DSCI 560 Lab 2 Report

Group Name: jjxxyy

Github: https://github.com/leo-ye0/dsci560-lab2

Demo Link: https://youtu.be/v2H5US_mBJg

John Tran 8897109598

Xiaohai Dong 1031572511

Yutao Ye 2448443089

Brain-storming Ideas

Xiaohai's ideas:

- Second hand trading group
    - Datasets:
        - Craigslist (CSV): A public trade platform, covering second-hand merchandise, cars, house rentals and so on. https://losangeles.craigslist.org/
        - Commercial second hand company websites: Focus on different categories items, more authoritative. https://www.therealreal.com/
        - Legal policies (PDF): Rules about second-hand trading, fraud detection, and buyers and sellers' responsibilities. https://www.ebay.com/help/policies/default/ebay-rules-policies?id=4205
    - Reasons: An AI assistant can help users post, search, manage, and make decisions more efficiently in a group chat. Users are able to compare prices, qualities, and more between the items in the group chat and the similar items on the internet.

Yutao's ideas:

- **FIFA World Cup 2026 fans chat**
    - Inspiration:
        - The 2026 World Cup is a perfect topic because it's the biggest single-sport event globally, guaranteeing a massive, engaged audience. Its unique format—the first with 48 teams hosted across three nations

(USA, Canada, Mexico)—creates many new points for discussion and analysis.
    - Fans can gather historical information and up to date information of the World Cup within group chat.
- Features:
    - **Pre-Tournament Phase**
        - **Qualification Data:** Live updates on which nations have qualified from each confederation.
        - **Squad Announcements & Injuries:** Track official national team squad selections and key player injuries.
        - **Venue & Schedule Info:** Provide details on host cities (like nearby Los Angeles), stadium information, tickets information, and the official match schedule once it's released.

    - **During-Tournament Phase**
        - **Live Match Data:** Instant scores, goals, cards, and major events like VAR checks. This is essential for real-time conversations.
        - **Group Standings & Permutations:** Real-time updates on group tables and knockout round scenarios. The AI can answer, "What does Team USA need to do to advance?"
        - **Player Stats:** Track the Golden Boot race, top assists, and other key player metrics to fuel debates on individual performances.
        - **AI comments:** Provide AI perspective.

    - **Historical-Tournament Phase**
        - **Tournament Rules:** Explain the new 48-team format, group stage tie-breakers, and knockout rules. The agent acts as the expert on the new format.
        - **World Cup History:** Provide quick facts on past winners, famous matches, and historical records to add depth and settle trivia arguments.

- Example Datasets:
    - **Kaggle FIFA World Cup historical Data (csv)**

        - Provide pre-match analysis to predict.
        - Knowledge base for historical matches
        - Source: https://www.kaggle.com/datasets/abecklas/fifa-world-cup

    - **Reddit (ASCII text)**
        - Voice of the fans
        - Learn how to comment for AI
        - Source: https://www.reddit.com/r/worldcup/

    - **Official 2026 World Cup Documentation (PDF)**
        - Deep, specific knowledge.
        - Complex questions about the new 48-team format, tie-breaker rules, or specific stadium logistics that aren't available in simple web articles.
        - Source: https://digitalhub.fifa.com/m/18d857c3ec3e64f8/original/FIFA-World-Cup-2026-Preliminary-Competition_EN.pdf
        - https://inside.fifa.com/official-documents

John's ideas:

- Chat platform to discuss sports events
    - This is a good way to get more users, as it creates a community where users can interact with each other
    - Sports are something that a lot of people follow, so this is a good way to find more users
- Features:
    - Polling
        - This allows users to make polls to vote on various topics (e.g., who will win the next game, will a player be sold to a different team, who will win the league)
    - Calendar/Event
        - This allows users to create an event (e.g., watch party, reminders) and notify others
    - Leaderboard

        - This allows users to create a leaderboard to keep track of points scored (e.g., correct predictions)
- Example datasets:
    - NBA Rulebook (PDF)
        - We can train the chatbot on sports rulebooks to answer questions that users have when discussing certain plays and calls
        - Source: https://official.nba.com/wp-content/uploads/sites/4/2023/10/2023-24-NBA-Season-Official-Playing-Rules.pdf
    - List of teams (csv)
        - We can train the chatbot on datasets on information about the league, teams, players, etc.
        - Source: https://www.kaggle.com/datasets/wyattowalsh/basketball/data
            - This is a SQLite database, but data can be easily queried and extracted in the form of a CSV file.
    - NBA leaderboard (HTML)
        - We can train the chatbot on data about real-time rankings of teams
        - Source: https://www.espn.com/nba/standings

FIFA World Cup 2026 fans group chat

Reasoning:

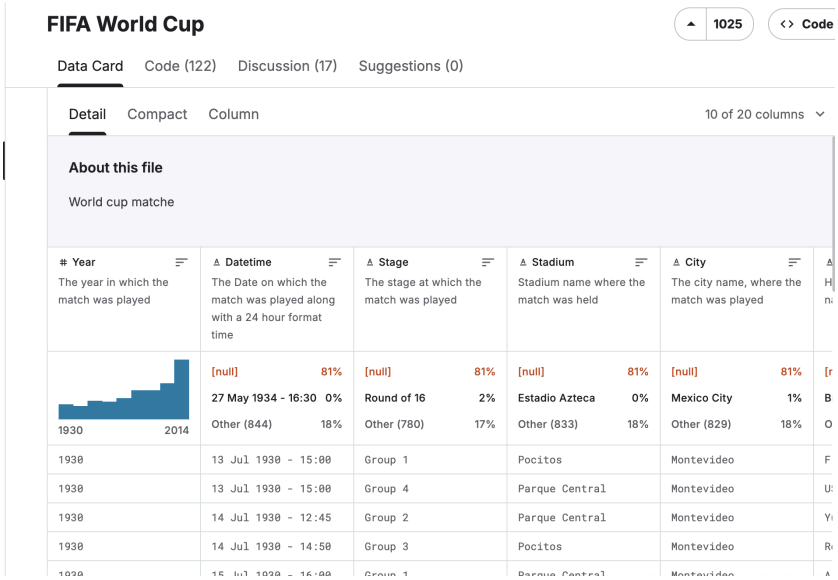
The 2026 World Cup is a great topic for our chat for three main reasons.

First, it is the biggest sports event in the world, so millions of fans will be excited to talk about it.

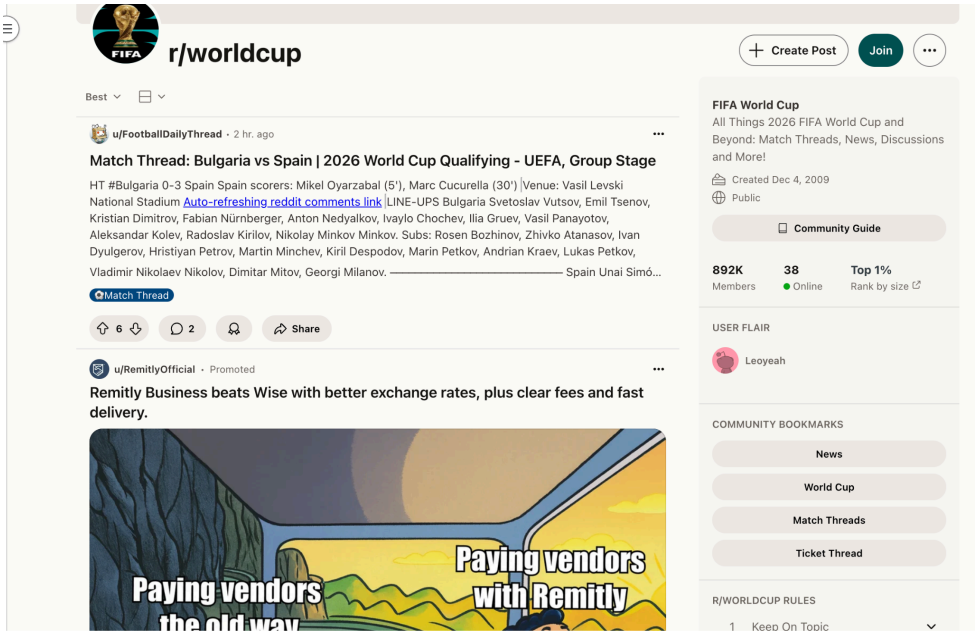
Second, this World Cup is special. It will be the first time with 48 teams, and the games will happen across three countries: the USA, Canada, and Mexico. This gives fans many new things to discuss. Fans may have many questions about the games, and they need a well knowledged assistant.

Finally, our chat will be a helpful guide for fans in all 16 host cities and all around the world. It will provide important info for places like Vancouver, Mexico City, and LA, such as game schedules and stadium details.

Example Datasets:

- **Kaggle FIFA World Cup historical Data (csv)**
  - Provide pre-match analysis to predict.
  - Knowledge base for historical matches
  - Source: https://www.kaggle.com/datasets/abecklas/fifa-world-cup
  - 

  - Description: The World Cups dataset shows all information about all the World Cups in history, while the World Cup Matches dataset shows all the results from the matches contested as part of the cups.

- **Reddit (ASCII text)**
  - Voice of the fans
  - Learn how to comment for AI
  - **Source:** https://www.reddit.com/r/worldcup/
  - 

  - **Description:** The **r/worldcup** subreddit is an online community on Reddit where soccer fans from around the globe come together to share news, media, and discussion about the FIFA World Cup and other major tournaments. It serves as a central hub for everything from match highlights and photos to debates about team performance and historical moments, with activity peaking during the actual tournament. It's essentially a dedicated space for fans to connect and engage with the biggest event in the sport. And Reddit has a helpful API to gather data.

- **Official 2026 World Cup Documentation (PDF)**
  - Deep, specific knowledge.
  - Complex questions about the new 48-team format, tie-breaker rules, or specific stadium logistics that aren't available in simple web articles.
  - Source: https://www.concacaf.com/media/uixf2p3u/concacaf-qualifiers_final-round-v1.pdf

- This pdf shows the schedule for the CONCACAF Qualification matches
- For this lab, I have chosen to extract the matchup information as well as which team is the home and away team

Data Collection

Kaggle Data (Xiaohai Dong)

The Python script downloads and processes the data for AI chat training.

Data Source:

https://www.kaggle.com/datasets/abecklas/fifa-world-cup

Data Collection:

- Downloads 3 CSV files (WorldCups.csv, WorldCupMatches.csv, and WorldCupPlayers.csv) using Kagglehub Python API from the Kaggle dataset.

```python
def download_worldcup_csv():
    # download dataset
    path = kagglehub.dataset_download("abecklas/fifa-world-cup")
    print("Dataset downloaded to:", path)

    df_names = {}
    # file names
    csv_files = [
        "WorldCups.csv",
        "WorldCupMatches.csv",
        "WorldCupPlayers.csv"
    ]

    output_dir = os.path.join(os.path.dirname(__file__), "../data")
    os.makedirs(output_dir, exist_ok=True)
```

Data Cleaning:

- In WorldCups.csv, convert the “Year” and “GoalScored” columns data type into integers for future data processing, and fill in the missing values with 0.
- In WorldCupMatches.csv, convert the “Year”, “Home Team Goals”, and “Away Team Goals” to integers.

```python
for filename in csv_files:
    csv_path = os.path.join(path, filename)
    if os.path.exists(csv_path):
        df = pd.read_csv(csv_path)
        if filename == "WorldCups.csv":
            df["Year"] = pd.to_numeric(df["Year"], errors="coerce").fillna(0).astype(int)
            df["GoalsScored"] = pd.to_numeric(df["GoalsScored"], errors="coerce").fillna(0).astype(int)

        elif filename == "WorldCupMatches.csv":
            df["Year"] = pd.to_numeric(df["Year"], errors="coerce").fillna(0).astype(int)
            df["Home Team Goals"] = pd.to_numeric(df["Home Team Goals"], errors="coerce").fillna(0).astype(int)
            df["Away Team Goals"] = pd.to_numeric(df["Away Team Goals"], errors="coerce").fillna(0).astype(int)
        output_path = os.path.join(output_dir, filename)
        df.to_csv(output_path, index=False)
        print(f"saved {filename} to {output_path}")
        df_names[filename] = df
    else:
        print(f"error: file {filename} not found")
```

Data Analysis:

- Shows the total tournaments, year range, and the top 3 winning countries in WorldCups.csv
- Shows the match count, average goals, highest-scoring matches, and top host cities in WorldCupMatches.csv
- Shows the total appearances, unique players, and the top 3 most frequent players in WorldCupPlayers.csv

```
if "WorldCups.csv" in dataframes:
    df = dataframes["WorldCups.csv"]
    print("\nWorld Cups Summary")
    print(f"Tournaments: {len(df)}")
    print(f"Years: {df['Year'].min()} - {df['Year'].max()}")
    print(f"Most Wins:\n{df['Winner'].value_counts().head(3).to_string()}")


if "WorldCupMatches.csv" in dataframes:
    df = dataframes["WorldCupMatches.csv"]
    df["TotalGoals"] = df["Home Team Goals"] + df["Away Team Goals"]
    print("\nMatch Stats")
    print(f"Total Matches: {len(df)}")
    print(f"Average Goals/Match: {df['TotalGoals'].mean():.2f}")
    print("Top Scoring Matches:")
    print(df.sort_values("TotalGoals", ascending=False)[[
        "Year", "Home Team Name", "Away Team Name", "TotalGoals"
    ]].head(3))
    print(f"Top Match Cities:\n{df['City'].value_counts().head(3).to_string()}")

if "WorldCupPlayers.csv" in dataframes:
    df = dataframes["WorldCupPlayers.csv"]
    print("\nPlayer Stats")
    print(f"Total Appearances: {len(df)}")
    print(f"Unique Players: {df['Player Name'].nunique()}")
    print("Most Frequent Players:")
    print(df['Player Name'].value_counts().head(3).to_string())
```

Output:

3 CSV files downloaded into the “data” folder.

A brief summary of these 3 data:

```
World Cups Summary
Tournaments: 20
Years: 1930 - 2014
Most Wins:
Winner
Brazil       5
Italy        4
Germany FR   3

Match Stats
Total Matches: 4572
Average Goals/Match: 0.53
Top Scoring Matches:
      Year Home Team Name Away Team Name  TotalGoals
94    1954        Austria    Switzerland          12
87    1954        Hungary     Germany FR          11
312   1982        Hungary    El Salvador          11
Top Match Cities:
City
Mexico City        23
Montevideo         18
Rio De Janeiro     18

Player Stats
Total Appearances: 37784
Unique Players: 7663
Most Frequent Players:
Player Name
RONALDO    33
KLOSE      32
OSCAR      28
```

Reddit Forum Data (Yutao Ye)

The script performs these data cleaning tasks for chat agent training:

Data Collection:

- Scrapes r/worldcup subreddit posts using Reddit API (praw)

- Extracts only essential fields: title, text, score, num_comments, top_comments

```
reddit = praw.Reddit(
    client_id="YOUR_CLIENT_ID",
    client_secret="YOUR_CLIENT_SECRET",
    user_agent="reddit_username"
)

def scrape_worldcup_data(limit=100):
    """Scrape posts and comments from r/worldcup for chatbot training data"""

    subreddit = reddit.subreddit("worldcup")
    posts_data = []

    # Get hot posts
    for post in subreddit.hot(limit=limit):
        post_info = {
            'title': post.title,
            'text': post.selftext,
            'score': post.score,
            'num_comments': post.num_comments
        }
```

Data Filtering:

- Removes bot/moderator messages that start with "Hello! Thanks for your submission"

- Collects up to 5 real user comments per post (checks 15 to filter out bots)

- Filters posts by engagement metrics (score, comment count)

- limit=None: Expand all comments (slower, more complete). limit=5: Expand up to 5 "MoreComments" objects. For our use case, limit=0 is good because we only want the top 5 comments for this simple lab.

```python
def scrape_worldcup_data(limit=100):
    """Scrape posts and comments from r/worldcup for chatbot training data"""

    subreddit = reddit.subreddit("worldcup")
    posts_data = []

    # Get hot posts
    for post in subreddit.hot(limit=limit):
        post_info = {
            'title': post.title,
            'text': post.selftext,
            'score': post.score,
            'num_comments': post.num_comments
        }

        # Get top comments for context
        post.comments.replace_more(limit=0)
        comments = []
        for comment in post.comments[:15]:  # Check more to filter out bots
            if hasattr(comment, 'body') and not comment.body.startswith('Hello! Thanks for your submission'):
                comments.append(comment.body)
                if len(comments) >= 5:  # Stop at 5 real comments
                    break

        post_info['top_comments'] = comments

        # Only keep posts with at least 2 real comments
        if len(comments) >= 2:
            posts_data.append(post_info)

    return pd.DataFrame(posts_data)
```

Data Quality Tools:

- analyze_data_quality() - Shows post count, text coverage, average scores
- Score filtering capability for high-quality content (Upvotes-Downvotes)
- Sample data preview for validation

```python
def analyze_data_quality(df):
    """Analyze scraped data for chatbot training suitability"""

    print(f"Total posts: {len(df)}")
    print(f"Posts with text: {len(df[df['text'].str.len() > 0])}")
    print(f"Average score: {df['score'].mean():.2f}")
    print(f"Average comments: {df['num_comments'].mean():.2f}")

    # Show sample data
    print("\nSample post titles:")
    for title in df['title'].head(5):
        print(f"- {title}")
```

Output:

Please refer to

https://github.com/leo-ye0/dsci560-lab2/blob/main/dsci560-lab2/data/worldcup_data.csv

- Single clean CSV file with conversation pairs:

- Posts (title + text) = Questions/Topics

- Comments = Responses/Answers

```
(venv) yutaoye@yutaoye-VMware20-1:~/Desktop/dsci560-lab2$ python scripts/data_exploration.py
Total posts: 100
Posts with text: 65
Average score: 96.73
Average comments: 70.92

Sample post titles:
- Match Thread: Bulgaria vs Spain | 2026 World Cup Qualifying - UEFA, Group Stage
- FIFA Launches Dynamic Pricing for 2026 World Cup: Tickets Start at $60, Could Soar to $6,730
- Match Thread: Luxembourg vs Northern Ireland | 2026 World Cup Qualifying - UEFA, Group Stage
- Match Thread: Slovakia vs Germany | 2026 World Cup Qualifying - UEFA, Group Stage
- Match Thread: Georgia vs Turkey | 2026 World Cup Qualifying - UEFA, Group Stage

Data saved to data/worldcup_data.csv
```

	title	text	score	num_comme	top_comments
2	FIFA Launches Dynamic Pr		252	171	['It‚Äôs price gouging not dynamic pricing', 'So the worst possible outcome and exact reason everyone fucking hates Ticketmaster. Awesome.', 'Reading all the comments from Americans, I think this W
3	Match Threa 86'		5	4	['GAWA', 'Green and White Army', 'Great run from Bradley there']
4	World cup ex I'm looking		3	86	['Realistically how many ‚Äúbig‚Äù teams are there? 10? 15 maybe? I get it‚Äôs fun to watch the top teams play, but there‚Äôs more than just France and Brazil when it comes to international soccer. \
5	Real Madrid, So I'm sure		5	3	['Unlikely that Rashford will feature much for England in the World Cup but you never know...\n\nObviously Trent and Bellingham play for Madrid and it is likely that one of them will feature, Kane too,
6	South Africa, South Africa		22	7	['It‚Äôs the same with Morocco who played Congo (a) and Niger (a) at home. South Africa is perhaps more understandable due to geographical proximity anyway but yes it‚Äôs very unfair. An argument
7	2026 FIFA W # Welcome		132	84	['Any news on prices of Cat 1, 2, 3 and if there will be a Cat 4 and who will qualify for it? Also I‚Äôd love to know how many matches is the maximum you‚Äôll be able to apply for. Please delete politica
8	Unpopular O I've heard		15	8	["Martinez being shit is a twitter/social media bantz brainrot meme. Just like the dunking on Harry Maguire a few years ago. Something not really based in reality but makes it easy for people to farm
9	US cities confront FIFA ov		162	49	['All that this will do is push celebrations into bars and out of major public venues.', 'Glad Chicago told them to kick rocks.', '> While Congress has approved $625 million in security funding nationwide, I
10	Which natior As the		32	28	['In terms of timing, the only team from Africa that can qualify after matchday 7, as far as I can tell, is Morocco. They have a 6 point difference over 2nd place Tanzania, who they have beaten twice and
11	Fifa Panini s Are the pani		19	12	['They‚Äôre a big thing in Argentina.', 'Brazil too', 'I have been living in Canada (not Canadian) for the last 4 years and no the only people who engage in exchanging stickers are Latinos', 'Big in the UK. (
12	Underrated t Now that		16	27	['Guatemala is my pick. Not only have they done pretty well the last couple of years (got to Gold Cup semifinal just a couple of months back), but Concacaf has a bunch of spots at play. 3 direct and 2 pl
13	‚ÄòMy Last‚Äô Lionel M		332	49	['Since 2030 is also in Argentina, the team is qualified already, so no qualify matches for Argentina until the WC2034 ;)', 'The fine print apart from the clickbait headline:\n\n>Inter Miami and Argentina
14	Curious if pe This should r		78	523	["Not in the same way, given that the US hasn't enslaved thousands of workers to build infrastructure. But I won't be watching the WC for the third time in a row.", 'Reddit moment', 'No. People are lit
15	When the Pitch Burns ‚Äì		33	31	['Yes, let‚Äôs pull the North American World Cup due to heat but keep awarding it to countries that are primarily deserts.', 'Time for Scotland to step up.\n\nThose blistering July days of 12-∞c mean it
16	For the first time in 52 ye		817	285	['This is like when at work they don‚Äôt give raises or bonuses but then they take away snacks from the snack room. Big misstep.', "It's insane that the World Cup TV rights in Argentina are only worth
17	Traveling By Train Betwee		42	25	['It was possible like over 100 years ago too, no?', 'No way it will be feasible at all', 'No way in hell this could happen by then.', 'Lmfao üò≠ that shits taking 30 days to get across', 'For everyone readin
18	The Rondo: Debating the (		6	4	["He got called up for the friendlies. I imagine he's on a short leash.", 'Tbf. He doesnt really have much talent to work with', "It must have a MLS quota in Pochettino's contract"]
19	Why is the A What is this		22	27	['i get confused why CAF hasnt made a system like AFC where the qualifiers are for both WC and Asian Cup. Africans teams play different qualifiers for both tournaments.', "It's because CAF runs their
20	Among Austi One thing I		38	142	['In Australia, ‚ÄúSoccer,‚Äù has a larger proportion when it comes to (non-cricket) ball sports by virtue of the fact that there‚Äôs also a big geographical rift between Rugby League (NRL aka ‚ÄòFooty,‚Äô
21	I had this wil Just		12	24	['I can imagine nowhere less suited to hosting a world cup than Cambodia, holy shit.', 'Let me quickly throw cold water to this idea. 2007 Asian Cup. Join hosts of Indonesia, Malaysia, Thailand, Vietnam
22	IYO, what co This merges		4	8	['The Special One: Mourinho!', 'I‚Äôm happy with Martinez. We‚Äôre playing some attractive football again and he‚Äôs doing a good job of blending youth and veteran leadership. Why change?', 'Bro se
23	The United K Y‚Äôall		0	20	['I know what all these words mean but this doesn‚Äôt make any sense, starting mainly with the fact that the UK have 4 different national teams, so which county would benefit? Plus the Class of ‚Äò92
24	Donald Trump says Vladim		373	162	["Russia are banned, so fifa won't allow it.", "Russia is not even playing, what's the point in putin attending", 'we are still only at the beginning of this', 'Yeah... No thanks bro', 'The scene when Ukraine
25	"Can I keep it?": Trump as		123	44	['This dude is like gold goblin', 'Trump always says he dosen‚Äôt kid around so I‚Äôll take it as he wanted to keep it because he‚Äôs a child', 'Remember when Blatter left and people thought FIFA would
26	World Cup draw will be he		84	58	['Sphere in Vegas would‚Äôve been so good.', 'Of course Trump just couldn‚Äôt *stand* having it at Sphere in Vegas, you know, the place *they originally announced* the draw to be. Would rather have
27	Fun fact ‚Äì I These stats c		29	7	["It still pisses me off that the 2030 World Cup isn't gonna be in Uruguay", '1981 tournament was a good press moment by the militairy junta that ruled Uruguay', 'Oh yeah? Well, Canada is ranked 28th
28	FIFA has opened applicati		170	98	['FIFA should have to pay people. We‚Äôre all losing our shirts but rich guys want volunteers? Classic.', 'Make billions off of the event but need to ask for free help. So backwards', 'A massive money-ma
29	Wich Nation That's it. For		27	62	["I'd say Portugal (Bruno Fernandes, Vitinha, Neves, Silva), tied with Spain (Pedri, Rodri, Olmo, Ruiz)\n\nthen Argentina (Enzo, Mac Allister, Almada, De Paul)", 'Portugal. Argentina. Spain. Netherlands
30	Fifa consider holding Club		165	217	['Stupid decision that will make the tournament loose its appeal due to over-exposition. FIFA under Infantino became even more greedy than it was before', 'They‚Äôre really turning football to a 9-5 jo
31	FIFA World C Hey		22	32	["For the love of God, don't let this corrupt organization use you for free labor.", 'Just curious, what do you expect to be positive from this experience? Not judging, but it just seems like free labor for ar
32	To save time It should, to		9	6	['I did like that, but with afc every 4 years and no nations league they could use more games', 'No. Asia needs a version of Nations League. While this is exhausting for Europe, it is necessary for Asia.\n
33	Should we ju Is it?		19	7	['A Nations League is not even necessary for every confederation', "Should we pretend that something is something that's it's not? No.", '...what?', 'My hot take is that each federation should have a na
34	Where‚Äôs n I had		88	25	['It hasn‚Äôt started yet. Only hospitality is available at the moment', 'Be patient, it‚Äôs gonna take a while.', '"There will be several distinct ticket sales phases from the start of sales on 10 September
35	‚ÄòGet in now‚Äô: Soccer		63	145	['You are correct businesses absolutely want to steal as much of your money as possible because the WC is coming to town.', "They can't stick on ads during the game, it'll never take up", 'The sport of
36	What bids w Technically		0	73	['England! Absolutely worthy. I am 50 and wondering if I will ever see a WC in my country.', 'Croatia, Austria, Hungary\n\nDifferent cultures, still in Europe, great weather, mountains or beach, not too e
37	The 2030 W Made a vide		13	3	['No worries this looks good.', 'I was always passionate about this topic, since i was a kid, but the decision made felt bittersweet. Hope you can learn a bit more of the importance of this edition and the
38	Bold Take: Si That's it. I th		0	123	['That Spain team is maybe the most insanely unbeatable team I‚Äôve ever seen. They would control 85 to 90% possession and barely break a sweat.', 'Spain 2010 and France 2018 had the hardest sem

	title
1	FIFA Launches Dynamic Pricing for 2026 World Cup: Tickets Start at $60, Could Soar to $6,730
2	Match Thread: Luxembourg vs Northern Ireland | 2026 World Cup Qualifying - UEFA, Group Stage
3	World cup expansion has completely ruined the qualification process. Time to scrap the procession and insert more meaningful fixtures.
4	Real Madrid, Barcelona, Bayern and Inter players in World Cup final
5	South Africa’s unique home ground advantage in the qualifiers
6	2026 FIFA World Cup Ticket & Travel Information Thread
7	Unpopular Opinion: Martinez is not that bad as a NT coach
8	US cities confront FIFA over World Cup costs: Local organizers are already starting to cut back on one of the tournament's hallmarks.

text
86' #**Luxembourg 1-3 Northern Ireland** Luxembourg scorers: Aiman Dardari (30') Northern Ireland scorers: Jamie Reid (7'), Shea Charles (46'), Justin Devenny
I'm looking ahead to this week's World cup qualifiers and all I can see is a bunch of mismatch encounters across the entire global qualification. What a waste of tim
So I'm sure everyone knows about the Bayern and Inter players featuring in the final. However, the two Spanish giants do have a part too. In 6 out of 7 last editions,
South Africa were drawn in CAF Group C together with Nigeria, Rwanda, Zimbabwe and Lesotho. Well it turns out that Zimbabwe and Lesotho do not have FIFA-acc
Welcome to the 2026 FIFA World Cup™ Ticket & Travel Information Thread! Your complete guide for buying tickets, checking fixtures, and planning your epic Wor
I've heard from left to right that Roberto Martinez is too limited and a profitional talent-wasting coach. I, myself included, critized him for being stubborn, way too d
As the September window is upon us, 12 possible nations now have the opportunity to clinch their spot at the World Cup in this break according to FIFA thus joining

	score	num_comments	top_comments
	252	171	['It’s price gouging not dynamic pricing', 'So the worst possible outcome and exact reason every
	5	4	['GAWA', 'Green and White Army', 'Great run from Bradley there']
	3	86	['Realistically how many “big” teams are there? 10? 15 maybe? I get it’s fun to watch the top tea
	5	3	['Unlikely that Rashford will feature much for England in the World Cup but you never know...\n\r
	22	7	['It’s the same with Morocco who played Congo (a) and Niger (a) at home. South Africa is perha
your 2026 World Cup experience unforgettable!**	132	84	['Any news on prices of Cat 1, 2, 3 and if there will be a Cat 4 and who will qualify for it? Also I’d
	15	8	["Martinez being shit is a twitter/social media bantz brainrot meme. Just like the dunking on Har
	162	49	['All that this will do is push celebrations into bars and out of major public venues.', 'Glad Chicag
	32	28	['In terms of timing, the only team from Africa that can qualify after matchday 7, as far as I can t
	19	12	['They’re a big thing in Argentina.', 'Brazil too', 'I have been living in Canada (not Canadian) for th

CONCACAF Matchups Data (John Tran)

This script takes a PDF file of the CONCACAF qualification matches and converts them into a csv file with the team matchups.

Data Source

- https://www.concacaf.com/media/uixf2p3u/concacaf-qualifiers_final-round-v1.pdf
- 

- This pdf shows the schedule for the CONCACAF Qualification matches
- For this lab, I have chosen to extract the matchup information as well as which team is the home and away team
- Next steps: extracting dates, matchdays, groups, etc.

PDF Loading

```
def load_pdf(file):
    with pdfplumber.open(path) as pdf:
        page = pdf.pages[0]
        text = page.extract_text()
        print(f'Loading pdf: {file}\n')
    return text
```

- Loads in PDF file and extracts text from the page

Text-to-List Conversion

```python
# Filters the text to only get matchups between countries
def get_matchups(text, keywords):
    home_v_away = []

    for line in text.split('\n'):
        line = line.split()
        if 'V' in line:
            index = line.index('V')
            left, right = line[index - 1], line[index + 1]
            if left in keywords:
                home = keywords[left]
            else:
                home = left
            if right in keywords:
                away = keywords[right]
            else:
                away = right
            home_v_away.append([home, away])

    return home_v_away
```

```python
keywords = {'El':'El Savador', 'Savador':'El Savador',
        'Trinidad':'Trinidad & Tobago', 'Tobago':'Trinidad & Tobago',
        'Costa':'Costa Rica', 'Rica':'Costa Rica'}
```

- Add in **keywords** dictionary to help with clarity when extracting the matchups
    - Getting the first words between 'V' (stands for versus)
    - **Keywords** dictionary helps with clarity when extracting ('Salvador', 'V', 'Costa' will be interpreted as 'El Salvador V Costa Rica')
    - The team on the left of 'V' is the home team, and the team on the right of 'V' is the away team
    - Append the matchups into a list

List-to-Dataframe Conversion

```python
def matchups_to_df(matchups_list):
    df = pd.DataFrame(matchups_list, columns=['Home', 'Away'])
    return df
```

- The list is then converted into a dataframe, which is then exported into a CSV file
- The CSV file has columns 'Home' and 'Away'

Output

- A CSV file displaying matchups and which teams are the Home and Away team

	Home	Away
2	Suriname	Panama
3	Guatemala	El Savador
4	Bermuda	Jamaica
5	Trinidad & Tobago	Curacao
6	Nicaragua	Costa Rica
7	Haiti	Honduras
8	Panama	Guatemala
9	Salvador	Suriname
10	Jamaica	Trinidad & Tobago
11	Curacao	Bermuda
12	Costa Rica	Haiti
13	Honduras	Nicaragua
14	Salvador	Panama
15	Suriname	Guatemala

Discussion

Most chatbots feel flat because they don't really get what you're saying or feeling. Talking to them isn't like talking to a real fan. Real people can follow what's happening right now, share in the highs and lows, and actually know the details that make the game exciting. Most bots just can't keep up.

What Today's Chatbots Are Missing?

Real-Time Awareness: Most chatbots are stuck in the past. They don't know that today—September 5th, 2025—is huge for World Cup qualifying. They can't talk about the CONMEBOL matches tonight or how the CONCACAF standings could make or break the USA's chances. They miss the live drama.

Personality: A typical bot can tell you the score, but it won't feel the moment. If you say, "That was a nail-biter," it won't share your excitement. Fans speak with passion, jokes, and slang—chatbots don't.

Deep Knowledge: Regular bots know trivia, but not the details that matter. They might tell you who won the last World Cup, but if you ask, "How do tie-breakers work under the new 48-team format?" they won't have a clue. They don't have the depth of a real expert.

How Do We Fix That?

Authentic Personality (Reddit Data): By training on fan discussions from places like r/worldcup, the bot learns the voice of real fans. It picks up the slang, the jokes, the banter—so conversations feel natural and passionate, not robotic.

Expert Knowledge (FIFA PDFs + Kaggle Data): Official FIFA documents give it the rulebook-level understanding that even hardcore fans need to look up. Historical Kaggle data gives it a perfect memory—instant recall of match results, player stats, and records to settle any debate.

Real-Time Awareness (Live Feeds): With live data, it stays connected to what's happening now. It can talk about tonight's qualifiers as they unfold, not just yesterday's scores.

Put all this together, and the chatbot isn't just another information tool. It becomes a true companion for fans—someone who knows the game, feels the excitement, and can talk about it like they're sitting right next to you.

Github Commits

Commits on Sep 7, 2025

Update README with Kaggle download instructions ttx-0802 authored 2 hours ago	Verified	23a4c7a
Upload 3 csv files ttx-0802 authored 2 hours ago	Verified	a34ca62
Upload data_exploration_kaggle.py ttx-0802 authored 2 hours ago	Verified	6b7da3c
Update readme ttx-0802 authored 2 hours ago	Verified	b6d6b54
restructure pdf_conversion.py and added to README file johntusc committed 6 hours ago		65f2154

Commits on Sep 6, 2025

rearranged files johntusc committed yesterday		c404be0
added pdf_conversion.py, concacaf_matchups.csv, and concacaf_qualifiers.pdf johntusc committed yesterday		4ebee6b

Commits on Sep 5, 2025

mask info leo-ye0 committed 2 days ago		e3a0092
Delete data directory leo-ye0 authored 2 days ago	Verified	ed1f0a3
Delete demo_script.md leo-ye0 authored 2 days ago	Verified	9585385
Delete requirements.txt leo-ye0 authored 2 days ago	Verified	1d5df36
Delete scripts directory leo-ye0 authored 2 days ago	Verified	526784b
Add folde leo-ye0 committed 2 days ago		e60a10f
Restructure: move files to repository root leo-ye0 committed 2 days ago		1b482b2
Add demo script and readme leo-ye0 committed 2 days ago		524dd06

Commits on Sep 4, 2025

Remove sensitive api info leo-ye0 committed 3 days ago		d56c8d3
Remove low comment thread leo-ye0 committed 3 days ago		6c44909
Process reddit data leo-ye0 committed 3 days ago		0665867
raw reddit data exploration leo-ye0 committed 3 days ago		dda9394
Add data folder with .gitkeep leo-ye0 committed 3 days ago		f5660d6
Set up folders leo-ye0 committed 3 days ago		e6d81ae