Capstone Project 2: Analysis of Berkshire Hathaway Shareholder Letters Using Natural Language Processing (NLP) Techniques

Tom Halloin

SPRINGBOARD DATA SCIENCE CAREER TRACK

“Read 500 pages every day. That's how knowledge works. It builds up, like compound interest. All of you can do it, but I guarantee not many of you will do it.”

- Warren Buffett, when asked how to prepare for an investing career.

The Problem

Warren Buffett became one of the most successful investors of all time because of his ability to comprehend thousands of pages of data. Buffett is the exception to the rule. Most people don’t have the patience to read 500 pages a day, let alone comprehend it (and I am not the exception). Instead, I am going to use machine learning to summarize the equivalent of reading 500 pages, Warren Buffett’s annual letters to shareholders since 1977. This capstone will be an exploration of Berkshire Hathaway’s annual shareholder letters. Using Natural Language Processing, I will try to summarize the letters and create a topic model to discover hidden themes that are present across the letters. This model will try to extract meaning from these letters by identifying recurring themes or topics and their change from letter to letter.

The Dataset

The data comes from Berkshire Hathaway’s shareholder letters available on their website. The letters come in both HTML and PDF format. Using Python, I managed to scrape the text from annual letters dating back from 1977. See the code in the notebook linked here for more information. Note that Berkshire eventually caught on that I was a bot, and therefore denied me access to their reports. Fortunately, this was after I gathered all of my data, so while the code used to work, there is no way for it to work anymore without access. The letters are also saved as part of this Milestone Report.

Data Wrangling

Before doing any analysis on the letters, I needed to do some data wrangling to clean the output from the pages and PDFs the letters came from into useable text. Before any data wrangling, the files looked like the following sample.

```
<!-- Global site tag (gtag.js) - Google Analytics -->
<script async src="https://www.googletagmanager.com/gtag/js?id=UA-136883390-1"></script>
<script>
  window.dataLayer = window.dataLayer || [];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());

  gtag('config', 'UA-136883390-1');
</script>
<HTML>
<HEAD>
  <TITLE>Chairman's Letter - 1977</TITLE>
</HEAD>
<BODY>
<P ALIGN=CENTER>
<B>BERKSHIRE HATHAWAY INC.</B>
</P>
<PRE>

<I>To the Stockholders of Berkshire Hathaway Inc.:</I>

     Operating earnings in 1977 of $21,904,000, or $22.54 per
share, were moderately better than anticipated a year ago.  Of
these earnings, $1.43 per share resulted from substantial
realized capital gains by Blue Chip Stamps which, to the extent
of our proportional interest in that company, are included in our
operating earnings figure.  Capital gains or losses realized
directly by Berkshire Hathaway Inc. or its insurance subsidiaries
are not included in our calculation of operating earnings.  While
too much attention should not be paid to the figure for any
single year, over the longer term the record regarding aggregate
capital gains or losses obviously is of significance.
```

Most of the cleaning for these files included removing HTML, anything before and after the actual contents of the letter, and non-ascii characters. A series of regular expressions removed unnecessary characters in tables and newline characters to make the text more readable. I skipped removing stop words and numbers because I want the ending summaries to read as close to actual sentences as possible.

The end result was a series of paragraphs in documents looking like this:

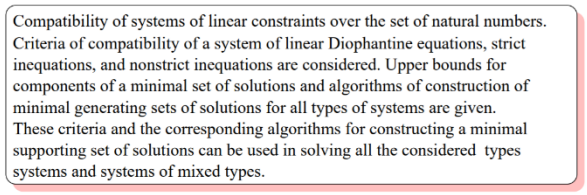
To the Stockholders of Berkshire Hathaway Inc.: Operating earnings in 1977 of $21,904,000, or $22.54 per share, were mod erately better than anticipated a year ago. Of these earnings, $1.43 per share resulted from substantial realized capital g ains by Blue Chip Stamps which, to the extent of our proportional interest in that company, are included in our operating ea rnings figure. Capital gains or losses realized directly by Berkshire Hathaway Inc. or its insurance subsidiaries are not i ncluded in our calculation of operating earnings. While too much attention should not be paid to the figure for any single year, over the longer term the record regarding aggregate capital gains or losses obviously is of significance.

Summarizing the Data

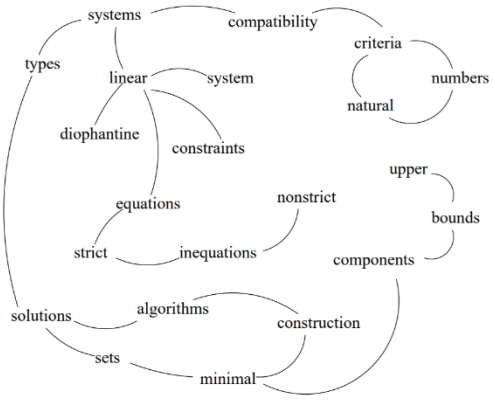
Word clouds provide a quick overview of text data. They take a group of words and create an image. The word cloud below represents the most common names of people and organizations featured in the annual letters. Berkshire appears multiple times, along with Charlie (Munger), and several of Berkshire's subsidiaries in Geico, Nebraska Furniture Mart (NFM), and NetJets.

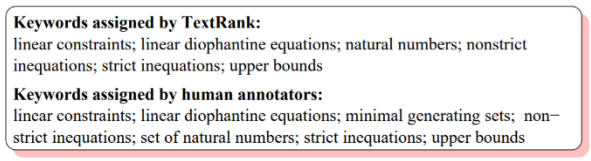
Word clouds provide a starting point for the actual "deliverable" – summaries of the letters. There are two ways to summarize data: *extractive* summarization and *abstractive* summarization. *Extractive* summarization selects the most important phrases from the text and extracts it to get the most important phrases. *Abstractive* summarization creates sentences from the text using deep learning and returns these new sentences as the summary. Computers are

really good for doing extractive summarization, but abstractive summarization is much harder because it involves creating new text to explain the document so that humans can understand it. There have been attempts with deep learning to create abstractive summaries, but that is beyond the scope of this project.



I will try three different extractive summarization methods: Textrank, Lexrank, and LSA. Textrank is a graph-based ranking model similar to Google's Pagerank algorithm. Each word links to other words based on the number of times the words appear near each other. The algorithm then assigns hidden weights to each of these links given how likely each of the linked words appears with each other. Suppose in the example to the right I start with "natural". Natural is linked with "criteria" and "numbers". Based on Textrank's weighting, however, it will determine that "natural numbers" is a better keyword than "natural criteria". The weighting itself can be a black box – see *Textrank: Bringing Order into Texts* by Rada Mihalcea and Paul Tarau for more details on the mathematical implementation.





See Textrank: Bringing Order into Texts by Rada Mihalcea and Paul Tarau, p. 4

In the example shown on the right, connected words are good examples of keywords that can summarize documents. Examples of summarization of the text on the right include "linear Diophantine equations", "upper bounds", and "natural numbers". Without any comprehension of the actual text, the graph shows that the text is at least about these three keywords. The algorithm will then look for sentences with these keywords to summarize the document.

Lexrank is very similar to Textrank. Both use the same graph-based weighting model, but differ in the method of generating weights. Lexrank uses a method of cosine similarity weighted by term frequency – inverse document frequency, otherwise known as TF-IDF, to weight terms that have more meaning instead of just looking at document frequencies. The end result of applying the Lexrank algorithm to a series of documents is a graph similar to the picture on the right here. Each circle is a document and the edges represent the cosine similarities between each of the documents. In this example, d1s1 (document 1, sentence 1) and d2s1 (document 2, sentence 1) have a high weighted cosine similarity and would be good candidates for summarizing the corpus of documents.

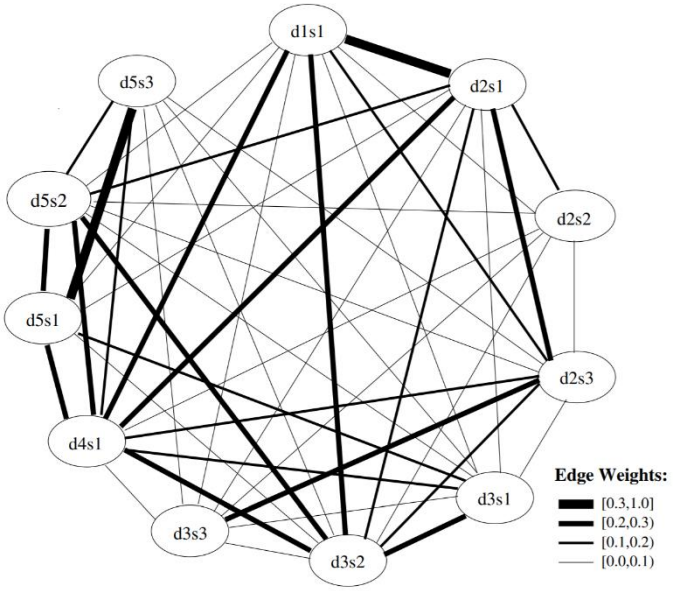
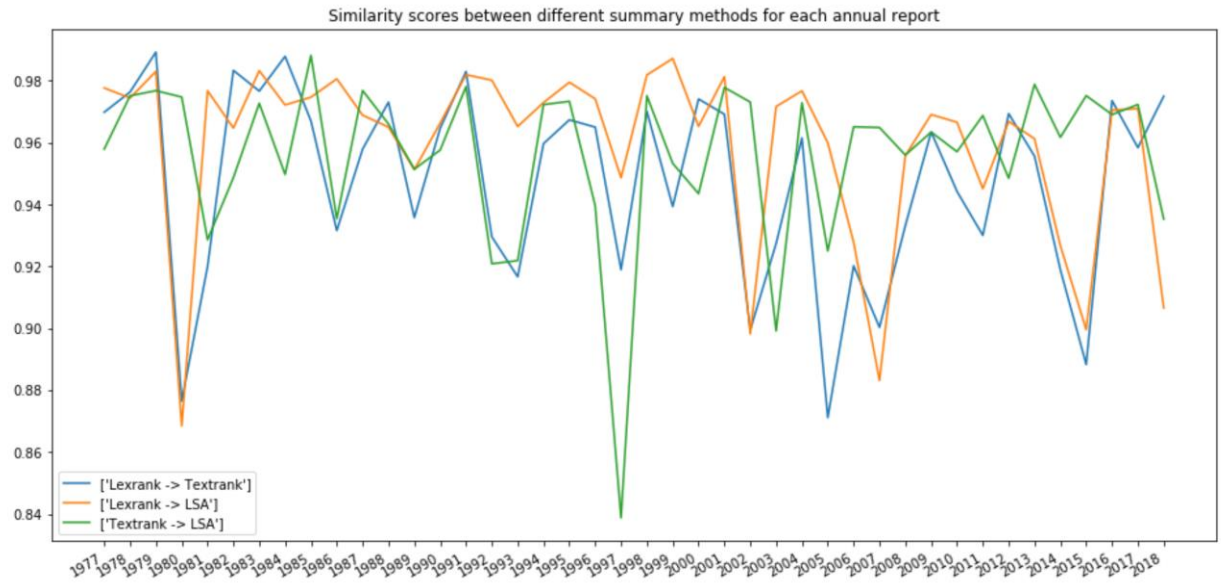


Figure 2: Weighted cosine similarity graph for the cluster in Figure 1.

See LexRank: Graph-based Lexical Centrality as Salience in Text Summarization by Gunes Erkan and Dragomir R. Radev for graph.

LSA, or *Latent Semantic Analysis*, analyzes similarities between the meanings of a series of documents and produces a series of concepts related to the documents. Like Lexrank and Textrank, LSA creates a term-document matrix consisting of word frequencies for each term in each document. Each word is a row while each document is a column. After creating this matrix, a linear algebra technique called singular value decomposition reduces the number of words, while keeping the same meaning in each of the documents. By reducing the number of words in each document, it is easier to find similarities between different documents. Sentences that keep the most words from the reduction are considered the “most important”. These reduced words and phrases also generate topic models that will be shown later.

For each annual report, I used each of these three techniques to summarize each year in the corpus by extracting the five most meaningful sentences according to its method. Then, I used the Spacy library to create a summary score to see how similar the summaries are to each other. Spacy has a large vocabulary of words that can capture similar phrases even if the words do not match, so the techniques should produce similar summaries according to Spacy. The below graph plots these similarity scores between the three different methods. Each line is a comparison between two of the methods for each year in the annual report. The divergences, as shown in the year 1997, provide a good contrast to see the different sentences each technique deems as important.



Here are the summaries for each method in 1997:

```
1 lexrank[20]
```

executed in 10ms, finished 19:21:29 2020-01-12

```
['Even though they are going to be net buyers of stocks formany years to come, they are elated when stock prices rise and de
pressedwhen they fall.',
 'Now assume that the dice will be thrown once a year; that you, the&amp;quot;bond-buyer,&amp;quot; agree to pay $50 million if a 12
appears; and thatfor &amp;quot;insuring&amp;quot; this risk you take in an annual &amp;quot;premium&amp;quot;of $1 million.',
 '&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp; At GEICO, we are paying in a waythat makes sense for both our owners and our manager
s.',
 '12/31/97        Shares  Company                      Cost      Market                                                (doll
ars in millions)    49,456,900  American Express Company     $1,392.7   $  4,414.0   200,000,000  The Coca-Cola Company
1,298.9     13,337.5    21,563,414  The Walt Disney Company         381.2      2,134.8    63,977,600  Freddie Mac
329.4      2,683.1    48,000,000  The Gillette Company            600.0      4,821.0    23,733,198  Travelers Group Inc.
604.4      1,278.6     1,727,765  The Washington Post Company      10.6        840.6     6,690,218  Wells Fargo & Company
412.6      2,270.9                Others                        2,177.1      4,467.2
Total Common Stocks          $7,206.9   $ 36,247.7                                                  &amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;
&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;Represents tax-basis cost which, in aggregate, is $1.8 billion less thanGAAP
cost.',
 'The fourremaining preferred commitments included two, Gillette and First EmpireState, that we had converted into common st
ock in which we had large unrealizedgains, and two others, USAir and Salomon, that had been trouble-prone.At times, the last
two had me mouthing a line from a country song: &amp;quot;Howcan I miss you if you wont go away?&amp;quot; &amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&
nbsp;&amp;nbsp; Since I delivered that report,all four holdings have grown significantly in value.']
```

```
1 textrank[20]
```

executed in 8ms, finished 19:21:29 2020-01-12

```
['&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp; Here are the growth rates of thetwo segments by decade:
Pre-tax Earnings Per Share                        Investments          Excluding All Income from  Decade Ending
Per Share                  Investments              1977                   24.6%                       27.6%      1987
26.5%                       24.1%      1997                  25.5%                    20.8% Annual Growth  Rate, 1967-19
97           25.6%                       24.2%&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp; During 1997, both parts of ourbusiness g
rew at a satisfactory rate, with investments increasing by $9,543per share, or 33.5%, and operating earnings growing by $29
6.43 per share,or 70.3%.',
 '&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp; Below are the new business andin-force figures for the last five years:
New Voluntary      Voluntary Auto      Years           Auto Policies    Policies in Force       1993              354,882
2,011,055           1994             396,217          2,147,549       1995              461,608           2,310,037
1996              617,669           2,543,699       1997              913,176           2,949,439&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nb
sp;&amp;nbsp; Of course, any insurer can growrapidly if it gets careless about underwriting.',
 '118.9     121.7       77.3      81.6  Sees Candies    58.6      51.9       35.0      30.8  Shoe Group     48.8      61.6
32.2      41.0  Purchase-Accounting Adjustments  (104.9)    (75.7)     (97.0)    (70.5)  Interest Expense(3)  (106.6)    (9
4.3)     (67.1)    (56.6)  Shareholder-Designated Contributions    (15.4)    (13.3)      (9.9)     (8.5)  Other    60.7
73.0       37.0      42.2                                                    Operating Earnings  1,715.7   1,221.4    1,
194.5     883.1 Capital Gains from Investments  1,111.9   2,484.5      707.1   1,605.5
Total Earnings - All Entities .$2,827.6  $3,705.9   $1,901.6  $2,488.6
&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;(1) From date of acquisition,December 23, 1996.&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;(2) I
ncludes StarFurniture from July 1, 1997.&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;(3) Excludes interestexpense of Finance Business
es.&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp;&amp;nbsp; Overall, our operating businessescontinue to perform exceptionally well, far outdoin
g their industry norms.We are particularly pleased that profits improved at Helzbergs after adisappointing 1996.',
 '(The dividends paid to us by these investees have been includedin the operating earnings itemized on page 11, mostly under
&amp;quot;InsuranceGroup: Net Investment Income.&amp;quot;)                                                          Berkshires Sha
re                                          of Undistributed                              Berkshires Approxi
mate     Operating Earnings Berkshires Major Investees   Ownership at Yearend(1)      (in millions)(2)  American Express Com
pany             10.7%                    $161The Coca-Cola Company                 8.1%                      216The Wa
lt Disney Company               3.2%                      65Freddie Mac                          8.6%
86The Gillette Company                  8.6%                       82The Washington Post Company          16.5%
30Wells Fargo & Company                 7.8%                      103
```

```
1 lsa[20]
```

executed in 6ms, finished 19:21:29 2020-01-12

```
['Whenever we buy into an industry whose leading participants arentknown to me, I always ask our new partners, &amp;quot;Are the
re any more athome like you?&amp;quot; Upon our purchase of Nebraska Furniture Mart in 1983,therefore, the Blumkin family told m
e about three outstanding furnitureretailers in other parts of the country.',
 'As was thecase with the Blumkins and Bill Child, I had no need to check leases, workout employment contracts, etc.',
 'That other possibility   synergygains   is usually unrealistic, since we expect acquirees to operate afterweve bought them
just as they did before.',
 'Coke continues to increase itsmarket dominance throughout the world, but, tragically, it has lost theleader responsible fo
r its outstanding performance.',
 'After his death, I read every one of themore than 100 letters and notes he had written me during the past nineyears.']
```

- Summarizing entire corpus

Milestone Report 2: Topic Modeling

Topic Modeling

- Rationale for topic modeling, how it differs from summarizing
    - Differs in output, output in TM is clusters and latent embedding.
    - TM works on entire corpus.
    - Summarization is for each individual document.
- Different types of topic modeling, tools used, expectation for results
- Applying topic modeling to corpus
- Measuring coherence of topic models

Recommendations