



**Computer Engineering Department**

**Natural Language Processing**

**Project Phase 1**

Yasaman Lotfollahi  
Ali Sedaghi

# Table of contents

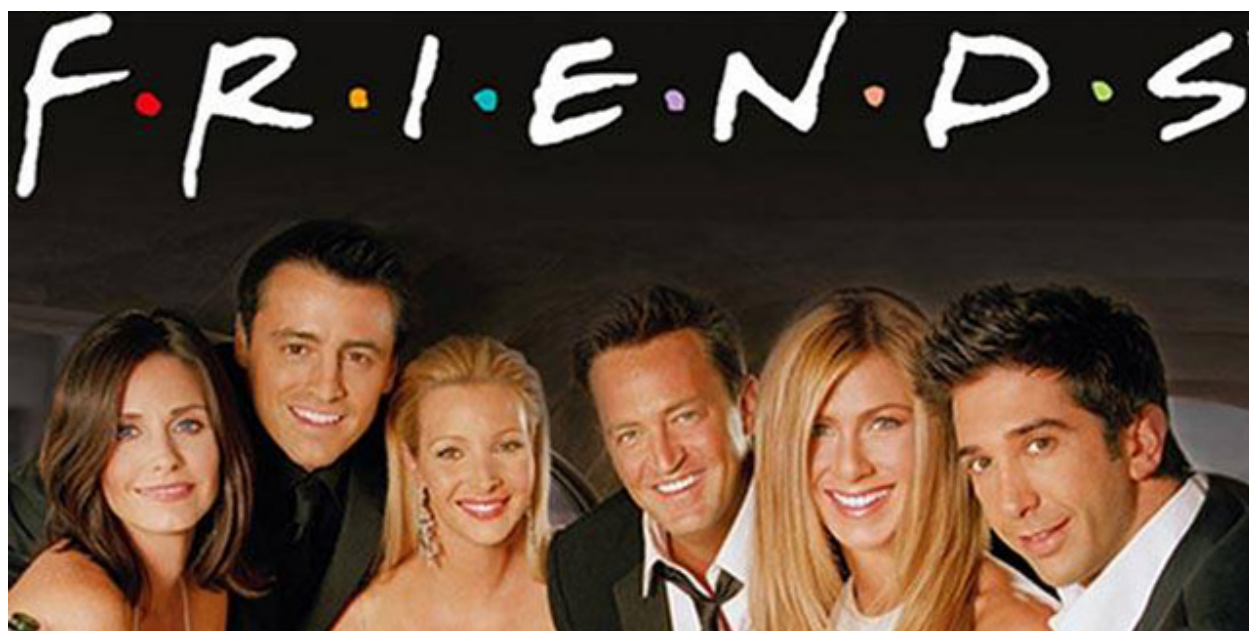
1	موضوع پروژه
1	انتخاب سریال
2	ریپوزیتوری گیت‌هاب
3	جمع‌آوری داده
4	ساختار داده خام
5	روش تفکیک جملات
5	روش تفکیک کلمات
5	پیش‌پردازش روی داده و تمیز کردن آن
5	مرحله اول: حذف وایت اسپیس‌های اضافه
6	مرحله دوم: تبدیل تمامی حروف به حروف کوچک
6	مرحله سوم: حذف کاراکترهای ویژه
6	مرحله چهارم: حذف کلمات کوتاه
6	مرحله پنجم: حذف Stop words
6	واحد برچسب‌گذاری
7	آمار داده‌ها قبل از پیش‌پردازش
7	ابر کلمات
7	هیستوگرام
8	آمار داده‌ها بعد از پیش‌پردازش
8	ابر کلمات
8	هیستوگرام
9	سایر دیتاست‌های آماده (استفاده نشده)

## موضوع پروژه

تشخیص دیالوگ کاراکتر در یک فیلم یا سریال

## انتخاب سریال

برای این امر از دیالوگ‌های سریال Friends استفاده شده است. این سریال سیتکام دارای 6 شخصیت اصلی به نام‌های Ross، Rachel، Joey، Chandler، Monica، Phoebe است. این شخصیت‌ها بیانگر کلاس‌ها هستند و در نهایت باید مدلی فراهم شود که با ورودی گرفتن یک دیالوگ تشخیص دهد آن دیالوگ مربوط به کدام شخصیت است. این سریال دارای 10 فصل و 236 قسمت است که تمامی قسمت‌های آن در دسترس جمع‌آوری شده آورده شده است.



## ریپوزیتوری گیت‌هاب

<https://github.com/ysmnlft/dialogue-prediction>

توضیحات لازم در فایل ReadMe این مخزن آورده شده است. این توضیحات شامل موارد زیر است:

- مسیر فایل‌ها: گزارش، دیتاست خام، دیتاست پیش‌پردازش شده، آمار دیتاست، گزارش پروژه
- توضیحاتی درباره دیتاست و جمع‌آوری آن
- نحوه اجرای پروژه: اجرای کراولر، پیش‌پردازش، استخراج آمار

### Dialogue Prediction

#### File directories

- Phase 1 report file
- Scripts data file
- Raw dialogues data file
- Preprocessed file step by step
- Cleaned dataset

#### Dataset

Friends tv series scripts used as dataset. Friends is an American television sitcom which aired on NBC from September 22, 1994, to May 6, 2004.

There are 7 main characters (classes) in this show:

- Ross
- Rachel
- Joey
- Chandler
- Monica
- Phoebe

The scripts are gathered from [Here](#).

### How to run

#### Requirements

Python packages must be installed:

```
pip install -r requirements.txt
```

#### Crawler

To run crawler and gather/update dataset:

```
cd src/crawler
scrapy crawl scripts -t csv -o ../../data/raw/scripts.csv
scrapy crawl dialogues -t csv -o ../../data/raw/dialogues.csv
```

#### Preprocessing

- Step 1: Remove white spaces
- Step 2: Lowercase all letters
- Step 3: Remove special characters
- Step 4: Remove short words
- Step 5: Remove stopwords

```
cd src/preprocessing
python preprocessor.py
```

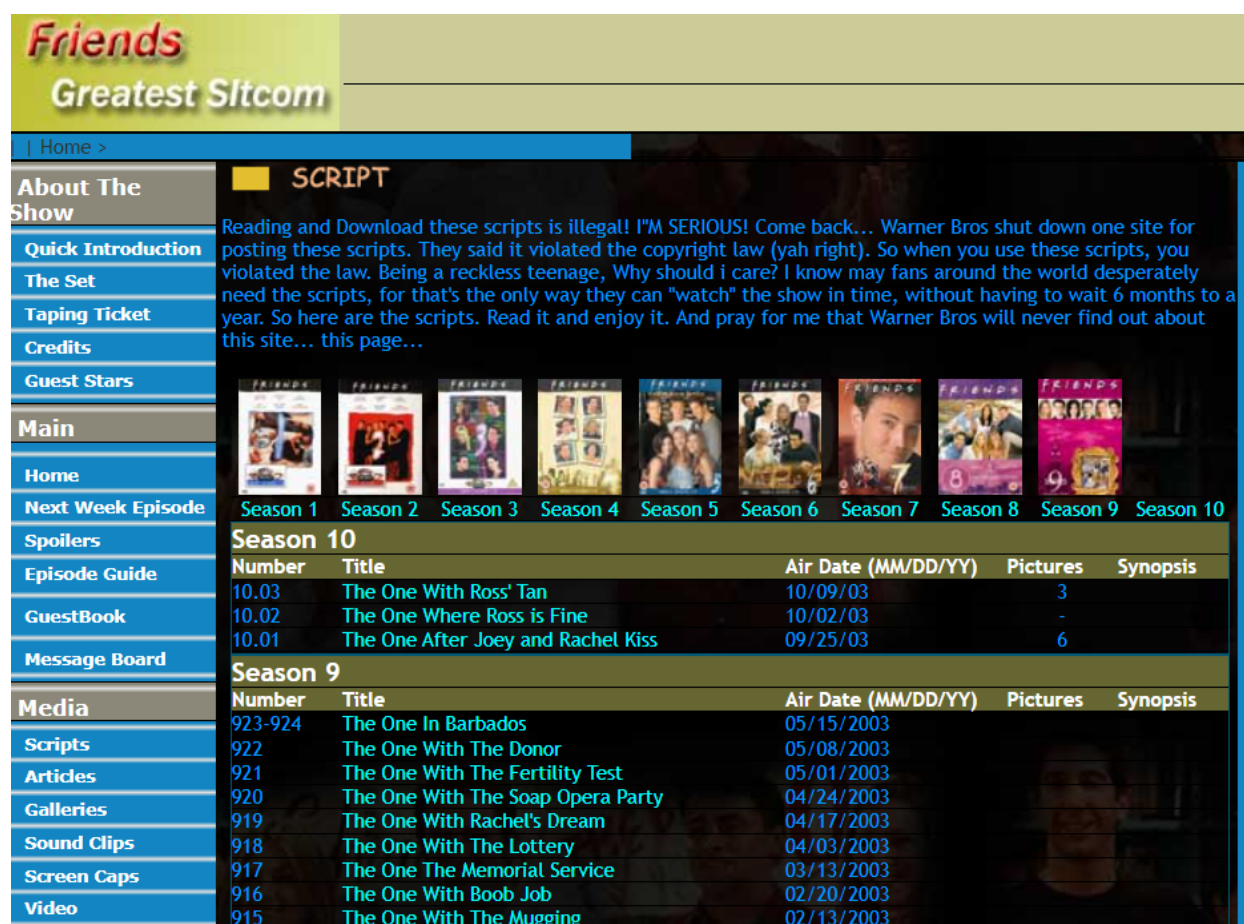
## جمع‌آوری داده

برای جمع‌آوری داده نیازمند پیکره‌ای بودیم که مشخص کند هر دیالوگ توسط کدام شخصیت گفته شده است. ابتدا سعی کردیم از زیرنویس‌های این سریال این امر را محقق سازیم اما مشکل این روش نامشخص بودن شخصیت‌ها بود.

روش دیگر نمایشنامه و متن سریال بود که در لینک زیر وجود داشت:

[https://www.oocities.org/friends\\_greatestsitcom/script.htm](https://www.oocities.org/friends_greatestsitcom/script.htm)

از پکیج Scrapy برای Crawl کردن روی این لینک استفاده شده است.



**Friends Greatest Sitcom**

| Home >

**SCRIPT**

Reading and Download these scripts is illegal! I'M SERIOUS! Come back... Warner Bros shut down one site for posting these scripts. They said it violated the copyright law (yah right). So when you use these scripts, you violated the law. Being a reckless teenage, Why should i care? I know may fans around the world desperately need the scripts, for that's the only way they can "watch" the show in time, without having to wait 6 months to a year. So here are the scripts. Read it and enjoy it. And pray for me that Warner Bros will never find out about this site... this page...

Season 1 Season 2 Season 3 Season 4 Season 5 Season 6 Season 7 Season 8 Season 9 Season 10

**Season 10**

Number	Title	Air Date (MM/DD/YY)	Pictures	Synopsis
10.03	The One With Ross' Tan	10/09/03	3	
10.02	The One Where Ross is Fine	10/02/03	-	
10.01	The One After Joey and Rachel Kiss	09/25/03	6	

**Season 9**

Number	Title	Air Date (MM/DD/YY)	Pictures	Synopsis
923-924	The One In Barbados	05/15/2003		
922	The One With The Donor	05/08/2003		
921	The One With The Fertility Test	05/01/2003		
920	The One With The Soap Opera Party	04/24/2003		
919	The One With Rachel's Dream	04/17/2003		
918	The One With The Lottery	04/03/2003		
917	The One The Memorial Service	03/13/2003		
916	The One With Boob Job	02/20/2003		
915	The One With The Mugging	02/13/2003		

**Media**

- Scripts
- Articles
- Galleries
- Sound Clips
- Screen Caps
- Video

## ساختار داده خام

فایل‌های استخراج شده توسط کراولر درون پوشه data/raw وجود دارند. این پوشه دارای دو فایل زیر می‌باشد:

**فایل scripts.csv:** این فایل شامل اطلاعات هر قسمت می‌باشد. ستون‌های آن به صورت زیر می‌باشد:

- season\_num
- episode\_num
- script\_link
- script\_title

	A	B	C	D	E	F	G	H
1	season_num	episode_num	script_link	script_title				
2	10	10.03	http://ww	The One With Ross' Tan				
3	10	10.02	http://ww	The One Where Ross is Fine				
4	10	10.01	http://ww	The One After Joey and Rachel Kiss				
5	9	923-924	http://ww	The One In Barbados				
6	9	922	http://ww	The One With The Donor				

**فایل dialogues.csv:** این فایل شامل دیالوگ‌های سریال می‌باشد. ستون‌های آن به صورت زیر می‌باشد:

- person
- dialogue

که اولی بیانگر کلاس دیالوگ می‌باشد.

	A	B	C	D	E	F	G	H
1	person	dialogue						
2	chandler	So, you and Rachel tonight, uh?						
3	joey	Yeah. It's actually our first official date						
4	chandler	Wow! So tonight may be the night! You're nervous?						
5	joey	Naa, no. This is the part I'm actually good at.						
6	chandler	What must it be like not to be crippled by fear and self-loathing.						

## روش تفکیک جملات






با توجه به این که هر رکورد دیتاست یک جمله دیالوگ است (اسکرپت اصلی) نیازی به تفکیک جملات نبود.

## روش تفکیک کلمات

از یکی از Tokenizer های درون NLTK به نام Treebank Word Tokenizer برای این امر استفاده شده است. توضیحات و دلیل استفاده از آن در تمرین شماره 1 به صورت کامل مورد بررسی قرار گرفت.

## پیش‌پردازش روی داده و تمیز کردن آن

پنج مرحله پیش‌پردازش روی دیتاست و مرحله به مرحله صورت می‌گیرد. نتیجه هر مرحله درون یک فایل جداگانه درون مسیر data/preprocessed موجود است.

Name	Date modified	Type	Size
 1.white_spaces.csv	5/23/2022 9:27 PM	Microsoft Excel Co...	671 KB
 2.lower_case.csv	5/23/2022 9:27 PM	Microsoft Excel Co...	671 KB
 3.special_chars.csv	5/23/2022 9:27 PM	Microsoft Excel Co...	615 KB
 4.short_words.csv	5/23/2022 9:27 PM	Microsoft Excel Co...	540 KB
 5.stop_words.csv	5/23/2022 9:27 PM	Microsoft Excel Co...	399 KB

## مرحله اول: حذف وایت اسپیس‌های اضافه

در این مرحله با استفاده از پکیج Regular Expression درون پایتون تمامی اسپیس‌ها، تب‌ها، خطوط جدید و ... با یک تک اسپیس جایگزین می‌شوند.

## مرحله دوم: تبدیل تمامی حروف به حروف کوچک

با استفاده از تابع lower پایتون تمامی کاراکترهای درون دیالوگ‌ها به حرف کوچک تبدیل می‌شوند.

## مرحله سوم: حذف کاراکترهای ویژه

در این مرحله با استفاده از پکیج RE پایتون تمامی کاراکترهای ویژه مانند [] ( ) ' " و ... حذف می‌شود.

## مرحله چهارم: حذف کلمات کوتاه

در این مرحله تمامی کلماتی که طول آن‌ها کمتر از 3 حرف می‌باشد حذف می‌شود.

## مرحله پنجم: حذف Stop words

با استفاده از مجموعه Stopwordهای درون NLTK تمامی این کلمات را از درون دیالوگ‌ها حذف می‌کنیم.

## واحد برچسب‌گذاری

با توجه به این که در این تسک سعی داریم دیالوگ‌ها را به یک شخصیت نگاشت دهیم، به ازای هر دیالوگ یک برچسب زدیم که بیانگر شخصیت گوینده آن دیالوگ است.



تعداد جملات: 9311

	Ross	Rachel	Joey	Chandler	Monica	Phoebe	All
Words count	112588	98972	88060	108566	93656	92032	593874
Tokens count	3596	3140	3029	3552	3125	3063	12457

[illegible]

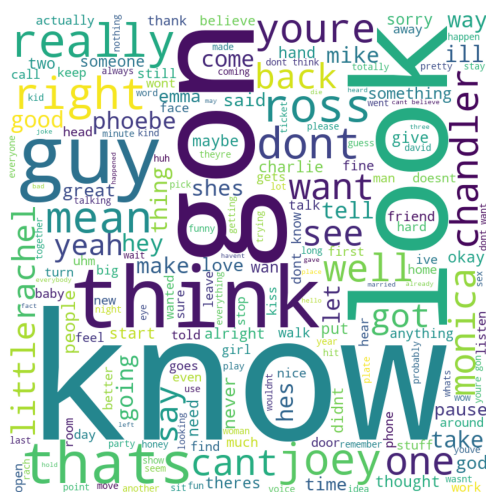
Word	Count (approx.)
i	3350
you	3100
the	2750
to	2450
a	2200
and	1600
that	1100
is	900
of	850
in	800
word	750
just	750
have	750
my	700
i'm	680
it	680
with	650
for	650
not	620
so	580
know	580

## آمار داده‌ها بعد از پیش‌پردازش

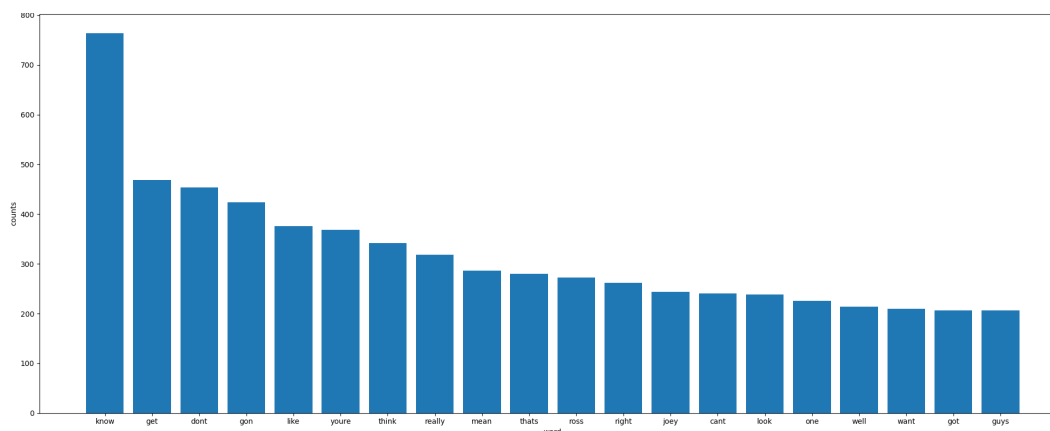
تعداد جملات: 8811

	Ross	Rachel	Joey	Chandler	Monica	Phoebe	All
Words count	63210	54034	49838	61056	51678	51238	331054
Tokens count	2408	2080	2013	2405	2087	2023	8319

ابر کلمات



هیستوگرام



**نکته:** این دو نمودار برای تک تک شخصیت‌ها موجود می‌باشد.

## سایر دیتاست‌های آماده (استفاده نشده)

در میان جست‌وجو یک دیتاست پیدا شد که مربوط به دیالوگ درون فیلم‌ها و شخص گوینده آن دیالوگ بود. در این دیتاست همچنین مخاطب آن دیالوگ نیز وجود داشت. لینک آن در ادامه آورده شده است:

[https://www.cs.cornell.edu/~cristian/Cornell\\_Movie-Dialogs\\_Corpus.html](https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html)