IBM ANNUAL REPORT TEXTUAL ANALYSIS

DATA620: Spring 2021
Assignment 12.1
Theodore Fitch
Dr. Majed Al-Ghandour

Executive Summary:

I performed textual analysis on the IBM Annual Reports from 1999, 2000, 2010, 2019, and 2020 looking for trends in order to create business value. These Annual Reports are written to the IBM shareholders to describe the financials, the business changes, and the technology changes. I found that the reports have become 9x longer and have become twice as complex. I also confirmed that longer words tend to be used less frequently than shorter words. The most common words stayed consistent year over year in their relative order. The top 200 words have a few interesting trends which reflect internal and external changes for IBM (like the move from selling products to services/software, and the move towards internet/remote based systems). The 10 most common words were used 10x more on average in 2020 than in 1999. IBM should consider: decreasing the complexity of these reports using a table of contents, standardizing terms/adding abbreviations, splitting up this report, or adding more organizational sections; creating a live version of yearly financial data so that shareholders may see IBM's software showcased while comparing year over year data; and creating a 5-minute version of this report so that the audience can get a high level view of everything and then jump to the regions of annual report which interest them the most.

Introduction:

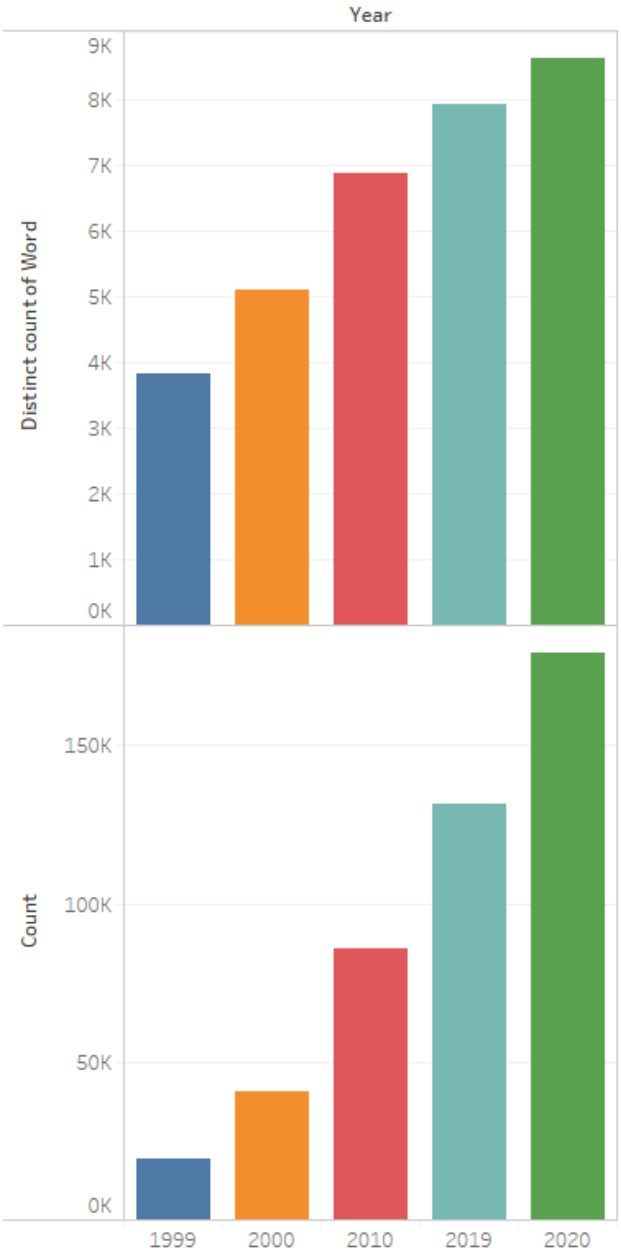
Since its inception in 1911, IBM has delivered unfathomable value to the world including ATMs, floppy discs, SQL, and UPC barcodes (Lu, 2016). IBM went public 1981 and releases an "Annual Report" for its shareholders to tell the story of its quantitative growth that year. These reports typically show KPIs (revenue, net income, dividends, assets, etc.), year over year growth, projected growth for the coming years, unique stories of IBM's (or a subsidiary's) growth, and

where they expect to be growing towards in the coming year. Thus, they primarily contain words related to financials, business, and technology. I analyzed these reports in particular because I appreciate IBM's wide array of products and because I wanted to find trends in the words used to provide insight on how to make these reports better in the future. Annual shareholder reports naturally can feel dry and boring – alienating the audience from reading them. So, finding ways to make the reports better is vital to keeping shareholders informed.

The annual reports from IBM for 1999, 2000, 2010, 2019, and 2020 were extracted from IBM's website (IBM 1999 Annual Report, n.d.)(IBM 2000 Annual Report, n.d.)(IBM 2010 Annual Report, n.d.)(IBM 2019 Annual Report, n.d.) (IBM 2020 Annual Report, n.d.). Originally, I was just going to use 2000, 2010, and 2020; however, I added 1999 and 2019 in order to see 1. changes in those years (where big technological changes were occurring) and 2. to confirm that reports at the decade mark weren't substantially longer than reports at a non-decade mark. Each report was converted from .PDF to .TXT format. A *Python* (3.7) script was written using *Spyder* (3.3.6) to: extract the words from each .TXT file; and to make a list of 1. unique words found in each report, 2. how many times each word was repeated, 3. what year that word was found, and 4. the character length of that word. Lastly, two .CSV files were made, one having all data and one having only the top 200 words from each year (1,000 words total). This was done so that analyses could be performed on all words and on just the top 200 words (the first .CSV was used for **Figure 1**, and the second .CSV was used for **Figure 2-5**). These .CSV files were explored using *Tableau* (2020.4.0 64-bit).

Dataset Insights and Visualizations:

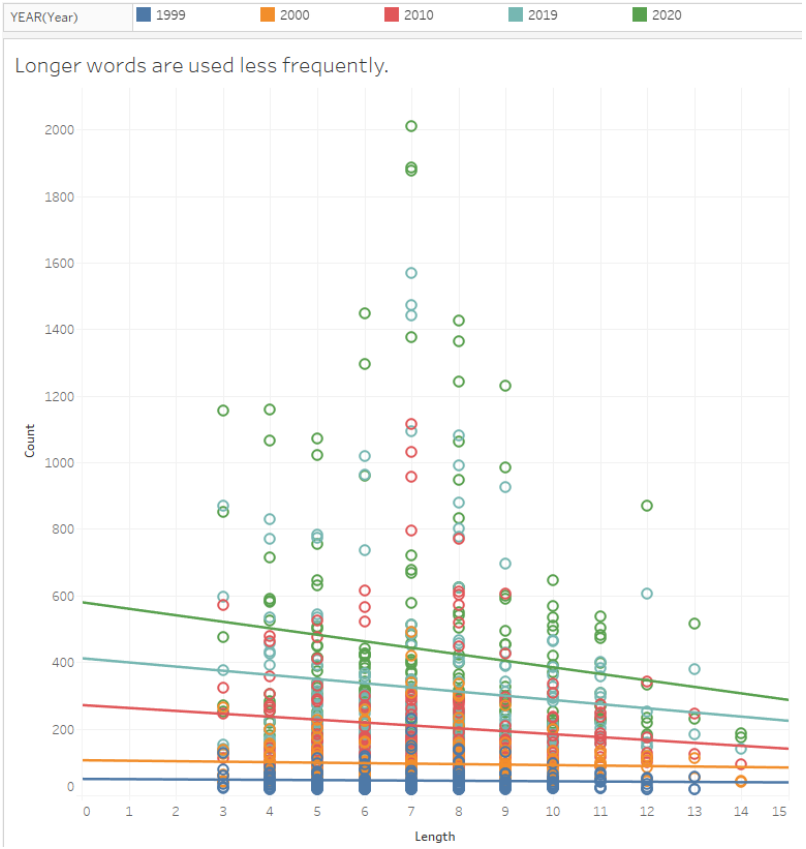
Figure 1. IBM Annual Reports grew 9x longer over 20 years and twice as complex in the language they use.



Beginning in 1999, 19,031 total words were used in the annual report with 3,808 distinct words. However, by 2020 179,145 words were used with 8,606 distinct words being used. This phenomenal shift likely reflects a few things. First, the complexity of the business has increased extensively to the point where executives think it is necessary to use 9 times the number of words to describe that year's activities. This is also reflected in the language being used becoming increasingly complex. The 2020 report used over double the number of distinct words from the 1999 report. This suggests that more topics are being spoken about. As previously

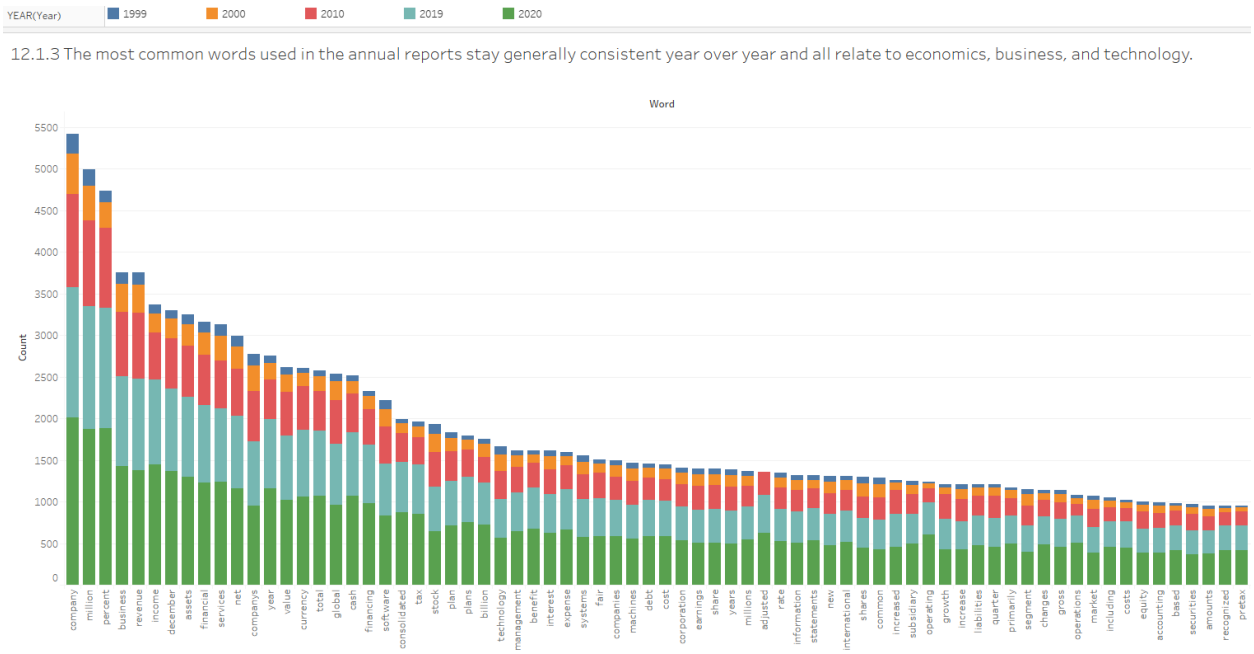
mentioned, these reports discuss the company's financials as well as business stories (past and projected).

Figure 2. Longer words are used less frequently in the IBM annual reports.



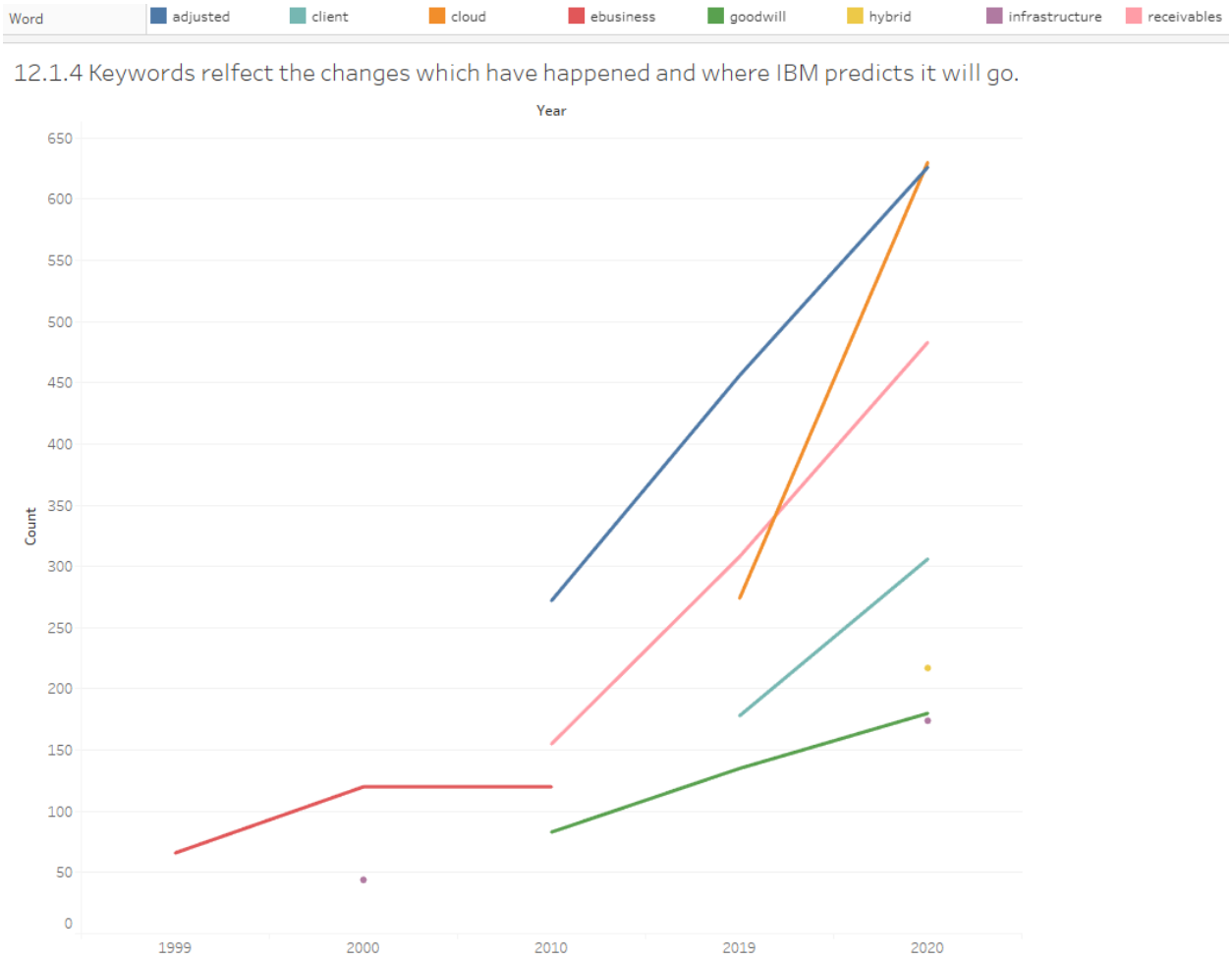
Intuitively, it makes sense that longer words would be used less frequently than short words. This was explored by comparing the count of the 200 most common words for each year versus the length of each word. Each year shows a downward trend between the variables, meaning our intuition is correct quantitatively. It's important to note the Python program written removed any common, simple words (e.g. "the", "to", "and"). Therefore, there are no 1 or 2 length words recorded in this graph. The average character length did not significantly vary year over year.

Figure 3. The most common words of IBM annual reports stay consistent over a 20-year period.



The 5 most common words across all years were “company”, “million”, “percent”, “business”, and “revenue”. These along with the next 67 pictured in **Figure 3** generally stay consistent over the 20-year timespan analyzed (there are some outliers like “adjusted” which will be addressed in **Figure 4**). Most words clearly reflect the general theme of the annual reports concerning economics, business, and technology (since effectively this is an accountability report to the shareholders).

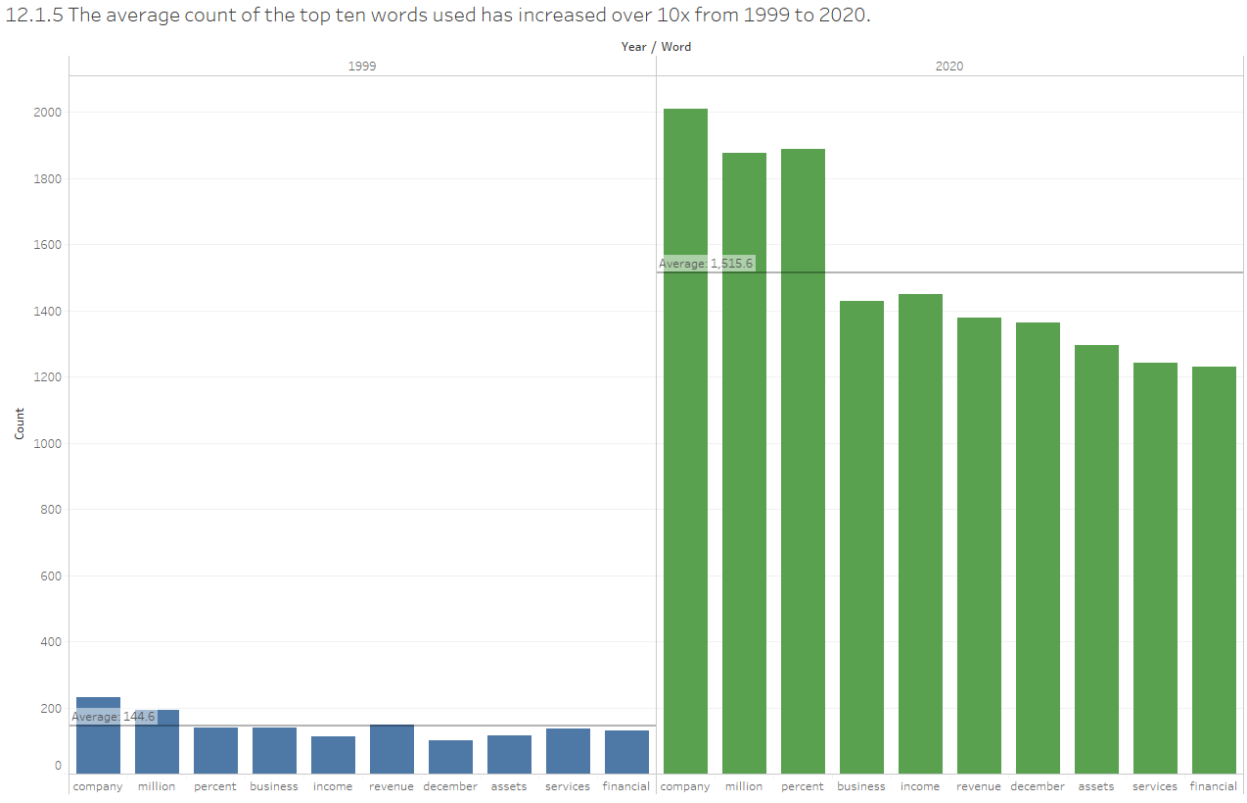
Figure 4. Keywords reflect the changes which have happened internally and externally and how IBM predicts the landscape will continue to change.



A few keywords being used reflect major changes which have taken place within IBM as well as in the broader world. Each of these words were chosen based on year over year changes observed in **Figure 3** (when all words were viewed not just the top 72 as shown in that figure). "Adjusted" saw a substantial increase over 2010 to 2020. While this is also used in the 1999 and 2000 reports, it is used far less frequently. This reflects the increase in revenue streams IBM has internationally since "adjusted" was almost always referring to money being "adjusted for currency" since there would be a fee to convert money into dollars (and thus change the reportable result). In recent years, IBM has generated many new software which are subscription

based. This explains the meteoric rise of “client” and “receivables” since these have become sudden new income streams. Next, as technology has exponentially moved towards being more internet based as the default modus operandi, words like “e-business” have been replaced by “cloud”. “E-business” was likely considered ahead of the curve in the early 2000’s, but in the 2020’s it is a term of the past. Like many other major technology companies most of IBM’s business is now e-business de facto and has moved from many physical products to mainly services/software (Rodriguez et al., 2017). It is also vital to point out the sudden increase in “hybrid” and “infrastructure” reflecting the business practice shift that has occurred with the onset of COVID-19. Where it was once the norm to have all workers in the officer and a cloud was a nice-to-have, it is commonplace that many folks are working from home, and now using a cloud-based system is a necessity. IBM’s new strategy into the 2020’s is to create a “hybrid cloud” where customers don’t need to reinvent their whole IT infrastructure. Finally, “goodwill” has become increasingly more used reflecting the increase in IBM’s number and complexity of acquisitions. It is important to point out that “goodwill” is also used in the 1999 and 2000 reports but is only used a few times (6 and 17 times respectively).

Figure 5. The average count of the top ten words used increased over 10x from 1999 to 2020.



Comparing the averages of the top ten words from 1999 versus 2020 shows a tenfold increase. These words each make sense since these reports are primarily financials reporting documents. It follows that these common financial words would remain the same year over year and would be used many times over in these reports. December at first was surprising to see; however, since the IBM fiscal year runs January to December, these reports frequently refer to December 31st as the closing day of the year. As previously mentioned, the reports have become 9 times longer from 1999 to 2020. Logically, it follows that the most frequent words would be used slightly more than less common words. While the stories and trends will change year over year, the words used to describe the financial summary will remain the same.

Conclusion:

Thus far, I have shown that: the annual reports have become increasingly more complex over time (both in length and the textual descriptions provided); the most common words have become 10x more common from 1999 to 2020; longer words tend to be used less commonly than shorter words; a few unique words show how IBMs business model changed from being products based to service/software based (and towards online/remote being de facto); and that the top 72 words used year over year have remained relatively consistent in their ranking for how often they are used. Hence, the natural recommendations from this would be:

- The report drafters should put commensurate effort into ordering, organizing, and sectioning the reports so that their audience can quickly jump to and read the sections they are interested in. None of the reports analyzed contained a table of contents which they desperately need. This will help readers more easily navigate the reports since they have become so complex.
- The report writers should consider standardizing terms and creating internal abbreviations early in the report, so they do not need to repeat themselves multiple times. For example, if "adjusted for currency" was changed to "AFC", this would save approximately 11,268 characters per report along with the reading time of the audience.
- IBM should consider making a dynamic version of this annual report where their shareholders can play with the graphs to compare year over year data (regionally, temporally, organizationally, etc.). This would allow IBM to showcase some of their new capabilities/software while presenting their glowing results.

- IBM should plan on preventing these reports from becoming more complex over time. At the current rate, the 2040 report will be approximately 810,000 words (with around 16,000 distinct words). Thus, IBM must consider ways to decrease the complexity of these reports. This may include making obvious section breaks, splitting this report into multiple reports, or adding an index.
- Lastly, the annual report should also be summarized in video format to give a 5-minute story of this 150-page document. This would save shareholders much precious time and allow them to explore the areas of the report which interest them the most based on the video summary.

References:

IBM 1999 Annual Report (n.d.). Retrieved March 31th, 2021, from IBM website: https://www.ibm.com/annualreport/assets/past-reports/1999-ibm-annual-report.pdf

IBM 2000 Annual Report (n.d.). Retrieved March 31th, 2021, from IBM website: https://www.ibm.com/annualreport/assets/past-reports/2000-ibm-annual-report.pdf

IBM 2010 Annual Report (n.d.). Retrieved March 31th, 2021, from IBM website: https://www.ibm.com/annualreport/assets/past-reports/2010-ibm-annual-report.pdf

IBM 2019 Annual Report (n.d.). Retrieved March 31th, 2021, from IBM website: https://www.ibm.com/annualreport/assets/past-reports/2019-ibm-annual-report.pdf

IBM 2020 Annual Report (n.d.). Retrieved March 31th, 2021, from IBM website: https://www.ibm.com/annualreport/assets/past-reports/2020-ibm-annual-report.pdf

Lu, Y. (2016). 20. 2011: IBM. In *Inside the Investments of Warren Buffett* (pp. 230-246). Columbia University Press.

Rodríguez, P., Haghighatkhah, A., Lwakatare, L. E., Teppola, S., Suomalainen, T., Eskeli, J., ... & Oivo, M. (2017). Continuous deployment of software intensive products and services: A systematic mapping study. *Journal of Systems and Software*, *123*, 263-291.