Extracting **keywords** from text

Lavanya Sharan  
Insight Data Science Fellow

with URX
Kendrick Lamar, Hip-Hop’s Newest Old-School Star

By Dave Turner
Last Updated: February 3rd, 2015

“Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect... You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back here, it means a lot to us,” said an elder.
Kendrick Lamar, Hip-Hop’s Newest Old-School Star
By Dave Turner
Last Updated: February 3rd, 2015

“Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect... You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back here, it means a lot to us,” said an eager...
Kendrick Lamar, Hip-Hop’s Newest Old-School Star

By Dave Turner
Last Updated: February 3rd, 2015

“Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect... You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back here, it means a lot to us,” said an eager
Kendrick Lamar, Hip-Hop’s Newest Old-School Star
By Dave Turner
Last Updated: February 3rd, 2015

“Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect... You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back here, it means a lot to us,” said an eager...
Kendrick Lamar, Hip-Hop’s Newest Old-School Star

By Dave Turner
Last Updated: February 3rd, 2015

“Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect. . . . You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back here, it means a lot to us,” said an eager

URX
Kendrick Lamar, Hip-Hop’s Newest Old-School Star
By Dave Turner
Last Updated: February 3rd, 2015

“Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect... You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back here, it means a lot to us,” said an eager...
Kendrick Lamar, Hip-Hop’s Newest Old-School Star

By Dave Turner
Last Updated: February 3rd, 2015

“Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect. . . . You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back here, it means a lot to us,” said an eager
Keywords matching ad

Kendrick Lamar, Hip-Hop's Newest Old-School Star

By Dave Turner
Last Updated: February 3rd, 2015

"Everybody just wants to have fun, be with the scene," Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. "Certain people get backstage, people that you would never expect... You ain't with the media! You ain't into music! You ain't into sports! You're just here." The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West's Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. "Hey, man, thank you again, appreciate the access back here, it means a lot to us," said an eager...

URX
Kendrick Lamar, Hip-Hop’s Newest Old-School Star
By Dave Turner
Last Updated: February 3rd, 2015

“Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect... You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back here, it means a lot to us,” said an eager...
A machine learning model for keyword extraction

Content → Candidate keywords → Feature extraction → Keyword Classifier → Keywords
A machine learning model for keyword extraction

Content → Candidate keywords → Feature extraction → Keyword Classifier → Keywords

Kendrick Lamar, Hip-Hop's Newest Old-School Star
By Pete Tavano
Last Updated: February 3rd, 2019

"Everybody just wants to have fun, be with the scene," Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. "Certain people get backdrops, people that you would never expect... You ain't with the media, you ain't into music! You ain't into sports! You're just here." The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West's "Yeezus" tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly three people. "Hey, man, thank you again, appreciate the access back..."
A machine learning model for keyword extraction

Content → **Candidate keywords** → **Feature extraction** → **Keyword Classifier** → Keywords

**Brooklyn**
A machine learning model for keyword extraction

Content → Candidate keywords → Feature extraction → Keyword Classifier → Keywords

**Term length**
- Brooklyn

**Wikipedia freq**

**TF-IDF score**

...
A machine learning model for keyword extraction

- **Content** → **Candidate keywords** → **Feature extraction** → **Keyword Classifier** → **Keywords**

**Example:**

**Term length**
- **Brooklyn**
- **Wikipedia freq**
- **TF-IDF score**
- **Logistic regression**
A machine learning model for keyword extraction

**Content** → **Candidate keywords** → **Feature extraction** → **Keyword Classifier** → **Keywords**

**Term length**
- Brooklyn

**Wikipedia freq**
- Logistic regression

- TF-IDF score

**Crowd500**
- 500 news articles
- Human-annotated keywords
- 9:1 training-test split
A machine learning model for keyword extraction

<table>
<thead>
<tr>
<th>Content</th>
<th>Candidate keywords</th>
<th>Feature extraction</th>
<th>Keyword Classifier</th>
<th>Keywords</th>
</tr>
</thead>
</table>

**Term length**
- Brooklyn
- Wikipedia freq
- TF-IDF score

**Logistic regression**
A machine learning model for keyword extraction

**Content** $\rightarrow$ **Candidate keywords** $\rightarrow$ **Feature extraction** $\rightarrow$ **Keyword Classifier** $\rightarrow$ **Keywords**

- **Term length**
- **Wikipedia freq**
- **TF-IDF score**

...$\rightarrow$ **Logistic regression** $\rightarrow$ **P(keyword) = 0.81**

---

*Example keyword: Brooklyn*

"Everybody just wants to have fun, be with the scene," Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. "Certain people get back there, people that you would never expect... You ain't with the media, you ain't into music. You ain't into sports. You're just here." The rapper, now 27, had just finished his set at the opening act on this stretch of Kanye West's Yeezus tour, and he was sitting low in a sonata chair in his trademark black hoodie surrounded by exactly three people. "Hey man, thank you again, appreciate the access back..."
A machine learning model for keyword extraction

Content → Candidate keywords → Feature extraction → Keyword Classifier → Keywords

Term length

Wikipedia freq

TF-IDF score

Logistic regression

P(keyword) = 0.81

Brooklyn

"Everybody just wants to have fun, be with the scene," Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. "Certain people get backlash, people that you would never expect... You ain't with the media. You ain't into music. You ain't into sports. You're just here." The rapper, now 27, had just finished his set at the opening act on this stretch of Kanye West's 'Nas tour, and he was sitting low in a armchair in his trademark black hoodie surrounded by exactly three people. "Hey man, thank you again, appreciate the access back..."
A machine learning model for keyword extraction

Content → Candidate keywords → Feature extraction → Keyword Classifier → Keywords

Term length
Wikipedia freq
TF-IDF score

Brooklyn

Logistic regression

P(keyword) = 0.81
State-of-the-art performance on Crowd500

![Bar chart showing top-15 keywords with Precision, Recall, and F-1 score for Random, AlchemyAPI, and Algorithm.]
Kendrick Lamar, Hip-Hop’s Newest Old-School Star

By Dave Turner
Last Updated: February 3rd, 2015

‘Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect. . . . You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back
Kendrick Lamar, Hip-Hop’s Newest Old-School Star

By Dave Turner
Last Updated: February 3rd, 2015

‘Everybody just wants to have fun, be with the scene,” Kendrick Lamar said when we met in his cramped quarters inside the Barclays Center in Brooklyn last fall. “Certain people get backstage, people that you would never expect. . . . You ain’t with the media! You ain’t into music! You ain’t into sports! You’re just here.” The rapper, now 27, had just finished his set as the opening act on this stretch of Kanye West’s Yeezus tour, and he was sitting low in an armchair in his trademark black hoodie surrounded by exactly those people. “Hey, man, thank you again, appreciate the access back
## Kendrick Lamar

### Tour History 2010 - 2015

**On Tour | 4 Upcoming Tour Dates**

<table>
<thead>
<tr>
<th>TOUR DATES</th>
<th>PAST DATES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Venue</td>
</tr>
<tr>
<td>Jul 11, 2014</td>
<td>Winston Far,</td>
</tr>
<tr>
<td>Jun 28, 2014</td>
<td>Mid-America Center</td>
</tr>
<tr>
<td>Jun 21, 2014</td>
<td>Scottrade Center</td>
</tr>
<tr>
<td>Jun 20, 2014</td>
<td>BBVA Compass Stadium</td>
</tr>
<tr>
<td>Jun 14, 2014</td>
<td>Joe Louis Arena</td>
</tr>
<tr>
<td>Jun 12, 2014</td>
<td>BB&amp;T Center</td>
</tr>
<tr>
<td>Jun 07, 2014</td>
<td>Gorge Amphitheatre</td>
</tr>
<tr>
<td>Jun 06, 2014</td>
<td>Fiddler’s Green Amphitheater</td>
</tr>
<tr>
<td>Jun 05, 2014</td>
<td>Optimus Primavera Sound</td>
</tr>
<tr>
<td>May 31, 2014</td>
<td>Primavera Sound Festival</td>
</tr>
<tr>
<td>May 30, 2014</td>
<td>Xfinity Center</td>
</tr>
</tbody>
</table>
Keyword finder: automatic keyword extraction from text

October 14, 2015  Cameron Grace

Dr. Lavanya Sharan, PhD, studies human visual perception from a behavioral and a computational perspective at M.I.T. As part of the Insight Data Science Fellows Program, she worked with URX to help us tackle an important business question: how can we use context to match publisher page content with advertisers? This gives an overview of her impressive findings and deliverables in collaboration with URX.

Problem

For my project, I partnered with URX, a San Francisco-based startup that helps advertisers build brand discovery experiences. The problem URX wants to solve is the following: how to match advertisers to content providers in a context-specific way? For example, if the content consists of a news article about Kendrick Lamar, a hip-hop musician, a matching ad would link to Kendrick Lamar's latest album. One way to accomplish this matching is to extract keywords from the news article, use those keywords to search a database of advertisers, and then serve the best matching ad.
About me

Keywords: Lavanya, image recognition, foodie
Keyword classifier

In-sample: 65%, out-of-sample: 65%, chance: 50%

"Brooklyn"

- Term frequency
- TF-IDF score
- Wikipedia frequency
- Term length
- Capitalized?
- Position in page
- Spread in page
- Named entity?
- Noun phrase?
- Ngram?

Logistic regression model
Simple heuristics beat complex features

- Term Length
- Wikipedia frequency
- TF-IDF
- Noun Phrase
- Named Entity
- Capitalized
- N-gram
- Term Freq
- Spread
- First Occurrence

Regression coefficient

-2  0  2  4  6  8  10  12  14
All stages of model required for performance

Top-15 keywords

- Precision
- Recall
- F-1 score

Fraction

Random  Candidates  Classifier  Both
Beats AlchemyAPI for a range of parameters
Beats AlchemyAPI for a range of parameters