Oportunidades y limitaciones del análisis de texto con R:

la discografía de Taylor Swift

Ariane Aumaitre
Instituto Universitario Europeo
14/11/2019

¿Por qué análisis de texto con R?

- Muchos datos hoy en día pueden ser extraídos de fuentes de texto
- Por ejemplo, en ciencias sociales:
    - *Debates parlamentarios, programas electorales, noticias de periódico, tweets...*
- Automatizando el proceso, el volumen de información que podemos analizar es mucho mayor
- R hace este proceso muy sencillo, especialmente si aplicamos herramientas de tidy data.

¿Por qué Taylor Swift?

- "Artist of the decade" según los AMAs
- Tiene 10 Grammys
- Más de 50 millones de discos vendidos
- Su nuevo disco, Lover, es increíble ♡

... y porque conozco *muy bien* sus canciones.

Temas de esta presentación

- Idea general de cómo funciona tidytext
- Ejemplos de preguntas que podemos contestar mediante análisis de texto
- ¿Cuáles son las limitaciones de esta técnica?

Slides, código y tutorial: https://github.com/aaumaitre/taylor_swift

Tidytext

- Un paquete que permite aplicar principios de "tidy data" a un corpus de texto
- Una observación por fila, una variable por columna
- En el caso del texto -> cada observación es la unidad que queramos analizar (palabras? frases?)
- Organizar nuestro texto así nos permite utilizar herramientas de tidyverse para el análisis

Preparando los datos

Esta es la base de datos descargada de Genius:

```
head(tay)
```

```
## # A tibble: 6 x 5
##   track_title track_n  line lyric                                   album
##   <chr>         <int> <int> <chr>                                   <chr>
## 1 Tim McGraw        1     1 He said the way my blue eyes shined     Taylor S~
## 2 Tim McGraw        1     2 Put those Georgia stars to shame tha~ Taylor S~
## 3 Tim McGraw        1     3 "I said, \"That's a lie\"Just a boy ~ Taylor S~
## 4 Tim McGraw        1     4 That had a tendency of gettin' stuck  Taylor S~
## 5 Tim McGraw        1     5 On backroads at night                 Taylor S~
## 6 Tim McGraw        1     6 And I was right there beside him all~ Taylor S~
```

Preparando los datos (II)

Con la función **unnest_tokens()** reducimos la columna **lyric** a nuestra unidad de estudio: las palabras.

```
library(tidytext) #Tidytext package
#Tokenizing our data:
tay_tok <- tay%>%
  #word is the new column, lyric the column to retrieve the information from
  unnest_tokens(word, lyric)
head(tay_tok)
```

```
## # A tibble: 6 x 5
##   track_title track_n  line album        word
##   <chr>         <int> <int> <chr>        <chr>
## 1 Tim McGraw        1     1 Taylor Swift he
## 2 Tim McGraw        1     1 Taylor Swift said
## 3 Tim McGraw        1     1 Taylor Swift the
## 4 Tim McGraw        1     1 Taylor Swift way
## 5 Tim McGraw        1     1 Taylor Swift my
## 6 Tim McGraw        1     1 Taylor Swift blue
```

Análisis exploratorio

Las palabras más repetidas no nos dan mucha información... 😭

```
tay_tok %>%
  count(word, sort = TRUE) %>%
  head()
```

```
## # A tibble: 6 x 2
##   word      n
##   <chr> <int>
## 1 you    1812
## 2 i      1519
## 3 the    1161
## 4 and    1109
## 5 me      646
## 6 to      639
```

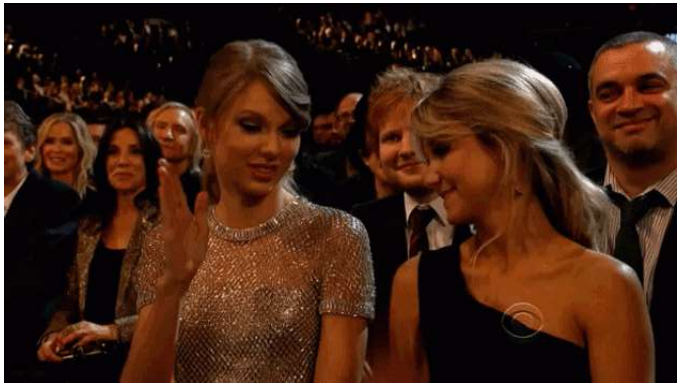
Análisis exploratorio (II)

- Una forma de resolver este problema es eliminar las llamadas **palabras vacías**, que no aportan significado
-

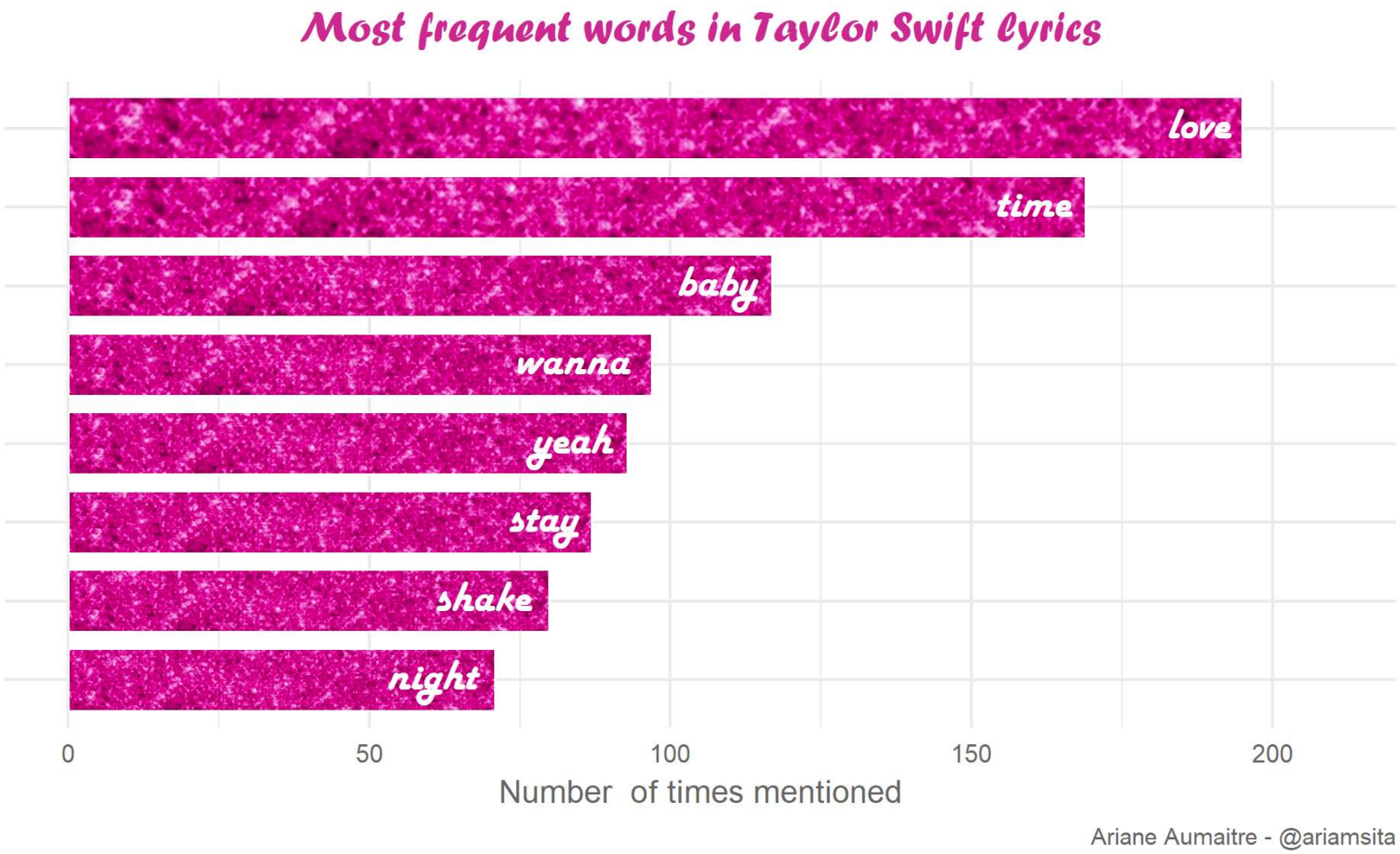
  ¿Cómo? Usando **anti_join()** y el diccionario **stop_words**

```
tidy_taylor <- tay_tok %>%
anti_join(stop_words)
tidy_taylor %>%
count(word, sort = TRUE) %>% head()
```

```
## # A tibble: 6 x 2
##   word      n
##   <chr> <int>
## 1 love    195
## 2 time    169
## 3 baby    117
## 4 ooh     108
## 5 wanna    97
## 6 yeah     93
```



¡Ahora sí!


Most frequent words in Taylor Swift lyrics
love
time
baby
wanna
yeah
stay
shake
night
0
50
100
150
200
Number of times mentioned
Ariane Aumaitre - @ariamsita

Aplicaciones: análisis de sentimiento

- El análisis de sentimiento nos permite aproximar las connotaciones subjetivas de los textos analizados
- ¿Predominan emociones positivas o negativas?
- Se lleva a cabo mediante diccionarios que aplicamos a nuestro texto
- Tidytext contiene tres diccionarios: *AFINN*, *bing* y *nrc*. Es importante elegir un diccionario que *tenga sentido* para el texto con el que estamos trabajando.
- Funciones que utilizaremos: `get_sentiments()` e `inner_join()`

La función get_sentiments()

```
tidy_taylor%>%
  inner_join(get_sentiments("bing"))
```

```
## # A tibble: 2,074 x 6
##    track_title track_n  line album        word      sentiment
##    <chr>         <int> <int> <chr>        <chr>     <chr>
##  1 Tim McGraw        1     2 Taylor Swift shame     negative
##  2 Tim McGraw        1     3 Taylor Swift lie       negative
##  3 Tim McGraw        1     4 Taylor Swift stuck     negative
##  4 Tim McGraw        1     9 Taylor Swift favorite  positive
##  5 Tim McGraw        1    12 Taylor Swift happiness positive
##  6 Tim McGraw        1    24 Taylor Swift hard      negative
##  7 Tim McGraw        1    25 Taylor Swift nice      positive
##  8 Tim McGraw        1    27 Taylor Swift favorite  positive
##  9 Tim McGraw        1    30 Taylor Swift happiness positive
## 10 Tim McGraw        1    40 Taylor Swift favorite  positive
## # ... with 2,064 more rows
```

La función get_sentiments()

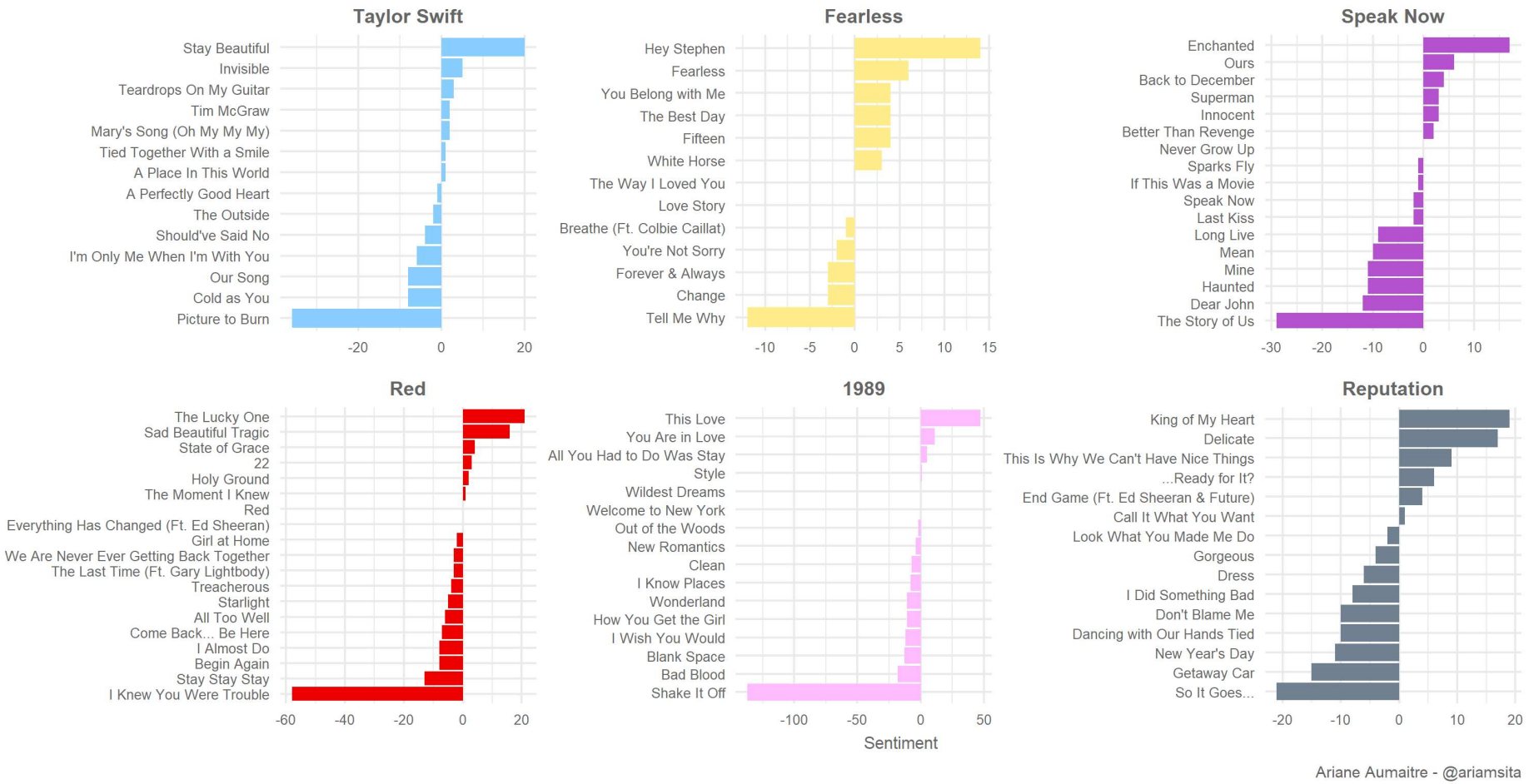
```
tidy_taylor%>%
  inner_join(get_sentiments("bing"))%>%
  count(album, track_title, sentiment)
```

```
## # A tibble: 183 x 4
##    album track_title                sentiment     n
##    <chr> <chr>                      <chr>     <int>
##  1 1989  All You Had to Do Was Stay negative      1
##  2 1989  All You Had to Do Was Stay positive      6
##  3 1989  Bad Blood                  negative     28
##  4 1989  Bad Blood                  positive     10
##  5 1989  Blank Space                negative     32
##  6 1989  Blank Space                positive     19
##  7 1989  Clean                      negative     19
##  8 1989  Clean                      positive     12
##  9 1989  How You Get the Girl       negative     11
## 10 1989  I Know Places              negative     10
## # ... with 173 more rows
```

La función get_sentiments()

```
tidy_taylor%>%
  inner_join(get_sentiments("bing"))%>%
  count(album, track_title, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = positive - negative)%>%
  arrange(desc(sentiment))
```

```
## # A tibble: 94 x 5
##    album       track_title                    negative positive sentiment
##    <chr>       <chr>                             <dbl>    <dbl>     <dbl>
##  1 1989        This Love                            14       61        47
##  2 Red         The Lucky One                         2       23        21
##  3 Taylor Swi~ Stay Beautiful                        1       21        20
##  4 Reputation  King of My Heart                      5       24        19
##  5 Reputation  Delicate                             11       28        17
##  6 Speak Now   Enchanted                             3       20        17
##  7 Red         Sad Beautiful Tragic                 14       30        16
##  8 Fearless    Hey Stephen                           2       16        14
##  9 1989        You Are in Love                       5       16        11
## 10 Reputation  This Is Why We Can't Have Nice ~      10       19         9
## # ... with 84 more rows
```


Taylor Swift's songs ranked by sentiment
Taylor Swift
Stay Beautiful
Invisible
Teardrops On My Guitar
Tim McGraw
Mary's Song (Oh My My My)
Tied Together With a Smile
A Place In This World
A Perfectly Good Heart
The Outside
Should've Said No
I'm Only Me When I'm With You
Our Song
Cold as You
Picture to Burn
-20
0
20
Fearless
Hey Stephen
Fearless
You Belong with Me
The Best Day
Fifteen
White Horse
The Way I Loved You
Love Story
Breathe (Ft. Colbie Caillat)
You're Not Sorry
Forever & Always
Change
Tell Me Why
-10
-5
0
5
10
15
Speak Now
Enchanted
Ours
Back to December
Superman
Innocent
Better Than Revenge
Never Grow Up
Sparks Fly
If This Was a Movie
Speak Now
Last Kiss
Long Live
Mean
Mine
Haunted
Dear John
The Story of Us
-30
-20
-10
0
10
Red
The Lucky One
Sad Beautiful Tragic
State of Grace
22
Holy Ground
The Moment I Knew
Red
Everything Has Changed (Ft. Ed Sheeran)
Girl at Home
We Are Never Ever Getting Back Together
The Last Time (Ft. Gary Lightbody)
Treacherous
Starlight
All Too Well
Come Back... Be Here
I Almost Do
Begin Again
Stay Stay Stay
I Knew You Were Trouble
-60
-40
-20
0
20
1989
This Love
You Are in Love
All You Had to Do Was Stay
Style
Wildest Dreams
Welcome to New York
Out of the Woods
New Romantics
Clean
I Know Places
Wonderland
How You Get the Girl
I Wish You Would
Blank Space
Bad Blood
Shake It Off
-100
-50
0
50
Sentiment
Reputation
King of My Heart
Delicate
This Is Why We Can't Have Nice Things
...Ready for It?
End Game (Ft. Ed Sheeran & Future)
Call It What You Want
Look What You Made Me Do
Gorgeous
Dress
I Did Something Bad
Don't Blame Me
Dancing with Our Hands Tied
New Year's Day
Getaway Car
So It Goes...
-20
-10
0
10
20
Ariane Aumaitre - @ariamsita

Cuidado con...

- Ironía
- Dobles sentidos
- La elección del diccionario
- Es importante comparar resultados

Comparando diccionarios

```
tidy_taylor%>%
  inner_join(get_sentiments("bing"))%>%
  filter(sentiment == "positive")%>%
  count(album, track_title, sentiment, sort = TRUE)%>%
    select(track_title, n)
```

```
## # A tibble: 89 x 2
##    track_title                              n
##    <chr>                                <int>
##  1 This Love                               61
##  2 Sad Beautiful Tragic                    30
##  3 Delicate                                28
##  4 End Game (Ft.&amp;nbsp;Ed&amp;nbsp;Sheeran & Future)    26
##  5 King of My Heart                        24
##  6 The Lucky One                           23
##  7 Gorgeous                                22
##  8 Stay Beautiful                          21
##  9 Enchanted                               20
## 10 Blank Space                             19
## # ... with 79 more rows
```

```
tidy_taylor%>%
  inner_join(get_sentiments("nrc"))%>%
  filter(sentiment == "positive")%>%
  count(album, track_title, sentiment, sort = TRUE)%>%
  select(track_title, n)
```

```
## # A tibble: 93 x 2
##    track_title              n
##    <chr>                <int>
##  1 This Love               59
##  2 Don't Blame Me          48
##  3 Never Grow Up           33
##  4 Starlight               28
##  5 Clean                   26
##  6 You Are in Love         26
##  7 Love Story              26
##  8 Blank Space             25
##  9 Sad Beautiful Tragic    25
## 10 The Lucky One           25
## # ... with 83 more rows
```

Comprender los resultados (I)

```
tidy_taylor%>%
  inner_join(get_sentiments("bing"))%>%
  filter(sentiment == "positive")%>%
  count(sentiment, track_title, word, sort = TRUE)%>%
  select(track_title, word, n)
```

```
## # A tibble: 379 x 3
##    track_title                          word          n
##    <chr>                                <chr>     <int>
##  1 This Love                            love         52
##  2 King of My Heart                     whoa         16
##  3 Delicate                             delicate     15
##  4 You Are in Love                      love         15
##  5 This Is Why We Can't Have Nice Things nice        14
##  6 Gorgeous                             gorgeous     13
##  7 Stay Beautiful                       beautiful    13
##  8 The Lucky One                        lucky        13
##  9 Sad Beautiful Tragic                 beautiful    12
## 10 Clean                                clean        11
## # ... with 369 more rows
```

Comprender los resultados (II)

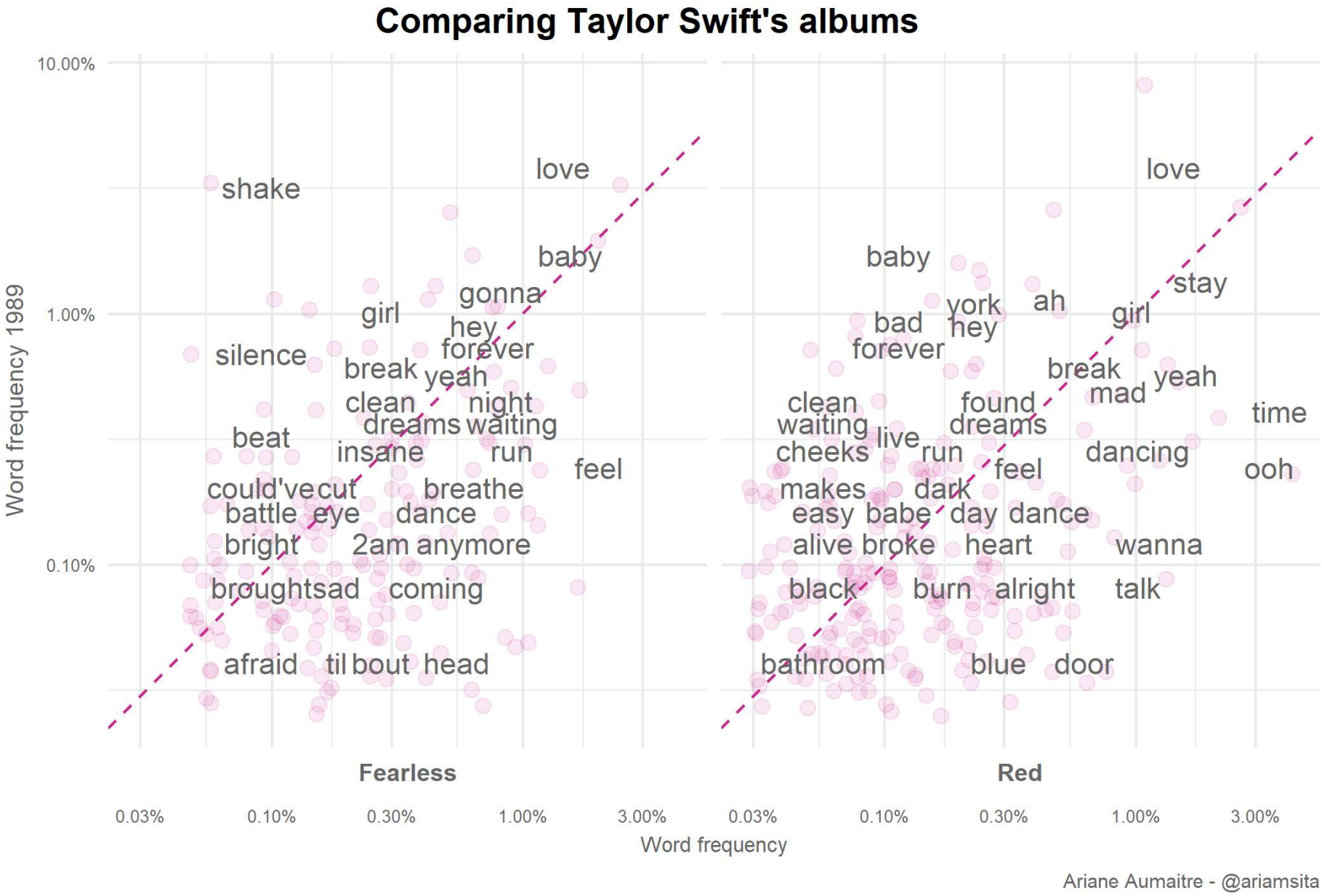
```
tidy_taylor%>%
  inner_join(get_sentiments("bing"))%>%
  filter(sentiment == "negative")%>%
  count(sentiment, track_title, word, sort = TRUE)
```

```
## # A tibble: 596 x 4
##    sentiment track_title             word        n
##    <chr>     <chr>                   <chr>   <int>
##  1 negative  Shake It Off            shake      78
##  2 negative  I Knew You Were Trouble trouble    32
##  3 negative  Shake It Off            fake       18
##  4 negative  Shake It Off            hate       16
##  5 negative  Bad Blood               bad        15
##  6 negative  Shake It Off            break      15
##  7 negative  Stay Stay Stay          mad        15
##  8 negative  I Did Something Bad     bad        10
##  9 negative  Picture to Burn         burn        9
## 10 negative  Wonderland              lost        9
## # ... with 586 more rows
```

Otras aplicaciones

- Identificar tendencias temporales
- Buscar similitudes o diferencias entre textos
- Topic modelling
- network analysis

Ejemplo: relación entre discos



Ejemplo 2: relación entre canciones