

Homework 5: Web Scraping

Important Terms

The way that I created my important terms was using TF-IDF. I calculated the TF-IDF for every word in every document. Then I sorted the TF-IDF weights for each document. Then I took the 5th root of the length of the set of tokens and round it to determine how many words to retrieve for that document. After, I simply add the N top terms of that document to the whole important terms list. I do this for every document. Which ultimately gets me my important terms list.

My prediction of the output of top terms include: lakers, los, angeles, lebron, james, anthony, davis, kobe, bryant, nba, basketball, finals, championship, shaq.

The following is the output from my program. The program is concerned with URLs about the Lakers and has a limit of anything greater than or equal to 25 URLs. Finally, keep in mind that not all URLs will be processed to text because not URLs will contain relevant info.

Output:

Printing the URL list in the database...

```
1: https://www.complex.com/sports/2020/10/fans-make-petition-to-not-give-kyle-kuzma-a-ring-if-lakers-win-nba-finals
2: https://www.espn.com/nba/story/\_/id/30041351/miami-heat-bam-adebayo-goran-dargic-ruled-game-3-nba-finals-los-angeles-lakers
3: https://www.foxnews.com/sports/lakers-anthony-davis-colin-kaepernick-jersey-nba-finals
4: https://www.cbssports.com/nba/news/lakers-vs-heat-score-live-nba-finals-updates-as-jimmy-butler-depleted-miami-look-for-much-needed-game-3-win/live/
5: https://clutchpoints.com/lakers-news-lebron-james-becomes-second-all-time-in-nba-playoff-assists/
6: https://secure.nba.com/article/2020/10/03/heat-daunting-challenge-lakers-finals%3Fcollection%3Dnews
7: https://www.washingtonpost.com/sports/2020/10/04/lakers-heat-nba-finals-game-3-live-updates/
8: https://www.cbssports.com/nba/news/nba-finals-mvp-race-making-the-case-for-lebron-james-and-anthony-davis-as-lakers-approach-2020-championship/
9: https://www.cbssports.com/nba/news/lakers-lebron-james-passes-john-stockton-for-second-most-playoff-assists-in-nba-history/
10: https://www.espn.com/nba/story/\_/id/30027392/los-angeles-lakers-lebron-james-anthony-davis-reach-shaq-kobe-heights-win
11: https://www.nba.com/lakers/
12: https://en.wikipedia.org/wiki/Los\_Angeles\_Lakers
13: http://www.nba.com/lakers
14: https://en.wikipedia.org/wiki/Los\_Angeles\_Lakers
15: https://www.cbssports.com/nba/news/lakers-vs-heat-score-live-nba-finals-updates-as-jimmy-butler-depleted-miami-look-for-much-needed-game-3-win/live/
16: https://www.espn.com/nba/team/\_/name/lal/los-angeles-lakers
17: https://www.instagram.com/lakers/%3Fhl%3Den
18: https://bleacherreport.com/los-angeles-lakers
19: https://en.wikipedia.org/wiki/Los\_Angeles\_Lakers
20: https://sports.yahoo.com/nba/teams/la-lakers/
21: https://twitter.com/intent/tweet?text=Fans Make Petition for Kyle Kuzma Not to Receive a Ring If Lakers Win NBA Finals
22: https://www.complex.com/tag/los-angeles-lakers
23: https://clutchpoints.com/lakers-news-fans-signing-petition-saying-kyle-kuzma-shouldnt-get-a-ring-if-l-a-wins-title/
24: https://twitter.com/intent/tweet?text=Fans Make Petition for Kyle Kuzma Not to Receive a Ring If Lakers Win NBA Finals
```

Note: The initial link is not included as it was a link to a google search for 'lakers' and it does not have relevant info we can extract besides the links.

Printing my prediction for important terms...

- 1: lakers
- 2: los
- 3: angeles
- 4: lebron
- 5: james
- 6: anthony
- 7: davis
- 8: kobe
- 9: bryant
- 10: nba
- 11: basketball
- 12: finals
- 13: championship
- 14: shaq

Printing the actual important terms in the database...

1: boston	
2: celtics	22: accessibility
3: much	23: kuzma
4: pf	24: postseason
5: butler	25: guys
6: ring	26: factset
7: doc	27: bubble
8: percent	28: stockton
9: green	29: issue
10: screens	30: cbs
11: duo	31: west
12: six	32: injury
13: provided	33: shaq
14: defeated	34: said
15: https	35: hall
16: head	36: davis
17: continues	37: nbafinals
18: ad	38: road
19: deserves	39: data
20: petition	40: career
21: turner	41: season

Note: A lot of these words makes sense. For example, ad is often used to refer to Anthony Davis (current player of the Lakers), the Boston Celtics are the rivals of the Lakers, and many other words are names of former/current players of the Lakers or another team in the nba.

Knowledge Base

The way I created my knowledge base was to take all the important words from the collection of documents and create a dictionary. The key would be the word and the value will be a list. This list will have every sentence containing that specific word from all documents. For example, if we have 20 documents and each document has two sentences containing “nba” and if nba is an important word then it will have a list of 40 elements each element corresponding to a sentence that contains nba.

This knowledge base is then stored into a database file. Each important term will have its own table and the entries of the table will be the sentences containing the term from all documents. For example, if we have 41 important terms, then the database file will have 41 tables. So, if the word “nba” is an important term with 40 sentences, then there will be a table called “nba” with 40 entries each corresponding to one sentence.

The following is the output of my knowledge base from the important words shown above.

boston:

- 1: with the franchise nickname “showtime”, given for their flashy, run and gun style of play in the 1980s, the los angeles lakers have the second most nba championships in history with 16, only one behind their rival, the boston celtics.
- 2: jerry west’s los angeles lakers then advanced to the nba finals six times in the 1960s, only to lose each meeting to the boston celtics.
- 3: as the mid 1980s bred one of the best rivalries in league history against larry bird and the boston celtics, pat riley’s lakers won five nba championships in nine years.
- 4: [9] the lakers are one of the most successful teams in the history of the nba, and have won 16 nba championships, the second-most behind the boston celtics.
- 5: los angeles lost in the finals to boston in seven games again in 1966, this time by two points.
- 6: [58] the lakers and celtics again met in the finals, and los angeles had home court advantage against boston for the first time in their rivalry.
- 7: [60] boston won the series in seven games however.[61] and earned their 11th nba championship in 13 seasons.
- 8: in the 1983-84 season, los angeles went 54-28, and played boston in the finals for the first time since 1969.
- 9: [110] boston won three of the next four to win the title and send los angeles's record to 0-8 in finals series against the celtics.
- 10: [112] the team won the title in the boston garden, becoming the only visiting team to ever win an nba championship there.
- 11: [117] the lakers defeated boston in the first two games of the finals, and the teams split the next four games, giving los angeles their second championship in three seasons.
- 12: [115] the series was highlighted by johnson's running "baby hook" shot to win game four at boston garden with two seconds remaining.
- 13: [119] at the lakers' championship celebration in los angeles, coach riley brashly declared that los angeles would repeat as nba champions,[115] which no team had done since the 1968-69 boston celtics.
- 14: the lakers also traded rick fox and gary payton to boston, for chris mihm, marcus banks, and chucky atkins before the 2004-05 season.
- 15: in the finals, the lakers played the boston celtics for the 12th time.
- 16: the rivalry between the boston celtics and the lakers involves the two most storied basketball franchises in national basketball association (nba) history.
- 17: [316] they also used to hold the record for most wins at home in the regular season (going 36-5 in 1971-72, then 37-4 in both 1976-77 and 1979-80) before the boston celtics set the current record of 40-1 in the 1985-86 season.

celtics:

- 1: with the franchise nickname “showtime”, given for their flashy, run and gun style of play in the 1980s, the los angeles lakers have the second most nba championships in history with 16, only one behind their rival, the boston celtics.
- 2: jerry west’s los angeles lakers then advanced to the nba finals six times in the 1960s, only to lose each meeting to the boston celtics.
- 3: as the mid 1980s bred one of the best rivalries in league history against larry bird and the boston celtics, pat riley’s lakers won five nba championships in nine years.
- 4: [9] the lakers are one of the most successful teams in the history of the nba, and have won 16 nba championships, the second-most behind the boston celtics.
- 5: led by hall of famers elgin baylor and jerry west, los angeles made the nba finals six times in the 1960s, but lost each series to the celtics, beginning their long and storied rivalry.
- 6: [34] after upsetting the hawks in six games in the division finals, they returned to the nba finals, but were swept by the celtics, beginning their long rivalry.
- 7: [47] the lakers, however, lost to the celtics by three points in overtime of game seven.
- 8: los angeles won 53 games in 1962-63, behind baylor's 34.0 ppg and west's 27.1 ppg[50] but lost in the nba finals in six games to the celtics.
- 9: [51] they lost again to celtics in the finals however, this time in five games.
- 10: [54] after dropping to 36 wins and losing in the first round of the 1967 nba playoffs, they lost in the finals to the celtics again in 1968.
- 11: [58] the lakers and celtics again met in the finals, and los angeles had home court advantage against boston for the first time in their rivalry.
- 12: [63] in 1970, west won his first scoring title at 31.2 ppg, the team returned to the finals, and for the first time in 16 years, they did not have to face the celtics; instead playing the new york knicks, who defeated them 4-3.
- 13: [110] boston won three of the next four to win the title and send los angeles's record to 0-8 in finals series against the celtics.
- 14: in the nba finals, the celtics were again the lakers' final hurdle.
- 15: [111] the lakers, behind 38-year-old finals mvp abdul-jabbar, recovered to defeat the celtics in six games.
- 16: [119] at the lakers' championship celebration in los angeles, coach riley brashly declared that los angeles would repeat as nba champions,[115] which no team had done since the 1968-69 boston celtics.
- 17: [144][145] the lakers defeated the portland trail blazers in the first round of the playoffs as o'neal scored 46 points in game 1 against the trail blazers, marking the highest single-game playoff scoring output by a laker since jerry west scored 53 against the celtics in 1969.
- 18: [207][208] in the playofffs, they defeated the nuggets in four games, the jazz in six, and the defending champion spurs in five, but lost to the celtics in six games in the nba finals.
- 19: in the finals, the lakers played the boston celtics for the 12th time.

20: they rallied back from a 3-2 disadvantage in the series and erased a 13-point deficit in the fourth quarter of the seventh game to defeat the celtics.
21: the rivalry between the boston celtics and the lakers involves the two most storied basketball franchises in national basketball association (nba) history.
22: the rivalry had been less intense since the retirements of magic johnson and larry bird in the early 1990s, but in 2008 it was renewed as the celtics and lakers met in the finals for the first time since 1987, with the celtics winning the series 4-2.
23: the two teams have won the two highest numbers of championships, the celtics 17, the lakers 16; together, the 33 championships account for almost half of the 73 championships in nba history.
24: the all time series record for the lakers vs celtics is 204-162 with the celtics being the forerunners.
25: [316] they also used to hold the record for most wins at home in the regular season (going 36-5 in 1971-72, then 37-4 in both 1976-77 and 1979-80) before the boston celtics set the current record of 40-1 in the 1985-86 season.

much:

1: lakers vs. heat score: live nba finals updates as jimmy butler, depleted miami look for much-needed game 3 win - cbssports.com
2: the lakers faced a tougher challenge from the heat in game 2, but the combination of james and anthony davis was simply too much for a miami team missing bam adebayo and goran dragic.
3: the two best players in the series are now, if only silently, competing against one another as much as they are competing against miami.
4: davis is covering so much ground and taking so much away that virtually anything inside the arc that isn't a dunk has been difficult.
5: this is especially important to davis' candidacy because james backers will argue that much of what lebron does isn't visible in a box score.
6: that isn't where much of lebron's game 1 production came from.
7: speaking of individual creation, it should be noted that james has to work much harder for his points than davis does.
8: while the heat spent much of game 2 playing zone defense, in butler, jae crowder and andre iguodala, they still at least have a variety of wing defenders to credibly throw at james.
9: there isn't much of a difference between nine and 10 assists, but that round number helps hammer home what a difference there is in their playmaking.
10: they respected his contributions, but some, such as worthy and scott, had grown tired of his intense practices and felt he tried to take too much credit for the team's successes.
11: after much speculation, head coach phil jackson returned for the 2010-11 season.
12: the lakers' offense isn't running quite as well as it did in the first two games - lebron james leads with 16 points, six assists and six rebounds - and anthony davis, who had three fouls in the first half, isn't nearly as much of a force as he was friday.

pf:

1: the broadcast helpfully points out that the lakers are 0-for-7 this quarter.
2: the lakers would lose 43 games in 1992-93 under randy pfund, their first losing season since 1976.
3: [130] during the 1993-94 season, pfund was fired during the season that would result in the lakers failing to make the playoffs for the first time since 1976.
4: anthony davis, pf
5: anthony davis, pf
6: anthony davis, pf
7: anthony davis, pf
8: anthony davis, pf

butler:

1: lakers vs. heat score: live nba finals updates as jimmy butler, depleted miami look for much-needed game 3 win - cbssports.com
2: rondo-kcp-kuzma-morris-davis vs. nunn-butler-robinson-crowder-olnyk
3: butler is killing the lakers right now.
4: 8-0 lakers run but jimmy butler remains very comfortable going at green one-on-one.
5: lebron tried to get switched onto herro but butler switched back quickly to eliminate the mismatch
6: oof, rough pass by butler.
7: 19-8-7 for butler in 24 minutes.
8: he's even taken some turns defending jimmy butler.
9: lebron is not davis defensively, but he is still quite good and deserves credit for taking the butler assignment for the bulk of the first two games.
10: while the heat spent much of game 2 playing zone defense, in butler, jae crowder and andre iguodala, they still at least have a variety of wing defenders to credibly throw at james.
11: during the 2004 off-season, the team entered a rebuilding phase when o'neal was traded to the miami heat for lamar odom, brian grant, caron butler, and a first-round draft pick.
12: [194] the team also traded caron butler and chucky atkins to washington for kwame brown and laron profit.
13: jimmy butler explains that the heat miss bam adebayo's presence, but they have to take caution and not make his injury worse by bringing him back too soon.
14: dragic was miami's second-leading scorer in the regular season behind jimmy butler, averaging 16.2 points off the bench.
15: the los angeles lakers and lebron james take on the miami heat and jimmy butler in game 3 of the nba finals on sunday at advenhealth arena.
16: the lakers' starters have continued to turn in sub-par performances in game 2 but miami's bench gets outplayed whenever jimmy butler is off the floor, so the outcome is a slim lead for the heat at halftime.
17: speaking of butler, he's once again doing everything for miami.
18: it didn't help that the heat didn't score while jimmy butler took a three-minute rest; the all-star forward leads miami with eight points and three assists.
19: miami: duncan robinson has gone from skinny d-iii player to starter in the nba finals | heat's injuries threatening to rob the nba finals of any drama | in an era of lebron james and superteams, jimmy butler created his own maverick model

percent:

1: in nba history, 24 players have technically shot over 60 percent from the field and 50 percent from behind the arc in the finals.
2: miami is shooting just 33.3 percent in the non-restricted area portion of the paint in the finals, and only 26.1 percent in the mid-range.
3: davis has six assists, but only 11 potential assists, meaning 54.5 percent of his possible assists have been converted into points.
4: his conversion rate is just under 43 percent.
5: only 47.8 percent of lebron's field goals have been assisted in this series.
6: davis has had 61.5 percent of his field goals assisted, and that gap is likely to widen as the series progresses.
7: only 28.9 percent of lebron's regular-season field goals were assisted.
8: [78] wilt chamberlain, playing in his final season, again led the league in rebounding and set the still standing nba record for field-goal percentage at 72.7%.
9: after re-acquiring derek fisher, the lakers started the 2007-08 season with a 25-11 record, before andrew bynum, their center who was leading the league in field-goal percentage, went out for the year due to a knee injury in mid-january.
10: note: gp = games played, w = wins, l = losses, % = winning percentage;
11: [312] highest field-goal percentage for a season at 54.5% (1984-85),[313] and highest road winning percentage at 0.816 (1971-72).
12: [55] they also hold records for having (into the 2009-10 season) the most wins (3,027), the highest winning percentage (61.9%), and the most nba finals appearances (31).
13: when you're in the elite percentile of being a competitor in this league, get to this stage and have unfortunate injuries ...

Note: This is not the complete knowledge base. The complete knowledge base with a limit of around 25 links is too big to fit all in this document. These are just some tables in the whole knowledge base database. If you would like to see the whole knowledge, go to the print_results functions and uncomment the if statement in the last for loop. This will print the entire KB.

Sample Dialog 1

Hello, I am a chat-bot that loves to talk about the Los Angeles Lakers. Tell me what is your name?

My name is Alfredo.

Hello Alfredo. I do sure hope you are a fan of the Lakers. Are you a fan?

Yes, most definitely, born and raised in Los Angeles.

Awesome! That means we can talk about the Lakers all day long. Do you know the Celtics?

Of course, the rivals of the Lakers.

Did you know: los angeles won 53 games in 1962-63, behind baylor's 34.0 ppg and west's 27.1 ppg but lost in the nba finals in six games to the Celtics?

No, I did not know that.

I thought you said you were a fan of the Los Angeles Lakers. Are you a bandwagon?

No! I was born and raised in Los Angeles.

Okay, let me test you one more time. Did you know: with the franchise nickname "showtime", given for their flashy, run and gun style of play in the 1980s, the los angeles lakers have the second most nba championships in history with 16, only one behind their rival, the boston celtics?

Yes, I did know that. I hope they win another chip this year so it can be tied.

Ahh. So maybe you aren't a bandwagon ...

Sample Dialog 2

Hello, I am a chat-bot that loves to talk about the Los Angeles Lakers. Tell me what is your name?

My name is Alfredo.

Hello Alfredo. I do sure hope you are a fan of the Lakers. Are you a fan?

No, the Los Angeles Clippers convinced me to switch.

I respect that, but let me tell you something about the greatest basketball team, with the franchise nickname "showtime", given for their flashy, run and gun style of play in the 1980s, the los angeles lakers have the second most nba championships in history with 16, only one behind their rival, the boston celtics. Can you believe that?

No

Well believe it ...

Quick Note

With the initial URL being a google search for the lakers, this means each run could potentially produce different results. Especially as the lakers are currently in the NBA Finals as of the making of this document. This means the google search is often being updated with new articles and information regarding the basketball team.

If the program is run multiple times and the database files created, are not deleted, then the URL database list will most likely increase with each run regardless on the limit. This because each run will have its set of URLs (abiding to the limit) and it will be added to the database file. However, the database file could potentially have the URLs from the previous run if the file was not deleted. It will get rid of any duplicates and add the new URLs to the database. This was done intentionally to have a large database of URLs from not just one run but multiple runs. If you would like to have URLs for that specific run, then after each run you will have to delete the database file containing the URLs.

The same thing applies to the knowledge base database file. Each run will produce different important terms and ultimately different sentences will be added to the current run knowledge base. This knowledge base will then be stored to the database file that could potentially have the previous knowledge base in it if it was not deleted. If there are repeated important terms then no table is created, duplicates sentences aren't added to the table, and only new sentences are added to the table. This is, again, done intentionally to build a bigger knowledge base after each run. If you would like a knowledge base only for that run, then after each run delete the knowledge base database file.

However, the exception is the important terms database file. After each run, if there is an important terms database file it will delete it and create a new file for the current run. This is because the important terms are local to the current run containing a set of documents. The next run might contain different documents and thus potentially change the list of important terms. Thus, we do not want to combine them because they are a reflection of their own set of documents and each run could potentially have different documents. In this case, the important terms database file does not need to be deleted the program takes care of it.

If you have any questions, concerns, or comments, please feel free to contact me at my email: axm180023@utdallas.edu

I would gladly explain any parts of my program. Thank you.