Comparing Keywords in CNN and BBC

Camila Lívio

October 2, 2023

1 Introduction

The data were retrieved from the open source, online community platform Kaggle. Specifically, I used the CNN and BBC news data sets.

1.1 Dataset Summary

- Corpus size: CNN = over 3 million words; BBC = over 200K words.
- Languages: en-US and en-GB
- Date range for CNN texts: 2011-08-24 17:54:07 UTC - 2022-03-21 11:40:28 UTC
- Date range for BBC texts: Fri, 01 Apr 2022 00:06:38 GMT - Wed, 31 May 2023 23:32:52 GMT

1.2 Objectives

(1) Compare and contrast the distribution of frequent words, bigrams and keywords in two news outlets, namely CNN and BBC;

(2) Isolate and observe the distribution of key terms such as *film*, *adaptation* and *literature* in the corpora;

(3) Briefly discuss the data and the computational techniques involved in the analysis;

(4) Raise awareness to a computational approach to text analysis.

2 Exploring the CNN data set

As a first look at the data, let's observe the distribution of frequent words:

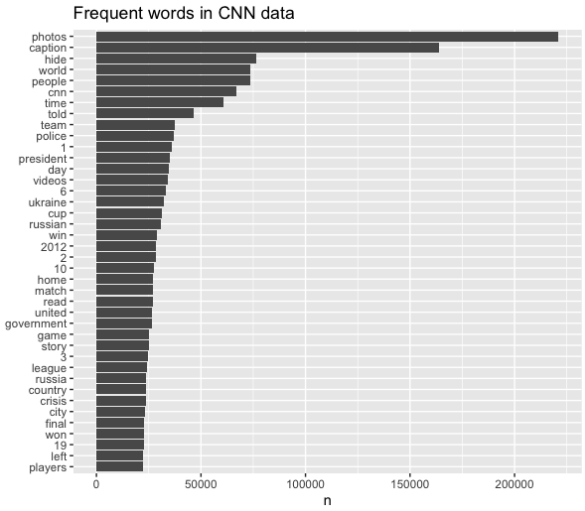

Figure 1: Frequent words in CNN data

To better understand the relationship between words, let's take a computational approach and extract bigrams: a sequence of two words that frequently appear together, such as in *go to* and *like it*. The list below shows the first 10 lines from our search:

```
   bigram            n
   <chr>         <int>
 1 of the       223250
 2 in the       196807
 3 to the        95390
 4 at the        79918
 5 on the        79237
 6 hide caption  76004
 7 for the       66455
 8 in a          60023
 9 to be         48771
10 and the       46017
```

Lots of function words! There are some tricks that we can do to observe more meaningful bigrams:

```
 word1  word2        n
   <chr>  <chr>    <int>
 1 hide   caption  76004
 2 photos photos   27871
 3 world  cup      18396
 4 told   cnn      17301
 5 photos crisis   14830
 6 covid  19       11206
 7 prime  minister 10752
 8 488    photos    8766
 9 168    photos    8517
10 euro   2012      7372
#  3,051,060 more rows
```

Now that I tokenized (split) the data into bigrams, it is possible to isolate key terms that interest us to observe what words appear before and after the word we are focusing on. Let's first analyze the words that most frequently precede the term *film* in the CNN data set:

```
      word1             n
   <chr>         <int>
```

```
 1 cannes           47
 2 short            40
 3 documentary      34
 4 feature          31
 5 bond             30
 6 winning          27
 7 horror           24
 8 venice           24
 9 british          20
10 gothenburg       18
11 hollywood        18
12 adult            17
13 american         17
14 popular          17
15 2013             15
16 islam            14
17 international    13
18 ridiculous       13
19 language         12
20 silent           12
```

How about words preceding the term *adaptation*?

```
     word1         n
   <chr>     <int>
1 tv            2
2 ambitious     1
3 bbc           1
4 climate       1
```

Unfortunately, there were no results for the keyword *literature*. Next, we can easily isolate the contexts in which our key terms appear, even though our data set contains millions of words.

```
   Keyword-in-context with 25 matches.
  [text35, 277]               Charlie Sheen in the 1984 | film | ’ Red Dawn.’
  [text35, 430]                 Chevy Chase in the 1985 | film | ’ Spies Like Us’As for
  [text35, 750]                 the 1980s came not from | film | or TV, but Sting’s
  [text44, 343]                      " Unfaithful," the | film | that this movie most closely
   [text46, 50]            the the 2022 British Academy | Film | Awards ( BAFTAs ) on
  [text55, 367]                  ," a stark independent | film | about troubled teens. Read
   [text57, 23]     back from acting.While promoting her | film | " Lost City,"
  [text59, 294]               Brooding and serious, the | film | caters to those weaned on
   [text60, 84]                           Bao" ), the | film | tells the story of 13-year-old
  [text60, 238]                    Out," another Pixar | film | that did an inordinately good
  [text60, 336]                  best Pixar fare, the | film | operates on multiple levels,
  [text64, 415]             with introducing a" popular | film | " category in 2018,
  [text64, 759]  animated, documentary and international | film | categories.The Academy’s ongoing efforts to
  [text82, 235] legend.Now immortalized in the Hollywood | film | " King Richard" --
  [text82, 265]                      by Will Smith, the | film | shows how Richard Williams catapulted
  [text82, 541]                       Venus. And in the | film | , it goes into that
 [text104, 182]         Sabourin has recently released a | film | with Patagonia that explores their
 [text104, 263]               and sexual assault -- the | film | " They/ Them"
 [text104, 284]                  community." Making the | film | was hard in a lot
 [text104, 396]                           Them," is a | film | about Lor Sabourin, an
 [text104, 734]              different pronouns. In the | film | , they describe how they
 [text104, 807]                        said." With the | film | , it definitely brings up
 [text104, 877]            is really powerful about the | film | [... ]
 [text104, 973]               The process of making the | film | was an intimate affair,
 [text104, 996]                  . They didn’t want the | film | to make generalizations about the
```

How does this term compare to the use of the term *movie* in the data set?

```
   Keyword-in-context with 25 matches.
 [text1, 2309]              tired, doesn’t watch a | movie | or look at a phone
  [text35, 72]                one kind of Cold War | movie | during that period, but
 [text35, 174]                       ," a 1983 TV | movie | considered so provocative that the
```

```
[text35, 201]                of the content, the | movie | drew a massive audience --
[text35, 843]                an idea to become a | movie | or TV show, it's
 [text44, 61]              that genre. While the | movie | falls apart toward the end
[text44, 346]               " the film that this | movie | most closely resembles in tone
[text44, 545]              watching, even if the | movie | seems destined to generate its
[text54, 190]            colors they wore in the | movie | . Kudrow currently stars in
[text55, 381]          was 14 while shooting the | movie | , discusses her discomfort with
[text55, 425]   notes that Manson referenced the | movie | when they first met.Documented with
[text55, 451]         look at children raised on | movie | and TV sets, an
  [text57, 9]           Sandra Bullock has a new | movie | coming out, but she's
[text59, 399] informed director Tim Burton's 1989 | movie | . Yet even with its
[text59, 550]                 before." While the | movie | is assured of a big
 [text60, 33]    a bright and appealing animated | movie | somewhat surprisingly headed directly to
[text60, 434]                is puzzling, with a | movie | that's qualitatively in the conversation
[text61, 429]         fare, including the recent | movie | " The Humans" --
[text63, 264]                 " ). Best animated | movie | :" The Mitchells vs
[text63, 277]      Machines." Best international | movie | :" The Hand of
[text63, 347]                 " ). Best animated | movie | :" Flee."
[text63, 400]                 2 ). Best animated | movie | :" Encanto,"
 [text64, 35]        network ABC and promote the | movie | business. And after record-low
[text64, 493]             that the Oscars and the | movie | business face has stemmed from
[text64, 659]         a nominee for best animated | movie | , which took off once
```

Before we turn to the BCC news data set, let's visualize in a graph the words that are frequently associated with *film* in the CNN data:

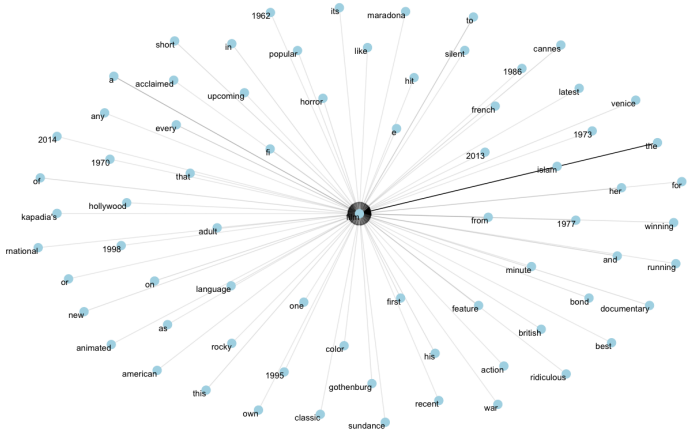

Figure 2: Words frequently associated with *film* in CNN data

3 Exploring the BBC data set

Let's start by observing the distribution of very frequent words in the BBC news data set. Inspect the scale in the x-axis to observe how frequent these words are in the BBC corpus. Do you see any patterns or trends? How does Figure 3 compare with Figure 1?

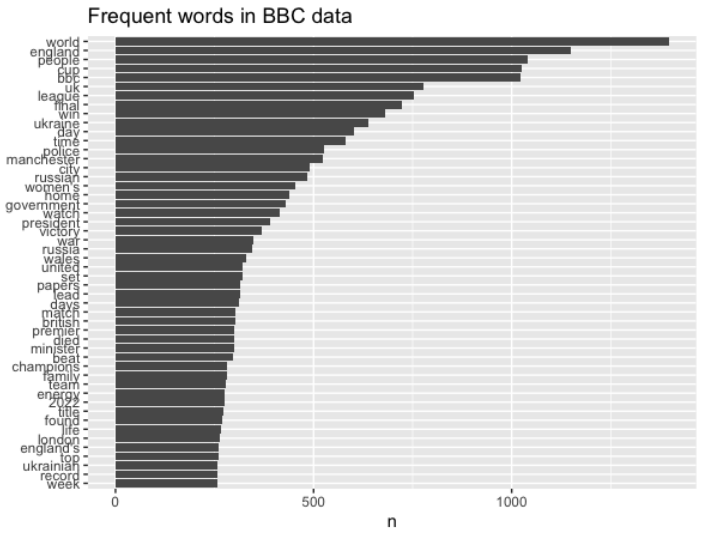

Figure 3: Frequent Words in the BBC news data

3.1 Inspecting bigrams

As we have seen in the first part of our analysis, observing individual words can tell us some trends, but it is only in combination with other words that we can make sense of meaning. Computationally, manipulating and analyzing bigrams can give us further insight about the data. What do you observe in the list of most frequent bigrams below? (Hint: a lot of uninteresting words!)

```
# A tibble: 165,286 × 2
   bigram        n
   <chr>     <int>
 1 of the     2294
 2 in the     2209
 3 at the     1039
 4 to the      827
 5 in a        822
 6 world cup   814
 7 for the     766
 8 on the      765
 9 the first   551
10 the uk      525
```

After cleaning stop words:

```
  word1      word2        n
   <chr>      <chr>    <int>
 1 world      cup        814
 2 premier    league     281
 3 manchester united     208
 4 prime      minister   204
 5 manchester city       198
 6 champions  league     179
 7 women's    world      167
 8 bbc        sport      165
 9 front      pages      154
10 rishi      sunak      131
```

Like we did before, let's target bigrams that contain some key words: *film* and *literature*. First, let's isolate all the bigrams that contain the word *film* in the second slot. In other words, in the list

below the column named "word1" corresponds to the words that precede our target word in the data set.

word1	n
<chr>	<int>
1 adult	2
2 animated	2
3 documentary	2
4 howard's	2
5 indian	2
6 nominated	2
7 short	2
8 tv's	2
9 1981	1
10 1997	1
11 1999	1
12 2006	1
13 academy	1
14 anderson's	1
15 anime	1

Unfortunately, the same procedure for the word *literature* did not return any results. After inspecting bigrams, let's see some full examples of the word *film* in context.

```
    > head(bbc_kwic, 20)
Keyword-in-context with 20 matches.
  [text147, 16]  chance to join Pixar's latest | film | Turning Red.
  [text189, 11]      the track from the Disney | film | while sheltering in a Ukraine
  [text235, 11] Jones in the Bafta-winning BBC | film | Marvellous, which is now
   [text923, 3]            The Oscar-nominated | film | draws on the director's childhood
   [text924, 2]                       Animated | film | Jujutsu Kaisen 0 is a
   [text931, 3]                    Apple TV's | film | about the hearing daughter of
   [text965, 4]            The pioneering LGBT | film | made during the Balkans conflict
  [text1004, 3]                      The 1997 | film | based in Sheffield is being
  [text1022, 3]                    Apple TV's | film | about the hearing daughter of
  [text1223, 3]                       Has the | film | star's on-stage slap done permanent
  [text1766, 3]                         A new | film | tells the story of a
 [text1917, 13]       in his industry with new | film | Toxic.
  [text1942, 5]               The makers of a | film | whose cinematographer was shot dead
 [text1967, 17]            her to make her new | film | .
 [text1969, 21]            on the" pan-Indian" | film | .
  [text2266, 5]                   A mystery, a | film | crew and a trip to
 [text2266, 19] part of Downton Abbey's second | film | outing.
 [text3119, 10]          an Oscar for the 1981 | film | theme, and also wrote
 [text3442, 22]             that works" in the | film | .
  [text3949, 5]         The lawsuit claims the | film | studio did not have the
```

Let's visualize in a graph the words that are frequently associated with *film* in the BBC data:

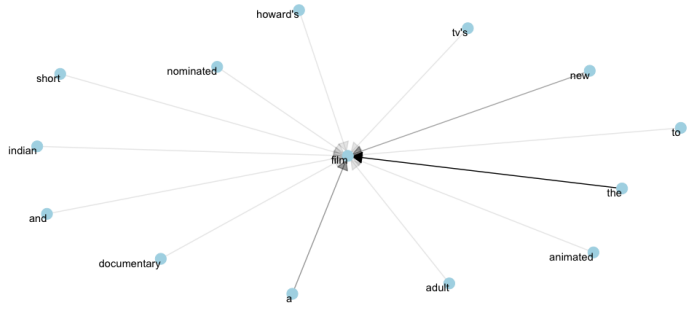

Figure 4: Words frequently associated with *film* in BBC data