What is text mining?
What is text mining?

The process of distilling actionable insights from text
Text mining workflow

1 - Problem definition & specific goals

2 - Identify text to be collected

3 - Text organization

4 - Feature extraction

5 - Analysis

6 - Reach an insight, recommendation or output
Semantic parsing vs. bag of words

Steph Curry missed a tough shot.

- **sentence**
- **noun phrase**
  - Steph Curry
  - missed a tough shot
- **verb phrase**
  - missed
  - a
  - tough
  - shot
- **named entity**
  - Steph Curry
Let’s practice!
Getting started
Building our first corpus

> # Load corpus
> coffee_tweets <- read.csv("coffee.csv", stringsAsFactors = FALSE)

> # Vector of tweets
> coffee_tweets <- coffee_tweets$text

> # View first 5 tweets
> head(coffee_tweets, 5)
[1] "@ayyytylerb that is so true drink lots of coffee"
[2] "RT @bryzy_brib: Senior March tmw morning at 7:25 A.M. in the SENIOR lot. Get up early, make yo coffee/breakfast, cus this will only happen ?"
[3] "If you believe in #gunsense tomorrow would be a very good day to have your coffee any place BUT @Starbucks Guns+Coffee=#nosense @MomsDemand"
[5] "RT @slaredo21: I wish we had Starbucks here... Cause coffee dates in the morning sound perff!"
Let’s practice!
Cleaning and preprocessing text
# Common preprocessing functions

<table>
<thead>
<tr>
<th>TM Function</th>
<th>Description</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>tolower()</td>
<td>Makes all text lowercase</td>
<td>Starbucks is from Seattle.</td>
<td>starbucks is from seattle.</td>
</tr>
<tr>
<td>removePunctuation()</td>
<td>Removes punctuation like periods and exclamation points</td>
<td>Watch out! That coffee is going to spill!</td>
<td>Watch out That coffee is going to spill</td>
</tr>
<tr>
<td>removeNumbers()</td>
<td>Removes numbers</td>
<td>I drank 4 cups of coffee 2 days</td>
<td>I drank cups of coffee days ago.</td>
</tr>
<tr>
<td>stripWhiteSpace()</td>
<td>Removes tabs and extra spaces</td>
<td>I like coffee.</td>
<td>I like coffee.</td>
</tr>
<tr>
<td>removeWords()</td>
<td>Removes specific words (e.g. &quot;the&quot;, &quot;of&quot;) defined by the data scientist</td>
<td>The coffee house and barista he visited were nice, she said hello.</td>
<td>The coffee house barista visited were nice, said hello.</td>
</tr>
</tbody>
</table>
Preprocessing in practice

> # Make a vector source: coffee_source
> coffee_source <- VectorSource(coffee_tweets)

> # Make a volatile corpus: coffee_corpus
> coffee_corpus <- VCorpus(coffee_source)

> # Apply various preprocessing functions
> tm_map(coffee_corpus, removeNumbers)
> tm_map(coffee_corpus, removePunctuation)
> tm_map(coffee_corpus, content_transformer(replace_abbreviation))
Another preprocessing step: word stemming

```r
> # Stem words
> stem_words <- stemDocument(c("complicatedly", "complicated", "complication"))

> stem_words
[1] “complic” “complic” “complic”

> # Complete words using single word dictionary
> stemCompletion(stem_words, c("complicate"))
  complic  complic  complic  "complicate"  "complicate"  "complicate"

> # Complete words using entire corpus
> stemCompletion(stem_words, my_corpus)
```
Let’s practice!
The TDM & DTM
### TDM vs. DTM

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>...</th>
<th>Term N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tweet 2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tweet 3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Term M</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Term Document Matrix (TDM)**

<table>
<thead>
<tr>
<th>Term 1</th>
<th>Term 2</th>
<th>Term 3</th>
<th>...</th>
<th>Term M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet 1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Tweet 2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tweet 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Tweet N</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Document Term Matrix (DTM)**

```r
> # Generate TDM
> coffee_tdm <- TermDocumentMatrix(clean_corp)

> # Generate DTM
> coffee_dtm <- DocumentTermMatrix(clean_corp)
```
Word Frequency Matrix (WFM)

> # Load qdap package
> library(qdap)

> # Generate word frequency matrix
> coffee_wfm <- wfm(coffee_text$text)
INTRO TO TEXT MINING: BAG OF WORDS

Let’s practice!