**Web Crawler Report**
**Shreya Valaboju, Soham Mukherjee**
**CS 4395.001**

**7. In a doc: (1) describe how you created your knowledge base, include screenshots of the knowledge base, and indicate your top 10 terms;**

Our top 10 terms were prominent basketball players for the Dallas Mavericks. These were:

**1: doncic**
**2: irving**
**3: wood**
**4: bertans**
**5: powell**
**6: bullock**
**7: kleber**
**8: ntilikina**
**9: mcgee**
**10: pinson**

We wanted to create a knowledge base out of these players as relevant statistics on them would probably be most relevant to someone who wants to use the chatbot. Here is a screenshot of the unpickled data:

```
"Luka_Doncic": {
    "name": "Luka Doncic",
    "Position": "Point guard / Shooting guard",
    "League": "NBA",
    "Born": "(1999-02-28) February 28, 1999 (age 24)Ljubljana, Slovenia",
    "Listed height": "6 ft 7 in (2.01 m)",
    "Listed weight": "230 lb (104 kg)",
    "NBA draft": "2018 / Round: 1 / Pick: 3rd overall",
    "Playing career": "2015-present",
    "2015-2018": "Real Madrid",
    "2018-present": "Dallas Mavericks",
    "Men's basketball": "Medals~~~~~~~~Men's basketball~~~Representing  Slovenia~~~EuroBasket~~~~2017 Turkey"
},
```

Here, we can see a JSON object for each player, which has relevant metrics and statistics such as Name, Position, League, etc. Each JSON dictionary object is printed out in the program as such:

```
Knowledge Base:
{'name': 'Luka Doncic', 'Position': 'Point guard / Shooting guard', 'League': 'NBA', 'Born': '(1999-02-28) February 28, 1999 (age 24)Ljubljana, Slovenia', 'L
{'name': 'Kyrie Irving', 'Position': 'Point guard', 'League': 'NBA', 'Born': '(1992-03-23) March 23, 1992 (age 30)Melbourne, Victoria, Australia', 'Nationali
{'name': 'Christian Wood', 'Position': 'Power forward', 'League': 'NBA', 'Born': '(1995-09-27) September 27, 1995 (age 27)Long Beach, California  U.S.', 'Lis
{'name': 'Davis Bertans', 'Position': 'Power forward', 'League': 'NBA', 'Born': '(1992-11-12) 12 November 1992 (age 30)Valmiera, Latvia', 'Listed height': '
{'name': 'Dwight Powell', 'Position': 'Center / Power forward', 'League': 'NBA', 'Born': '(1991-07-20) July 20, 1991 (age 31)Toronto, Ontario, Canada', 'Lis
{'name': 'Reggie Bullock', 'Position': 'Small forward / Shooting guard', 'League': 'NBA', 'Born': '(1991-03-16) March 16, 1991 (age 31)Baltimore, Maryland, U
{'name': 'Maxi Kleber', 'Position': 'Power forward', 'League': 'NBA', 'Born': '(1992-01-29) 29 January 1992 (age 31)Wurzburg, Germany', 'Listed height': '6 f
{'name': 'Frank Ntilikina', 'Position': 'Point guard', 'League': 'NBA', 'Born': '(1998-07-28) 28 July 1998 (age 24)Ixelles, Belgium', 'Nationality': 'French
{'name': 'JaVale McGee', 'Position': 'Center', 'League': 'NBA', 'Born': '(1988-01-19) January 19, 1988 (age 35)Flint, Michigan, U.S.', 'Listed height': '7 ft
{'name': 'Theo Pinson', 'Position': 'Shooting guard / Small forward', 'League': 'NBA', 'Born': '(1995-11-05) November 5, 1995 (age 27)Greensboro, North Carol
```

Here, we can see the cleaned data with the most frequent terms:

```
cleaned_url2.txt top 25 terms:  [('dončić', 0.07712632895574992), ('tripledouble', 0.014823896454247176), ('madrid', 0.014797670687675778), ('euroleague', 0
cleaned_url3.txt top 25 terms:  [('green', 0.09427178390258574), ('australian', 0.0275451174841964), ('sydney', 0.018417014384012064), ('arizona', 0.0159312
cleaned_url4.txt top 25 terms:  [('hardaway', 0.07952754781892786), ('ten', 0.023588850664289014), ('michigan', 0.02332808069355217), ('big', 0.013829515530
cleaned_url5.txt top 25 terms:  [('hardy', 0.0922267412614758), ('g', 0.028555669980884858), ('nevada', 0.028100243278655363), ('consensus', 0.0218948369746
cleaned_url6.txt top 25 terms:  [('holiday', 0.1287969024883801), ('aaron', 0.011364432572504126), ('huskies', 0.011364432572504126), ('derrick', 0.00757628
cleaned_url7.txt top 25 terms:  [('irving', 0.06134419678627819), ('cavaliers', 0.013187704080264906), ('irvings', 0.010519968293055643), ('points', 0.00594
cleaned_url8.txt top 25 terms:  [('kleber', 0.1401367367219827), ('bundesliga', 0.023356122786997117), ('tg', 0.015570748524664746), ('würzburg', 0.01557074
cleaned_url9.txt top 25 terms:  [('lawson', 0.14986845454989817), ('twoway', 0.016391808579215628), ('carolina', 0.015616382020796406), ('canada', 0.0143684
cleaned_url10.txt top 25 terms:  [('mcgee', 0.09339516396985341), ('mcgees', 0.011708473011710794), ('dunk', 0.009366778409368635), ('wizards', 0.0078558211
cleaned_url11.txt top 25 terms:  [('morris', 0.07899903487825295), ('marcus', 0.02004456906126968), ('twins', 0.014474216938420747), ('brothers', 0.01403119
cleaned_url12.txt top 25 terms:  [('ntilikina', 0.11289525766470672), ('strasbourg', 0.036126482452706145), ('french', 0.022579051532941343), ('sig', 0.0180
cleaned_url13.txt top 25 terms:  [('pinson', 0.1445306739768881), ('pinsons', 0.0262783043594342), ('tar', 0.023352713592307194), ('nets', 0.020889367085368
cleaned_url14.txt top 25 terms:  [('powell', 0.11710651332271113), ('allacademic', 0.016729501903244445), ('canadian', 0.013900675567606088), ('earned', 0.0
cleaned_url15.txt top 25 terms:  [('wood', 0.14405959972238275), ('delaware', 0.018588335448049383), ('march', 0.012921499688514478), ('waived', 0.011096243
cleaned_url16.txt top 25 terms:  [('wright', 0.13900675567606088), ('twoway', 0.019822652235330526), ('buffaloes', 0.018588335448049383), ('billups', 0.0185
```

We decided to use players as the top 10 terms in lieu of terms at random because those terms would provide the most relevant information to someone who is a fan of the Mavericks.

**(2) write up a sample dialog you would like to create with a chatbot based on your knowledge base**

**Input**: Tell me about Luka
**Output**: Sure, what do you want to know about him?
**Input:** Where is Luka from?
**Output:** Luka is from Ljubljana, Slovenia.
**Input:** What position does Luka play?
**Output:** Point guard / Shooting guard
**Input:** What players do you have info on?
**Output:** Sure, here's a list of players:

- **Doncic**
- **Irving**
- **Wood**
- **Bertans**
- **Powell**
- **Bullock**
- **Kleber**
- **Ntilikina**
- **Mcgee**
- **Pinson**