

# Comparing Famous Athletes



---

Skyler Cho, Kevin Koh, Jackson Winslow



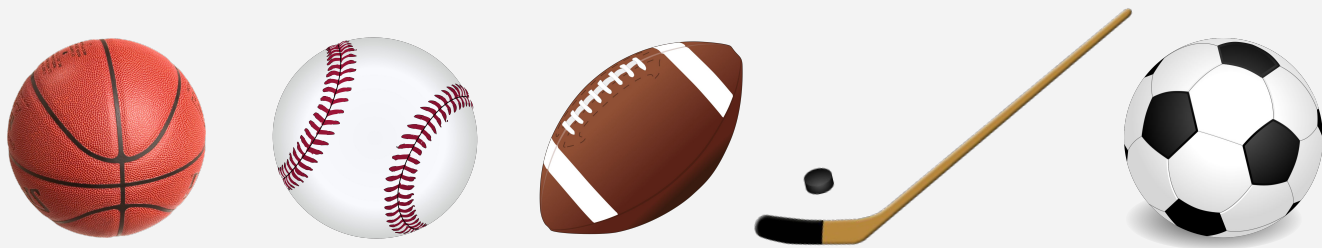
# 01

## Motivation

---

# Why compare athletes?

- We love sports + We love Text Processing  $\Rightarrow$  Opportunity to utilize class materials on what we love!
- What similarities or differences can we find about top players in different sports?
- Legendary players from the 5 biggest sports in the US



# Why compare athletes?



LeBron James  
(American Basketball Player)



Lionel Messi  
(Argentine Soccer Player)



Mike Trout  
(American Baseball Player)



Patrick Mahomes  
(American Football Player)



Sidney Crosby  
(Canadian Ice Hockey Player)

# Why compare athletes?

- Project Goals
  - Are there any differences in how the media describe these athletes?
  - Is current performance reflected in the way media talk about these athletes?
  - Are different words/adjectives used to describe young vs. old athletes?
  - Does the athlete's team's winning record influence the words/adjectives used to describe the athlete?





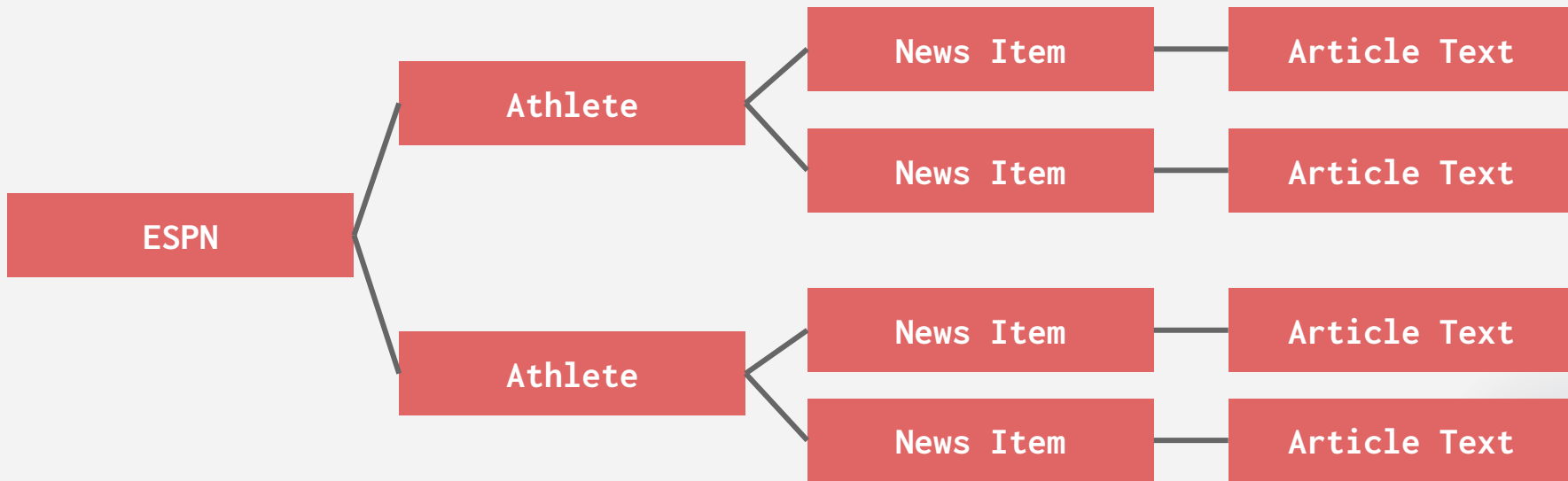
# 02

## Data

---

# Sourcing

- We decided to target **ESPN**, a major sports news source for American sports
- Relevant endpoints were discovered by exploring network activity while browsing relevant pages on the site



# Procedure

<https://api-app.espn.com/allsports/apis/v1/now...>

- Using Axios in Node.js, we scraped the last 200 relevant news items for each athlete
- Filtered out any items that weren't articles (videos, podcasts)
- Saved the associated internal link for each article
- Fetched the article text for each saved link





# Statistics

ESPN historically has not paid  
as much attention to the NHL



	Messi	Lebron	Trout	Crosby	Mahomes
Articles	114	67	93	26	85
Tokens	233,613	60,329	94,955	23,903	120,312
Types	7,554	6,323	8,505	3,498	8,552

Total Tokens: **533,112**

# Pre-Processing

Before bigrams/collocations:

- Downcased all words
- Tokenized
- Lemmatized
- Curated robust stoplist

After bigrams/collocations:

- Removed stopwords from tokenlist



# 03

## Method

---

# Bigrams

- Created list of bigrams
- Collected top 50 bigrams
- Selected top 3 bigrams not containing a stopword

(Examples/results shown later)

# Collocations/Word Clouds

- Collocations
  - Create finder object from lemmatized tokenlist
  - Apply frequency filter of 2 to finder
  - Collected 10 best based on PMI
  - No insights were drawn from collocation findings
- Word Clouds
  - Create frequency distribution of 20 most common tokens
  - Create word clouds based on lemmatized tokenlist excluding stopwords
  - No insights were drawn from word cloud findings

# TF-IDF/Heat Maps

- **Adjectives only** tokenlist added
  - Used Python package spaCy
  - Utilized built-in spaCy part-of-speech NLP techniques to identify adjectives
- TF-IDF (for full and adjectives list)
  - Used **sklearn** package to get TF-IDF vectors excluding “english” stopwords
- Heat Map (for full and adjectives list)
  - Used **altair** package to draw heat map from TF-IDF results



# 04

## Results & Analyses

---

# Bigrams

- From the list of lemmatized words
- Top 3 bigrams without stop words


Messi	Lebron	Trout	Crosby	Mahomes
inter miami	los angeles	home run	stanley cup	super bowl
world cup	lebron james	los angeles	sidney crosby	kansa city
league cup	james said	mike trout	year signed	last season


 = major competition



# Bigrams

Messi	Lebron	Trout	Crosby	Mahomes
inter miami	los angeles	home run	stanley cup	super bowl
world cup	lebron james	los angeles	sidney crosby	kansa city
league cup	james said	mike trout	year signed	last season

 = won major competition

 = did not win major competition

# Bigrams

- Won the Super Bowl LVII (57) on February 12, 2023

Messi	Lebron	Trout	Crosby	Mahomes
inter miami	los angeles	home run	stanley cup	super bowl
world cup	lebron james	los angeles	sidney crosby	kansa city
league cup	james said	mike trout	year signed	last season

- Won League Cup on August 19, 2023
- Won World Cup on December 18, 2022

# Bigrams

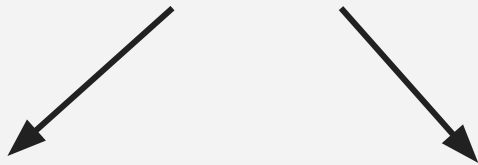
Messi	Lebron	Trout	Crosby	Mahomes
inter miami	los angeles	home run	stanley cup	super bowl
world cup	lebron james	los angeles	sidney crosby	kansa city
league cup	james said	mike trout	year signed	last season



- Did not win the Stanley Cup
- One Source: “were going to do something dramatic in the offseason after missing the Stanley Cup playoffs for the first time in 16 seasons”

# Bigrams

- Trout has never won a major professional competition
- LeBron last won a major championship in 2020



Messi	Lebron	Trout	Crosby	Mahomes
inter miami	los angeles	home run	stanley cup	super bowl
world cup	lebron james	los angeles	sidney crosby	kansa city
league cup	james said	mike trout	year signed	last season

# TF-IDF (all lemmatized words)

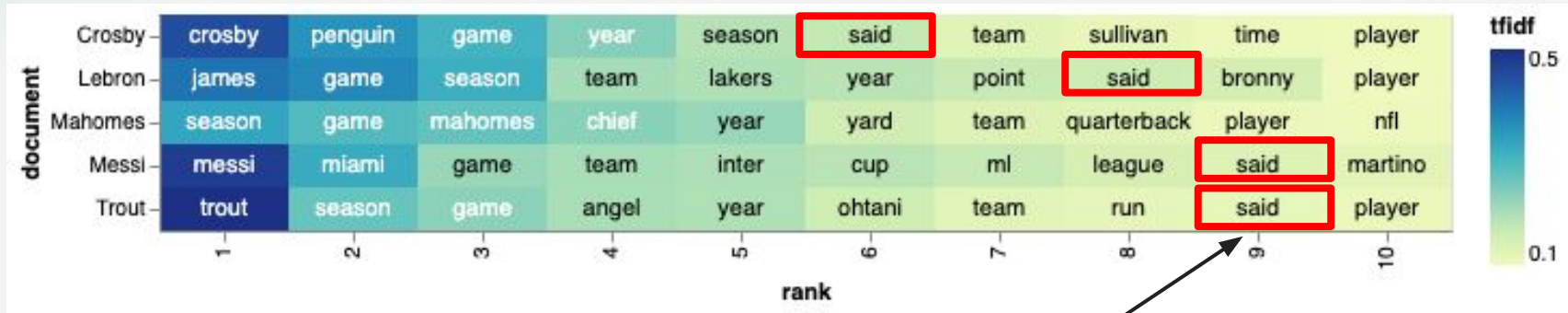


# TF-IDF (all lemmatized words)



- "team" is an important word for all players
- The best players are judged on their team's success

# TF-IDF (all lemmatized words)



- "said" is an important word for almost all players
- Generally, the best players are interviewed the most

# TF-IDF (all lemmatized words)



- “season” is an important word for all players, except Messi
- Those who played for the whole season have the word “season”
- Past year, Messi was only important for
  - League Cup
  - World Cup
  - Basically, non existent in the league



# TF-IDF (adjectives)



# TF-IDF (adjectives)



- "best" is an important word for all players
- Articles describe them as the "best"

# TF-IDF (adjectives)



- Those who are in the twilight of the career are described as “old”
- Youngest of the “old” is 36
- Trout and Mahomes are still relatively in their peak (28 and 32 year old respectively)

# TF-IDF (Lebron)

## TF-IDF (all lemmatized words)

Lebron	james	game	season	team	lakers	year	point	said	bronny	player
--------	-------	------	--------	------	--------	------	-------	------	--------	--------

## TF-IDF (adjectives)

Lebron	durant	best	old	big	second	cardiac	regular	usc	high	able
--------	--------	------	-----	-----	--------	---------	---------	-----	------	------

# TF-IDF (Lebron)

## TF-IDF (all lemmatized words)



## TF-IDF (adjectives)



- All red boxes are words related to “**Bronny** James” (Lebron’s son)
- He attends **USC**, but had **cardiac** arrest recently

# TF-IDF (Lebron)

## TF-IDF (all lemmatized words)



## TF-IDF (adjectives)



- Shows how Lebron is not only a superstar athlete, but also a public figure
- His personal/private life is also very important



# 05

## Conclusion & Reflection

---

# Conclusion

Discoveries:

- Surprised to see the importance of Bronny and his life
- Interesting to see how “season” is not important to Messi
- Those who are almost ready to retire = “old”
- Importance of “said” for most athletes



## Future Work

- Analyze more athletes in order to find more important words in TF-IDF (low score)
- Analyze more sources other than ESPN

## Addressing Problems

- No public API's:
  - Use our scraping technique on other sources

# Thank you!

Any Questions?