الغد', 'ولا', 'تهلك', 'بلا', '', 'إخوان']

STEMMING

['', سبق, 'اخ', 'على', 'عتب', 'وإذ', 'الغد', 'ولا', 'تهل', 'بلا', 'اخو']

REMOVING STOP WORDS

['عتب', 'سبق', 'الغد', 'تهل', 'بلا']

Most frequent 30 Words

Qalab means in Heart.

DATA CLASSIFICATION.

MULTI-CLASS CLASSFCATION

Can we build an ML classifier to predict poem labels?

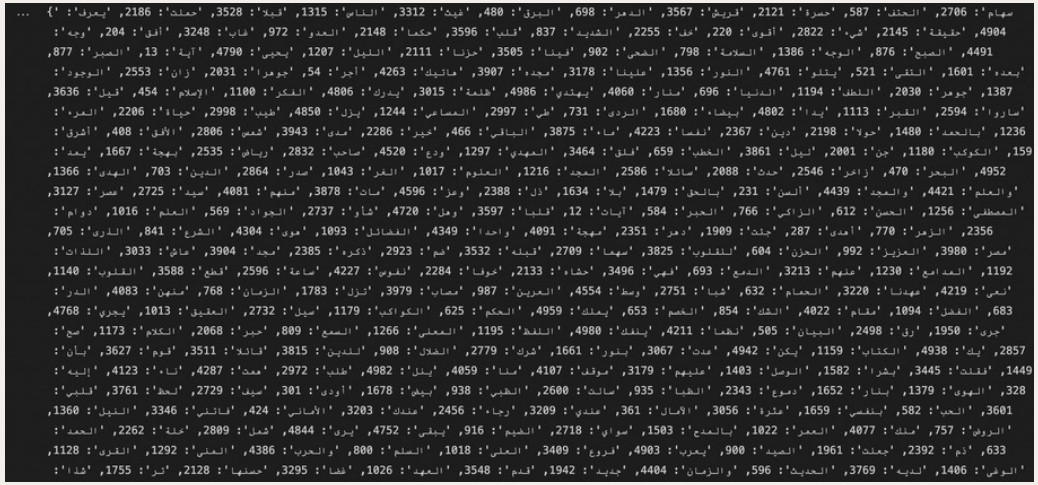
The labels are: 'غزل' ,'شوق' ,'حزن' ,'رومنسيه' ,'رثاء' ,'مدح' ,'هجاء' ,'فراق' ,'عتاب'

1. Text Encoding
2. Build and Evaluate Classification Models
3. Testing

MULTI-CLASS CLASSFCATION

1 Text Encoding

I used Count Vectorizer to create a Bag of Words and encode them.




```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

MULTI-CLASS CLASSFCATION

2 Build and Evaluate Classification Models

	precision	recall	f1-score	support
حزن	0.30	0.25	0.37	275
رثاء	0.39	0.47	0.33	479
رومنسيه	0.35	0.41	0.31	226
شوق	0.49	0.49	0.50	253
عتاب	0.19	0.17	0.23	279
غزل	0.32	0.30	0.34	260
فراق	0.31	0.25	0.41	187
مدح	0.55	0.61	0.51	293
هجاء	0.26	0.22	0.31	218
accuracy			0.37	2470
macro avg	0.37	0.35	0.35	2470
weighted avg	0.36	0.37	0.36	2470

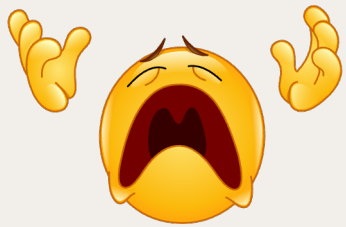
We can notice here we have low Accuracy which means the model is underfitting

MULTI-CLASS CLASSFCATION

2 **Build and Evaluate Classification Models**

even after trying different model the problem still there!!

	Model	Score
2	Random Forest	0.363158
1	Multinomial Naive Bayes	0.359109
0	Logistic Regression	0.333603
3	LinearSVC	0.301619



MULTI-CLASS CLASSFCATION

3 Testing

```
#qoute
test_text = ['لعمري لقد أوهيت قلبي عن العزا وطأطأت رأسي والفؤاد كئيب لقد قصمت مني قناة صليبة ويقصم عود النبع وهو صليب ']
# حزن
# encoding
test_vector = count_vector.transform(test_text)
test_vector = test_vector.toarray()

## Perform and Evaluate the Model
text_predict_class = encoder.inverse_transform(RandomModel.predict(test_vector))
print(test_text)
print(text_predict_class)
```

```
['لعمري لقد أوهيت قلبي عن العزا وطأطأت رأسي والفؤاد كئيب لقد قصمت مني قناة صليبة ويقصم عود النبع وهو صليب ']
['رومنسيه']
```

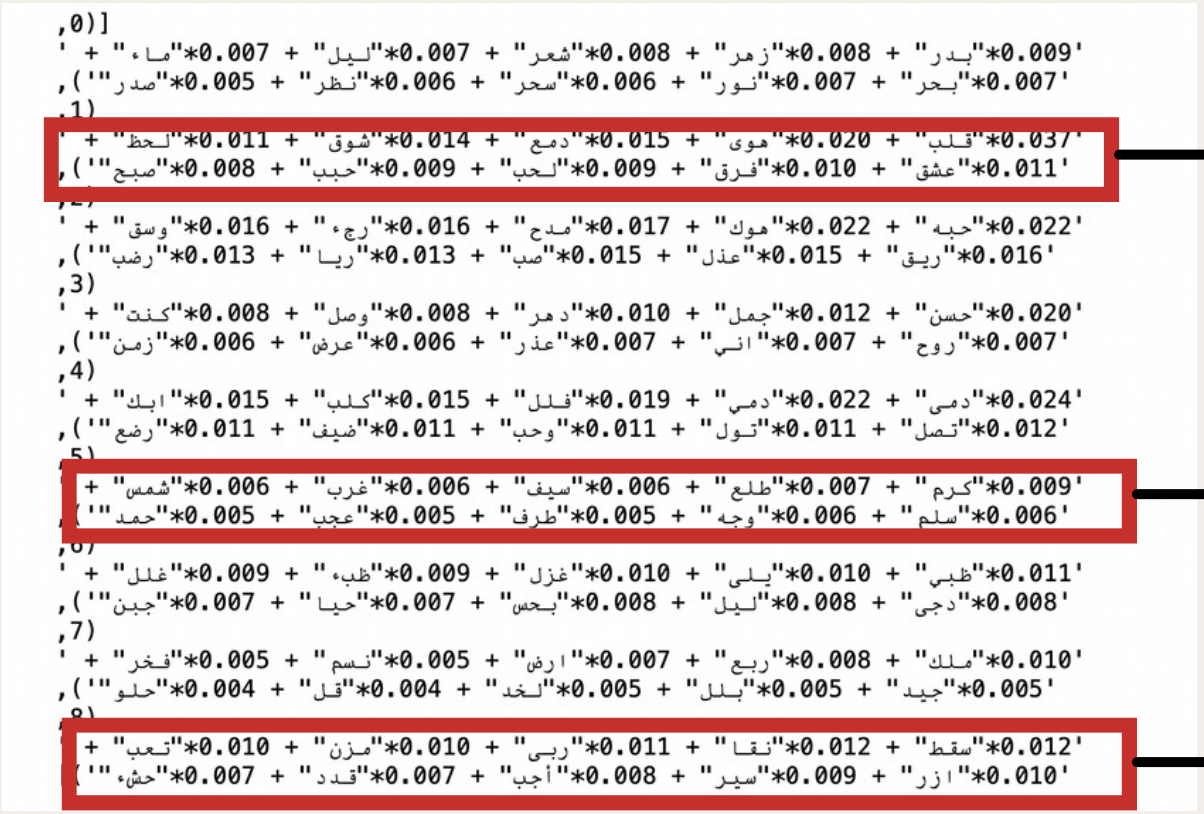
```
#qoute
test_text = ['إن وترت قلبك الهموم فما مثل انتصار بالناي والوتر وشادن حيرت لواحظه ألحاظ عين الغزال بالحور']
# غزل
# encoding
test_vector = count_vector.transform(test_text)
test_vector = test_vector.toarray()

## Perform and Evaluate the Model
text_predict_class = encoder.inverse_transform(RandomModel.predict(test_vector))
print(test_text)
print(text_predict_class)
```

```
... ['إن وترت قلبك الهموم فما مثل انتصار بالناي والوتر وشادن حيرت لواحظه ألحاظ عين الغزال بالحور']
['غزل']
```

TOPIC MODELING

How to predict labels based on words clustering?



TOPIC MODELING

How to predict labels based on words clustering?

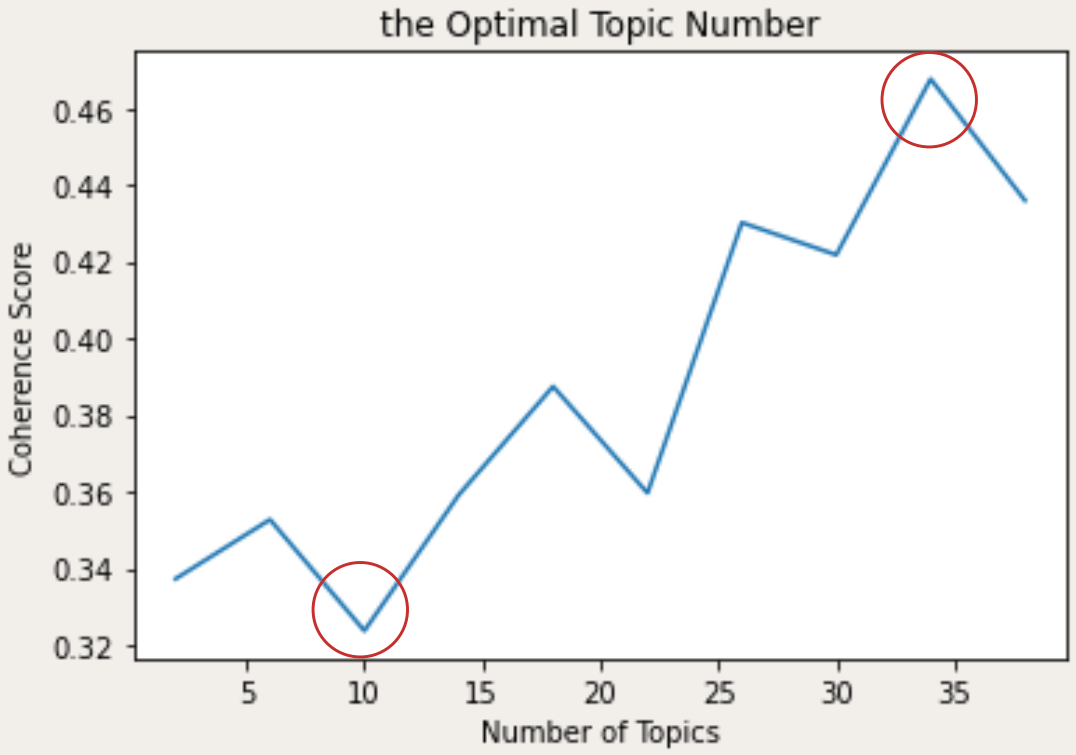
TOPIC MODELING

How to predict labels based on words clustering?

Evaluation:

The Number of Topics we used: **9**
Coherence Score: 0.29

The Optimal Topic Number: **34**
Coherence score: 0.46



CONCLUSION

CHALLENGES AND LIMITATION

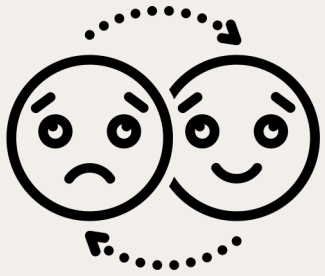
SMALL DATASET

TIME LIMITATION

NEW TOPICS

FUTURE WORK

SENTIMENT ANALYSIS



EXPAND THE DATASET

IMPROVE THE CLASSIFICATION MODEL

KEY TAKEAWAYS

BE PATIENT!

ALWAYS TEST YOUR LIMITS.

GOOGLE AND **YOUTUBE** ARE YOUR BEST FRIENDS.

DR. RICK SCAVETTA

Instructor

LUJAIN FELEMBAN

Assistant Instructor

ACKNOWLEDGMENT

REFERENCES:

- **How to Create an LDA Topic Model in Python with Gensim:**
https://www.youtube.com/watch?v=TKjjlp5_r7o
- **Misk Skills Data Science Book:**
http://www.ylz.ncx.mybluehost.me/scavetta.academy/misk/11_nlp/materials/code/_build/html/04_nlp_unsupervised_learning.html
- **NLTK book Chapter1 Language Processing and Python:**
https://www.nltk.org/book/ch01.html
- **Count Vectorizer vs TFIDF Vectorizer | Natural Language Processing:**
https://www.linkedin.com/pulse/count-vectorizers-vs-tfidf-natural-language-processing-sheel-saket/
- **Product-Categorization-NLP:**
https://github.com/aniass/Product-Categorization-NLP/blob/master/Text_analysis.ipynb